NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

MASTER OF SCIENCE
“INFORMATION AND COMMUNICATION TECHNOLOGIES”

M.Sc. Thesis

MEASURING INFLUENCE OF USERS AND PHOTOS REGARDING A SPECIFIC TOPIC IN FLICKR

Ploutarchos E. Liosis

SUPERVISORS: **Tsalgatidou Afroditi,** Associate Professor
Koutrouli Eleni, Postdoctoral Researcher

ATHENS

SEPTEMBER 2024

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΤΡΗΣΗ ΕΠΙΡΡΟΗΣ ΧΡΗΣΤΩΝ ΚΑΙ ΦΩΤΟΓΡΑΦΙΩΝ ΣΧΕΤΙΚΑ ΜΕ ΣΥΓΚΕΚΡΙΜΕΝΟ ΘΕΜΑ ΣΤΟ FLICKR

Πλούταρχος Ε. Λιόσης

Επιβλέπουσες: **Τσαλγατίδου Αφροδίτη,** Αναπληρώτρια Καθηγήτρια
Κουτρούλη Ελένη, Μεταδιδακτορική Ερευνήτρια

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2024

M.Sc. Thesis

Measuring influence of users and photos regarding a specific topic in Flickr

Ploutarchos E. Liosis

S.N.: M1506

SUPERVISORS: **Tsalgatidou Afroditi,** Associate Professor
Koutrouli Eleni, Postdoctoral Researcher

September 2024

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέτρηση επιρροής χρηστών και φωτογραφιών
σχετικά με συγκεκριμένο θέμα στο Flickr

Πλούταρχος Ε. Λιόσης
Α.Μ.: M1506

Επιβλέπουσες: **Τσαλγατίδου Αφροδίτη,** Αναπληρώτρια Καθηγήτρια
Κουτρούλη Ελένη, Μεταδιδακτορική Ερευνήτρια

Σεπτέμβριος 2024

ABSTRACT

This M.Sc. dissertation presents a comprehensive approach to influence detection and suggestions regarding a specific topic, within the context of the photo-sharing platform Flickr.

After a quick review of three of the most used filtering techniques, specifically content-based filtering, collaborative filtering and hybrid filtering, we present a solution that consists of a MySQL database, a backend service layer - built on top of Flickr's APIs and the database - implemented using the Spring Boot Framework, and an Angular web application that provides end-users with an interface to interact with the web servers.

A key part of the proposed system is the recommendation engine, which identifies related tags that will compose a specific topic, using similarity criteria such as Euclidean distance, Cosine similarity and Weighted Measure Similarity formulas.

Our ultimate goal is to provide actionable insights into the key drivers of influence within the platform, identifying trends and influential photos and users, while creating a scalable, data-driven system that can adapt to a wide range of topics.

SUBJECT AREA: Recommendation Systems, Influence Metrics, Social Networks

KEYWORDS: recommendation systems, influence, social networks

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία παρουσιάζει μία ολοκληρωμένη προσέγγιση για τον εντοπισμό της επιρροής σχετικά με ένα συγκεκριμένο θέμα, στα πλαίσια της πλατφόρμας διαμοιρασμού φωτογραφιών Flickr.

Αρχικά αναλύονται τρεις από τις πιο γνωστές τεχνικές φιλτραρίσματος, που είναι το φιλτράρισμα βάσει περιεχομένου, το συνεργατικό φιλτράρισμα και το υβριδικό φιλτράρισμα.

Στη συνέχεια, προτείνεται μια λύση που αποτελείται από μια βάση δεδομένων MySQL, ένα backend σύστημα και μια εφαρμογή που παρέχει στους τελικούς χρήστες την κατάλληλη διεπαφή για αλληλεπίδραση με το υπόλοιπο σύστημα. Το backend σύστημα έχει αναπτυχθεί πάνω από τις διεπαφές προγραμματισμού εφαρμογών (Application Programming Interfaces - APIs) της πλατφόρμας Flickr και τη βάση δεδομένων και έχει υλοποιηθεί σε Spring Boot Framework με τη χρήση της Java, ενώ η εφαρμογή είναι γραμμένη σε Angular.

Βασικό μέρος του προτεινόμενου συστήματος αποτελεί η μηχανή συστάσεων, η οποία προσδιορίζει τις σχετικές ετικέτες που θα συνθέσουν ένα συγκεκριμένο θέμα, χρησιμοποιώντας κριτήρια ομοιότητας όπως η Ευκλείδεια απόσταση, η ομοιότητα συνημίτονου και σταθμισμένες μετρικές ομοιότητας.

Τελικός μας στόχος είναι να παρέχουμε αξιόπιστα αποτελέσματα και πληροφορίες σχετικά με τους βασικούς παράγοντες επιρροής της πλατφόρμας, εντοπίζοντας τις τάσεις και τους πιο επιδραστικούς χρήστες και φωτογραφίες, δημιουργώντας παράλληλα ένα σύστημα, που βασίζεται στα δεδομένα και μπορεί να επεκταθεί σε ένα ευρύ φάσμα θεμάτων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Συστήματα Συστάσεων, Υπολογισμός Επιρροής, Κοινωνικά Δίκτυα

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: συστήματα συστάσεων, επιρροή, κοινωνικά δίκτυα

CONTENTS

LIST OF FIGURES

PREFACE

This thesis was developed for the M.Sc. program in the Department of Informatics and Telecommunications at the National and Kapodistrian University of Athens, in the area of Information and Communication Technologies.

In this project, we present a Recommendation System, based on Weighted Hybrid and Collaborative filtering techniques, to discover related tags and topics. Additionally, we combine multiple influence factors to measure influence scores and identify the most influential users and photos in Flickr.

Our solution integrates several modern technologies, including Java, Spring Boot for the service layer, Angular for the front-end, and MySQL as the database, to effectively process, analyze, and display insights related to user and content influence.

1. INTRODUCTION

Influence in social networks is a concept that has garnered significant attention in research due to its relevance in understanding user impact and topic popularity. [1] [2] Influence can be viewed through various metrics, such as the popularity a post or user generates, the engagement it attracts, or even the potential impact on future actions. Different platforms measure influence differently; for example, Twitter uses follower count as a simple metric, while other systems incorporate additional information like the number of posts, shares, or likes to provide a more comprehensive picture.

Identifying influential users and trending topics is particularly valuable for advertisers and content creators, as it allows them to effectively target popular users and topics to reach wider audiences. In social networks, influence is often tied to current discussions represented by tags / hashtags, which are indicators of trending topics that frequently evolve based on public interest. The spread of news, for instance, can be amplified by influential users, as their reach can lead to quicker dissemination among followers.

The present research deals with Flickr, a photo management and sharing application, widely known for its services. With millions of users worldwide, Flickr has become a source of inspiration for photographers of all levels, providing a space to showcase their work, engage with fellow enthusiasts, and discover creative perspectives. A photograph in most cases is still relevant for months or years after the date it has been taken. A travel photograph, for example, as a travel tip and trick, does not have "expiration date" and can be influential indefinitely.

Within this community, certain photos or posts can become inspirational and influential, gathering significant attention and recognition. Metrics like views, favorites, and comments serve as indicators of a post's influence, helping photographers gauge the impact and reach of their work while also fostering a sense of validation, support and inspiration within the Flickr community.

1.1 Thesis Main Objective

The purpose of this thesis is to develop a comprehensive system to assess and quantify the influence of users and photos within Flickr, specifically focusing on targeted topics defined by specific tags.

Initially, the process starts with the collection of a robust dataset related to a chosen tag, which includes associated photos, user interactions, comments, and favorites. Leveraging collaborative filtering techniques, the system then identifies tags that are most similar to the original tag, combining these into a broader topic for more extensive analysis. This enables the definition of a topic as a collection of closely related tags, encompassing all relevant content and interactions. Subsequent steps involve calculating the influence of individual users and identifying the most influential photos within these topics. By evaluating how users interact with these tags and the level of engagement photos receive, the proposed system aims to pinpoint key influencers and the most influential photos.

1.2 Thesis Structure

The rest of this thesis has been structured as follows:

The following section is about influence in digital era, internet evolution through the last decades and social media platform usage.

The third section presents a quick review of three of the most used filtering techniques, specifically content-based filtering, collaborative filtering and hybrid filtering.

In the fourth section we define our proposed methodology in order to find similar tags that act as a topic and find the most influential users and photos regarding this topic using specific influence factors and metrics.

In the fifth section we analyze Flickr APIs' features and architecture, and how we can integrate with them in order to collect the necessary data.

In the sixth section we present our system solution that consists of a MySQL database, a backend service layer - built on top of Flickr's APIs and the database - implemented using the Spring Boot Framework, and an Angular web application that provides end-users with an interface to interact with the web servers.

In the seventh section we present our evaluation results with many valuable statistics and influence scores, to assess various aspects of the proposed approach and its effectiveness.

Finally, in the last section, eighth, we conclude on the solution we have proposed, with some possible future directions.

2. INFLUENCE ON DIGITAL WORLD

If we want to specify what influence means and understand the way it is used on the internet, we have to go back several years and take a brief look at the state that internet was, as well as its evolution in the following years. The advent of the internet has revolutionized the way we connect, communicate, and consume information, leading to a significant shift in the dynamics of influence even though it initially served as an information-sharing platform, with limited avenues for user participation and interactivity.

2.1 How did the Internet evolve?

In the late 1990s and early 2000s, the Internet began to gain popularity as more people gained access to it. It was already an important part of everyday life for a significant percentage of them. During this period, websites served as the primary means of information dissemination. News websites, forums and internet communities were constantly being introduced, gaining popularity in a short period of time. Influencers emerged in niche communities and relied on their websites to share content and engage with their audience. While these early influencers had a limited reach compared to today, they laid the foundation for the concept of online influence.

In the 21st century a transition was completed from traditional industries to a new economic domination of information and communications technology. Sooner or later, most companies recognized the potential of the internet. The emerging digital transformation and the exposure to new users, forced them to try to exploit their competitive advantage. For this reason, they started focusing their business plans on this brand-new communication channel and tried to get their customers in the same way. This new internet era, also known as Web 2.0, gave rise to new techniques of brand building and targeting audience through influencer marketing.

Figure 1 Web 2.0

2.2 Web 2.0, a new digital era

Even this ‘strange’ term, Web 2.0, was a marketing trick, in a way. There is no clear definition of this, as technological evolution could not be divided into versions as a matter of fact. In most cases, after all, technology does not change, only the way we perceive it and use it changes. This term was coined simply to mark the beginning of this new digital era where everyone would be connected. As Tsekeris and Katerelos (2012, p. 233) [3] concisely puts it, “It is vividly shown that what mostly defines Web 2.0 and differentiates it from Web 1.0 is the explosion of user-generated content (a fundamental bottom-up process). This amazingly reinforces social dynamics and provides a stable flow of unpredictable creativity, innovation and adaptation.”. It is noteworthy that this era brought about a democratization of influence, enabling anyone with internet access to build an online presence and connect with others.

The evolution in the way the internet works did not happen overnight, but it was a surprise and the result of a long process. At the beginning it was about the available volume of information. Material began to be created for any kind of information that could interest a large - or even small - number of people. Search engines increased, improved and became smarter, producing better results and suggestions at a very fast pace. One of the first industries that entered the digital age en masse was media. Taking advantage of the immediacy offered by the internet, they created their online presence, transmitting news and information in real time.

Apart from the above-mentioned, ordinary users without necessarily being familiar with technology - as was the case in the past - began to create communities, quite reminiscent of forums as we know them today, where anyone could write or ask something. That was just the beginning. After that, blogs, websites, personal pages on social networks - most of them completely free - were all made available to internet users to capture their ideas, their thoughts, whatever they like and what they do not.

2.3 Influence and influencers

Around the mid-2000s, it was the period when digital technologies started to play a prominent role in shaping up and regulating the behaviors, performances, standards, etc., of societies, communities, organizations, and individuals. Thus, the terms "influence" and "influencer" came to the fore or to put it correctly, they were redefined to characterize the digital age in which digital technologies are used in almost every aspect of life.

While this phase marked a significant shift in online participation, as users became active contributors rather than passive consumers, it was the rise of social media platforms that truly transformed the landscape of online influence. Platforms like Facebook, Twitter, YouTube, and Instagram, which emerged in the late 2000s and early 2010s, provided individuals with powerful tools to amplify their reach and engage with larger audiences. These platforms allowed users to share a wide range of content, from personal updates to news articles, photos, and videos. As a result, influencers started to gain significant following, often centered around specific topics or areas of interest.

Over time, social media has become deeply integrated into our daily lives, impacting various aspects of society. Influencers, bloggers, and content creators have emerged as influential figures, built large followings and shaped trends in areas such as fashion, beauty, travel, and more. Brands have recognized the power of social media influencers to sway consumer behavior and have increasingly leveraged their reach for marketing purposes. The influence of social media continues to grow as these platforms refine their algorithms to deliver personalized content to users based on their interests and engagement patterns.

Today, the social media Influence is being used in ways that shape business, innovation, politics, world culture, careers, and more. Social media has facilitated rapid dissemination of information and mobilization of like-minded individuals, leading to the organization of protests, social movements, and even revolutions.

2.4 Social media platform usage

Facebook – a top tier social media platform – remains one of the most widely used social media sites. It is notable that almost a quarter of the world's population is now on Facebook. [4] In the U.S., whereas only five percent of Americans used a social media network in 2005, that number is the last five years over 70 percent. [5]

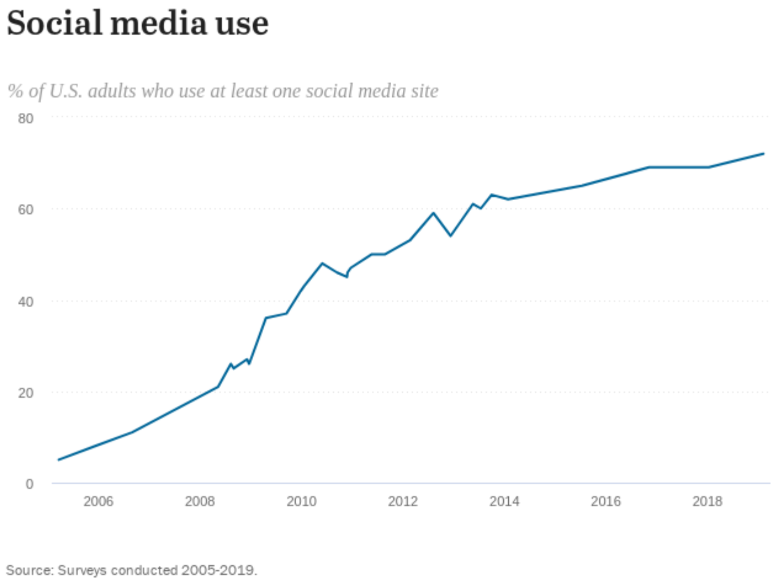


Figure 2 Social media use over time

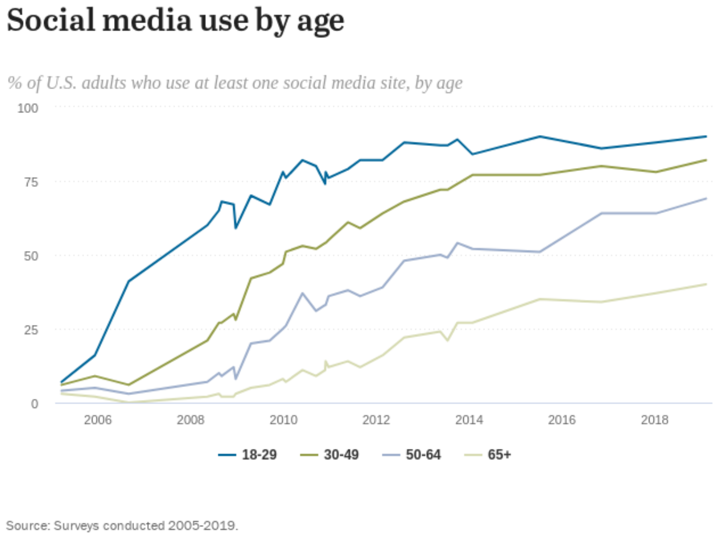


Figure 3 Social media use over time by age

As has already been mentioned by Simplilearn in their article [6] “In comparison to other media, the influence of social media in political campaigns has increased tremendously. Social networks play an increasingly important role in electoral politics — first in the ultimately unsuccessful candidacy of Howard Dean in 2003, then in the election of the first African American president in 2008, and again in the Twitter-driven campaign of Donald Trump.” The New York Times also reports [7] that “The election of Donald J. Trump is perhaps the starkest illustration yet that across the planet, social networks are helping to fundamentally rewire human society.”

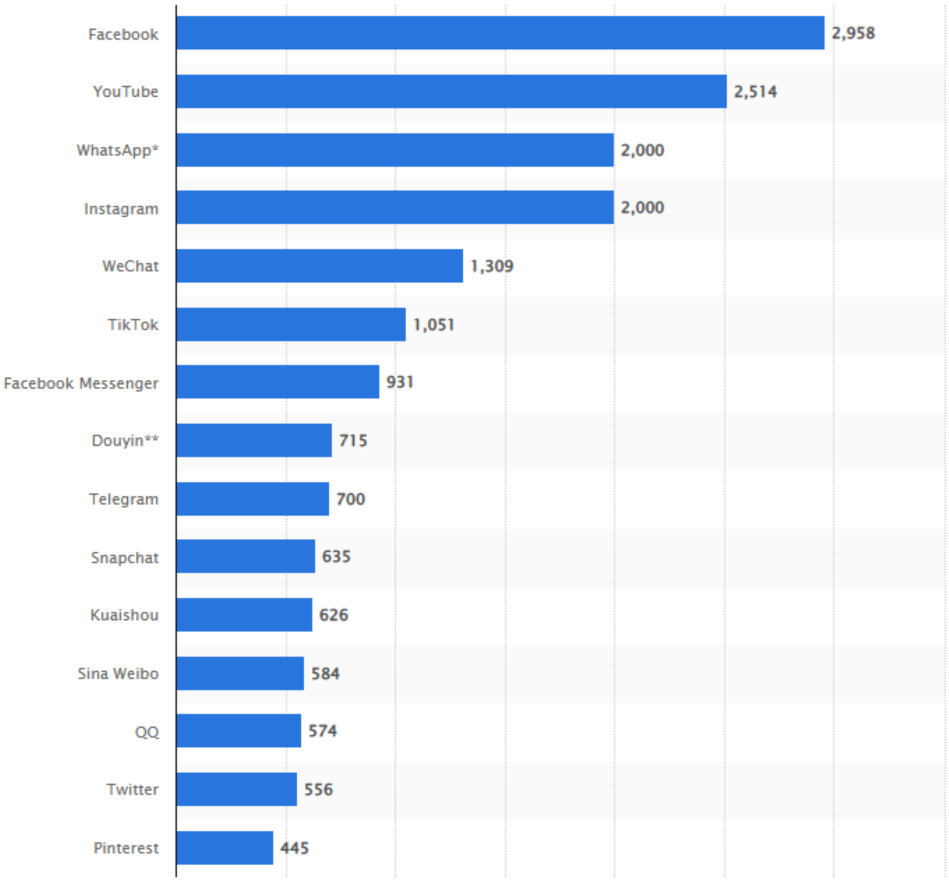


Figure 4 Most popular social networks worldwide as of January 2023, ranked by number of monthly active users (in millions) [8]

It is important to note that while the internet and social media have opened new avenues for influence, they have also raised concerns about privacy, fake news, and the manipulation of public opinion. As we continue to navigate the ever-evolving digital landscape, understanding the dynamics of influence on the web becomes crucial for individuals, organizations, and societies.

But, at the end of the day, information is power, and the positive impact of social media is huge and overcome these drawbacks. Without a way to spread the information, people cannot make use of its power. Sharing is about getting people to see and respond to content. If the content is still relevant and the need for information still exists, it’s always valuable for any individual or organization to use social media to keep publishing.

3. RECOMMENDATION SYSTEMS AND FILTERING TECHNIQUES

The billions of new electronic devices of all kinds that we produce every year, the rapid increase of internet users, the ever-increasing information we share online as well as the advent of the Internet of Things (IoT) fuel this exponential growth of data which we observe evolving, especially in the last decade. This rapid increase in the amount of published information or data and the effects of this abundance, which we sometimes refer to as information explosion [9], is one of the biggest challenges of our time in many fields of Information Technology such as Data Science and Data Analytics.

With the ever-growing volume of available data online, the problem of managing all this information becomes more difficult, which can lead to information overloading. Techniques to gather knowledge from an overabundance of digital information have existed since the 1970s. Search engines, such as Google, Bing etc. and general information retrieval systems, have partially solved this problem in their fields, however, personalization and prioritization of information, where a system filters and maps the available content to user's interests, were absent for years. This has increased the demand for recommendation systems more than ever before. A recommender system, as a subclass of data and information filtering, seeks to predict the ratings and/or the preferences a target user will give in an item, such as a product, movie, picture, song, etc.

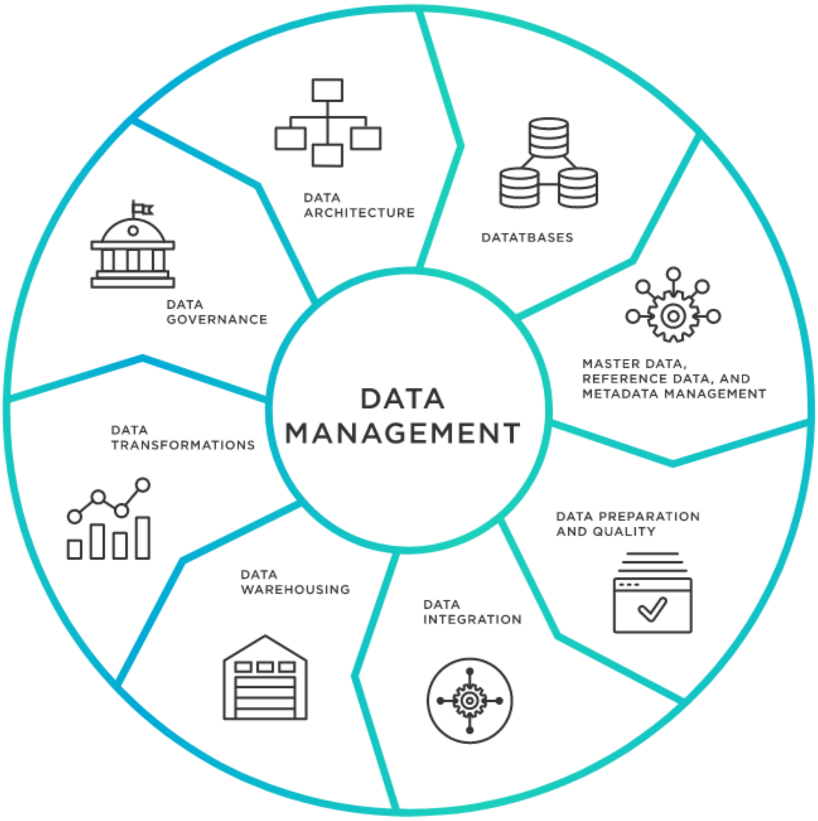


Figure 5 Data management use cases

3.1 Data management and recommendations

Data Management comprises all disciplines related to managing data as a valuable resource. Topics that we are interested in data management include but not limited to data architecture, data integration and data cleansing. If the data is not well defined, the data will be misused in applications. If the process is not well defined, it is impossible to meet user needs.

While some companies are good at collecting data, they are not managing it well enough to make sense of them. Simply collecting data is not enough. As PwC points out in an article about 'Create Value from Data' [10], "Data is not, despite the headlines, the new oil. To continue the analogy, most data is hidden, polluted, un-processable and too expensive to extract in a meaningful format." Sometimes it doesn't matter how much data you have if you do not have a clearly defined plan for how to use them and what value you want to produce.

Every data science or machine learning algorithm, every data model, and every data analysis requires carefully prepared, well cleaned and accurate data. Data professionals would agree that upwards of 50% of the job consists of managing the data [11] and cleaning and organizing data is the least enjoyable part of their work. Getting the right data from various sources, finding accurate data mining patterns, cleaning and organizing data, etc. prove that added value from data is not a given or an established fact, but it needs to be created!

Consequently, there is a big need to use efficient and accurate recommendation techniques within a system that we want to provide us with relevant and dependable recommendations for users.

3.1.1 Information collection phase

Information collection or data collection is a methodical process of gathering and analyzing specific information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate the results. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed.

Regardless of the field of study or preference for defining data (quantitative or qualitative), accurate data collection is essential to maintain research integrity. A formal data collection process is necessary as it ensures that the data gathered are both defined and accurate. This way, subsequent decisions based on arguments embodied in the findings are made using valid data. [12] The process provides both a baseline from which to measure and in certain cases an indication of what to improve.

To support the observation of errors in the information collection process we have to maintain data integrity. There are two approaches that may protect data integrity and secure scientific validity of study results invented by Craddick, Crawford, Rhodes, Redican, Rukenbrod and Laws in 2003:

- Quality assurance – all actions carried out before data collection

Its main focus is prevention, which is primarily a cost-effective activity to protect the integrity of data collection. Standardization of protocol best demonstrates this cost-effective activity, which is developed in a comprehensive and detailed procedures manual for data collection. The risk of failing to identify problems and errors in the research process is evidently caused by poorly written guidelines.

- Quality control – all actions carried out during and after data collection

Since quality control actions occur during or after the data collection all the details are carefully documented. There is a necessity for a clearly defined communication structure as a precondition for establishing monitoring systems. Uncertainty about the flow of information is not recommended as a poorly organized communication structure leads to lax monitoring and can also limit the opportunities for detecting errors. Quality control is also responsible for the identification of actions necessary for correcting faulty data collection practices and minimizing such future occurrences. A team is more likely to not realize the necessity to perform these actions if their procedures are written vaguely and are not based on feedback or education.

But can collecting such a large amount of data from a website be practical? Nowadays, many researchers are interested in obtaining new kinds of data directly from the web. When done manually, this approach is prone to human error. When automated, it may violate a website's terms of service. Instead, researchers typically rely on other programmatic methods like REST APIs [13].

A REST API (also known as RESTful API) is an application programming interface (API or web API) that conforms to the constraints of REST architectural style and allows for interaction with RESTful web services. REST stands for representational state transfer and was created by computer scientist Roy Fielding. [14]

An API is a back-end interface through which third-party developers may connect new add-ons to an existing service. An API is a set of definitions and protocols for building and integrating application software. It's sometimes referred to as a contract between an information provider and an information user - establishing the content required from the consumer (the call) and the content required by the producer (the response). You can think of an API as a mediator between the users or clients and the resources or web services they want to get.

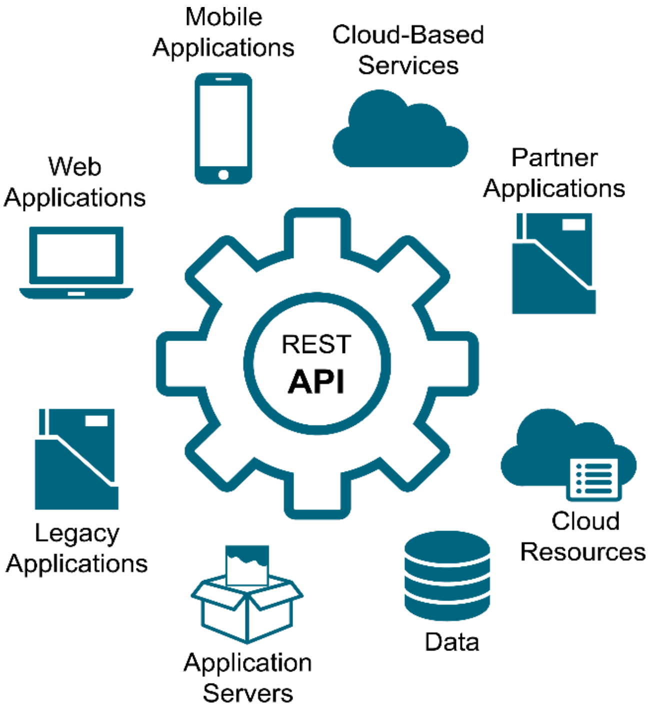


Figure 6 REST APIs

REST APIs communicate via HTTP (Hypertext Transfer Protocol) requests [15] to perform standard database functions like creating, reading, updating, and deleting records (also known as CRUD) within a resource. For example, a REST API would use a GET request to retrieve a record, a POST request to create one, a PUT request to update a record, and a DELETE request to delete one. All HTTP methods can be used in API calls. [16]

For example, the API design for a weather service could specify that the user supplies a zip code, and that the producer replies with a 2-part answer, the first being the high temperature, and the second being the low.

Many major companies, as well as government agencies, have created public APIs. Public APIs are open to anyone and can be used without restrictions. These organizations want to provide easy access to their data to encourage developers to use and extend their platforms with third party applications. There are also private APIs that are only accessible by authorized users and may be subject to usage restrictions.

3.1.2 Data Transformation phase

In the field of Information Technology, Data Transformation is the process of converting data from one format to another, typically from the format of a source or legacy system into the required format of a destination system. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration and application integration. [17]

In the data transformation stage, a series of rules or functions are applied to the extracted data in order to prepare it for loading into the end target. Tools and technologies used for data transformation can vary widely based on the format, structure, complexity, and volume of the data being transformed.

An important function of transformation is data cleansing, which aims to pass only "proper" data to the target. Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. [18] Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. This challenge occurs when different systems interact in the relevant systems' interfacing and communication. Character sets that may be available in one system may not be so in others.

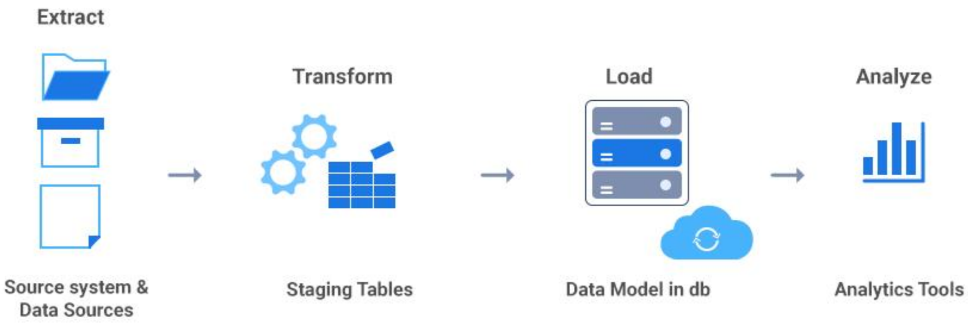


Figure 7 ETL Process (Extract - Transform - Load)

In any case, one or more of the following transformation types may be required to meet the business and/or the technical needs of an information collection and data integration process:

- Selecting only certain columns to load
- Translating coded values - if the source system codes male as "1" and female as "2", but the warehouse codes male as "M" and female as "F"
- Encoding free-form values - mapping "Male" to "M"
- Deriving a new calculated value - sale_amount = qty * unit_price
- Sorting and/or ordering the data based on a list of columns to improve search performance
- Splitting a column into multiple columns - converting a comma-separated list, specified as a string in one column, into individual values in different columns
- Looking up and validating the relevant data from tables or referential files
- Applying any form of data validation - failed validation may result in a full rejection of the data, partial rejection, or no rejection at all

3.1.3 Recommendation phase

Recommendation or Prediction phase has to do with the ability to predict and make useful suggestions to users, by considering their profile, preferences and/or actions during interaction with an application or website. These recommendations follow as the consequence of a large amount of data that we have collected, analyzed and rated during all the above data processes.

Recommendation systems - or recommender systems - have impacted on our lives in many ways. It is not an exaggeration to say that they have even redefined the way we think and/or act. One example of this impact is how our online shopping procedures and experiences are being redefined. As we browse through various products in a website, a core Recommendation System offer recommendations of other products we might be interested in. Similarly, in a movie recommendation system, it might suggest similar movies based on genre, actors, directors, or other attributes that the user has shown a preference for in the past.

Regardless of the perspective — service provider or end-user, recommendation systems have been immensely beneficial. And all this information explosion is the driving force behind Recommendation systems. Typical Recommendation systems cannot complete their work without sufficient data and plenty of users' data, as for the previous example, past purchases, items viewed, browsing history, users’ past ratings and feedback for the Recommendation systems to provide relevant and effective recommendations. In a nutshell, even the most advanced Recommenders cannot be effective without them.

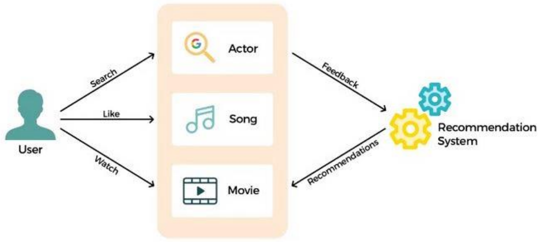


Figure 8 Recommendation Systems

3.2 Recommendation filtering techniques

Recommendation filtering techniques are used to filter the data to make them compatible with the standard recommendation system model which includes the three main concepts user-items-ratings.

In mathematical terms, a recommender system and its tasks are defined as follows:

- A set of users *U* including the target user *T*
- A set of items *I* that are to be filtered to make targeted and useful recommendations to *T*
- Learn a task to predict the likeliness of another item *I* to *T*, based on the *T*'s past preferences and interaction data

Several different techniques used for filtering by Recommendation Systems. This thesis' objective is to present a quick review of three of the most used of such techniques. Content-based Filtering, Collaborative Filtering and Hybrid filtering are the most well-known of such techniques too.

Focusing on the big picture, the choice of the appropriate filtering technique should be based on the type of data or information which will be filtered. Content-based filtering analyzes the characteristics of items and user preferences to make recommendations. It emphases on matching the features or attributes of items with the user's profile. On the other hand, collaborative filtering uses users’ behavior, preferences of similar users, and interactions in addition to item attributes. It assumes that users who have shown similar tastes in the past are likely to have similar preferences in the future.

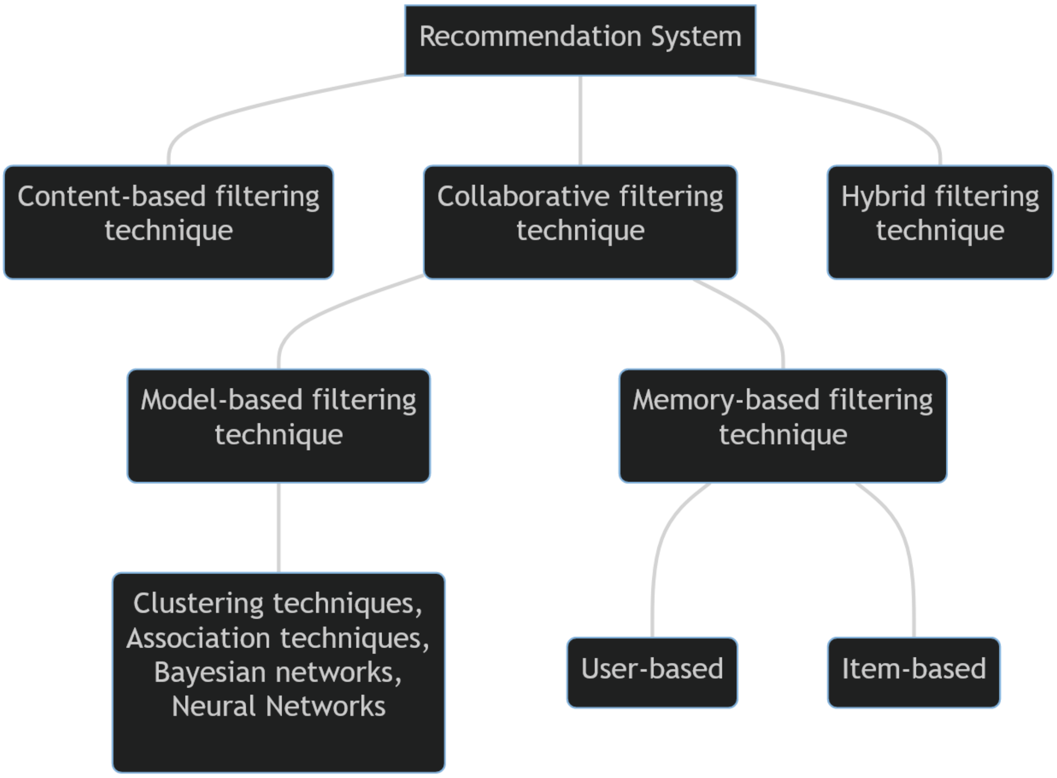

Recommendation System
Content-based filtering technique
Collaborative filtering technique
Hybrid filtering technique
Model-based filtering technique
Memory-based filtering technique
Clustering techniques, Association techniques, Bayesian networks, Neural Networks
User-based
Item-based

Figure 9 The main types of recommendation system techniques

3.2.1 Content-based filtering

Content-based filtering technique is a domain-dependent technique that emphasizes more on the analysis of the characteristics of an item in order to generate predictions. In other words, these algorithms try to predict and recommend items that are similar to those that a user liked in the past or is examining in the present. The objects of interest are defined by their associated features in a Content-based filtering system.

These techniques produce recommendations by learning the underlying model with either statistical analysis methods or machine learning approaches. Essentially, these methods use an item profile (for example a set of discrete characteristics, attributes and/or features) describing the item within the system.

To abstract the features of the items in the system, an item presentation algorithm is applied. Widely used algorithms are the tf–idf representation (also known as vector space representation) [22], Probabilistic models such as Bayesian Network Classifiers [23], Decision Trees [24] or Artificial Neural Networks [25] to model the relationship among different documents in an assortment of written texts.

In order to create a user's profile, these systems typically focus on two basic types of information:

- A model based on the user's preference.
- A history of the user's interaction with the recommender system.

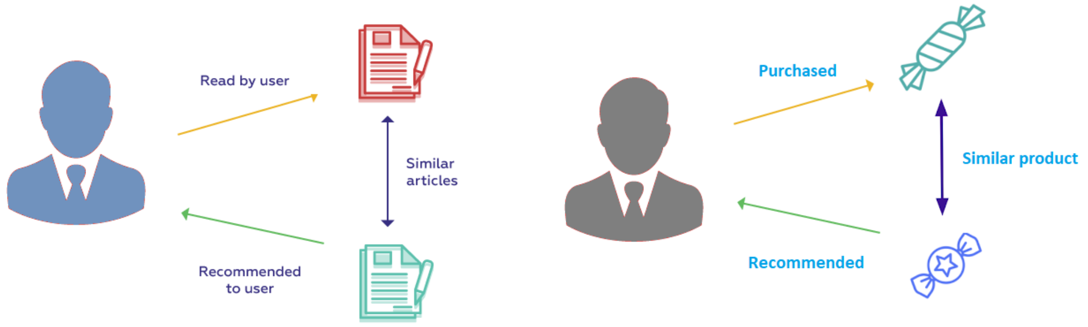


Figure 10 Content-Based Filtering technique

Once we know the preferences or the likings of the target user, the recommender system can predict and recommend similar items or products. Content-based filtering does not require other users' data and information during recommendations to the target user.

3.2.1.1 Pros, Cons and examples

Text recommendation systems like the newsgroup filtering system, for instance, use the words of their texts as features. [19] When documents such as web pages, publications and news are to be recommended, content-based filtering technique is the most effective. In this technique, recommendation is made based on the user profiles using features extracted from the content of the items the user has evaluated in the past. [20] [21]

Content-based filtering uses different types of models to find similarity between documents to generate meaningful and effective recommendations.

Items that are mostly related to the positively rated items are suggested to the user. A major disadvantage of this technique is the need to have a comprehensive and thorough knowledge and description of the features of the items in the profile.

In contrary to Collaborative filtering, the Content-based filtering is not so complex since only the analysis of the items that an individual has bought or seen is needed. In particular, can handle the cold-start problem, where new items have no or limited user data, by relying on item characteristics. This filtering technique does not need the profile of other users since they do not influence recommendation. In addition, if the user profile changes, it still has the potential to adjust its recommendations within a very short period.

Another major issue with content-based filtering is whether the system can learn user preferences from users' likes or actions regarding one content source and use them across another content type. When the system is limited to recommending content of the same type as the user is already using, the value from the recommendation system is significantly less than when other content types from other services can be recommended.

For example, recommending news articles based on browsing news is useful, but would be much more useful when music, movies, products, discussions etc. from different services can be recommended based on news browsing. To overcome this, most content-based recommender systems nowadays use some hybrid systems and algorithms.

3.2.2 Collaborative filtering techniques

Collaborative filtering explores techniques for matching individuals with similar interests and making recommendations on this basis. The inspiration behind collaborative filtering techniques comes from the idea that a person will often get better recommendations from someone with similar tastes and perceptions to themselves.

Collaborative filtering assumes that people who agreed in the past will agree in the future, and that they will like similar kinds of items as they liked in the past. The system generates recommendations using only information about rating profiles for different users or items. By locating peer users or items with a rating history like the current user or item, they generate recommendations using this neighborhood.

Collaborative filtering algorithms frequently need (1) the contribution of a target user, (2) a simple method to represent users' preferences/favorites, and (3) some algorithms that are capable of finding and matching people with similar interests.

In most cases, the standard procedure of a collaborative filtering system is the following:

- A target user expresses their preferences by rating items/products/interests of the system. These ratings can be considered as an approximate representation of the user's interest in the corresponding domain.
- The system matches this user's ratings against other users' and finds the people with the most "similar" tastes.
- With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user (presumably the absence of rating is often considered as the unfamiliarity of an item)

A key problem of collaborative filtering is how to combine and weigh the preferences of user neighbors. Sometimes, users can immediately rate the recommended items. As a result, the system gains an increasingly accurate representation of user preferences over time.

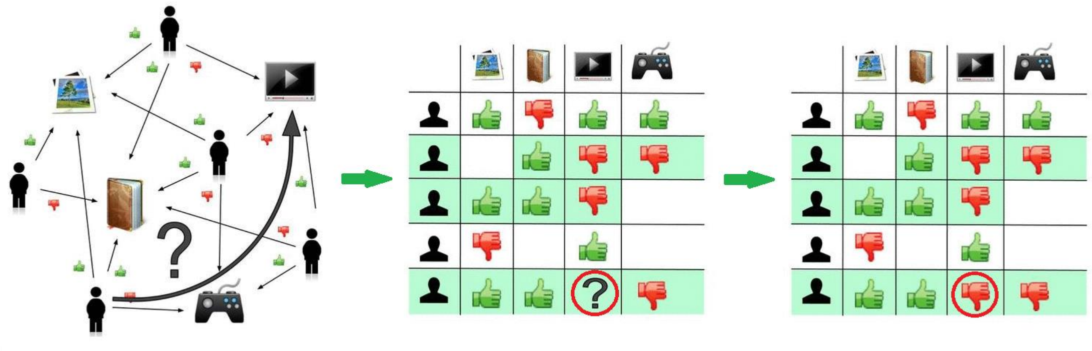


Figure 11 Collaborative Filtering technique based on a ratings system [26]

This figure gives us a simple example of predicting the user's rating using collaborative filtering. In the beginning, people rate different items (such as images, books, movies and games). After that, the system is making predictions about user's rating for another item that user has not rated yet. These predictions are calculated based on the existing ratings of other users, who have similar ratings to the target user. For instance, in this case the system has made a prediction that the target user won't like the movies.

3.2.2.1 Model-based filtering

In the Model-based filtering approaches, models are developed using various data mining or machine learning algorithms to make predictions of target user’s rating about unrated items. There are many model-based collaborative filtering algorithms, for instance, Bayesian networks, clustering models, latent semantic models such as singular value decomposition, probabilistic latent semantic analysis, multiple multiplicative factors, latent Dirichlet allocation and Markov decision process-based models. [27]

This approach potentially offers its benefits in both speed and scaling as the model can be created prior to the recommendation process, so it has a higher performance than the memory-based approach. Another advantage is that it can give an intuitive explanation of the recommendations it makes, so it’s easier for users to trust the system and accept the suggestions made to them.

Moreover, depending on the learning technique used to create the model, this approach can lead to higher prediction accuracy while reducing the sparsity problem, making it easy to apply to applications that consist of large data sets. This feature is, at the same time, a major drawback for this type of filtering, since the model, in order to be accurate, a large amount of data should be available.

Another major disadvantage of the model-based approach is that prediction results do not automatically adapt to data changes. Instead of that, the model has to be redesigned to reflect the updated data and information, and this process can be extremely costly.

3.2.2.2 Memory-based filtering

Memory-based filtering techniques consist of two basic methods, the user-based and the item-based. In user-based, if similar users with similar ratings for the same or similar items are found, then the target user’s rating for another item that the user has never interacted with can be predicted.

Collaborative filtering systems have several different methods, but systems commonly can be reduced to two steps:

- Look for users who share the same rating patterns with the target user whom the prediction is for.
- Use the ratings from all those like-minded users found in step one to calculate and predict the target user’s rating

For example, two users, A and B, have given similar ratings to some movies:

• A = Terminator:	7,	Predator:	6,	Robocop:	5
• B = Terminator:	7,	Predator:	6,	Robocop:	?

In this case, predicting the rating is a very easy process. This falls under the category of user-based collaborative filtering. A specific application of this is the user-based Nearest Neighbor algorithm.

On the other hand, item-based collaborative filtering, which was invented by Amazon.com in 1998 [28] and published in an academic conference in 2001 [29], is a technique of collaborative filtering for recommender systems based on the similarity between items calculated using people's ratings of those items. To put it simply, users who bought an x-item also bought a y-item. Item-based approaches proceed in an item-centric manner:

- Build an item-item matrix determining relationships between pairs of items
- Infer the tastes of the current user by examining the matrix and matching that user's data

3.2.2.3 Pros, Cons and examples

Collaborative filtering can offer unexpected recommendations by identifying patterns and similarities among users. CF is well-suited for handling large-scale datasets. It can efficiently process and leverage user-item interaction data to generate recommendations. As the amount of data grows, collaborative filtering models can still provide meaningful recommendations without significant performance degradation. But as CF relies on user data to generate recommendations, in certain cases, this data may contain sensitive or private information. Ensuring proper data anonymization and protection is crucial to maintain user privacy and address potential data spillover issues.

Collaborative filtering also can address the cold-start problem, which occurs when there is insufficient or no data available for new items or users. By leveraging the behavior of existing users, Collaborative filtering can make reasonable recommendations for new users or items without requiring extensive item information. On the other hand, it still faces challenges when dealing with new users who have not provided sufficient data for effective recommendation. Without historical user preferences, it can be difficult to find similar users or make accurate suggestions. CF struggles as well with the cold-start problem with new items that have no or limited user interaction data. It may take time for new items to accumulate enough data for accurate recommendations, potentially delaying their visibility and adoption.

3.2.3 Hybrid filtering

Hybrid approaches in recommendation systems combine content-based filtering and collaborative filtering methods to leverage the strengths of both techniques and improve recommendation accuracy and diversity.

Some very common hybrid approaches are the following:

1) Weighted Hybrid: In this approach, recommendations from both content-based and collaborative filtering models are combined using weights. In order to cover the drawbacks of each approach with the advantages of other approaches, both approaches can be combined with an approach known as hybrid technique. Every single model generates a set of recommendations, and the final list is formed by merging the results. The weights can be decided based on several factors, such as the reliability or the historical performance of each approach. But the results, in some cases, show that this technique may not really boost the performance up, but it helps to give prediction score for unrated movies that are impossible to be recommended by only using collaborative filtering. [30]

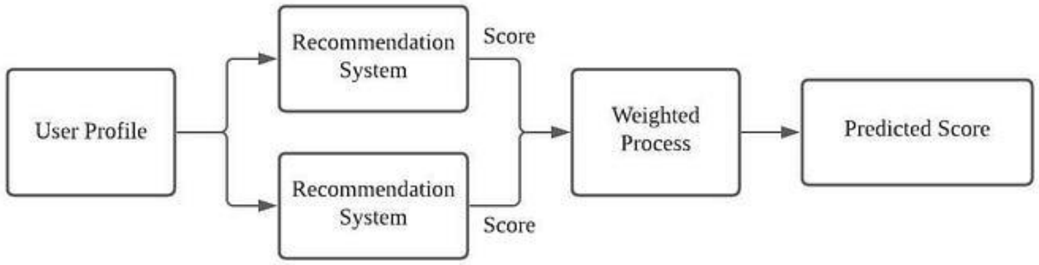


Figure 12 Weighted Hybrid Recommendation System

2) Switching Hybrid: Instead of combining the recommendations, the switching hybrid approach decides on either the content-based or collaborative filtering method based on certain conditions and / or user characteristics. For instance, when there is sufficient and accurate user data available, collaborative filtering might be used. However, if the user is new or the data is sparse, the system might switch to content-based filtering. The decision of which method to use can be based on rules, machine learning models, or hybrid learning algorithms.

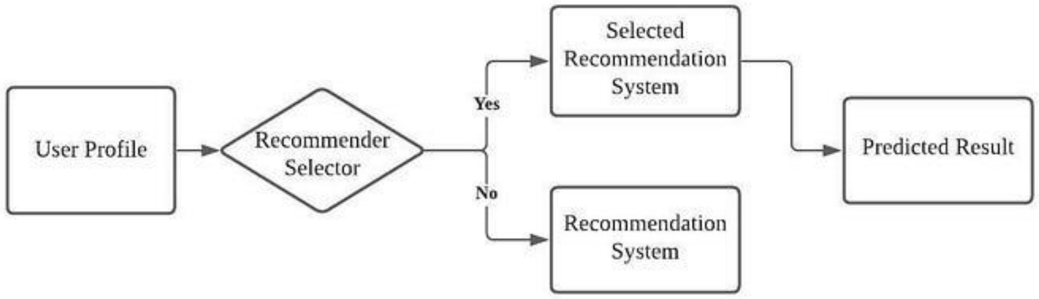


Figure 13 Switching Hybrid Recommendation System

3) Cascade Hybrid: The cascade hybrid approach defines a strict hierarchical structure recommendation system using one filtering method to pre-filter the items and then applying the other method to refine the recommendations. For example, content-based filtering can be used initially as primary RS to narrow down the item pool based on user preferences or item attributes. The remaining items are then passed to a secondary collaborative filtering method for further personalized recommendations. This approach can benefit from the precision of content-based filtering and the diversity of collaborative filtering.

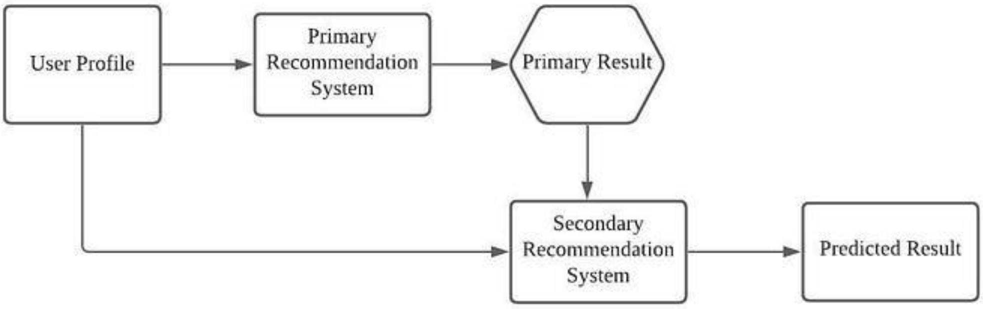

User Profile
Primary Recommendation System
Primary Result
Secondary Recommendation System
Predicted Result

Figure 14 Cascade Hybrid Recommendation System

4) Feature Combination Hybrid: This approach combines the features used in both content-based and collaborative filtering methods. The content-based filtering model considers item features, while the collaborative filtering model takes into account user-item interactions. The features from both methods are merged to create a unified feature representation, which is then used to generate recommendations. This approach can capture both the fundamental characteristics of items and the collaborative behavior of users.

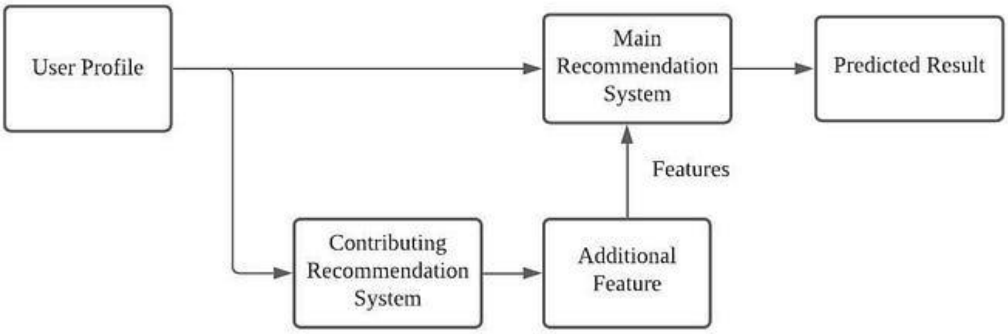

User Profile
Main Recommendation System
Predicted Result
Features
Contributing Recommendation System
Additional Feature

Figure 15 Feature Combination Hybrid Recommendation System

Hybrid approaches aim to overcome the limitations and constraints of individual filtering methods and provide more accurate, diverse, and personalized recommendations. The choice of a specific hybrid approach depends on the characteristics of the dataset, domain, available resources, and desired recommendation objectives.

4. MEASURING INFLUENCE IN SOCIAL NETWORKS

Today there is an abundance of information and rich social activity of many people through various social networks. Social networking services (SNS, also known as social networks or social media) are online platforms which people use to build social networks or with other people who share similar personal or career interests, activities, backgrounds or real-life connections. [32]

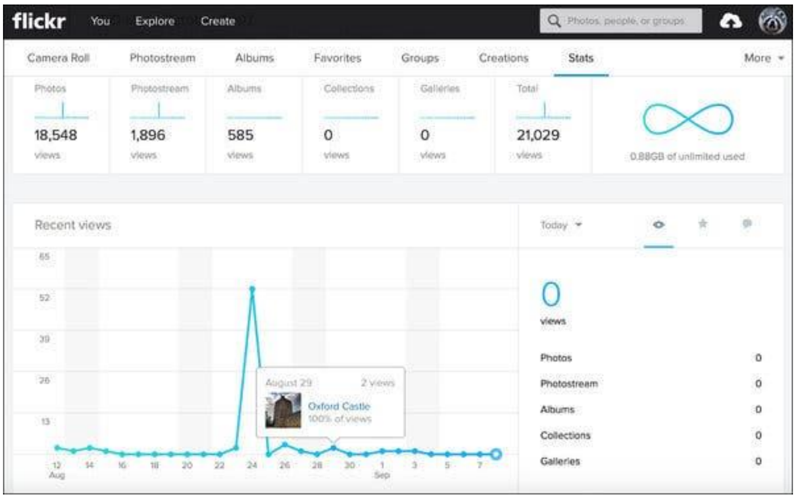
Social networking services vary in format and the number of features. However, most incorporate common features: [33] [34]

- Social networking services are Internet-based applications
- User-generated content (UGC), such as user-submitted digital photos, text posts, tags, online comments, likes and many more, is the core engine of social networking services
- Users create service-specific profiles for the site or app that are designed and maintained by the SNS organization
- Social networking services facilitate the development of online social networks by connecting a user's profile with those of other individuals or groups

All this online shared information creates opportunities for various kinds of exploitation in order to explore the following questions:

- Which is the trend, in terms of popular posted items, for a specific topic?
- Which members of social networks are experts or influential regarding a specific topic?

Measuring influence in social networks is a complex task that attempts to measure and quantify the impact an individual or entity has within a given social network. Various approaches and metrics have been developed to measure influence, each with its own strengths and limitations. It's important to note that measuring influence in social networks – most of the time - is not objective, and different metrics can yield different results. The interpretation of influence may vary depending on the goals and context of the analysis. Therefore, a combination of multiple metrics and qualitative assessments is often used to gain a more comprehensive understanding of influence within social networks.



Photos	Photostream	Albums	Collections	Galleries	Total
18,548 views	1,896 views	585 views	0 views	0 views	21,029 views

0.88GB of unlimited used

August 29 2 views
Oxford Castle
100% of views

0
views

Photos	0
Photostream	0
Albums	0
Collections	0
Galleries	0

Figure 16 Flickr Pro stats page

4.1 Influence of users and photos regarding a specific topic in Flickr

Our approach towards the above questions is based on using the tag functionality of social networks and the assumption that a topic is represented by one or more tags. Tagging allows you to structure your digital content by themes/categories/locations/trends etc., make them searchable, and give you a better interaction with other users. The social network we use is Flickr. Flickr is an image and short-video hosting service, and it is mainly popular with amateur and professional photographers.

The steps included in our approach are the following:

Step 1	Collect the appropriate data, photos and users, which have used a specific tag t_i
Step 2	Find the most similar tags related to tag $t_{i,}$ thus resulting in a new tag set which represents a topic *T*
Step 3	Collect the appropriate data, photos and users, which have used at least one of the tags that represent the topic
Step 4	Find the most influential users regarding the topic
Step 5	Find the most influential photos regarding the topic

More specifically, given a specific tag t_i, we collect all necessary information – photos, users, comments, favs (likes) etc. in the first step and then we estimate the most similar tags based on collaborative filtering techniques which take into consideration the level of usage of tags by users and the usage of common tags along with two tags in the second step. We define topic *T* as the set which contains all of them, along with t_i.

We then collect information regarding the specific topic, in the third step of our process, and get the following result sets:

- The set P_T of all photographs having at least one tag that belongs to *T*.
- The set *U* of all users which have posted and/or involved in at least one photo that belongs to *P*.
- Supplementary data sets such as tags *T*, comments *C* and favs *F* having involved in at least one photo that belongs to *P***.**

Finally, in the fourth and fifth steps, we use the above sets to find the most influential users belonging to *U* regarding their photos belonging to *P* and the most influential photos belonging to *P*, based on social activity-related criteria.

4.1.1 Collection of data regarding a specific tag

Collecting data through an Application Programming Interface (API) has become a popular method for obtaining structured and real-time data from various sources. APIs are sets of rules and protocols that allow different software applications to communicate with each other and exchange data. We first choose a specific tag t_i as the source tag of our research, and then collect the necessary datasets through the relevant Flickr APIs.

Collecting data through Flickr APIs and combining them, enables us to integrate valuable information into our system to make effective decisions. Maintaining data integrity between Flickr and our system is an important part of the process. The main reason is to support the observation of errors in the data collection process and as a result in the calculation process which follows in the next steps.

4.1.2 Find the most similar tags

In order to proceed with the second step, we use two criteria for accessing similarity:

a) The common tags that two distinct tags have been used with. If two tags t_i and t_j, have been used the same number times together with another tag t, this tag is related to the same level as the two tags t_i and t_j.

b) The level of their usage by users who have used them in common. If two users have used the two tags with similar frequency this means that they consider them as similar or of the same level of interest.

Specifically, we use (1) and (2) to estimate the Euclidean distance of two tags. We thus define two similarity measures for tags according to the two criteria and combine them in one.

$$sim_{euclidean}(t_i, t_j)_{comtags} = 1/(1 + \sqrt{\sum_{c=1}^{T} \left(r_{t_{i,l}} - r_{t_{j,l}}\right)^2}) \quad (1)$$

Where:

- t_i, t_j are two distinct tags,
- T is the set of <u>common tags</u> that the two tags have been used with t_i, t_j,
- $r_{t_{i,l}}, r_{t_{j,l}}$ are the numbers of photos which have used the tag c and have also used the tags t_i and t_j respectively, and
- $sim_{euclidean}(t_i, t_j)_{comtags}$ is the similarity of tags t_i, t_j regarding their common usage with other tags

$$sim_{euclidean}(t_i, t_j)_{user} = 1/(1 + \sqrt{\sum_{u=1}^{U} \left(r_{t_{i,u}} - r_{t_{j,u}}\right)^2}) \qquad (2)$$

Where:

- t_i, t_j are two distinct tags,
- U is the set of the users which have used the two tags in their photos,
- $r_{t_{i,u}}, r_{t_{j,u}}$ are the numbers of photos of user u which have used the tags t_i and t_j respectively, and
- $sim_{euclidean}(t_i, t_j)_{user}$ is the similarity of tags t_i, t_j regarding their common usage by users

We can also estimate the similarity between two tags according to the two criteria using the Cosine similarity measure.

$$Cosine\ Similarity(t_i, t_j) = \frac{\sum_{k=1}^{n}(r_{i,k}) * (r_{j,k})}{\sqrt{\left[\sum_{k=1}^{n}(r_{i,k})^2\right]} * \sqrt{\left[\sum_{k=1}^{n}(r_{j,k})^2\right]}} \qquad (3)$$

We will either use one of the above similarity measures (Cosine or Euclidean distance-based similarity) or will use average measures to define final similarity metrics for user-based similarity and common tags-based similarity.

Then, we will combine the two similarity measures to estimate the similarity between two tags.

$$WMS\ (t_i, t_j) = w_a * sim(t_i, t_j)_{comtags} + w_b * sim(t_i, t_j)_{user} \quad (4)$$

Where:

- w_a, w_b are the weights we use for the two kinds of similarity and $\mathrm{w}_a + \mathrm{w}_b = 1$

4.1.3 Collection of data regarding a specific topic

Before we calculate and present Influence metrics about users and the topic, we have to repeat the data collection process for each one of the selected common tags, and their related datasets, one at a time. This process allows us to create a wider set of data related to the specific topic of our research.

4.1.4 Find the most influential users regarding a specific topic

For each one of the users belonging to *U* we estimate an influence score regarding any (Top-N similar) tag of the tag set *T*, which is representative of a topic, based on weighted number of views / comments / favs of related photos and the weighted number of user's followers. This data used to represent a variety of criteria which we consider as important for determining influence. These criteria are presented in the rest of this section, along with the related metrics - formulas used for the estimation of a user's influence regarding a topic.

Estimating a User's Influence on a Topic

The criteria for estimating a user's influence regarding a specific tag are described below, together with the metrics which represent them.

Interest: The total number of views a user's photos - which has been assigned at least one tag t_i of a topic *T* - have got shows the interest of other users for the user's photos. We are interested in the interest level of u_i's photos containing t_i, compared to the general level of interest that photos representing *T* attracted. We thus use the following interest metric:

$Int(u_i, T)$ = the ratio of the number of views of photos of a specific user u_i containing at least one tag t_i of a topic *T*, to the total number of views of photos which represent *T*. This metric shows the relative interest of users to view u_i's photos compared to the total amount of view-interest that related photos generate.

$$Int(u_i, t_i) = \frac{number\ of\ views\ of\ u_i^{\prime}s\ photos\ which\ contain\ t_i}{number\ of\ views\ of\ all\ photos\ containing\ t_i} \quad (5)$$

Involvement: The total number of comments a user's photos - which have been assigned at least one tag t_i of a topic *T* - have got shows the level of involvement of other users for the user's photos. We are interested in the involvement level of u_i's photos containing t_i, compared to the general level of involvement of users with photos representing *T*. We thus use the following involvement metric:

$Inv(u_i, t_i)$ = the ratio of the number of comments of the photos of a specific user u_i's containing at least one tag t_i of a topic *T*, to the total number of comments which represent *T*. This metric shows the relative involvement of users with u_i's photos which contain t_i, compared to the total amount of involvement of users with photos used with.

$$Inv(u_i, t_i) = \frac{number\ of\ comments\ of\ u_i' s\ photos\ which\ contain\ t_i}{number\ of\ comments\ of\ all\ photos\ containing\ t_i} \qquad (6)$$

Preference: A user's influence can be measured by the number of her followers; the more friends a user has got, the more she is trusted / preferred. This metric shows the relative influence of users with u_i's photos which contain t_i, compared to the total influence of users with photos representing *T*.

$$P(u_i, t_i) = \frac{number\ of\ followers\ of\ u_i\ with\ photos\ containing\ t_i}{number\ of\ followers\ of\ all\ users\ with\ photos\ containing\ t_i} (7)$$

Endorsement: In today's social networks every post of a user can be endorsed by other users, e.g. by declaring they like it. The more users endorse a post, the more influence this post has over users. This gives us an insight into how valuable the user's opinion on some topics is. We are interested in the value of the user's opinion on a topic, in relation to the value of other users' opinions on that topic. For the endorsement metric we have used the following formula:

$E(u_i, t_i)$ = the ratio of the number of favorites that u_i's photos contain at least one tag t_i of a topic *T* have been assigned to the total number of favorites assigned to photos which represent *T*. This metric shows the relative endorsement in user's photos compared to the total endorsement for photos containing t_i.

$$E(u_i, t_i) = \frac{number\ of\ favorites\ of\ u_i i' s\ photos\ containing\ t_i}{number\ of\ favorites\ of\ all\ photos\ containing\ t_i} \qquad (8)$$

Finally, we are using an equally weighted measure to display the influence score of a user u_i concerning a topic *T*, as a combination of the result scores of the above four influence factors:

$$UserInf(u_i, t_j) = w_{Int} * Int(u_i, t_j) + w_{Inv} * Inv(u_i, t_j) + w_P * P(u_i, t_i) + w_E * E(u_i, t_j) \quad (9)$$

Where:

- $w_{Int}, w_{Inv}, w_P, w_E$ are the weights we assign to the factors described above,
- $w_{Int} + w_{Inv} + \ w_P + w_E = 1$, and
- $w_{Int} = \ w_{Inv} = \ w_P = w_E = \ \frac{1}{4}$

4.1.5 Find the most influential photos regarding a specific topic

For the tags belonging to the set T of similar tags, we find the photos which have used at least one of them (set $\mathrm{P_T}$). For each photo *p* belonging to the set P_T (which uses one or more tags of the set T of similar tags), we estimate its influence based on the number of views, comments and favs of the photo related to the total number of views, comments and favs respectively, of all photos belonging to P_T. We extract thus a list with recommended photos for a specific topic, i.e. the ones with the highest influence score. We use the formula below for estimating the influence of a photo comparatively to all photos related to a specific topic.

$$Inf(p) = w_{views} * \left(\frac{NoViews(p)}{P_T\ NoViews}\right) + w_{comments} * \left(\frac{NoComments(p)}{P_T\ NoComments}\right) + w_{favs} * \left(\frac{NoFavs(p)}{P_T NoFavs}\right) \quad (10)$$

Where:

- $No\ Views\ (p)$ is the number of views of photo *p*,
- $NoComments(p)$ is the number of comments of photo *p*,
- $NoFavs(p)$ is the number of favs on photo *p*,
- $P_T\ NoViews$ is the number of the views of all photos belonging to P_T,
- $P_T\ NoComments$ is the number of the comments of all photos belonging to P_T,
- $P_T\ NoFavs$ is the number of the favs of all photos belonging to P_T,
- w_{views}, $w_{comments}$, w_{favs} are the weights of the numbers of views, favs and comments respectively, related to the total number of topic related photos. We consider these photos as the most influential ones for a specific topic, since they are the ones which are mapped to the most related tags to the topic and are attributed to the highest social activity,
- Furthermore, $w_{views} + w_{comments} + w_{favs} = 1$, and
- $w_{views} = w_{comments} = w_{favs} = \frac{1}{3}$

5. WHAT IS FLICKR?

Flickr is an online photo-sharing platform and social media website. It was created in 2004 by Ludicorp, which was later acquired by Yahoo in 2005. Flickr constitutes a popular way for amateur and professional photographers to upload, host and share high-resolution photographs and short videos. One of the early versions of Flickr was a chat room named FlickrLive that lets users exchange photos in real-time. It was subsequently dropped as Flickr evolved away from the code base of Game Neverending. [36]

As of March 20, 2013, Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily, as reported by The Verge. [37] Over the years, Flickr has gone through several changes and ownership transitions. In 2018, it was acquired by SmugMug, a photography-focused company, which has since made efforts to revitalize the platform and cater to the needs of photographers and enthusiasts. The total number of public photos uploaded until Jan. 2018 was 6.47 billion as measured in early 2018, and it has decreased to 2.38 billion as measured in July 2019, due to a new account policy and / or some consecutive PRO subscription price changes. [38]

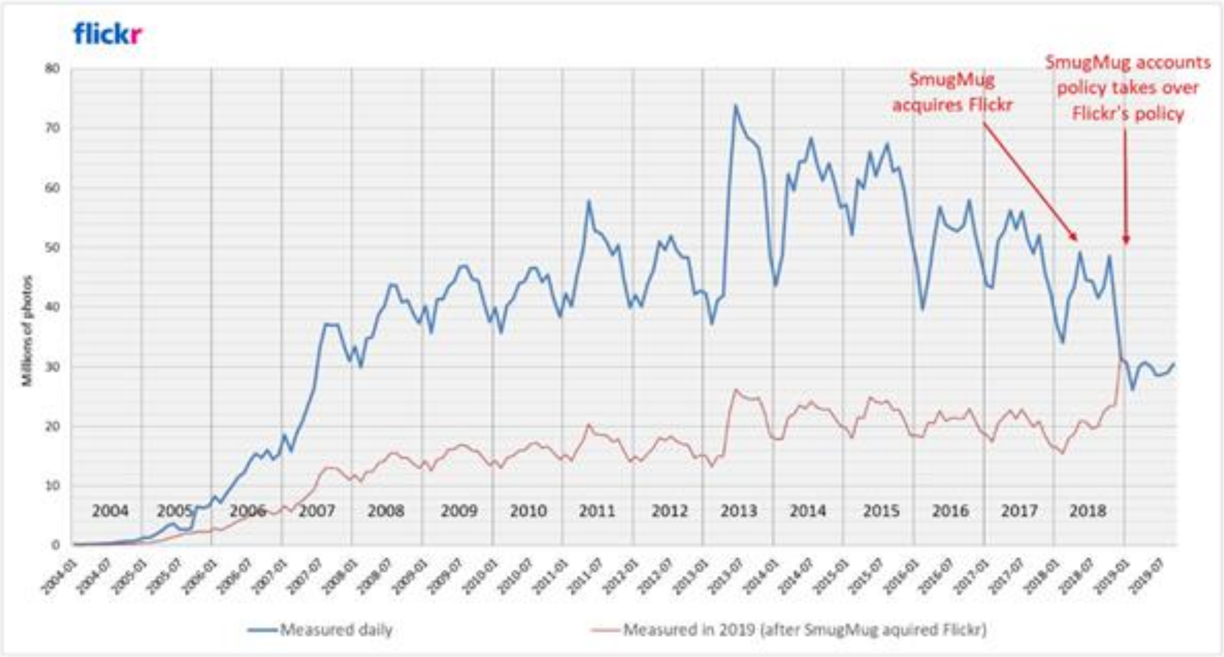
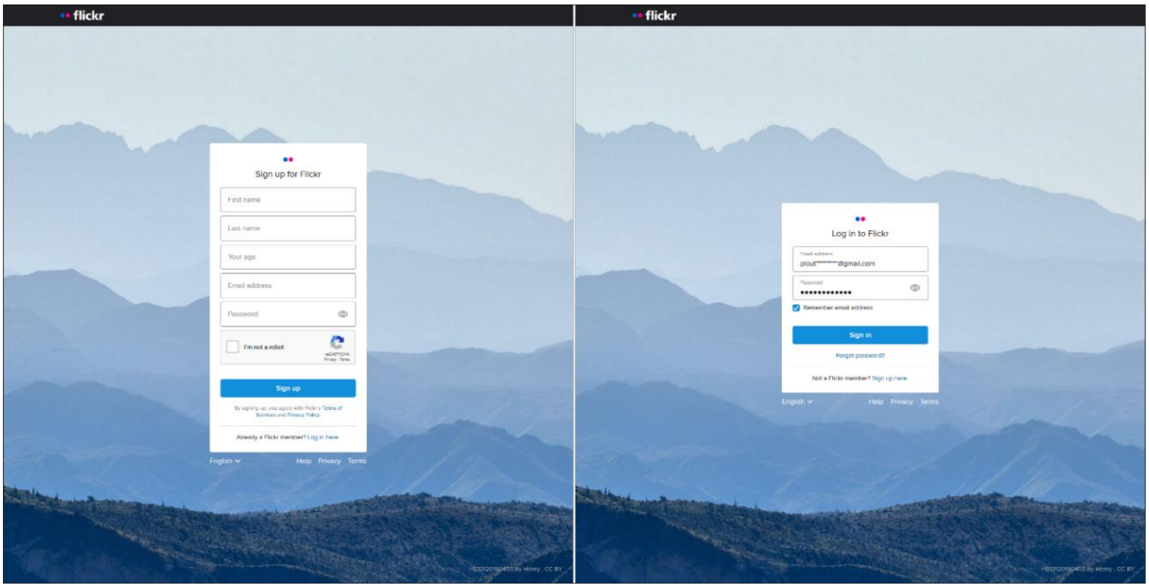


Figure 17 Millions of public photos uploaded per month [39]

Despite the criticisms, many people continue to use the service because of its seamless functionalities. Photos and videos can be accessed from Flickr without the need to register an account, but an account must be made to upload content to the site. Registering an account also allows users to create a profile page containing photos and videos that the user has uploaded and grants the ability to add another Flickr user as a contact (follower).

5.1 The App Garden

The App Garden [40] is a place where developers can showcase the applications they've created, and where you can find new ways to explore Flickr using home grown applications created by other Flickr members. If you want to create a new application, it is necessary to login in your Flickr account, navigate to The App Garden and click on Create an App menu item.



By signing up, you agree with Flickr's Terms of Services and Privacy Policy.

Already a Flickr member? Log in here

Email address: plout*********@gmail.com

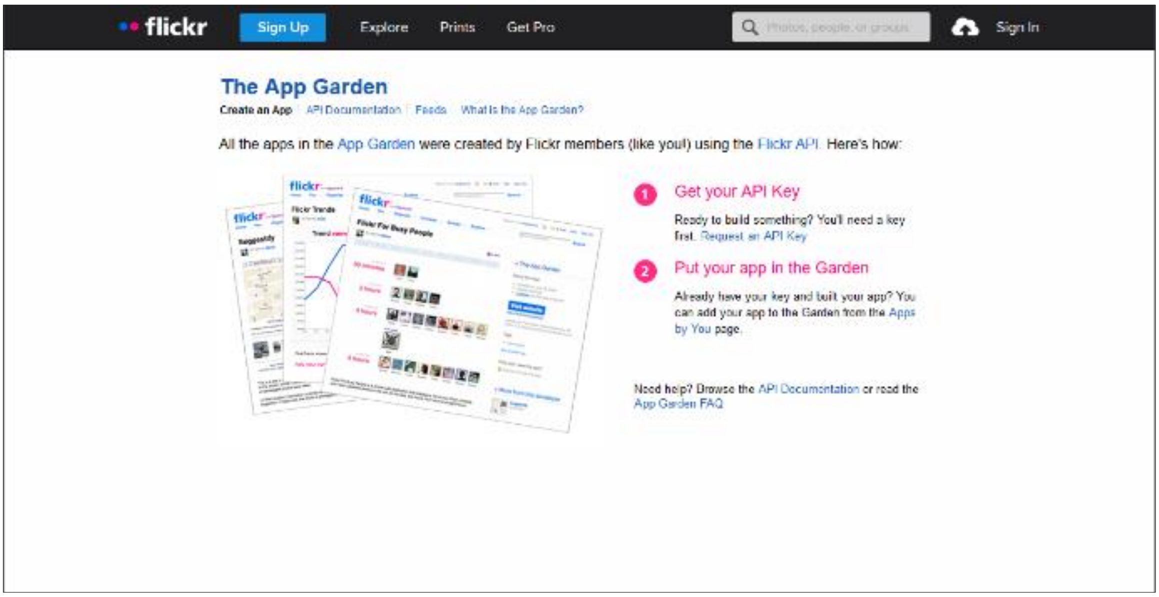
Password: ●●●●●●●●●●●●●

☑ Remember email address

Not a Flickr member? Sign up here

Figure 18 Sign up and login, so you can access Flickr API and The App Garden

First of all, you'll need an API key for your application. Some APIs use API keys for authorization. An API key is a token that a client provides when making API calls and it is supposed to be a secret that only the client and server know.



The App Garden

All the apps in the App Garden were created by Flickr members (like you!) using the Flickr API. Here's how:

1 Get your API Key

Ready to build something? You'll need a key first. Request an API Key

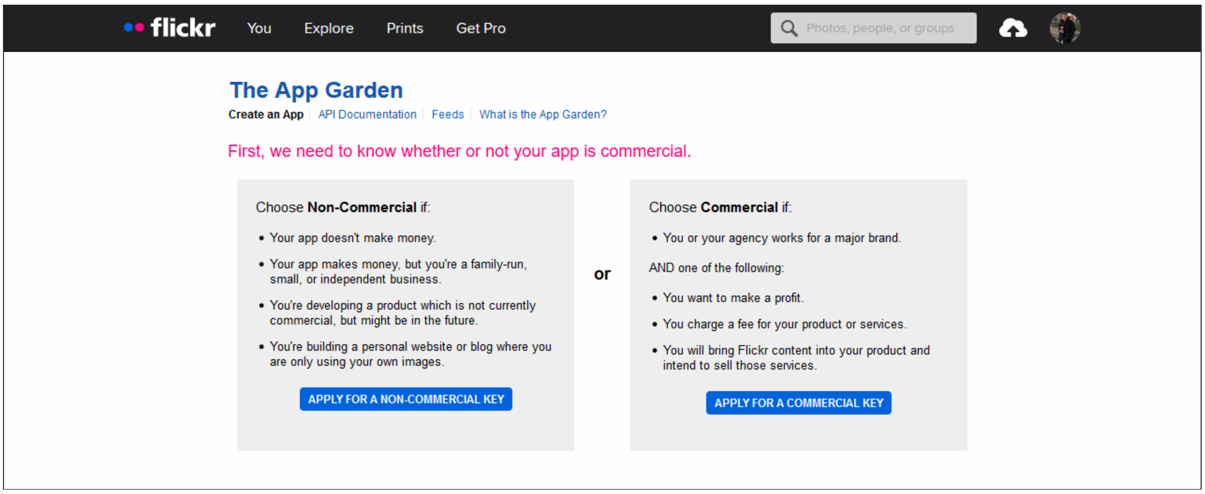
2 Put your app in the Garden

Already have your key and built your app? You can add your app to the Garden from the Apps by You page.

Need help? Browse the API Documentation or read the App Garden FAQ

Figure 19 The App Garden

When you request a new API Key, Flickr needs to know whether your app is commercial. The Flickr API is available free of charge for non-commercial use by developers and stakeholders of all kinds. Nevertheless, commercial use is possible by prior arrangement. In our case, we applied for a non-commercial key since our application does not make money and use only public photographs for research purposes.



The App Garden

Create an App | API Documentation | Feeds | What is the App Garden?

First, we need to know whether or not your app is commercial.

Choose **Non-Commercial** if:

- Your app doesn't make money.
- Your app makes money, but you're a family-run, small, or independent business.
- You're developing a product which is not currently commercial, but might be in the future.
- You're building a personal website or blog where you are only using your own images.

APPLY FOR A NON-COMMERCIAL KEY

or

Choose **Commercial** if:

- You or your agency works for a major brand.

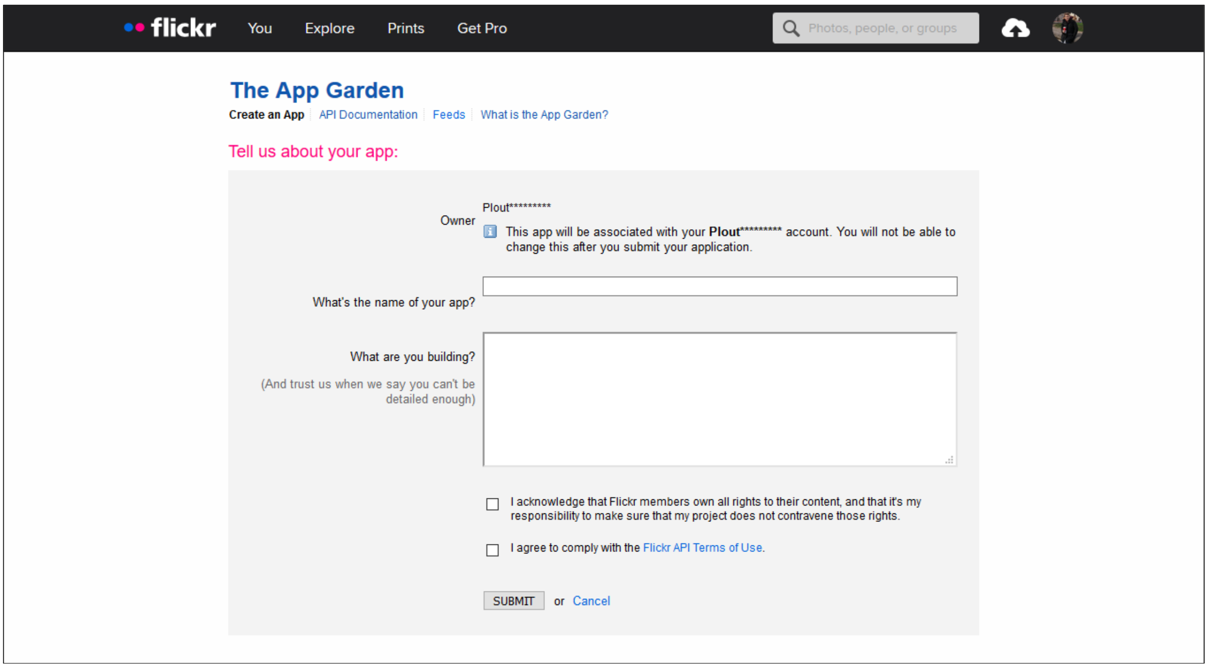
AND one of the following:

- You want to make a profit.
- You charge a fee for your product or services.
- You will bring Flickr content into your product and intend to sell those services.

APPLY FOR A COMMERCIAL KEY

Figure 20 Request an API Key

To continue with your application, you must choose a name for your app, complete a brief description and agree with the Flickr API Terms of Use.



The App Garden

Create an App | API Documentation | Feeds | What is the App Garden?

Tell us about your app:

Owner: Plout*********

This app will be associated with your **Plout*********** account. You will not be able to change this after you submit your application.

What's the name of your app?

What are you building?

(And trust us when we say you can't be detailed enough)

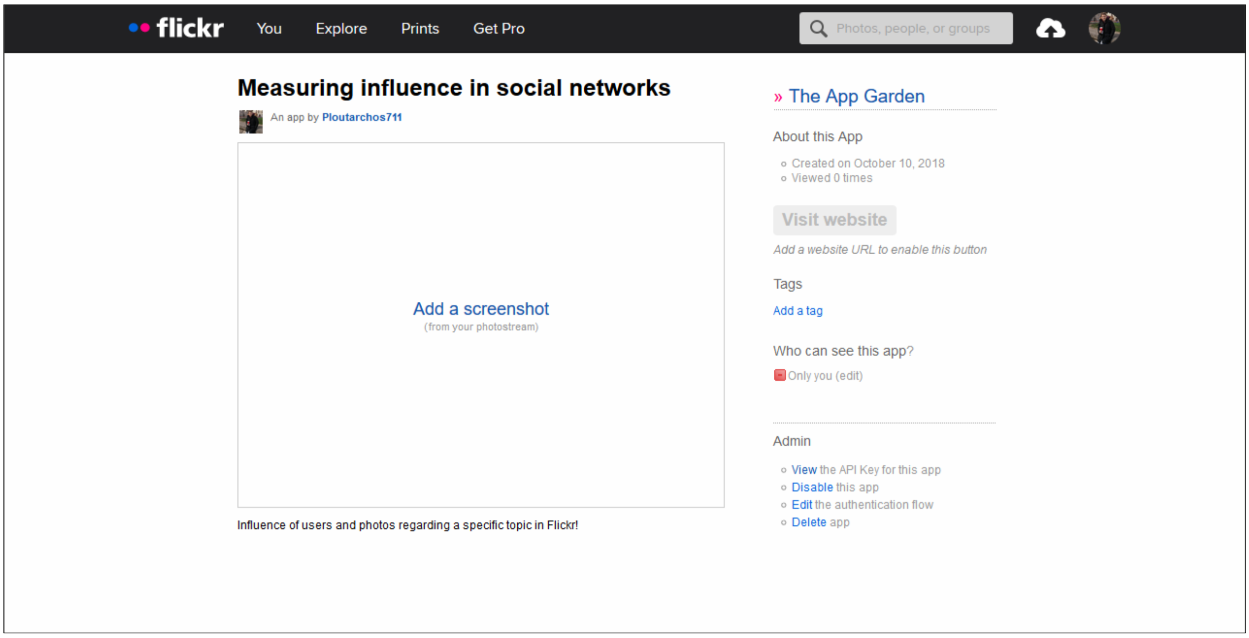
□ I acknowledge that Flickr members own all rights to their content, and that it's my responsibility to make sure that my project does not contravene those rights.

□ I agree to comply with the Flickr API Terms of Use.

SUBMIT or Cancel

Figure 21 Submit your App

When you receive the key ID and secret, you can start building your application using a large list of available API methods. Each API key has an app page that is private by default.



Measuring influence in social networks

An app by Ploutarchos711

Add a screenshot
(from your photostream)

Influence of users and photos regarding a specific topic in Flickr!

» The App Garden

About this App

- Created on October 10, 2018
- Viewed 0 times

Visit website

Add a website URL to enable this button

Tags

Add a tag

Who can see this app?

Only you (edit)

Admin

- View the API Key for this app
- Disable this app
- Edit the authentication flow
- Delete app

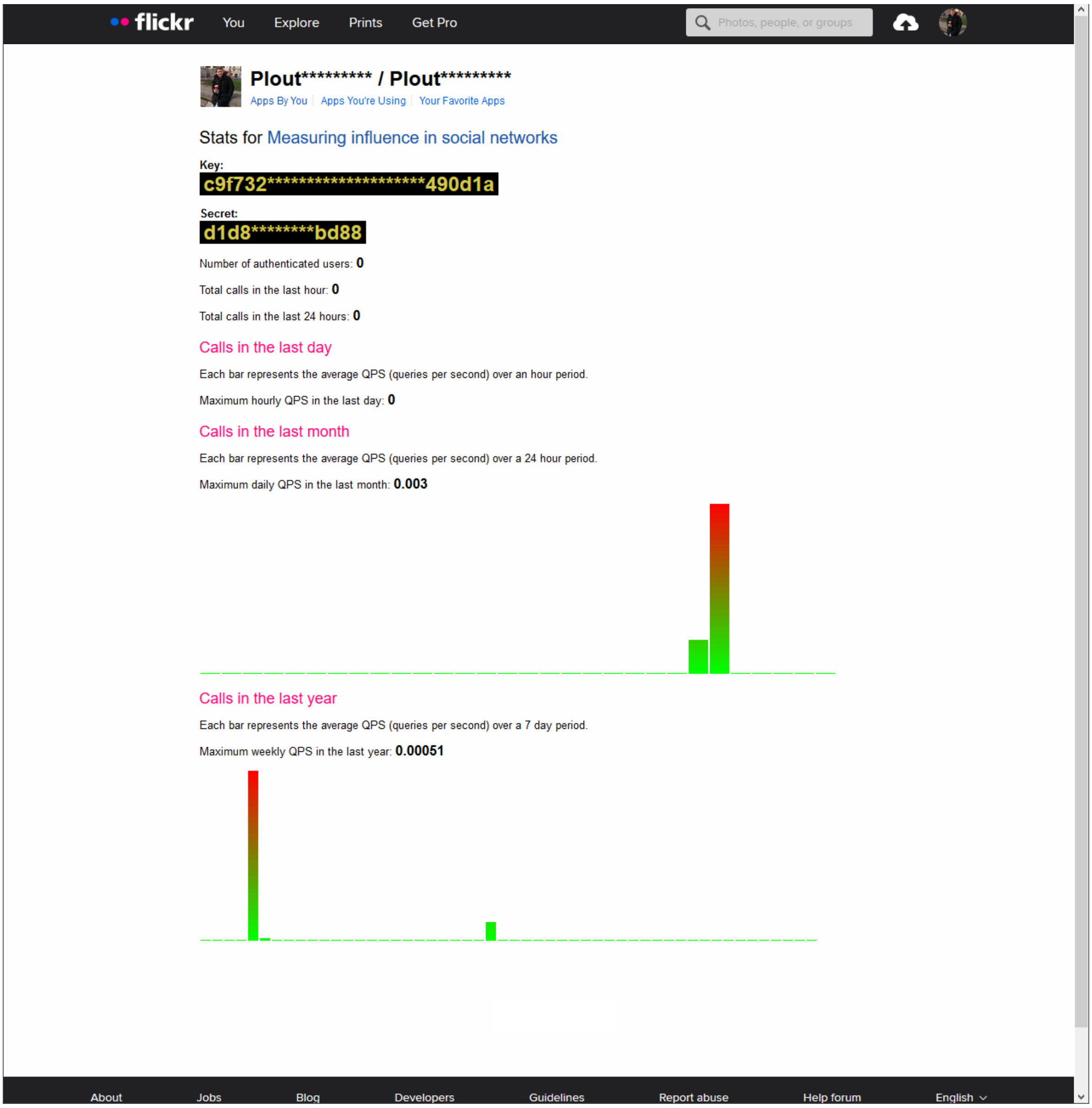
Figure 22 'Measuring influence in social networks' app page

There is only one application page for each API key, and it is not feasible to create one without a key. If you want to create multiple applications, you'd need to use a separate key to showcase each on its own app page. When you're ready to showcase your app, fill out the application page to let people know what it's about.

Many of Flickr's API methods require the user to be signed in. In the past Flickr used its own authentication API, but now, users should only be authenticated using the OAuth specification which is the industry standard. By using the OAuth standard, developers will provide in their applications a secure way for people to sign-in into their Flickr accounts with all the different account types that Flickr is supporting (Yahoo! ID, Google ID, Facebook). Flickr's OAuth flows work for web-applications, desktop apps and mobile applications as well.

However, as we said before, our application is a simple third-party application which does not make money and uses only public photographs for research purposes. In this case, users' consent and the implementation of an OAuth flow are not mandatory.

Finally, on your application page, you can see some typical API key statistics e.g., total calls in the last hour, total calls in the last 24 hours, number of authenticated users etc. Although they are quite limited, it is particularly important to monitor your application usage, due to the strict constraint that occurs in the number of the requests - 3600 queries per hour.



Plout********* / Plout*********

Apps By You | Apps You're Using | Your Favorite Apps

Stats for Measuring influence in social networks

Key:
c9f732********************490d1a

Secret:
d1d8********bd88

Number of authenticated users: **0**

Total calls in the last hour: **0**

Total calls in the last 24 hours: **0**

Calls in the last day

Each bar represents the average QPS (queries per second) over an hour period.

Maximum hourly QPS in the last day: **0**

Calls in the last month

Each bar represents the average QPS (queries per second) over a 24 hour period.

Maximum daily QPS in the last month: **0.003**

Calls in the last year

Each bar represents the average QPS (queries per second) over a 7 day period.

Maximum weekly QPS in the last year: **0.00051**

Figure 23 Api Key statistics

If your API Key and, therefore, your application has been disabled, it's likely that it was in violation of this constraint.

5.2 Flickr APIs

Flickr has an open Application Programming Interface (API). An API, practically speaking, simplifies programming by abstracting the underlying implementation and only exposing objects or actions the developer needs. It defines interactions between multiple software intermediaries and the kinds of calls or requests that can be made, how to make them, the data formats that should be used, the conventions to follow, etc.

Flickr API allows developers to build applications and services that interact with the Flickr platform. This has led to the integration of Flickr with various third-party tools, websites, and applications, further expanding its reach and functionality.

With billions of photos available, many of them along with valuable metadata such as tags, geolocation, EXIF data and many more, the Flickr API is a powerful tool for users and developers. Anyone can write their own program to access and remix Flickr data (like photos, video, tags, profiles or groups) in a totally new or different way.

In fact, the Flickr API is how you can access that data. Almost all the functionality that runs flickr.com is available through the API and is completely free to use as a service to all Flickr members as well as developers and other integrators, so they can create even more ways to interact with photos beyond flickr.com.

Here is a concise diagram of Flickr's API interactions:

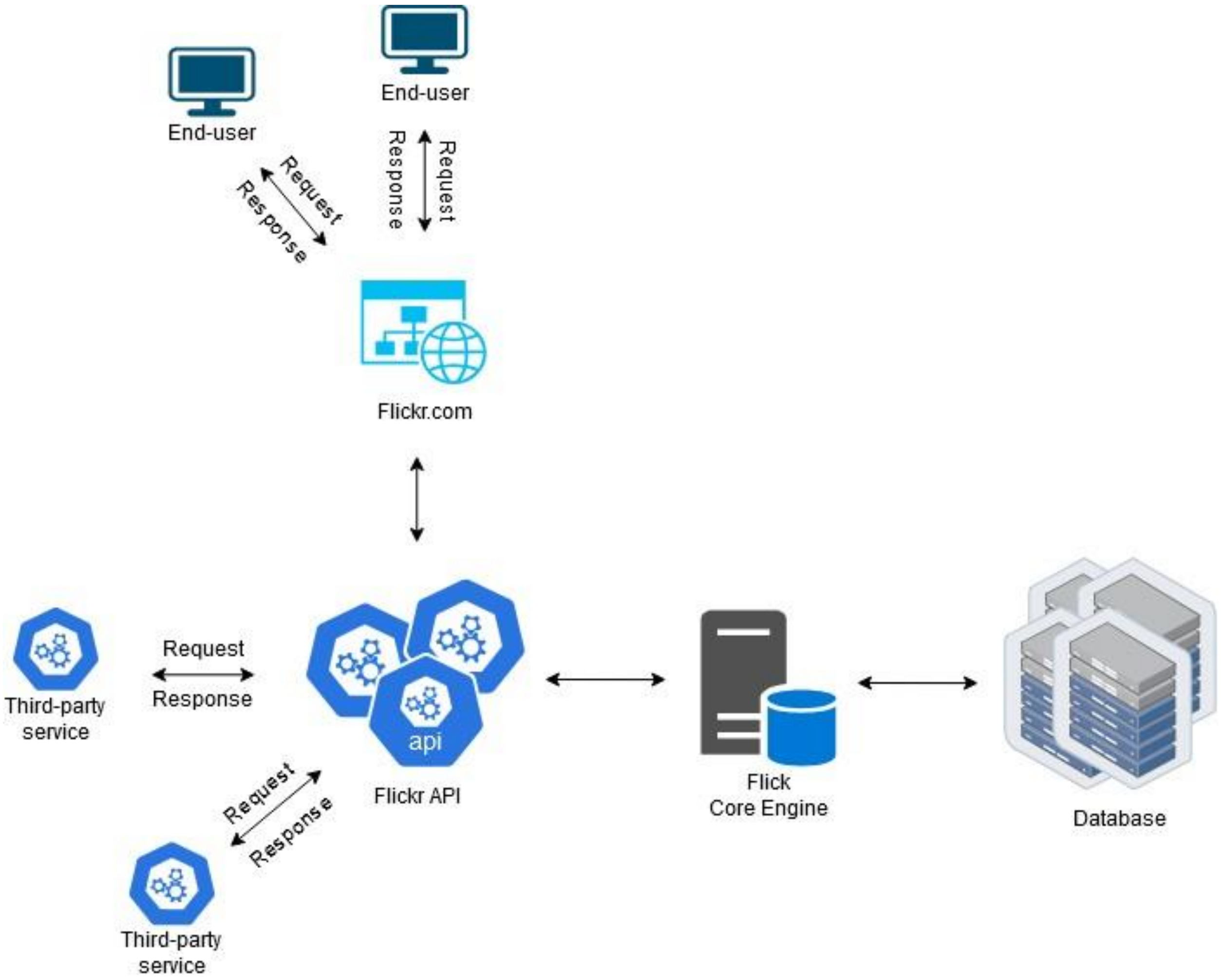


Figure 24 Flickr API high level diagram

5.3 Flickr APIs Methods

Flickr provides us with a plethora of APIs and their methods in a variety of domains including, but not limited to, People, Photos, Favorites and Comment. Most of them need users' consent. However, in the context of this research we use exclusively public photos and user information.

To perform an action using the Flickr API, as also mentioned in the API documentation, you need to select a calling convention, send a request to its endpoint specifying a method and some arguments, and will receive a formatted response.

The REQUIRED parameter method is used to specify the calling method.

The REQUIRED parameter api_key is used to specify your API Key.

The optional parameter format is used to specify a response format. However, in our case we use exclusively json format.

As for the encoding, the Flickr API expects all data to be UTF-8 encoded.

The following are the API methods that used in our application to integrate with Flickr APIs to collect the necessary information for our research.

5.3.1 flickr.photos.search

This method returns a list of photos matching some criteria. Only photos visible to the calling user will be returned. To return private or semi-private photos, the caller must be authenticated with 'read' permissions and have permission to view the photos. This method does not require authentication.

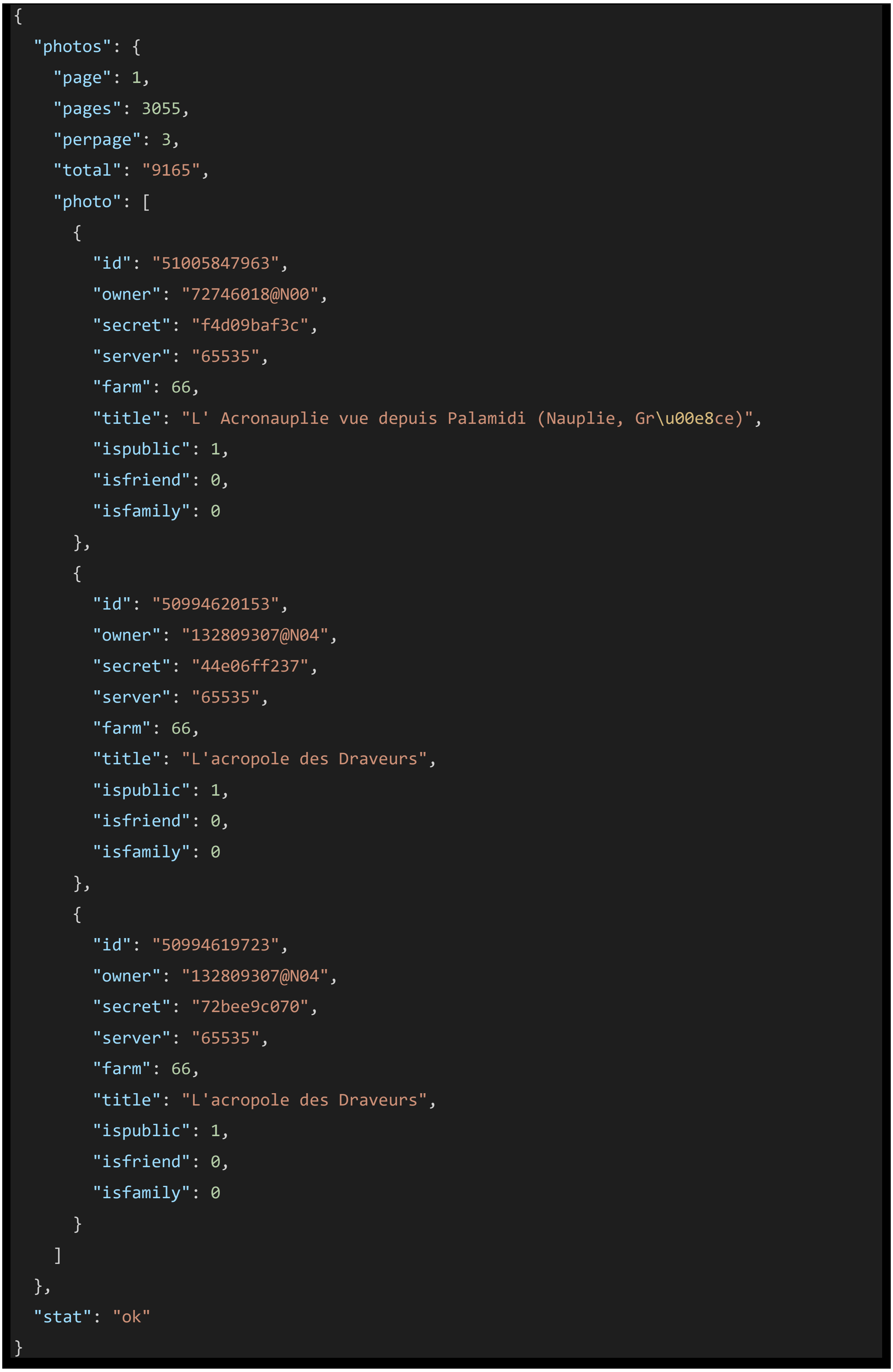
Unauthenticated requests will only return public photos!

HTTP Method: GET

Query Parameters:

Key	Value	Description
Method	flickr.photos.search	
format	json	(Optional)
api_key	{api_key}	Your API application key.
tags	{tag}	A comma-delimited list of tags. Photos with one or more of the tags listed will be returned.
per_page	{per_page}	(Optional) Number of photos to return per page. If this argument is omitted, it defaults to 100. The maximum allowed value is 500.
page	{page}	(Optional) The page of results to return. If this argument is omitted, it defaults to 1.

JSON response:



```
{
  "photos": {
    "page": 1,
    "pages": 3055,
    "perpage": 3,
    "total": "9165",
    "photo": [
      {
        "id": "51005847963",
        "owner": "72746018@N00",
        "secret": "f4d09baf3c",
        "server": "65535",
        "farm": 66,
        "title": "L' Acronauplie vue depuis Palamidi (Nauplie, Gr\u00e8ce)",
        "ispublic": 1,
        "isfriend": 0,
        "isfamily": 0
      },
      {
        "id": "50994620153",
        "owner": "132809307@N04",
        "secret": "44e06ff237",
        "server": "65535",
        "farm": 66,
        "title": "L'acropole des Draveurs",
        "ispublic": 1,
        "isfriend": 0,
        "isfamily": 0
      },
      {
        "id": "50994619723",
        "owner": "132809307@N04",
        "secret": "72bee9c070",
        "server": "65535",
        "farm": 66,
        "title": "L'acropole des Draveurs",
        "ispublic": 1,
        "isfriend": 0,
        "isfamily": 0
      }
    ]
  },
  "stat": "ok"
}
```

5.3.2 flickr.photos.getInfo

Get information about a photo. The calling user must have permission to view the photo. This method does not require authentication.

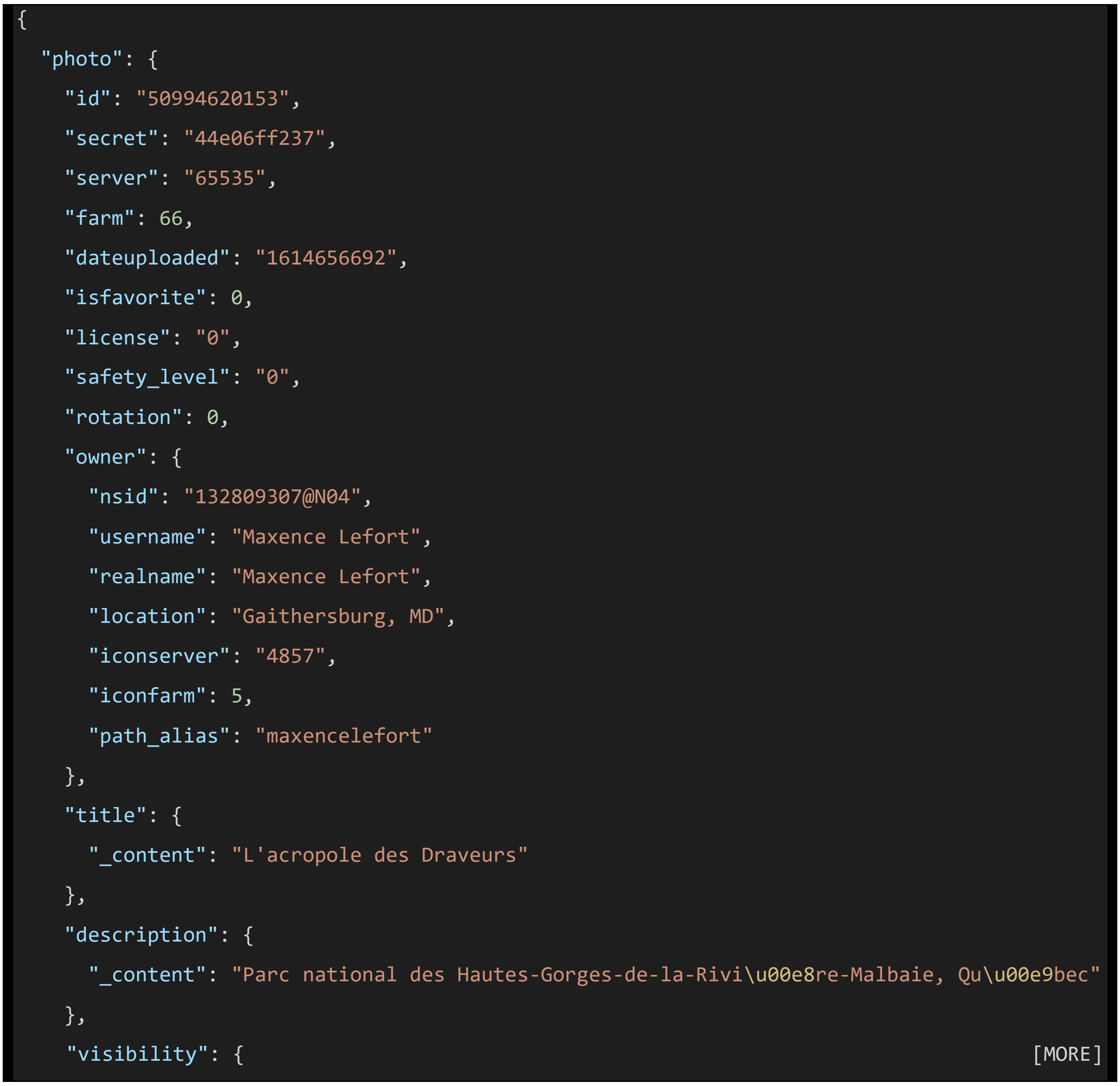
Unauthenticated requests will only return information in case of a public photos!

HTTP Method: GET

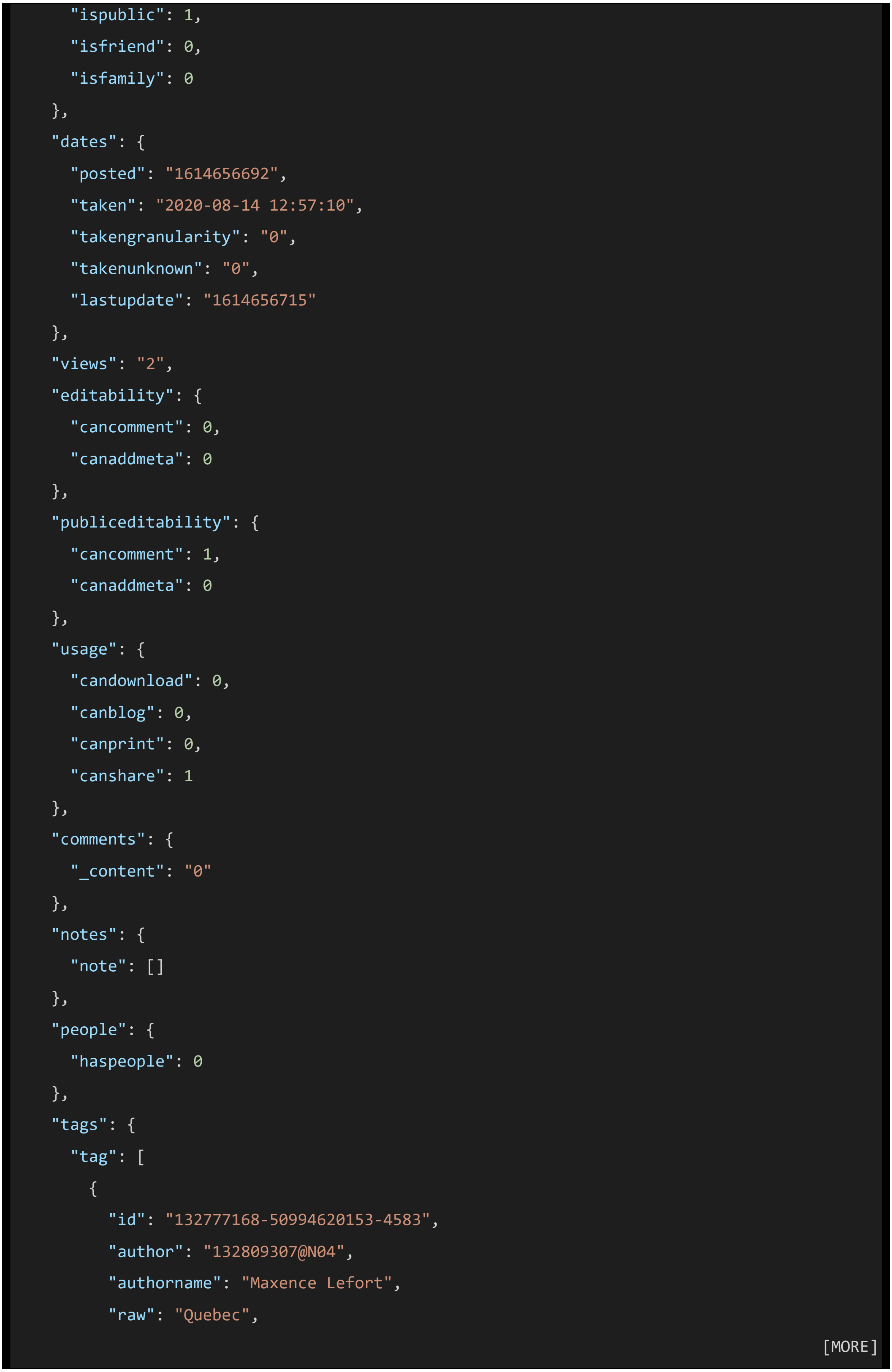
Query Parameters:

Key	Value	Description
method	flickr.photos.getInfo	
format	json	(Optional)
api_key	{api_key}	Your API application key.
photo_id	{photo_id}	The id of the photo to get information for.

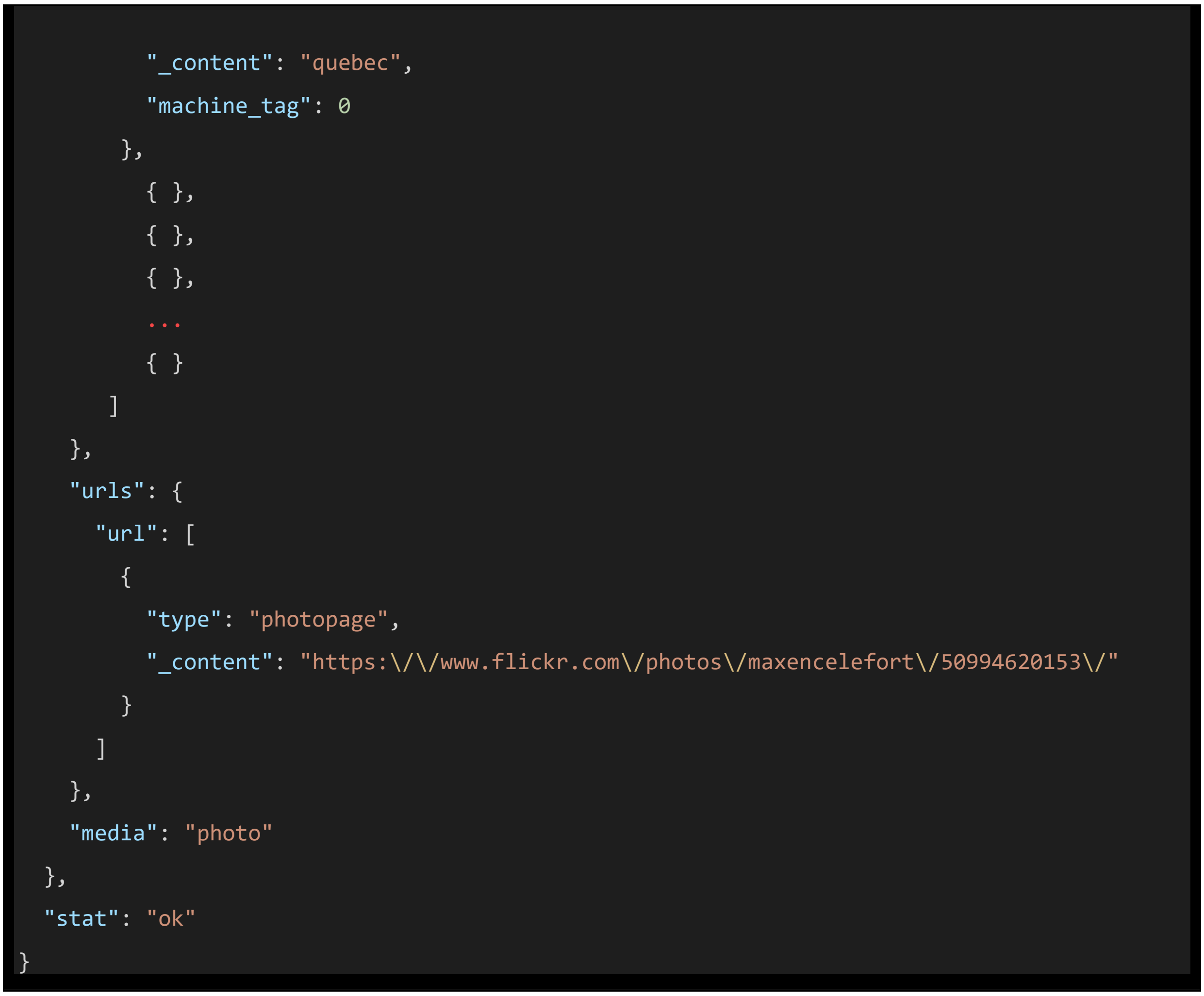
JSON response:



```
{
  "photo": {
    "id": "50994620153",
    "secret": "44e06ff237",
    "server": "65535",
    "farm": 66,
    "dateuploaded": "1614656692",
    "isfavorite": 0,
    "license": "0",
    "safety_level": "0",
    "rotation": 0,
    "owner": {
      "nsid": "132809307@N04",
      "username": "Maxence Lefort",
      "realname": "Maxence Lefort",
      "location": "Gaithersburg, MD",
      "iconserver": "4857",
      "iconfarm": 5,
      "path_alias": "maxencelefort"
    },
    "title": {
      "_content": "L'acropole des Draveurs"
    },
    "description": {
      "_content": "Parc national des Hautes-Gorges-de-la-Rivi\u00e8re-Malbaie, Qu\u00e9bec"
    },
    "visibility": {                                                                    [MORE]
```



```
      "ispublic": 1,
      "isfriend": 0,
      "isfamily": 0
    },
    "dates": {
      "posted": "1614656692",
      "taken": "2020-08-14 12:57:10",
      "takengranularity": "0",
      "takenunknown": "0",
      "lastupdate": "1614656715"
    },
    "views": "2",
    "editability": {
      "cancomment": 0,
      "canaddmeta": 0
    },
    "publiceditability": {
      "cancomment": 1,
      "canaddmeta": 0
    },
    "usage": {
      "candownload": 0,
      "canblog": 0,
      "canprint": 0,
      "canshare": 1
    },
    "comments": {
      "_content": "0"
    },
    "notes": {
      "note": []
    },
    "people": {
      "haspeople": 0
    },
    "tags": {
      "tag": [
        {
          "id": "132777168-50994620153-4583",
          "author": "132809307@N04",
          "authorname": "Maxence Lefort",
          "raw": "Quebec",
```



```
          "_content": "quebec",
          "machine_tag": 0
        },
          { },
          { },
          { },
          ...
          { }
       ]
    },
    "urls": {
      "url": [
        {
          "type": "photopage",
          "_content": "https:\/\/www.flickr.com\/photos\/maxencelefort\/50994620153\/"
        }
      ]
    },
    "media": "photo"
  },
  "stat": "ok"
}
```

5.3.3 flickr.photos.getFavorites

Returns the list of people who have favorited a given photo. The calling user must have permission to view the photo. This method does not require authentication.

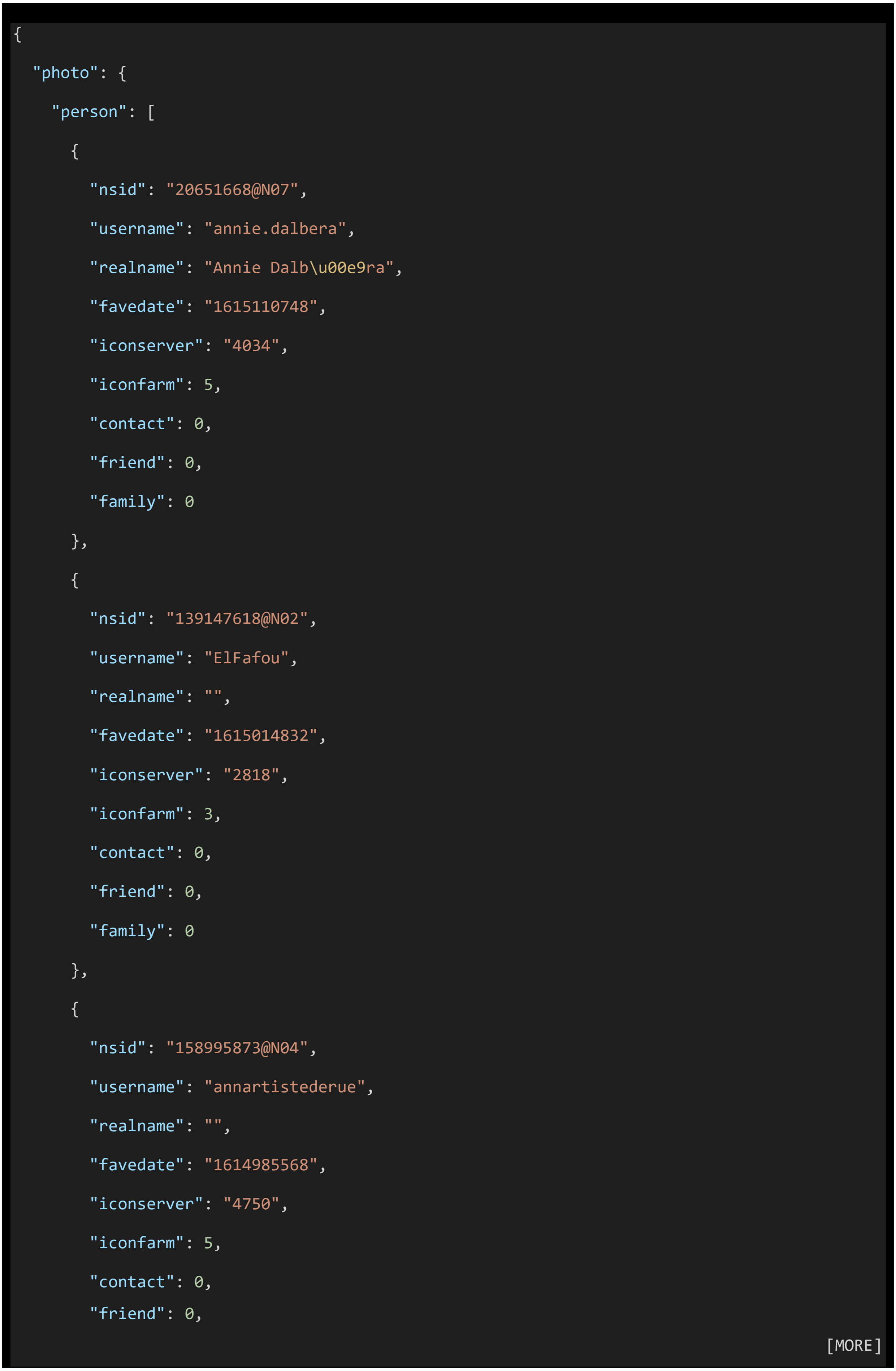
Unauthenticated requests will only return information in case of a public photos!

HTTP Method: GET

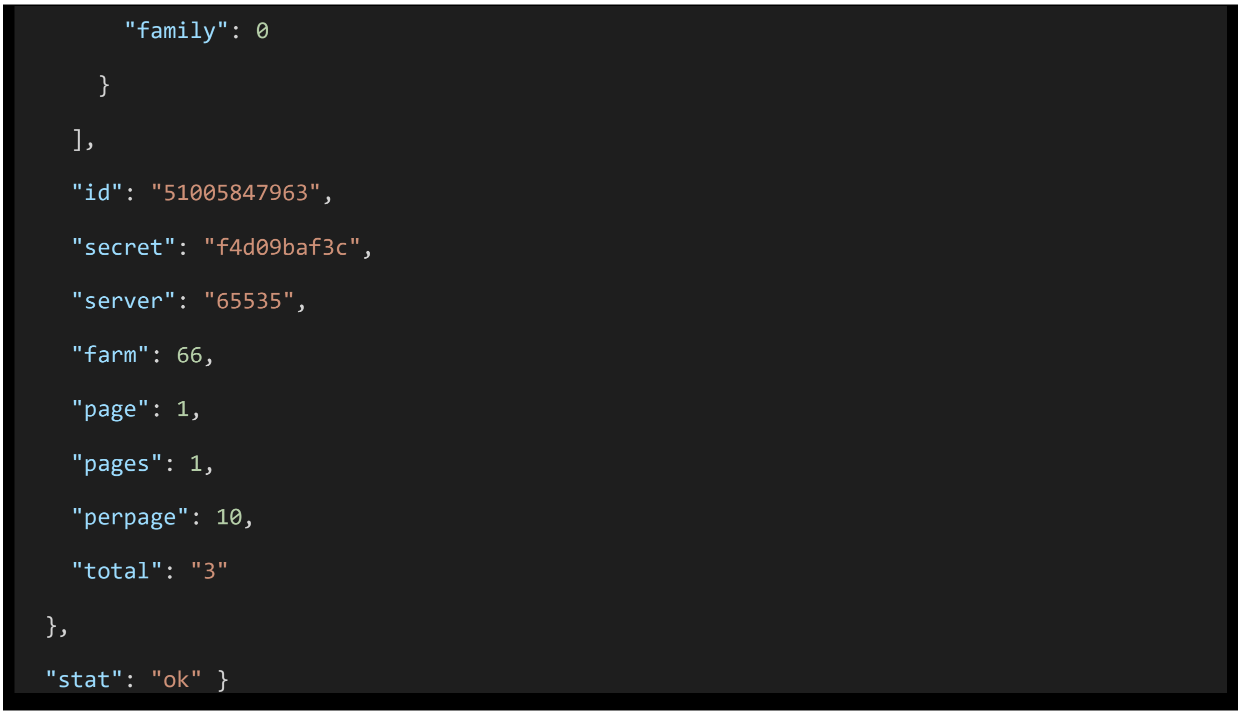
Query Parameters:

Key	Value	Description
method	flickr.photos.getInfo	
format	json	(Optional)
api_key	{api_key}	Your API application key.
photo_id	{photo_id}	The id of the photo to get information for.
per_page	{per_page}	(Optional) Number of photos to return per page. If this argument is omitted, it defaults to 100. The maximum allowed value is 500.
page	{page}	(Optional) The page of results to return. If this argument is omitted, it defaults to 1.

JSON response:



```
{
  "photo": {
    "person": [
      {
        "nsid": "20651668@N07",
        "username": "annie.dalbera",
        "realname": "Annie Dalb\u00e9ra",
        "favedate": "1615110748",
        "iconserver": "4034",
        "iconfarm": 5,
        "contact": 0,
        "friend": 0,
        "family": 0
      },
      {
        "nsid": "139147618@N02",
        "username": "ElFafou",
        "realname": "",
        "favedate": "1615014832",
        "iconserver": "2818",
        "iconfarm": 3,
        "contact": 0,
        "friend": 0,
        "family": 0
      },
      {
        "nsid": "158995873@N04",
        "username": "annartistederue",
        "realname": "",
        "favedate": "1614985568",
        "iconserver": "4750",
        "iconfarm": 5,
        "contact": 0,
        "friend": 0,
                                                                                    [MORE]
```



```
        "family": 0
      }
    ],
    "id": "51005847963",
    "secret": "f4d09baf3c",
    "server": "65535",
    "farm": 66,
    "page": 1,
    "pages": 1,
    "perpage": 10,
    "total": "3"
  },
  "stat": "ok" }
```

5.3.4 flickr.photos.comments.getList

Returns the comments for a photo. The calling user must have permission to view the photo. This method does not require authentication.

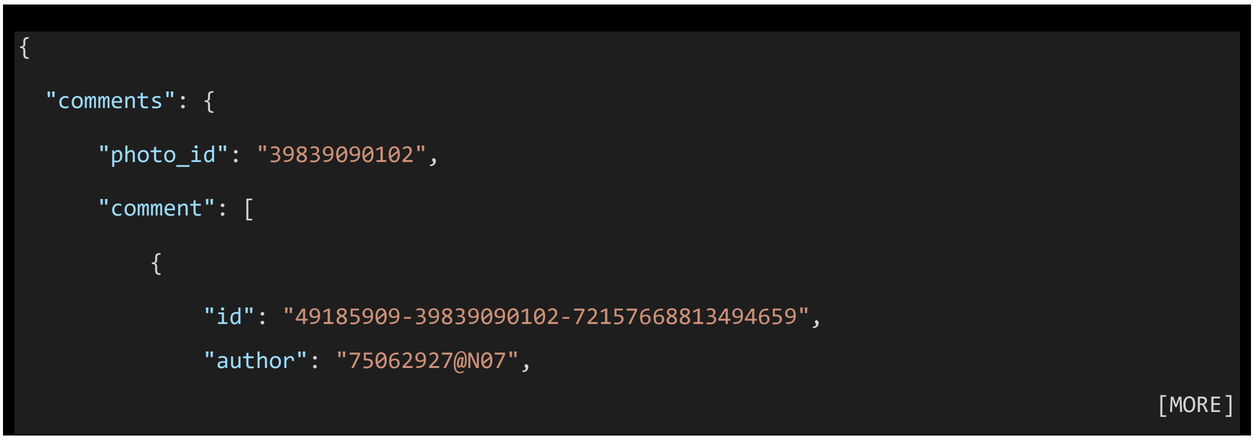
Unauthenticated requests will only return information in case of a public photos!

HTTP Method: GET

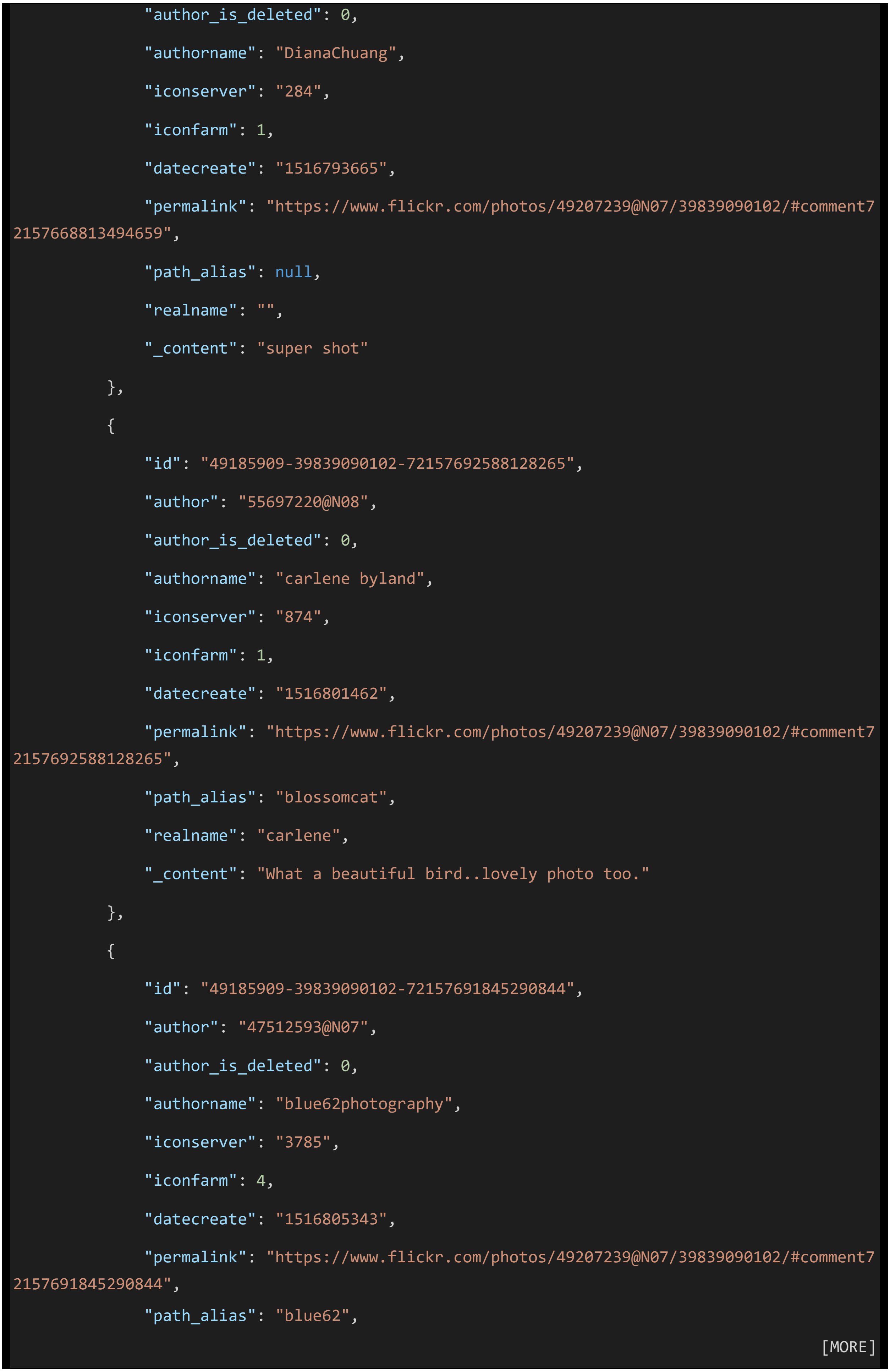
Query Parameters:

Key	Value	Description
method	flickr.photos.getInfo	
format	json	(Optional)
api_key	{api_key}	Your API application key.
photo_id	{photo_id}	The id of the photo to get information for.

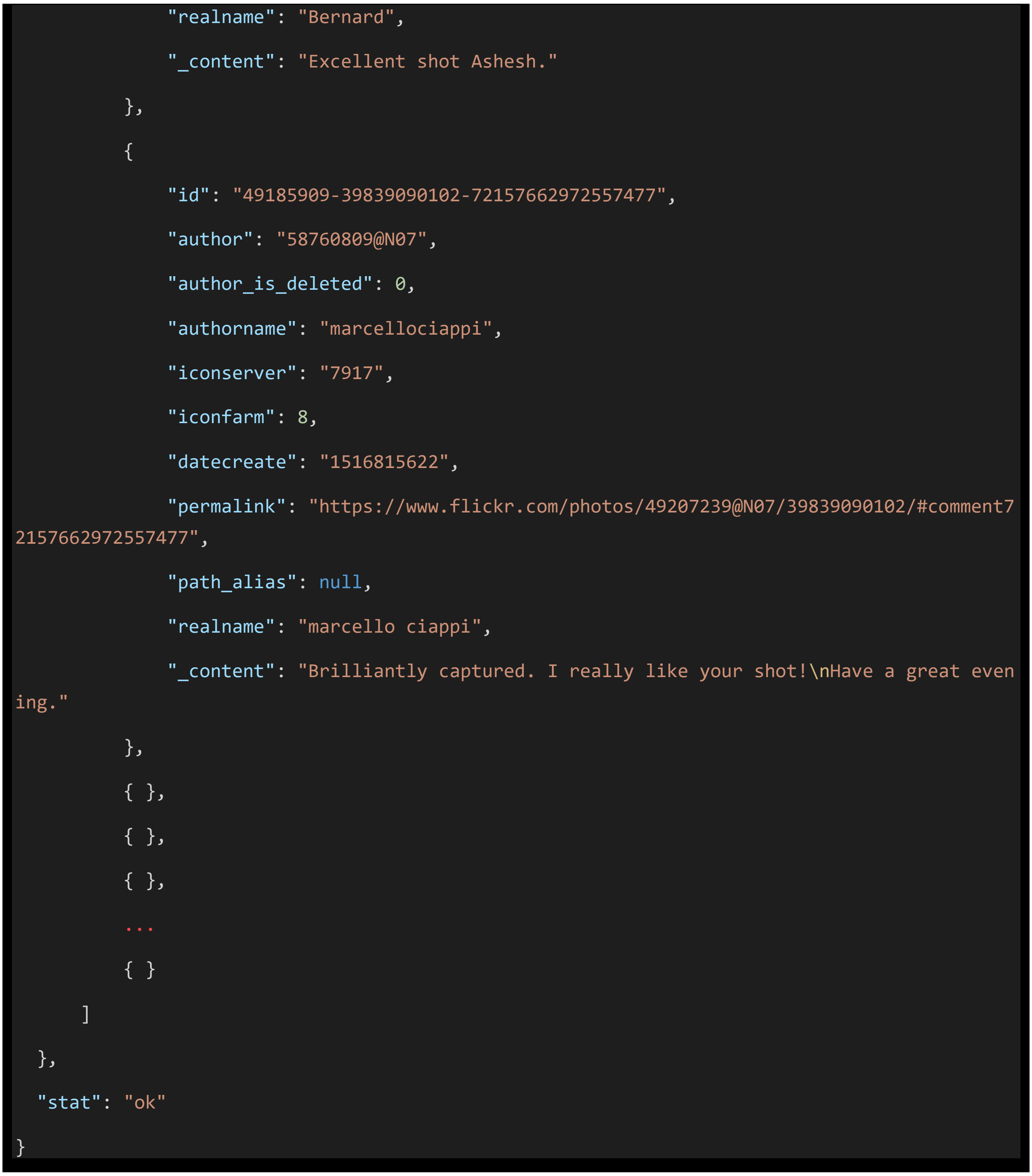
JSON response:



```
{
  "comments": {
      "photo_id": "39839090102",
      "comment": [
          {
              "id": "49185909-39839090102-72157668813494659",
              "author": "75062927@N07",
                                                                  [MORE]
```



```
            "author_is_deleted": 0,
            "authorname": "DianaChuang",
            "iconserver": "284",
            "iconfarm": 1,
            "datecreate": "1516793665",
            "permalink": "https://www.flickr.com/photos/49207239@N07/39839090102/#comment7
2157668813494659",
            "path_alias": null,
            "realname": "",
            "_content": "super shot"
        },
        {
            "id": "49185909-39839090102-72157692588128265",
            "author": "55697220@N08",
            "author_is_deleted": 0,
            "authorname": "carlene byland",
            "iconserver": "874",
            "iconfarm": 1,
            "datecreate": "1516801462",
            "permalink": "https://www.flickr.com/photos/49207239@N07/39839090102/#comment7
2157692588128265",
            "path_alias": "blossomcat",
            "realname": "carlene",
            "_content": "What a beautiful bird..lovely photo too."
        },
        {
            "id": "49185909-39839090102-72157691845290844",
            "author": "47512593@N07",
            "author_is_deleted": 0,
            "authorname": "blue62photography",
            "iconserver": "3785",
            "iconfarm": 4,
            "datecreate": "1516805343",
            "permalink": "https://www.flickr.com/photos/49207239@N07/39839090102/#comment7
2157691845290844",
            "path_alias": "blue62",
                                                                                [MORE]
```



```
              "realname": "Bernard",
              "_content": "Excellent shot Ashesh."
          },
          {
              "id": "49185909-39839090102-72157662972557477",
              "author": "58760809@N07",
              "author_is_deleted": 0,
              "authorname": "marcellociappi",
              "iconserver": "7917",
              "iconfarm": 8,
              "datecreate": "1516815622",
              "permalink": "https://www.flickr.com/photos/49207239@N07/39839090102/#comment72157662972557477",
              "path_alias": null,
              "realname": "marcello ciappi",
              "_content": "Brilliantly captured. I really like your shot!\nHave a great evening."
          },
          { },
          { },
          { },
          ...
          { }
      ]
  },
  "stat": "ok"
}
```

6. INFLUENCE METRICS SYSTEM ARCHITECTURE

The proposed solution consists of

- a MySQL database,
- a backend service layer - on top of Flickr's APIs and the database - implemented in Spring Boot Framework,
- and an Angular web client that provides end-users the interface to interact with the web servers.

Here is a concise diagram of our solution's interactions:

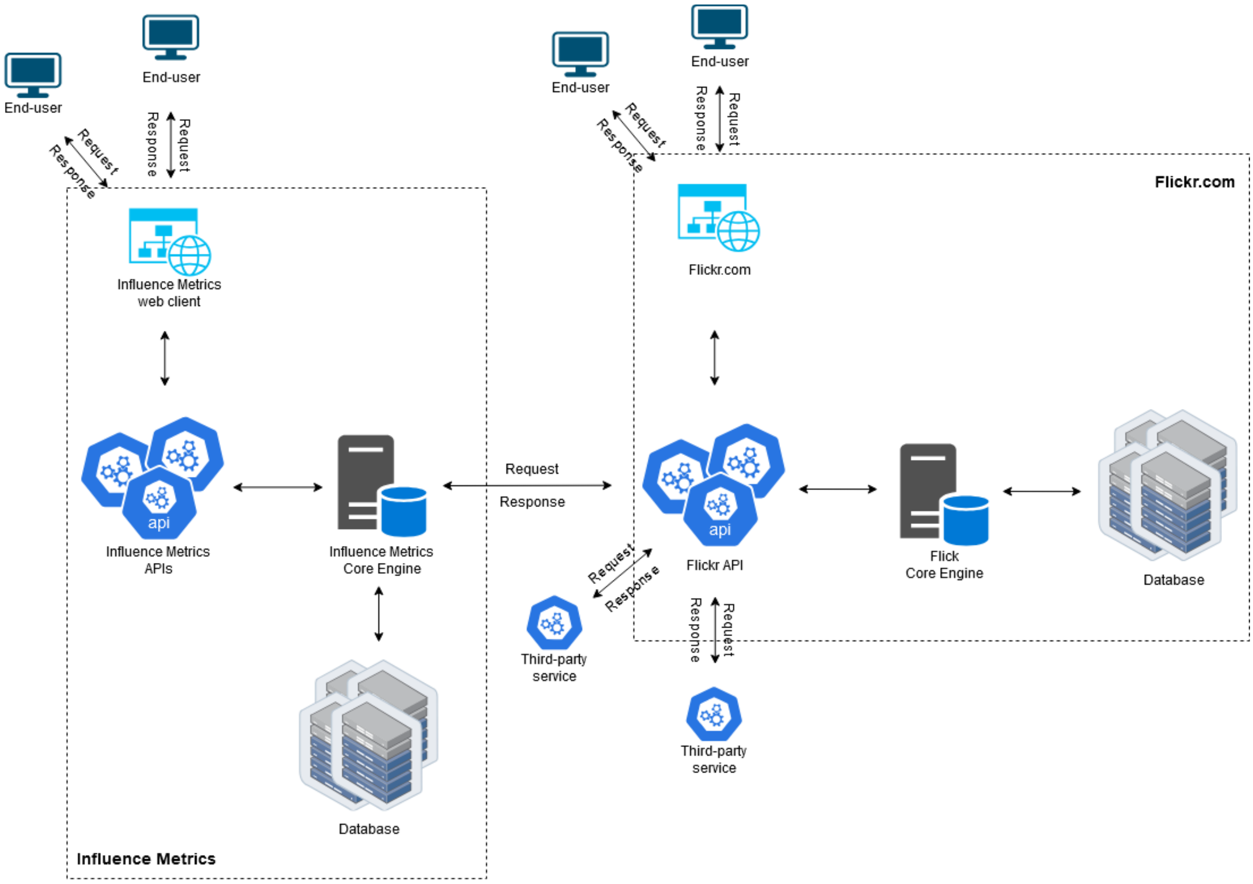


Figure 25 High level system architecture diagram

6.1 The service layer

The service layer defines application's boundary and its set of available operations. It is responsible for the core functionality of extracting, transforming, and loading (ETL) data from the Flickr APIs. It facilitates seamless data retrieval by extracting public photos, user information, and relevant metadata, then transforming this raw data into a structured format suitable for analysis.

This layer also manages the full set of CRUD (Create, Read, Update, Delete) operations, enabling efficient data management of large image datasets, metadata and user information, ensuring that the collected data is accurate. By analyzing photos metadata as well as user behavior, such as views, likes, comments, and follows, the system can evaluate trends and determine key influencers within the Flickr's community.

By choosing Spring Boot for our backend implementation, we were created a robust and scalable service layer that enhances the overall performance and maintainability of our solution. Spring Boot offers several advantages such as:

- Easy setup and rapid development
- High performance and reliability
- Native REST API support
- Excellent documentation and community

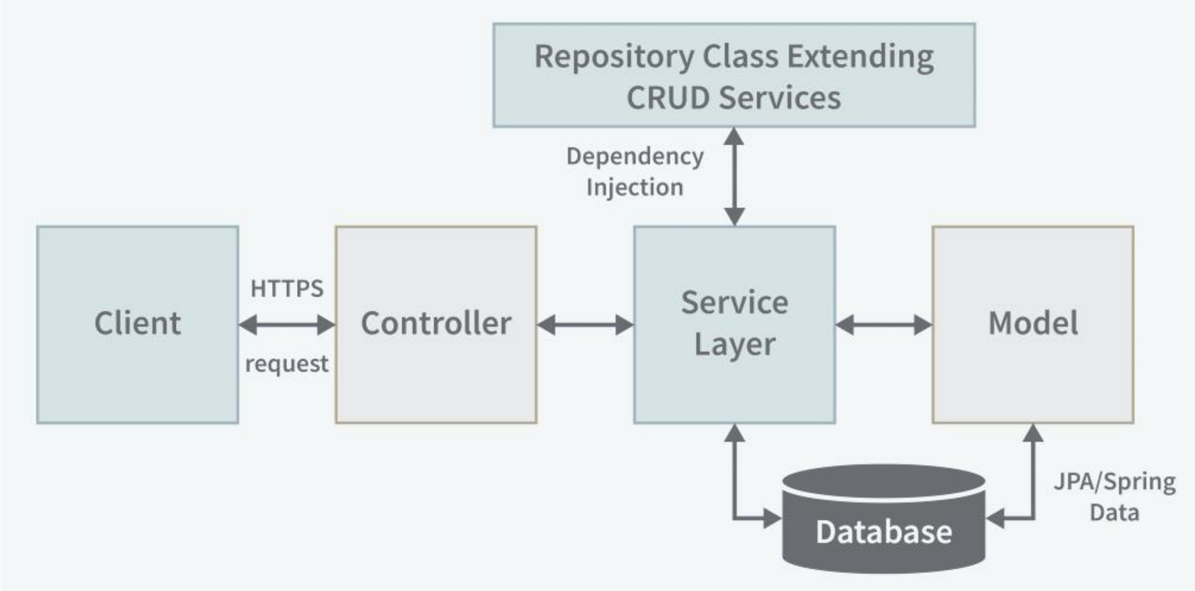
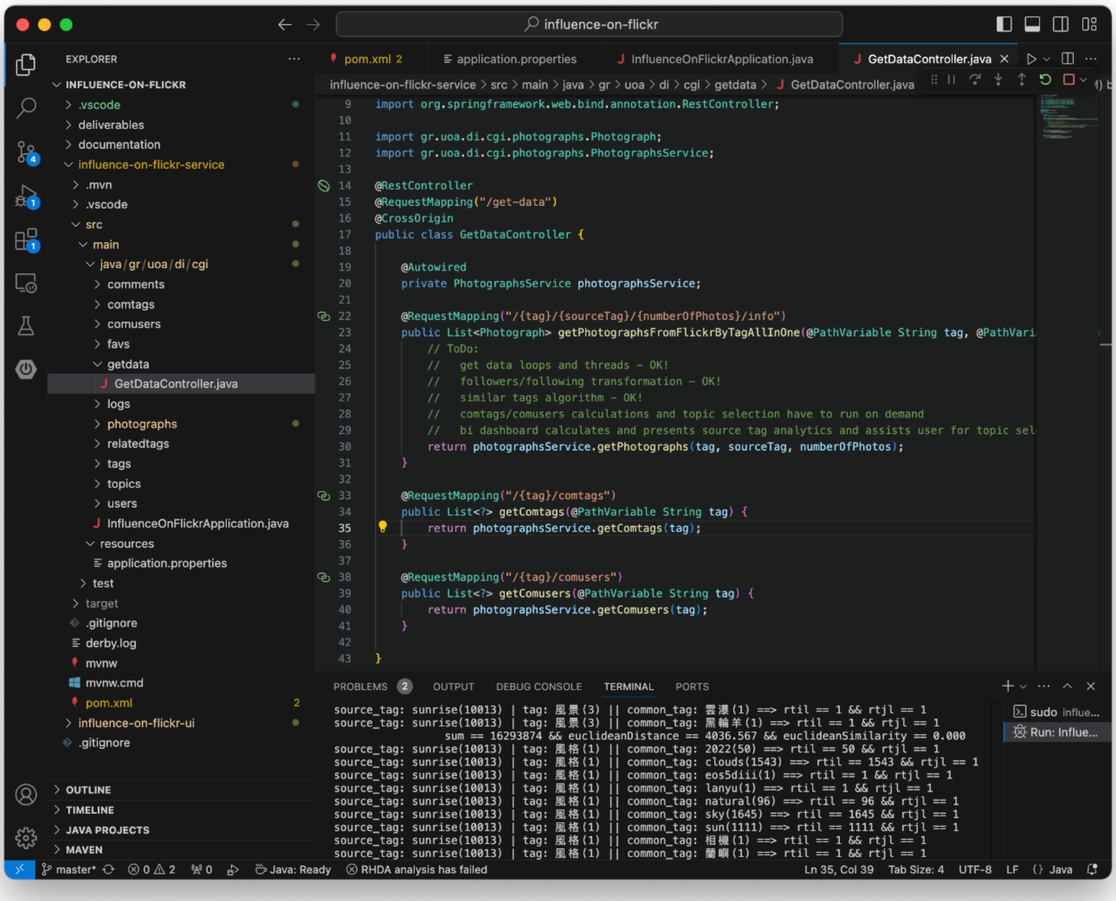


Figure 26 Back-end service layer architecture [41]

Spring Boot [42] simplifies the development process by providing pre-configured setups for common frameworks and dependencies, reducing the need for extensive configuration. This allows developers to focus on building the core functionality. It also natively supports building RESTful services, which is exactly what we need to interact with the Flickr API and manage CRUD operations. Its seamless integration with data serialization formats like JSON and XML makes it easy to handle API requests and responses, which is critical for applications needing to perform high-volume data operations and ETL processes.

Spring Boot has an extensive and well-maintained set of documentation, along with a large and active developer community. This means access to a lot of libraries and troubleshooting support, allowing for quick resolution of any issues that arose to us during development.



```java
import org.springframework.web.bind.annotation.RestController;

import gr.uoa.di.cgi.photographs.Photograph;
import gr.uoa.di.cgi.photographs.PhotographsService;

@RestController
@RequestMapping("/get-data")
@CrossOrigin
public class GetDataController {

    @Autowired
    private PhotographsService photographsService;

    @RequestMapping("/{tag}/{sourceTag}/{numberOfPhotos}/info")
    public List<Photograph> getPhotographsFromFlickrByTagAllInOne(@PathVariable String tag, @PathVari
        // ToDo:
        //   get data loops and threads - OK!
        //   followers/following transformation - OK!
        //   similar tags algorithm - OK!
        //   comtags/comusers calculations and topic selection have to run on demand
        //   bi dashboard calculates and presents source tag analytics and assists user for topic sel
        return photographsService.getPhotographs(tag, sourceTag, numberOfPhotos);
    }

    @RequestMapping("/{tag}/comtags")
    public List<?> getComtags(@PathVariable String tag) {
        return photographsService.getComtags(tag);
    }

    @RequestMapping("/{tag}/comusers")
    public List<?> getComusers(@PathVariable String tag) {
        return photographsService.getComusers(tag);
    }

}
```

```
source_tag: sunrise(10013) | tag: 風景(3) || common_tag: 雲瀑(1) ==> rtil == 1 && rtjl == 1
source_tag: sunrise(10013) | tag: 風景(3) || common_tag: 黑輪羊(1) ==> rtil == 1 && rtjl == 1
             sum == 16293874 && euclideanDistance == 4036.567 && euclideanSimilarity == 0.000
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: 2022(50) ==> rtil == 50 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: clouds(1543) ==> rtil == 1543 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: eos5diii(1) ==> rtil == 1 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: lanyu(1) ==> rtil == 1 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: natural(96) ==> rtil == 96 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: sky(1645) ==> rtil == 1645 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: sun(1111) ==> rtil == 1111 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: 相機(1) ==> rtil == 1 && rtjl == 1
source_tag: sunrise(10013) | tag: 風格(1) || common_tag: 蘭嶼(1) ==> rtil == 1 && rtjl == 1
```

Figure 27 Influence-on-flickr service running on Visual Studio Code Server

6.1.1 Influence Metrics APIs Methods

Our solution includes - along with all the other CRUD operations about photographs, favs, comments etc. - three basic API methods to get data from Flickr API and calculate influence. To perform an action using the Influence Metrics API, you need to send a request to its endpoint, specifying the HTTP method and some path parameters and then you will receive a formatted response.

As for the encoding, the Influence Metrics API expects all data to be UTF-8 encoded.

6.1.1.1 /get-data/{tag}/{sourceTag}/{numberOfPhotos}/info

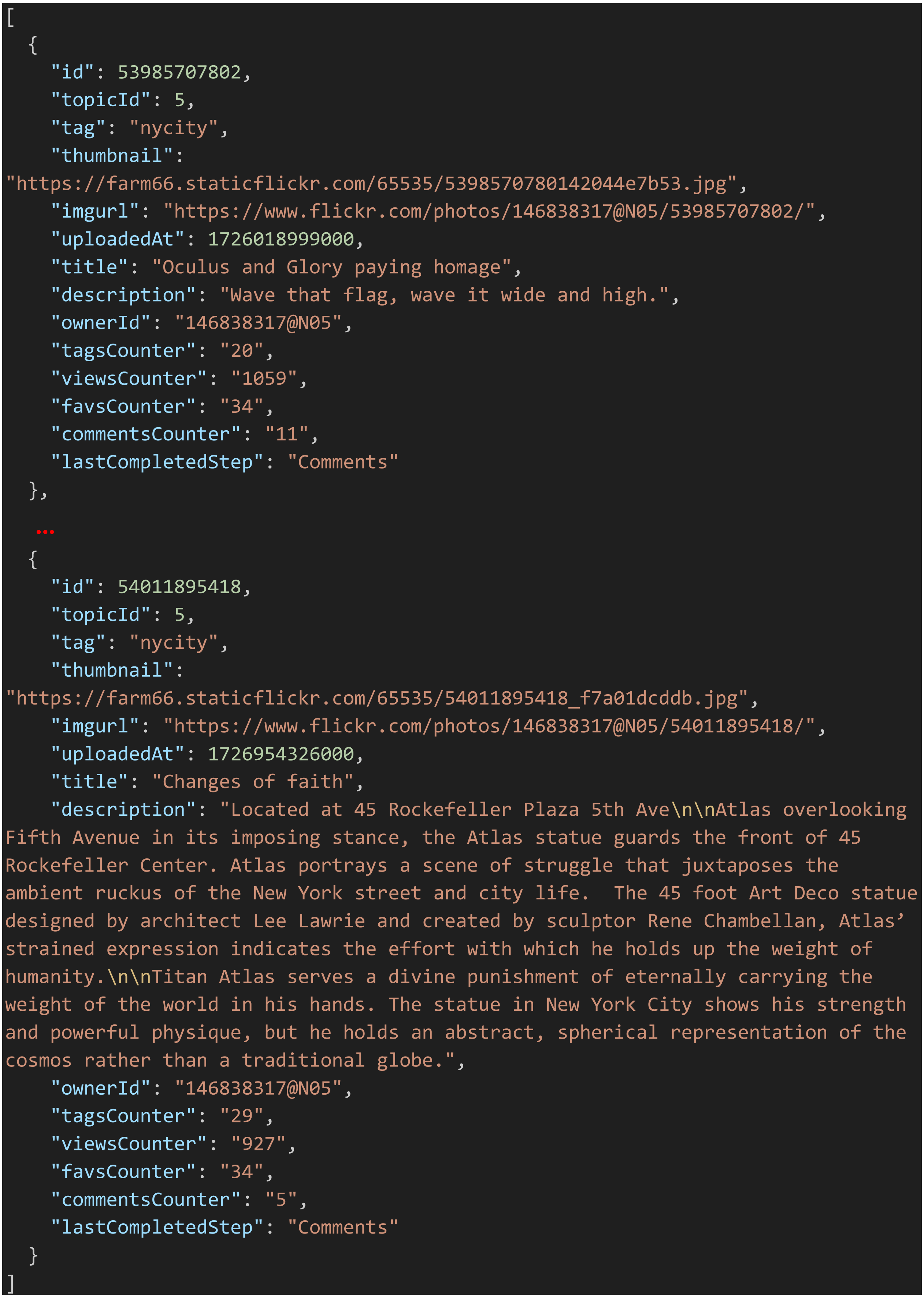
This method collects and returns a list of public photos matching some criteria.

HTTP Method: GET

Path Parameters:

Key	Description
tag	The requested tag
sourceTag	The topic associated with the requested tag
numberOfPhotos	The number of photos to collect

JSON response:



```
[
  {
    "id": 53985707802,
    "topicId": 5,
    "tag": "nycity",
    "thumbnail":
"https://farm66.staticflickr.com/65535/5398570780142044e7b53.jpg",
    "imgurl": "https://www.flickr.com/photos/146838317@N05/53985707802/",
    "uploadedAt": 1726018999000,
    "title": "Oculus and Glory paying homage",
    "description": "Wave that flag, wave it wide and high.",
    "ownerId": "146838317@N05",
    "tagsCounter": "20",
    "viewsCounter": "1059",
    "favsCounter": "34",
    "commentsCounter": "11",
    "lastCompletedStep": "Comments"
  },

   ...
  {
    "id": 54011895418,
    "topicId": 5,
    "tag": "nycity",
    "thumbnail":
"https://farm66.staticflickr.com/65535/54011895418_f7a01dcddb.jpg",
    "imgurl": "https://www.flickr.com/photos/146838317@N05/54011895418/",
    "uploadedAt": 1726954326000,
    "title": "Changes of faith",
    "description": "Located at 45 Rockefeller Plaza 5th Ave\n\nAtlas overlooking
Fifth Avenue in its imposing stance, the Atlas statue guards the front of 45
Rockefeller Center. Atlas portrays a scene of struggle that juxtaposes the
ambient ruckus of the New York street and city life.  The 45 foot Art Deco statue
designed by architect Lee Lawrie and created by sculptor Rene Chambellan, Atlas’
strained expression indicates the effort with which he holds up the weight of
humanity.\n\nTitan Atlas serves a divine punishment of eternally carrying the
weight of the world in his hands. The statue in New York City shows his strength
and powerful physique, but he holds an abstract, spherical representation of the
cosmos rather than a traditional globe.",
    "ownerId": "146838317@N05",
    "tagsCounter": "29",
    "viewsCounter": "927",
    "favsCounter": "34",
    "commentsCounter": "5",
    "lastCompletedStep": "Comments"
  }
]
```

6.1.1.2 /get-data/{tag}/comtags

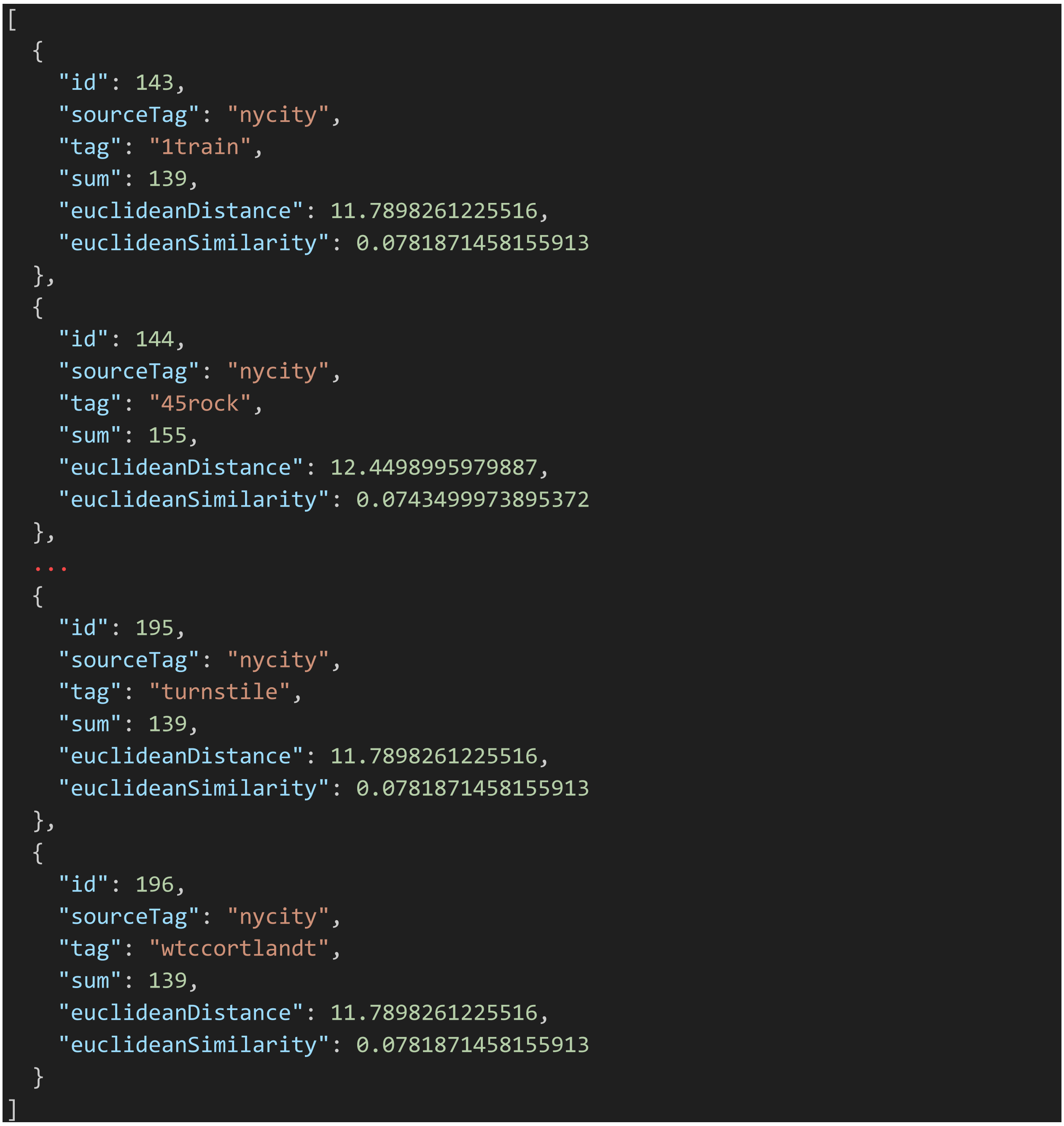
This method calculates the common tags that two distinct tags have been used with, estimates their Euclidean distance and the related similarity measure.

HTTP Method: GET

Path Parameters:

Key	Description
tag	The requested tag

JSON response:



```
[
  {
    "id": 143,
    "sourceTag": "nycity",
    "tag": "1train",
    "sum": 139,
    "euclideanDistance": 11.7898261225516,
    "euclideanSimilarity": 0.0781871458155913
  },
  {
    "id": 144,
    "sourceTag": "nycity",
    "tag": "45rock",
    "sum": 155,
    "euclideanDistance": 12.4498995979887,
    "euclideanSimilarity": 0.0743499973895372
  },
  ...
  {
    "id": 195,
    "sourceTag": "nycity",
    "tag": "turnstile",
    "sum": 139,
    "euclideanDistance": 11.7898261225516,
    "euclideanSimilarity": 0.0781871458155913
  },
  {
    "id": 196,
    "sourceTag": "nycity",
    "tag": "wtccortlandt",
    "sum": 139,
    "euclideanDistance": 11.7898261225516,
    "euclideanSimilarity": 0.0781871458155913
  }
]
```

6.1.1.3 /get-data/{tag}/comusers

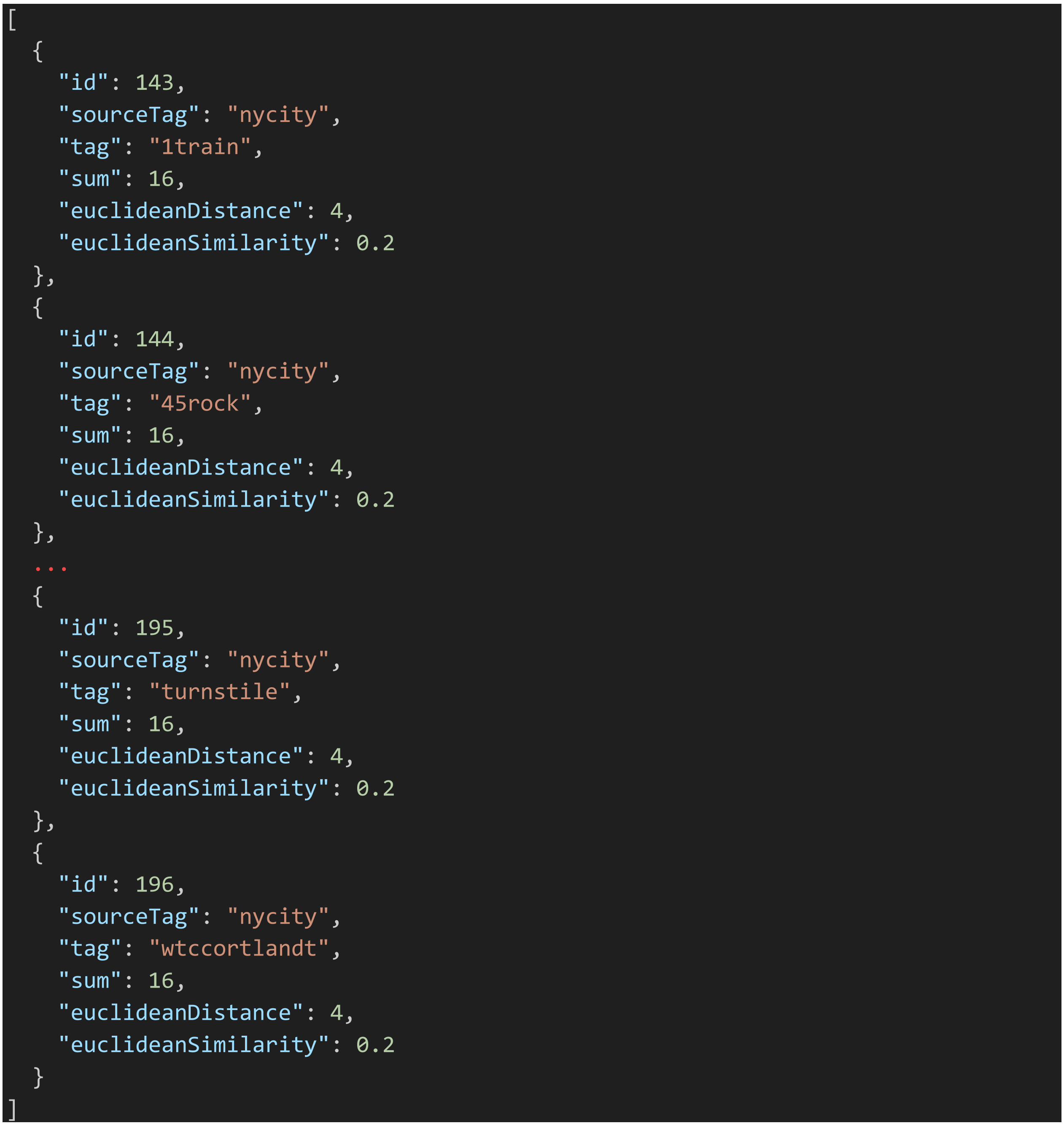
This method determines the level of usage of tags by users who have utilized them in common, estimates their Euclidean distance and the related similarity measure.

HTTP Method: GET

Path Parameters:

Key	Description
tag	The requested tag

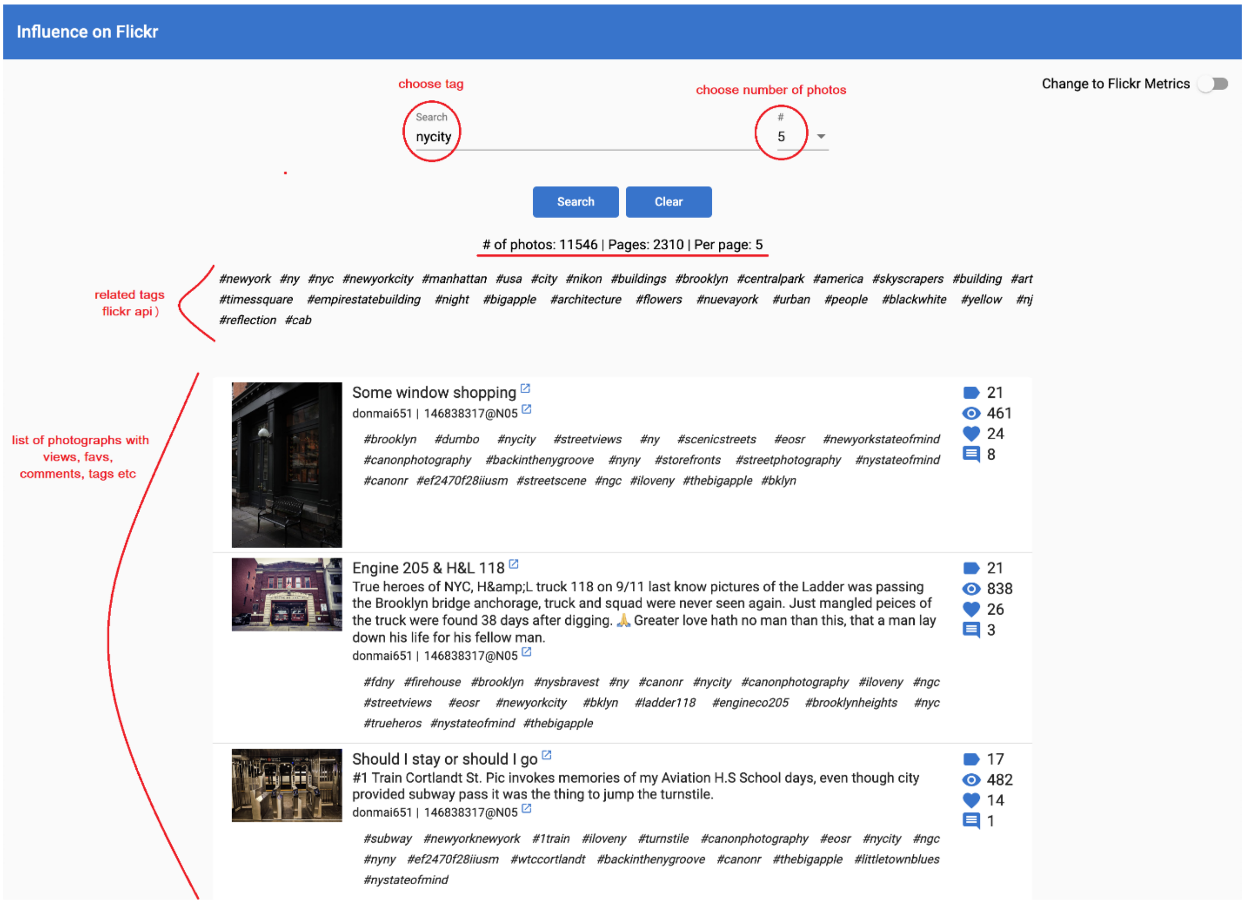
JSON response:



```
[
  {
    "id": 143,
    "sourceTag": "nycity",
    "tag": "1train",
    "sum": 16,
    "euclideanDistance": 4,
    "euclideanSimilarity": 0.2
  },
  {
    "id": 144,
    "sourceTag": "nycity",
    "tag": "45rock",
    "sum": 16,
    "euclideanDistance": 4,
    "euclideanSimilarity": 0.2
  },
  ...
  {
    "id": 195,
    "sourceTag": "nycity",
    "tag": "turnstile",
    "sum": 16,
    "euclideanDistance": 4,
    "euclideanSimilarity": 0.2
  },
  {
    "id": 196,
    "sourceTag": "nycity",
    "tag": "wtccortlandt",
    "sum": 16,
    "euclideanDistance": 4,
    "euclideanSimilarity": 0.2
  }
]
```

6.2 The Web Client

The web client of our solution is designed to enhance user experience by providing a clean, intuitive, and user-friendly interface. Built with a focus on simplicity and efficiency, the client allows users to seamlessly interact with the underlying back-end service layer for tasks such as viewing collected data, managing CRUD operations, and accessing key features such as photo and user influence metrics, and topic trends. This approach empowers users to easily and quickly search, filter, and engage with the content they are analyzing without needing extensive technical knowledge.



Search: nycity | #: 5

# of photos: 11546 | Pages: 2310 | Per page: 5

#newyork #ny #nyc #newyorkcity #manhattan #usa #city #nikon #buildings #brooklyn #centralpark #america #skyscrapers #building #art #timessquare #empirestatebuilding #night #bigapple #architecture #flowers #nuevayork #urban #people #blackwhite #yellow #nj #reflection #cab

Some window shopping
donmai651 | 146838317@N05
#brooklyn #dumbo #nycity #streetviews #ny #scenicstreets #eosr #newyorkstateofmind #canonphotography #backinthenygroove #nyny #storefronts #streetphotography #nystateofmind #canonr #ef2470f28iiusm #streetscene #ngc #iloveny #thebigapple #bklyn
🏷 21
👁 461
♥ 24
💬 8

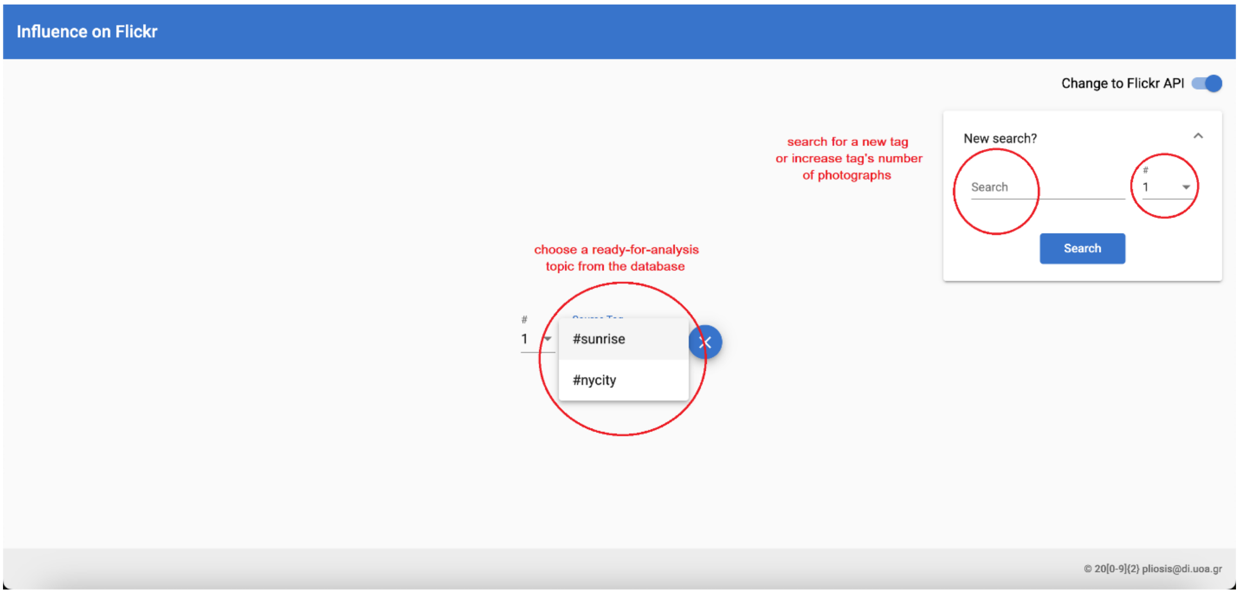
Engine 205 & H&L 118
True heroes of NYC, H&amp;L truck 118 on 9/11 last know pictures of the Ladder was passing the Brooklyn bridge anchorage, truck and squad were never seen again. Just mangled peices of the truck were found 38 days after digging. 🙏 Greater love hath no man than this, that a man lay down his life for his fellow man.
donmai651 | 146838317@N05
#fdny #firehouse #brooklyn #nysbravest #ny #canonr #nycity #canonphotography #iloveny #ngc #streetviews #eosr #newyorkcity #bklyn #ladder118 #engineco205 #brooklynheights #nyc #trueheros #nystateofmind #thebigapple
🏷 21
👁 838
♥ 26
💬 3

Should I stay or should I go
#1 Train Cortlandt St. Pic invokes memories of my Aviation H.S School days, even though city provided subway pass it was the thing to jump the turnstile.
donmai651 | 146838317@N05
#subway #newyorknewyork #1train #iloveny #turnstile #canonphotography #eosr #nycity #ngc #nyny #ef2470f28iiusm #wtccortlandt #backinthenygroove #canonr #thebigapple #littletownblues #nystateofmind
🏷 17
👁 482
♥ 14
💬 1

Figure 28 Simple user interface to explore Flickr APIs

Choosing Angular framework [43] with Angular Material UI component library [44] allows us to build a highly interactive, scalable, and maintainable web client that enhances user experience and supports the complex data interactions required by our solution. Angular offers several benefits as a web client programming framework such as:

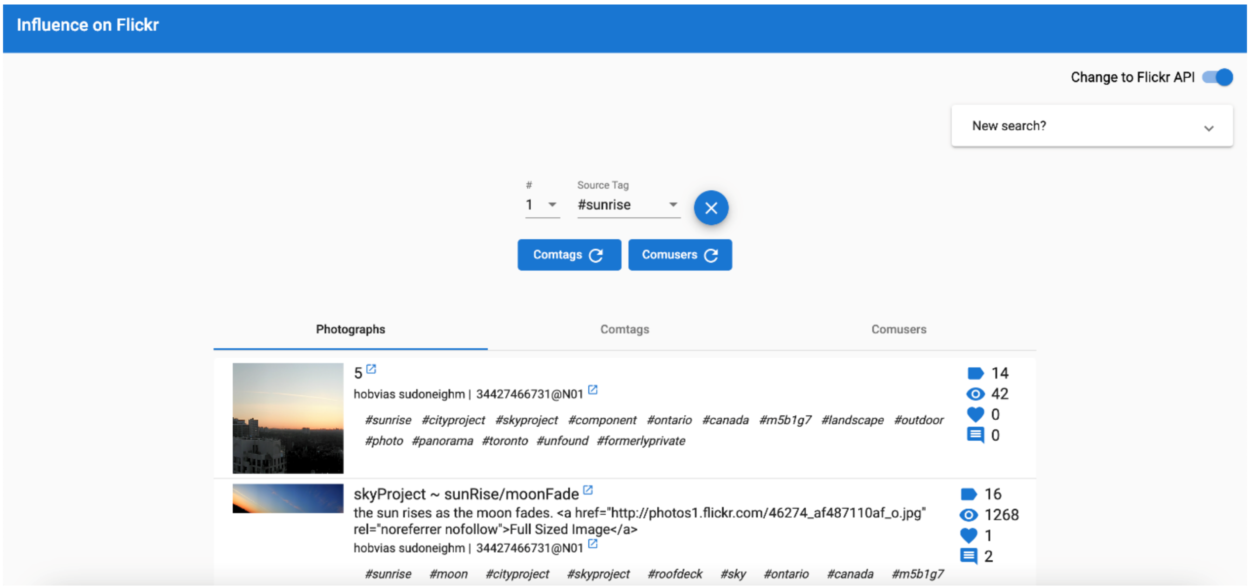
- ➢ Easy setup and rapid development
- ➢ Two-way data binding
- ➢ Component-based architecture
- ➢ Efficient Single Page Application (SPA) Support
- ➢ Built-in dependency injection



search for a new tag or increase tag's number of photographs

choose a ready-for-analysis topic from the database

Figure 29 Simple user interface to interact with Influence Metrics APIs 1/2



5
hobvias sudoneighm | 34427466731@N01
#sunrise #cityproject #skyproject #component #ontario #canada #m5b1g7 #landscape #outdoor #photo #panorama #toronto #unfound #formerlyprivate
14
42
0
0

skyProject ~ sunRise/moonFade
the sun rises as the moon fades. <a href="http://photos1.flickr.com/46274_af487110af_o.jpg" rel="noreferrer nofollow">Full Sized Image</a>
hobvias sudoneighm | 34427466731@N01
#sunrise #moon #cityproject #skyproject #roofdeck #sky #ontario #canada #m5b1g7
16
1268
1
2

Figure 30 Simple user interface to interact with Influence Metrics APIs 2/2

Angular is backed by Google and has a large, active community, offering extensive documentation, libraries, and third-party integrations. This support helps ensure that you can easily find solutions, resources, and tools. Angular's Command Line Interface (CLI) offers powerful tools for project initialization, building, testing, and deployment, making the development process faster and more efficient. This reduces the complexity of configuration, allowing you to focus on coding the core functionality.

Angular's two-way data binding automatically synchronizes the model and the view, reducing the need for manual updates and ensuring that changes in the UI are instantly reflected in the application's data, and vice versa. The component-based structure helps us to break down the UI into modular, reusable elements, promoting cleaner code and easier maintainability.

6.3 The Database

The database architecture for the application is built on a relational database, ensuring data is stored in well-structured tables with clear relationships between entities such as users, photos, tags, favs and comments. This approach provides consistency and integrity through the use of constraints, making it ideal for managing complex datasets.

Hibernate ORM (Object-Relational Mapping) is used to bridge the gap between the object-oriented service layer and the relational database. Hibernate automatically maps Java objects to database tables, simplifying CRUD operations and allowing developers to work with data in an object-oriented manner without writing extensive SQL queries.

We choose MySQL [45] as the solution’s relational database. MySQL is a popular choice for databases in web applications due to its reliability, performance, and ease of use. As a relational database, it supports structured data storage with robust features like foreign keys, indexing, and transactions, making it ideal for handling complex relationships and maintaining data integrity.

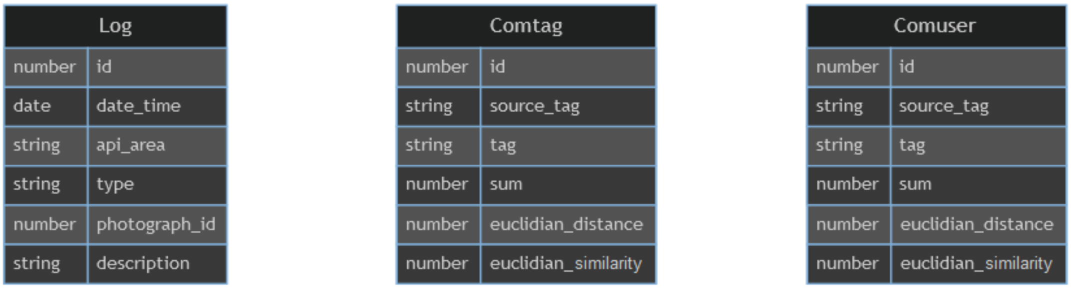
MySQL is also known for its strong community support, extensive documentation, and compatibility with a wide range of platforms, including Spring Boot and Hibernate ORM, which seamlessly integrate with it for data persistence. Additionally, MySQL offers features like replication, high availability, and performance optimization tools, ensuring your application can meet demanding data access and storage requirements while maintaining fast response times. Its open-source nature and commercial versions provide flexibility for both small-scale projects and enterprise-level solutions.

6.3.1 Entity-Relationship (ER) diagram

The first three tables are integral to logging activities and calculations for the recommendation engine.

System Log table is used to track interactions and operations within the system, providing a log of all different system events and user actions for debugging, reviewing, and monitoring purposes, that is also timestamped.

Comtag and comusers tables store the results of tag-based and user-based similarity calculations, using Euclidean distance to measure the similarity.



Log	
number	id
date	date_time
string	api_area
string	type
number	photograph_id
string	description

Comtag	
number	id
string	source_tag
string	tag
number	sum
number	euclidian_distance
number	euclidian_similarity

Comuser	
number	id
string	source_tag
string	tag
number	sum
number	euclidian_distance
number	euclidian_similarity

Figure 31 Calculated data stored in related data tables

Below is the entity-relationship diagram that represents the basic database schema for our solution, capturing the relationships between users, photos, tags, topics, and interactions like favorites and comments.

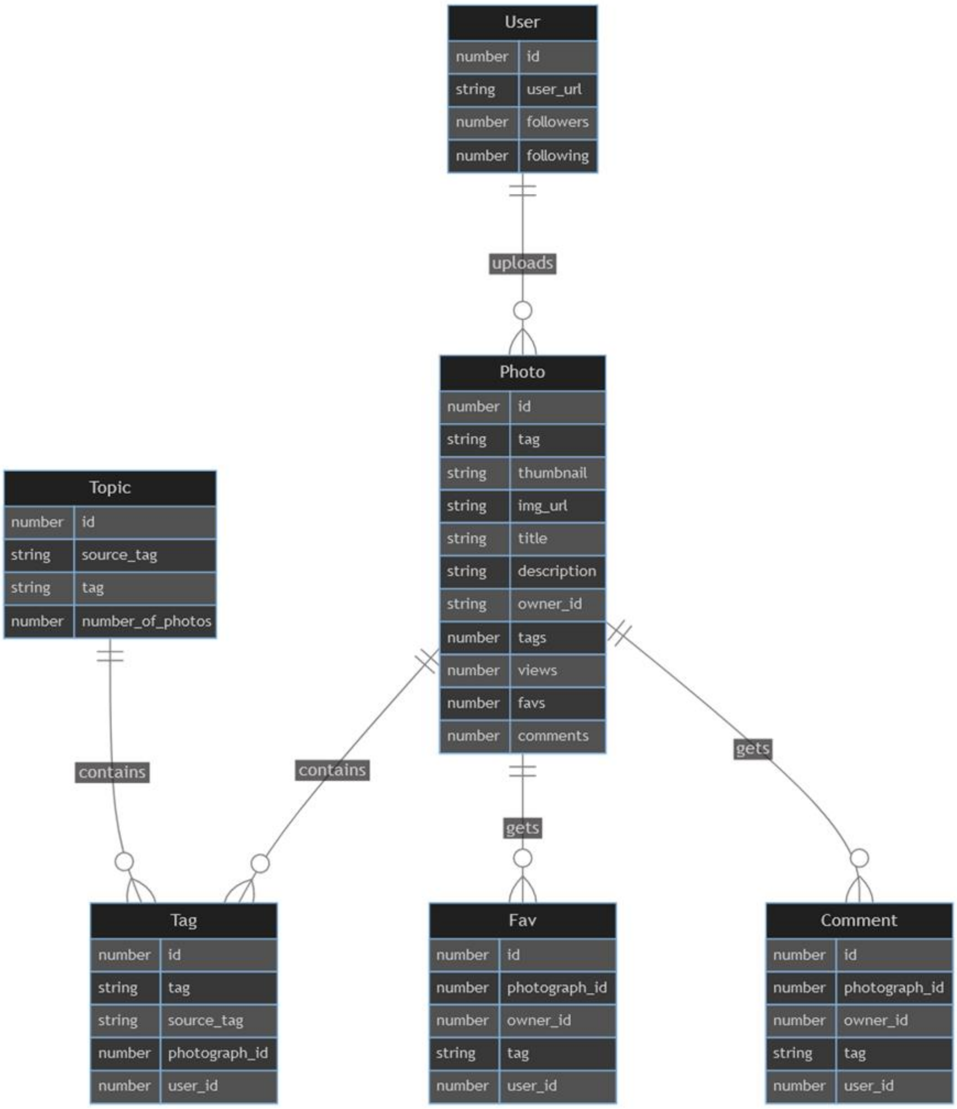


Figure 32 Solution's basic Entity Relationship (ER) Diagram

Relationships summary:

- *User to Photo*: A one-to-many relationship, where one user can upload many photos.
- *Photo to Tag*: A one-to-many relationship, where a photo can have multiple tags associated with it.
- *Photo to Fav*: A one-to-many relationship, where one photo can have multiple favorites.
- *Photo to Comment*: A one-to-many relationship, where one photo can have multiple comments.
- Topic to Tag: A one-to-many relationship, where one topic can have multiple tags that belong to it.

6.4 Recommendation Engine’s (RE) sequence diagrams

The sequence diagrams that follow illustrate our solution's interactions between end-users, web client, back-end service layer, Flickr APIs etc. and they describe how – and in what order – requests and responses are sent.

The *web client* operates as the interface for the *end-user* to search photos by tag and request additional calculations (common tags and users). The *back-end service layer* handles communication with *Flickr APIs*, manages the data (CRUD operations), and stores them in the *app database*.

After that, users can examine the data and perform on-demand calculations, with results displayed in a user-friendly format on the *web client* and/or a *related dashboard*.

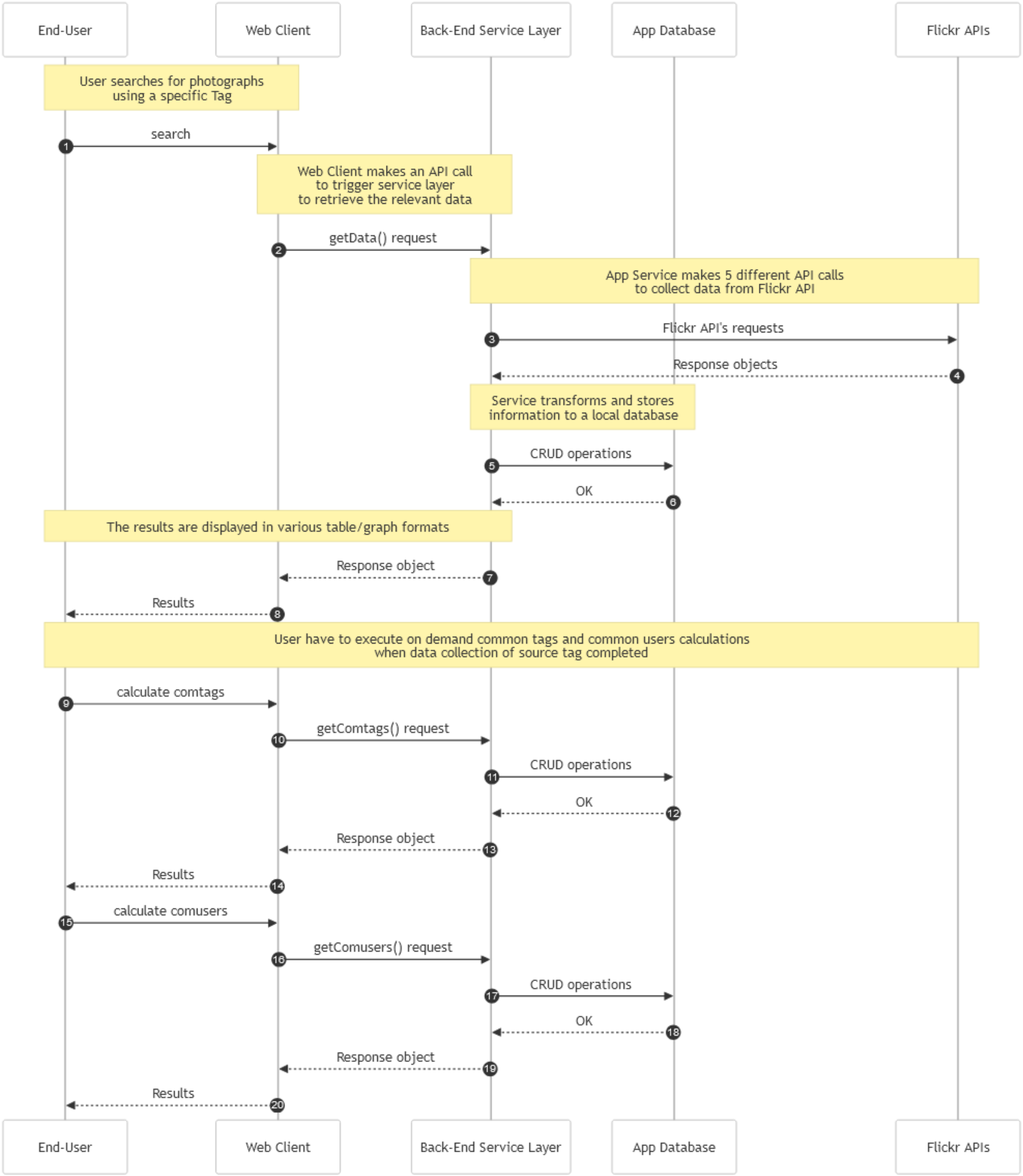

End-User
Web Client
Back-End Service Layer
App Database
Flickr APIs
User searches for photographs using a specific Tag
1 search
Web Client makes an API call to trigger service layer to retrieve the relevant data
2 getData() request
App Service makes 5 different API calls to collect data from Flickr API
3 Flickr API's requests
4 Response objects
Service transforms and stores information to a local database
5 CRUD operations
6 OK
The results are displayed in various table/graph formats
7 Response object
8 Results
User have to execute on demand common tags and common users calculations when data collection of source tag completed
9 calculate comtags
10 getComtags() request
11 CRUD operations
12 OK
13 Response object
14 Results
15 calculate comusers
16 getComusers() request
17 CRUD operations
18 OK
19 Response object
20 Results
End-User
Web Client
Back-End Service Layer
App Database
Flickr APIs

Figure 33 Data collection regarding a specific tag and RE’s calculations to find the most similar tags

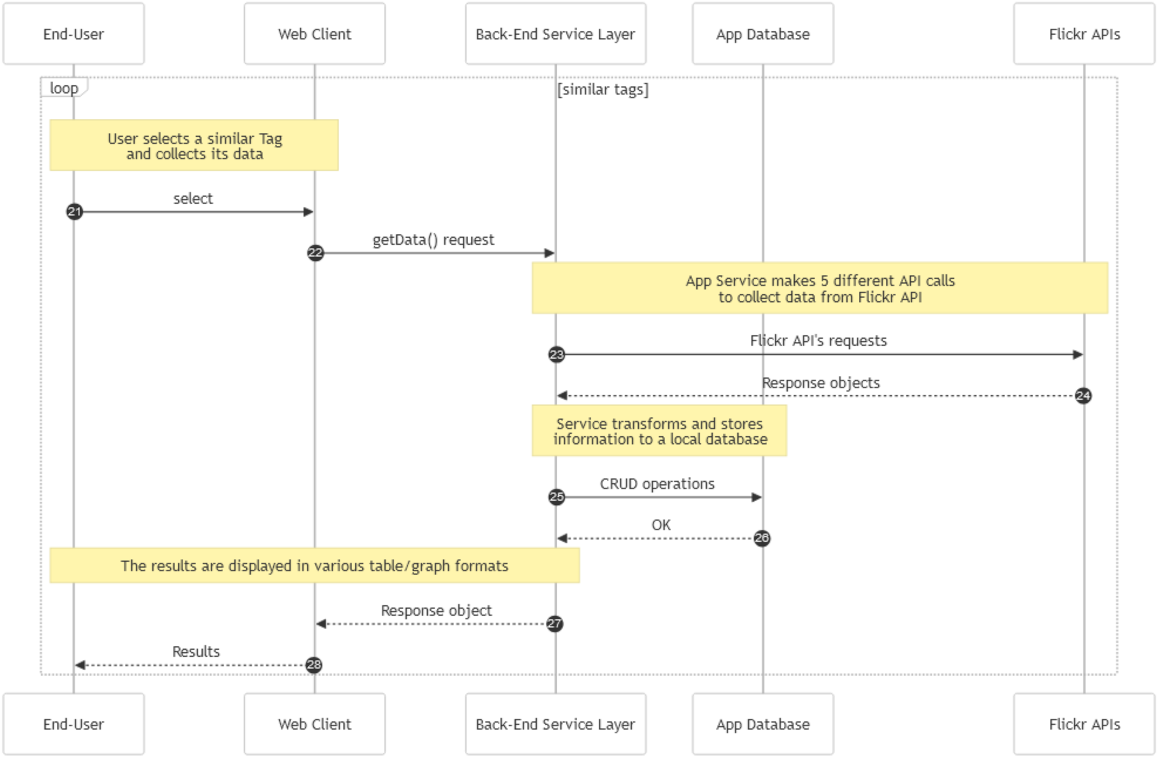
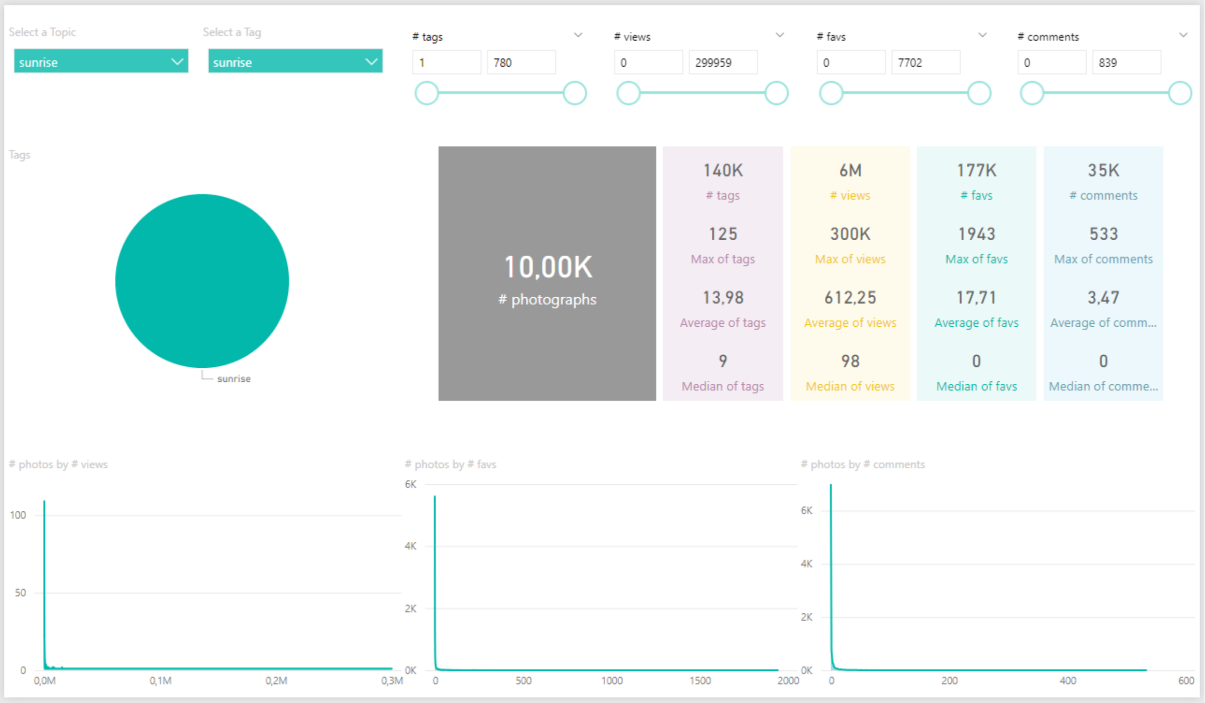


Figure 34 Data collection regarding a specific topic (multiple similar tags)

7. RESULTS AND EVALUATION

We first chose to search for the latest 10.000 public photos which include the tag *#sunrise* among others, and we collect all of them, along with their details.

The bellow dashboard offers a clear view of how frequently photos tagged with *#sunrise* are viewed, favorited, and commented on. It also highlights that only a small number of photos receive significantly higher interactions and engagement, while the majority receive less attention.



Select a Topic: sunrise

Select a Tag: sunrise

tags: 1 – 780

views: 0 – 299959

favs: 0 – 7702

comments: 0 – 839

10,00K
# photographs

140K # tags	6M # views	177K # favs	35K # comments
125 Max of tags	300K Max of views	1943 Max of favs	533 Max of comments
13,98 Average of tags	612,25 Average of views	17,71 Average of favs	3,47 Average of comm...
9 Median of tags	98 Median of views	0 Median of favs	0 Median of comme...

Figure 35 Insights related to the #sunrise tag

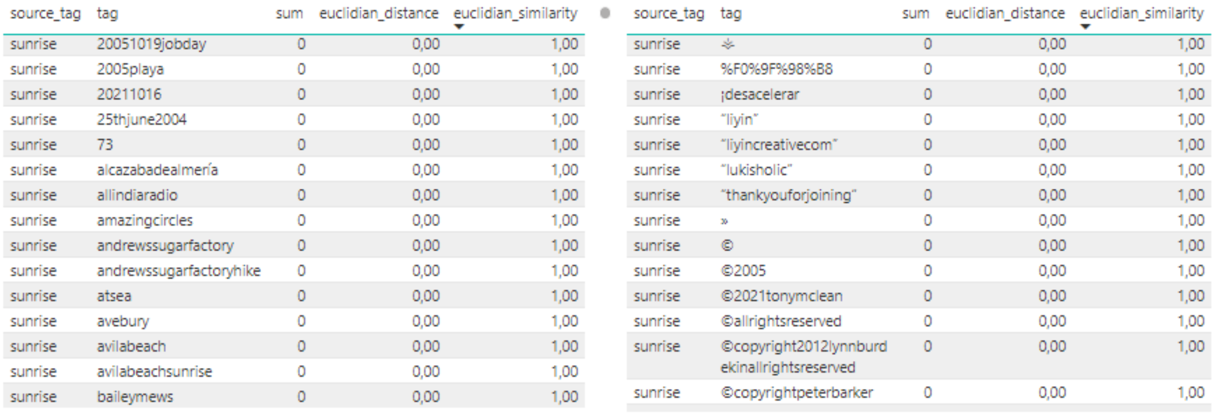
Some interesting statistics are the following:

- A total of ~140,000 tags included in the dataset, almost ~14 per image and the count of distinct tags are 17,749.
- Over 6 million total views across the dataset, the most viewed photo has 300,000, the average views are ~600 and median are only 98, which means significantly lower interest for more than half of dataset.
- Over 170,000 total favorites across the dataset, the most favorite photo gets 1,943 favorites, average favorites per photo are only ~17 and median are 0(!), which means zero endorsement for more than half of dataset.
- Over 35,000 total comments on photos, the most commented photo gets 533 comments, average comments per photo are only ~3.5 and median are 0(!) again, which means zero involvement for more than half of dataset.

The three bottom graphs show us the distribution of photos by the number of views, favorites, and comments. The histograms revealed a significant disparity in the interactions (views, favorites, and comments) across photographs. This phenomenon is often described as the long tail distribution [46].

The vast majority of photos appear to have very few interactions. This is the ‘long tail’ where a large number of entities exist, but they individually receive little engagement. Only a small number of photos receive a disproportionate amount of attention, favorites, and views. These photos dominate interactions.

The following tables display calculated results of the top 15 tags per case, based on the two criteria we selected in the second step of our process to determine similarity. As we can easily notice, we have numerous 'similar' tags, which are likely producing a false result. In the first table we have the common tags, where two distinct tags have been used the same number times together with our main tag, and in the second table we have the level of their usage by users who have tagged their photos with them in common.



source_tag	tag	sum	euclidian_distance	euclidian_similarity
sunrise	20051019jobday	0	0,00	1,00
sunrise	2005playa	0	0,00	1,00
sunrise	20211016	0	0,00	1,00
sunrise	25thjune2004	0	0,00	1,00
sunrise	73	0	0,00	1,00
sunrise	alcazabadealmería	0	0,00	1,00
sunrise	allindiaradio	0	0,00	1,00
sunrise	amazingcircles	0	0,00	1,00
sunrise	andrewssugarfactory	0	0,00	1,00
sunrise	andrewssugarfactoryhike	0	0,00	1,00
sunrise	atsea	0	0,00	1,00
sunrise	avebury	0	0,00	1,00
sunrise	avilabeach	0	0,00	1,00
sunrise	avilabeachsunrise	0	0,00	1,00
sunrise	baileymews	0	0,00	1,00

source_tag	tag	sum	euclidian_distance	euclidian_similarity
sunrise	⚘	0	0,00	1,00
sunrise	%F0%9F%98%B8	0	0,00	1,00
sunrise	¡desacelerar	0	0,00	1,00
sunrise	"liyin"	0	0,00	1,00
sunrise	"liyincreativecom"	0	0,00	1,00
sunrise	"lukisholic"	0	0,00	1,00
sunrise	"thankyouforjoining"	0	0,00	1,00
sunrise	»	0	0,00	1,00
sunrise	©	0	0,00	1,00
sunrise	©2005	0	0,00	1,00
sunrise	©2021tonymclean	0	0,00	1,00
sunrise	©allrightsreserved	0	0,00	1,00
sunrise	©copyright2012lynnburdekinallrightsreserved	0	0,00	1,00
sunrise	©copyrightpeterbarker	0	0,00	1,00

Figure 36 Similarity by common tags versus similarity by usage with similar frequency

The observation about the perfect similarity scores (1.00) and the lack of meaningful interaction is directly related to the nature of our dataset, which is solely based on the tag *#sunrise*.

Since our dataset in the first step collects only recent photos tagged with *#sunrise* this inherently limits the diversity and depth of data available for comparing *#sunrise* to other tags. This creates a homogeneous dataset where every photo has the common *#sunrise* tag. As a result, comparisons with other tags are limited to co-occurrences with *#sunrise* rather than the broader context of how these tags appear across all Flickr photos.

Many of the tags listed in the above tables, such as dates (*#25thjune2004*), and general terms (*#copyright*), are likely user-generated or tied to specific contexts that aren't necessarily indicative of photo content. Since these tags do not have broad applicability or strong conceptual links with *#sunrise* (or any other widely used tags like *#beach, #sun*, or *#sky*), their inclusion in the similarity calculations does not contribute significantly to meaningful patterns of co-occurrence. Instead, they might affect the dataset with highly specific, less relevant tags.

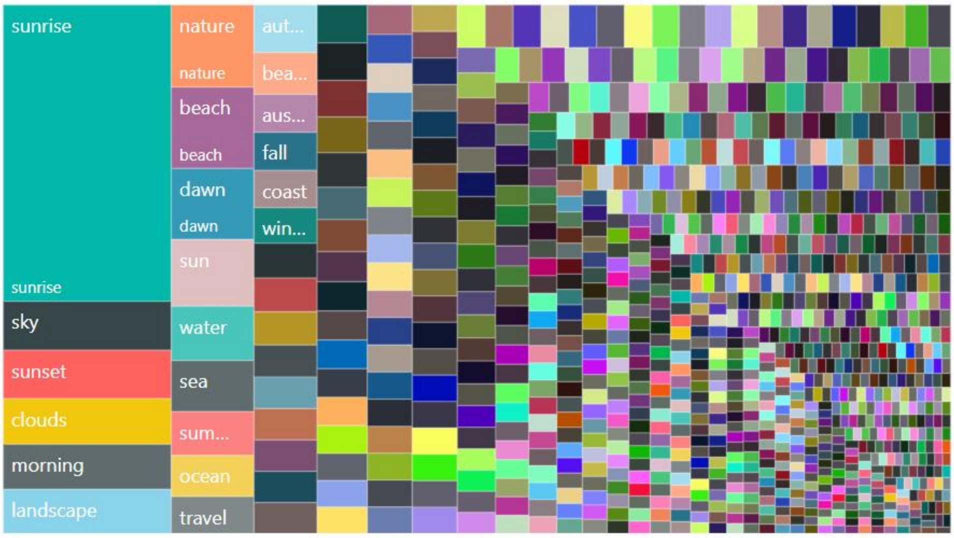


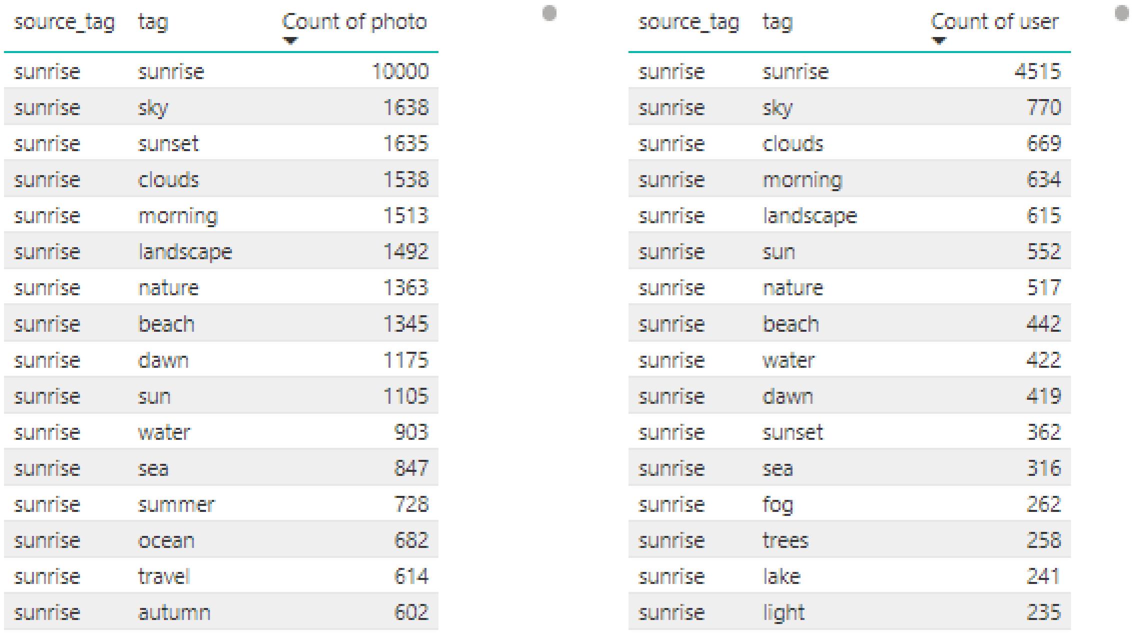
Figure 37 Tree map chart of most used tags

Tags that appear in only a few photos of our dataset or are used by just a small number of users are often outliers or highly specific to particular contexts. These tags do not contribute to meaningful content-based relationships and can skew the analysis by adding noise to the dataset.

For that reason, we suggest the following approach. By filtering out tags that appear in only one or two photos or are used by just a few users, we expect to:

- reduce irrelevant data,
- improve the accuracy by focusing on commonly used tags, and
- avoid situations where all tags have a perfect similarity score due to lack of diversity in the data.

Furthermore, removing infrequently used tags allows our system to focus on tags that have been applied to multiple photos by different users. These frequently occurring tags are more likely to represent meaningful concepts and will contribute to more accurate and useful similarity calculations.



source_tag	tag	Count of photo
sunrise	sunrise	10000
sunrise	sky	1638
sunrise	sunset	1635
sunrise	clouds	1538
sunrise	morning	1513
sunrise	landscape	1492
sunrise	nature	1363
sunrise	beach	1345
sunrise	dawn	1175
sunrise	sun	1105
sunrise	water	903
sunrise	sea	847
sunrise	summer	728
sunrise	ocean	682
sunrise	travel	614
sunrise	autumn	602

source_tag	tag	Count of user
sunrise	sunrise	4515
sunrise	sky	770
sunrise	clouds	669
sunrise	morning	634
sunrise	landscape	615
sunrise	sun	552
sunrise	nature	517
sunrise	beach	442
sunrise	water	422
sunrise	dawn	419
sunrise	sunset	362
sunrise	sea	316
sunrise	fog	262
sunrise	trees	258
sunrise	lake	241
sunrise	light	235

Figure 38 Tags by frequency of occurrence and by distinct users that used them

Popular tags that are used in multiple photos by different users are more likely to capture common themes or trends. For example, tags like *#sky, #clouds, #beach*, or *#sun* are likely to be applied by many users to different photos of *#sunrise* and reflect real-world relationships between visual content and tag usage. This approach will lead to a more robust and informative analysis.

Also, by applying equal weights, we ensured that each factor was treated with equal importance, resulting in a balanced similarity measure across the following elements, Co-occurrence frequency, User behavior, Frequency of usage and User engagement.

Based on the above information, we recommend combining the two similarity measures with the frequency of occurrence (at least X% of our dataset) and the distinct count of users who used them (at least Y% of distinct users) into a single, equally weighted measure for all related tags.

$$WMS\ (T) = w_a * sim(t_i, t_j)_{comtags} + w_b * sim(t_i, t_j)_{user} + w_c * \left(\frac{NoPhotos(T)}{TotalNoPhotos}\right) + w_d * \left(\frac{DistinctNoUsers(T)}{TotalDistinctNoUsers}\right)$$

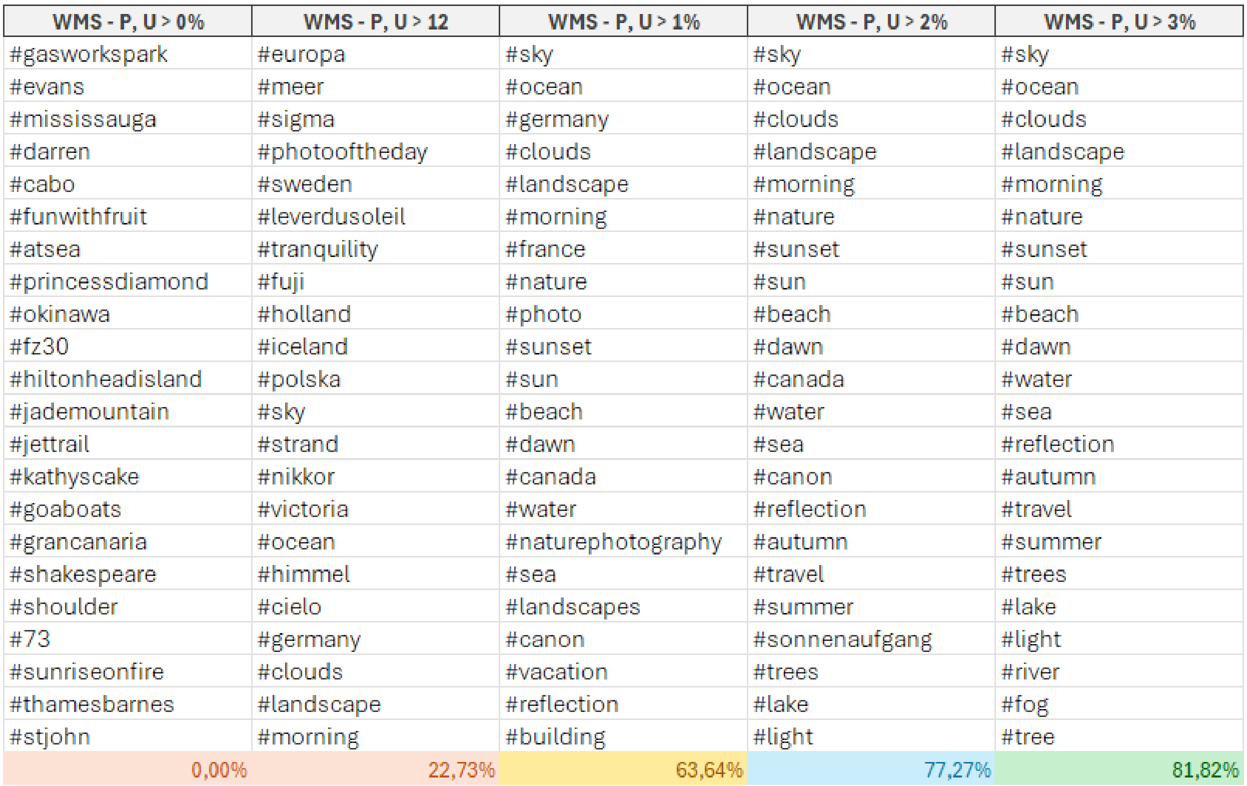
Where:

- $w_a,\ w_b,\ w_c,\ w_d$ are the weights we use for the four measures, with $w_a + w_b + w_c + w_d = 1$, and
- $w_a = w_b = w_c = w_d = \frac{1}{4}$

For evaluation purposes we will compare our results with Flickr APIs' related tags endpoint. This service provides a list of tags commonly associated with a given tag, in our case, *#sunrise*, based on clustered usage patterns. The related tags are listed in descending order, indicating those at the top.

Here are the Flickr APIs' related tags:
#morning #sky #clouds #sun #water #sea #orange #reflection #ocean #beach #silhouette #nature #tree #blue #dawn #lake #fog #mist #light #sunset #nikon #sand #trees #cloud #waves #red #early #landscape #canon #yellow #winter #pink #sol #boat #mountains #color

Consequently, we analyzed four different scenarios where we filtered out rare or niche tags that appeared in very few photos or were used by only a small number of people, and we found a list of tags which have the highest similarity to the source tag *T*.



WMS - P, U > 0%	WMS - P, U > 12	WMS - P, U > 1%	WMS - P, U > 2%	WMS - P, U > 3%
#gasworkspark	#europa	#sky	#sky	#sky
#evans	#meer	#ocean	#ocean	#ocean
#mississauga	#sigma	#germany	#clouds	#clouds
#darren	#photooftheday	#clouds	#landscape	#landscape
#cabo	#sweden	#landscape	#morning	#morning
#funwithfruit	#leverdusoleil	#morning	#nature	#nature
#atsea	#tranquility	#france	#sunset	#sunset
#princessdiamond	#fuji	#nature	#sun	#sun
#okinawa	#holland	#photo	#beach	#beach
#fz30	#iceland	#sunset	#dawn	#dawn
#hiltonheadisland	#polska	#sun	#canada	#water
#jademountain	#sky	#beach	#water	#sea
#jettrail	#strand	#dawn	#sea	#reflection
#kathyscake	#nikkor	#canada	#canon	#autumn
#goaboats	#victoria	#water	#reflection	#travel
#grancanaria	#ocean	#naturephotography	#autumn	#summer
#shakespeare	#himmel	#sea	#travel	#trees
#shoulder	#cielo	#landscapes	#summer	#lake
#73	#germany	#canon	#sonnenaufgang	#light
#sunriseonfire	#clouds	#vacation	#trees	#river
#thamesbarnes	#landscape	#reflection	#lake	#fog
#stjohn	#morning	#building	#light	#tree
0,00%	22,73%	63,64%	77,27%	81,82%

Figure 39 Most similar tags based on WMS with threshold in at least 0%, 12, 1%, 2% and 3% respectively

Our analysis includes the following thresholds for photos *P* and distinct users *U*:

a. WMS - P, U > 0% - **Match Rate**: 0.00%

   Includes the widest variety of tags, some of which appear highly specific to individual users or contexts (e.g., *#gasworkspark, #darren, #funwithfruit*). Most similar tags in this scenario are unlikely to be generally relevant, which results in no overlapping with Flickr's related tags.

b. WMS - P, U > 12 - **Match Rate**: 22.73%
   As the threshold increased, the tag set narrows and becomes more relevant to broader contexts (e.g., *#europa, #sigma, #fuji*). There's a modest improvement in overlap with Flickr’s related tags, reflecting a shift toward more general and widely used tags.

c. WMS - P, U > 1% - **Match Rate**: 63.64%
   Started to filter out user-specific or very niche tags. The resulting tags includes many nature and landscape-related terms (e.g., *#sky, #ocean, #clouds, #landscape, #morning*). This is a significant improvement in match rate, showing that the tags identified by WMS are increasingly aligned with Flickr’s related tags.

d. WMS - P, U > 2% - **Match Rate**: 77.27%
   The tags became even more focused on general nature, landscape, and weather-related terms (e.g., *#sun, #beach, #water, #sunset*). As the threshold increases, the tags become more generalized, and the overlap with Flickr's related tags improves further.

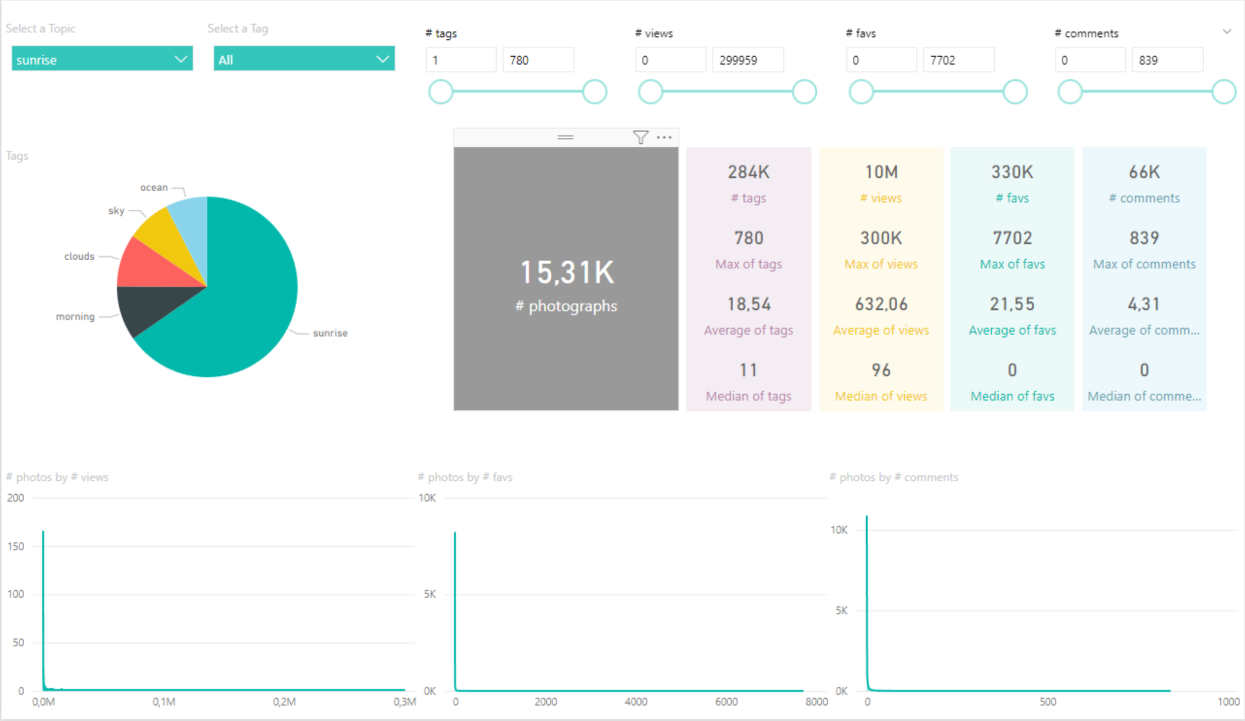
e. WMS - P, U > 3% - **Match Rate**: 81.82%
   The tags are almost entirely composed of widely applicable, commonly used tags related to nature, weather, and landscapes (e.g., *#ocean, #clouds, #sun, #reflection*). This is the highest match rate, indicating that these tags are the most similar to what Flickr’s API would recommend based on clustered usage analysis.

The match rate measures how well the tags identified in each scenario correspond with Flickr's related tags for "sunrise" (e.g., *#morning, #sky, #clouds, #sun*). As the thresholds become more restrictive, the relevance of the tags increases, resulting in a higher match rate with Flickr’s related tags. Lower thresholds (0%, >12 users) include many user-specific or niche tags, leading to poor overlap with general-use tags like those provided by Flickr’s API. Higher thresholds (1%, 2%, 3%) filter out these less-relevant tags, resulting in more general, widely used tags that have a much higher match rate.

The high match rates at 2% and 3% show that these thresholds are effective in aligning with general public usage patterns as identified by Flickr’s clustered analysis, making them suitable for robust, generalizable tag recommendations.

Before calculating and presenting the final influence metrics, we selected four additional tags (*#sky, #morning, #clouds* and *#ocean*) that match with the Flickr’s most related tags, to collect their data, which became part of our research on the sunrise topic.

Our dataset now includes over 15,000 photos in total, which is an increase of about 53%.



Select a Topic: sunrise

Select a Tag: All

# tags		# views		# favs		# comments	
1	780	0	299959	0	7702	0	839

15,31K
# photographs

284K	10M	330K	66K
# tags	# views	# favs	# comments
780	300K	7702	839
Max of tags	Max of views	Max of favs	Max of comments
18,54	632,06	21,55	4,31
Average of tags	Average of views	Average of favs	Average of comm...
11	96	0	0
Median of tags	Median of views	Median of favs	Median of comme...

Figure 40 Insights related to the #sunrise topic

Some updated statistics are the following:

- Total tags have more than doubled (from 140K to 284K). Average tags per photo increased from 13.98 to 18.54. Median tags per photo increased slightly from 9 to 11.
- Total views increased from 6M to 10M. Average views per photo increased slightly from 612.25 to 632.06. Median views per photo remained almost the same (98 to 96).
- Total favorites increased from 177K to 330K. Average favorites per photo increased from 17.71 to 21.55. Median favorites per photo increased from 0 to 4, showing a broader engagement.
- Total comments nearly doubled, from 35K to 66K. Average comments per photo increased from 3.47 to 4.31. Median comments per photo remained at 0, which means zero involvement for more than half of dataset, again.

The distributions in both datasets show a similar pattern: a small percentage of photos receive the majority of interactions (views, favorites, comments), while most photos have fewer interactions. More than 60% of the total photos capped simultaneously with a maximum value of 200 views, 5 favs and 5 comments. Across all metrics (views, favorites, comments), the distribution is heavily skewed toward the lower end.

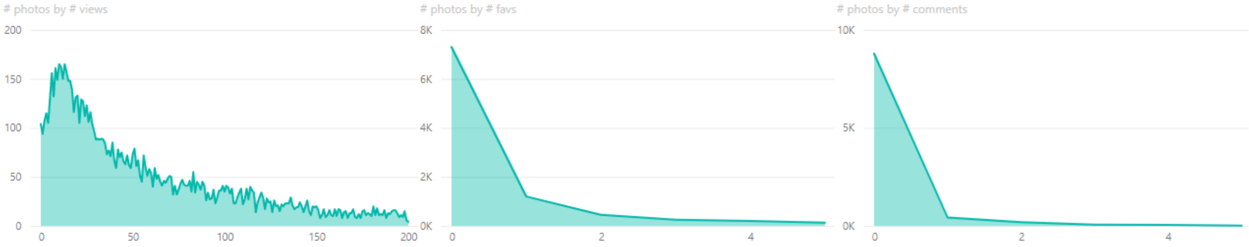
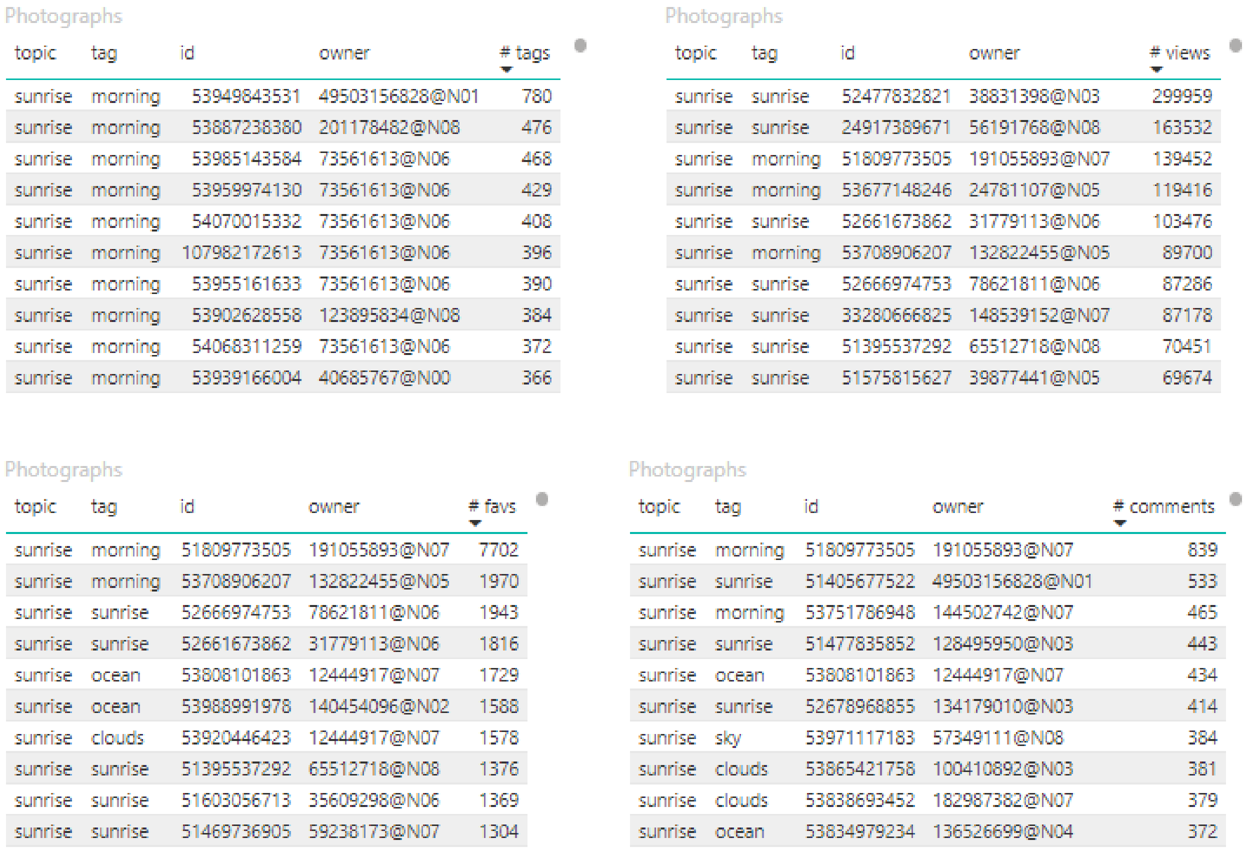


Figure 41 Distribution histograms

Leading photos and users regarding #sunrise topic



Photographs

topic	tag	id	owner	# tags
sunrise	morning	53949843531	49503156828@N01	780
sunrise	morning	53887238380	201178482@N08	476
sunrise	morning	53985143584	73561613@N06	468
sunrise	morning	53959974130	73561613@N06	429
sunrise	morning	54070015332	73561613@N06	408
sunrise	morning	107982172613	73561613@N06	396
sunrise	morning	53955161633	73561613@N06	390
sunrise	morning	53902628558	123895834@N08	384
sunrise	morning	54068311259	73561613@N06	372
sunrise	morning	53939166004	40685767@N00	366

Photographs

topic	tag	id	owner	# views
sunrise	sunrise	52477832821	38831398@N03	299959
sunrise	sunrise	24917389671	56191768@N08	163532
sunrise	morning	51809773505	191055893@N07	139452
sunrise	morning	53677148246	24781107@N05	119416
sunrise	sunrise	52661673862	31779113@N06	103476
sunrise	morning	53708906207	132822455@N05	89700
sunrise	sunrise	52666974753	78621811@N06	87286
sunrise	sunrise	33280666825	148539152@N07	87178
sunrise	sunrise	51395537292	65512718@N08	70451
sunrise	sunrise	51575815627	39877441@N05	69674

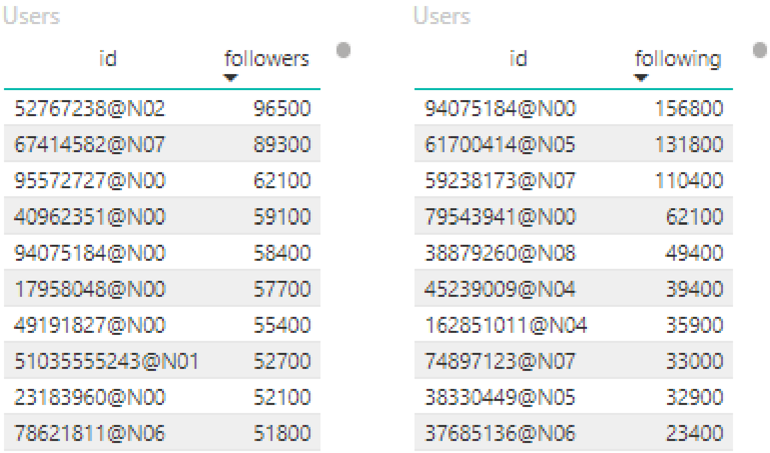
Photographs

topic	tag	id	owner	# favs
sunrise	morning	51809773505	191055893@N07	7702
sunrise	morning	53708906207	132822455@N05	1970
sunrise	sunrise	52666974753	78621811@N06	1943
sunrise	sunrise	52661673862	31779113@N06	1816
sunrise	ocean	53808101863	12444917@N07	1729
sunrise	ocean	53988991978	140454096@N02	1588
sunrise	clouds	53920446423	12444917@N07	1578
sunrise	sunrise	51395537292	65512718@N08	1376
sunrise	sunrise	51603056713	35609298@N06	1369
sunrise	sunrise	51469736905	59238173@N07	1304

Photographs

topic	tag	id	owner	# comments
sunrise	morning	51809773505	191055893@N07	839
sunrise	sunrise	51405677522	49503156828@N01	533
sunrise	morning	53751786948	144502742@N07	465
sunrise	sunrise	51477835852	128495950@N03	443
sunrise	ocean	53808101863	12444917@N07	434
sunrise	sunrise	52678968855	134179010@N03	414
sunrise	sky	53971117183	57349111@N08	384
sunrise	clouds	53865421758	100410892@N03	381
sunrise	clouds	53838693452	182987382@N07	379
sunrise	ocean	53834979234	136526699@N04	372

Figure 42 Top 10 photos by number of tags, views, favs and comments



Users

id	followers
52767238@N02	96500
67414582@N07	89300
95572727@N00	62100
40962351@N00	59100
94075184@N00	58400
17958048@N00	57700
49191827@N00	55400
51035555243@N01	52700
23183960@N00	52100
78621811@N06	51800

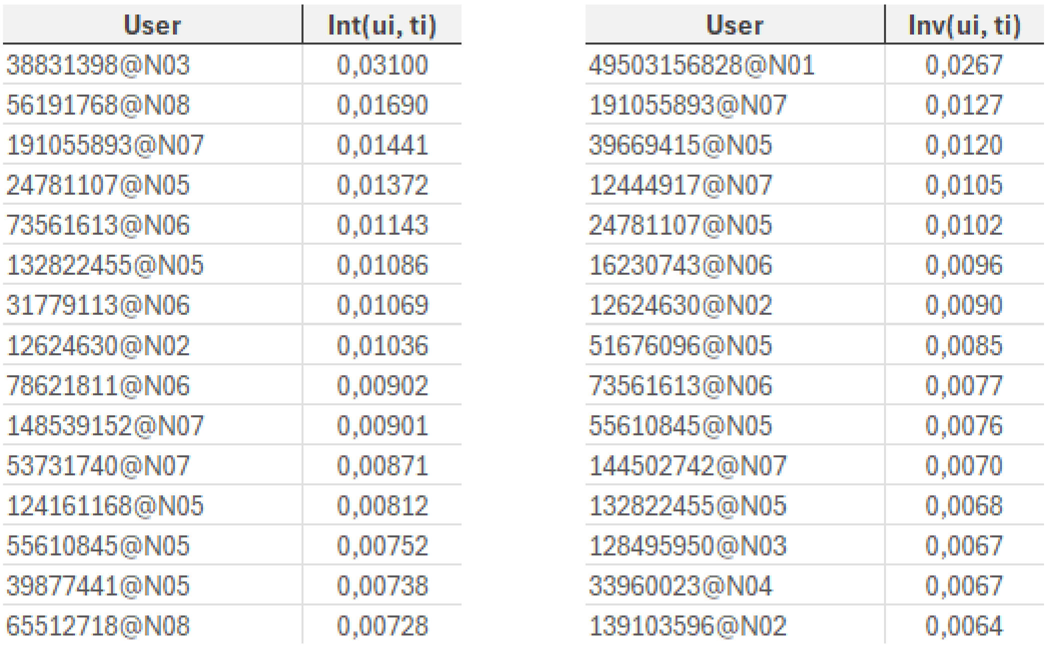
Users

id	following
94075184@N00	156800
61700414@N05	131800
59238173@N07	110400
79543941@N00	62100
38879260@N08	49400
45239009@N04	39400
162851011@N04	35900
74897123@N07	33000
38330449@N05	32900
37685136@N06	23400

Figure 43 Most-followed users

Most influential users regarding #sunrise topic

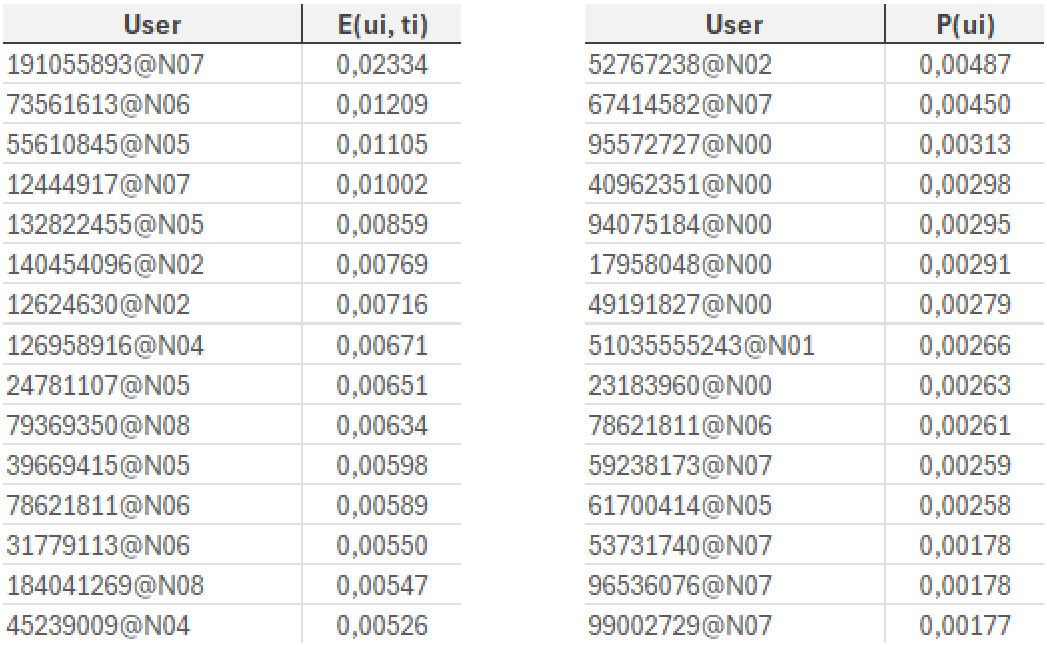
Here are the top results in the four different influence factors, as outlined in the third chapter of this thesis, along with the final Influence score for the most influential user.



User	Int(ui, ti)
38831398@N03	0,03100
56191768@N08	0,01690
191055893@N07	0,01441
24781107@N05	0,01372
73561613@N06	0,01143
132822455@N05	0,01086
31779113@N06	0,01069
12624630@N02	0,01036
78621811@N06	0,00902
148539152@N07	0,00901
53731740@N07	0,00871
124161168@N05	0,00812
55610845@N05	0,00752
39877441@N05	0,00738
65512718@N08	0,00728

User	Inv(ui, ti)
49503156828@N01	0,0267
191055893@N07	0,0127
39669415@N05	0,0120
12444917@N07	0,0105
24781107@N05	0,0102
16230743@N06	0,0096
12624630@N02	0,0090
51676096@N05	0,0085
73561613@N06	0,0077
55610845@N05	0,0076
144502742@N07	0,0070
132822455@N05	0,0068
128495950@N03	0,0067
33960023@N04	0,0067
139103596@N02	0,0064

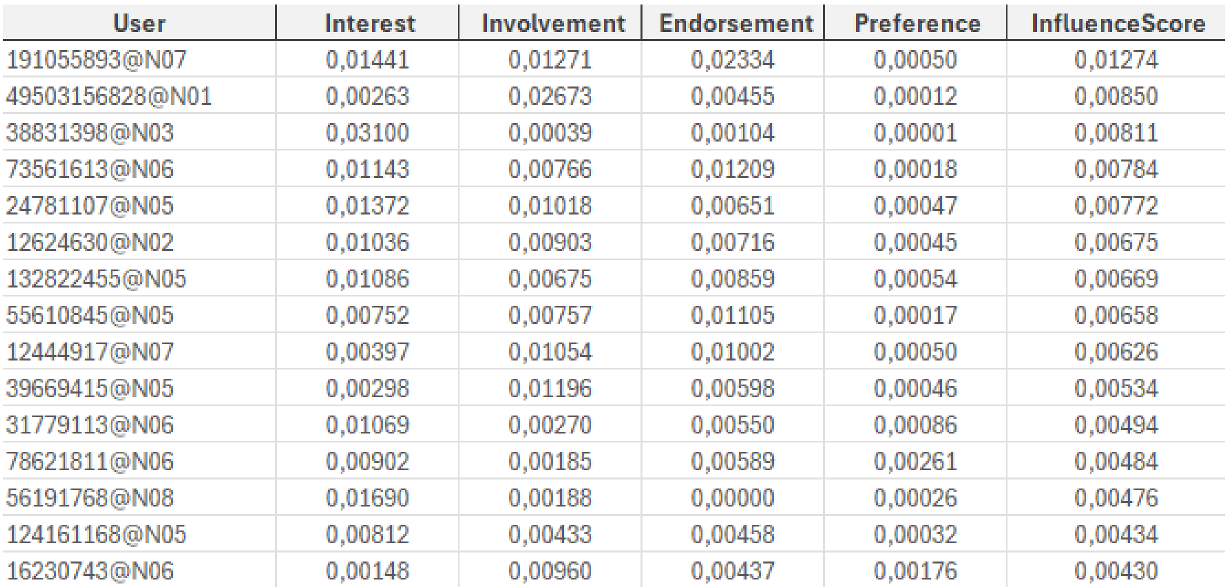
Figure 44 Top 15 Users by Interest and Involvement influence factors



User	E(ui, ti)
191055893@N07	0,02334
73561613@N06	0,01209
55610845@N05	0,01105
12444917@N07	0,01002
132822455@N05	0,00859
140454096@N02	0,00769
12624630@N02	0,00716
126958916@N04	0,00671
24781107@N05	0,00651
79369350@N08	0,00634
39669415@N05	0,00598
78621811@N06	0,00589
31779113@N06	0,00550
184041269@N08	0,00547
45239009@N04	0,00526

User	P(ui)
52767238@N02	0,00487
67414582@N07	0,00450
95572727@N00	0,00313
40962351@N00	0,00298
94075184@N00	0,00295
17958048@N00	0,00291
49191827@N00	0,00279
51035555243@N01	0,00266
23183960@N00	0,00263
78621811@N06	0,00261
59238173@N07	0,00259
61700414@N05	0,00258
53731740@N07	0,00178
96536076@N07	0,00178
99002729@N07	0,00177

Figure 45 Top 15 Users by Endorsement and Preference influence factors



User	Interest	Involvement	Endorsement	Preference	InfluenceScore
191055893@N07	0,01441	0,01271	0,02334	0,00050	0,01274
49503156828@N01	0,00263	0,02673	0,00455	0,00012	0,00850
38831398@N03	0,03100	0,00039	0,00104	0,00001	0,00811
73561613@N06	0,01143	0,00766	0,01209	0,00018	0,00784
24781107@N05	0,01372	0,01018	0,00651	0,00047	0,00772
12624630@N02	0,01036	0,00903	0,00716	0,00045	0,00675
132822455@N05	0,01086	0,00675	0,00859	0,00054	0,00669
55610845@N05	0,00752	0,00757	0,01105	0,00017	0,00658
12444917@N07	0,00397	0,01054	0,01002	0,00050	0,00626
39669415@N05	0,00298	0,01196	0,00598	0,00046	0,00534
31779113@N06	0,01069	0,00270	0,00550	0,00086	0,00494
78621811@N06	0,00902	0,00185	0,00589	0,00261	0,00484
56191768@N08	0,01690	0,00188	0,00000	0,00026	0,00476
124161168@N05	0,00812	0,00433	0,00458	0,00032	0,00434
16230743@N06	0,00148	0,00960	0,00437	0,00176	0,00430

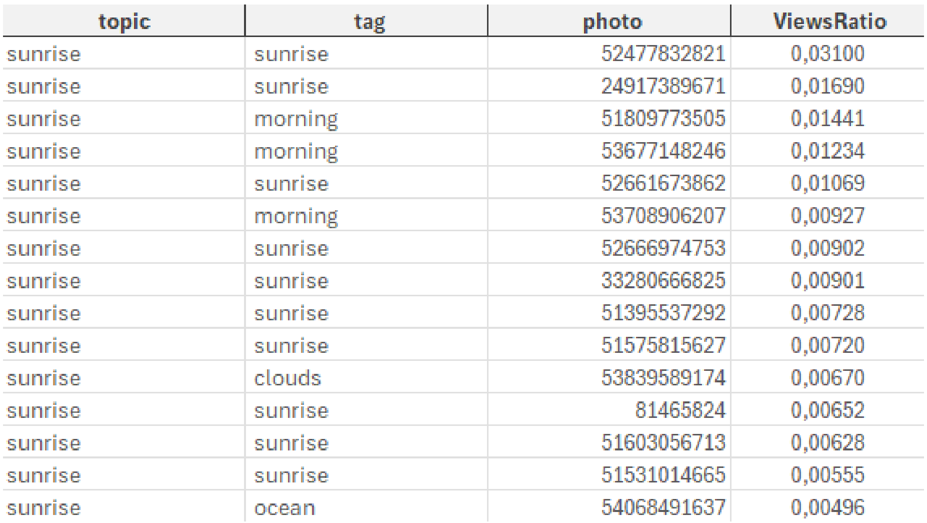
Figure 46 Most influential users regarding #sunrise topic

Key common users are:

- 191055893@N07: Top in Endorsement (0.02334) and high in Interest (0.01441) and Involvement (0.01271). This user is highly influential due to their consistent engagement.
- 49503156828@N01: Dominates Involvement (0.02673) and ranks reasonably in Preference (0.0127) and Interest (0.00263).
- 73561613@N06: High in Interest (0.01143) and Endorsement (0.01209).
- 24781107@N05: Strong in Interest (0.01372) and Involvement (0.01018), with moderate Endorsement.
- 38831398@N03, in contrast, stands out primarily due to their top position in Interest (0.03100), despite not ranking highly in the other influence factors!

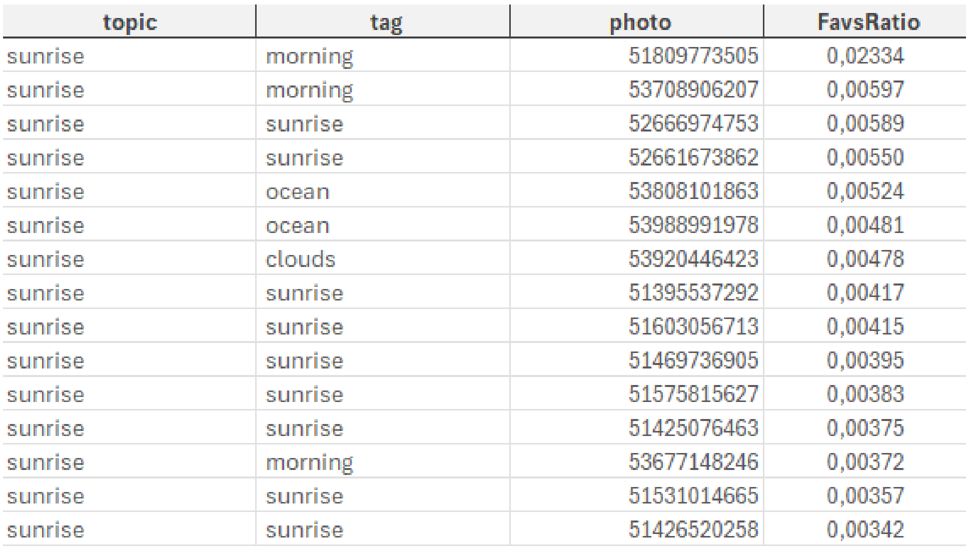
Most influential photos regarding #sunrise topic

Here are the top results in the three different influence factors, as outlined in the third chapter of this thesis, along with the final Influence score for the most influential photo.



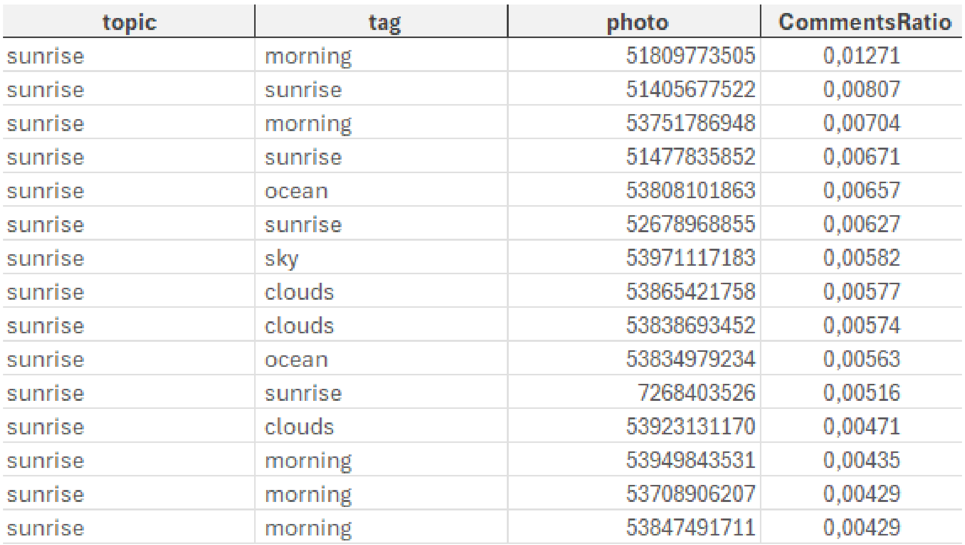
topic	tag	photo	ViewsRatio
sunrise	sunrise	52477832821	0,03100
sunrise	sunrise	24917389671	0,01690
sunrise	morning	51809773505	0,01441
sunrise	morning	53677148246	0,01234
sunrise	sunrise	52661673862	0,01069
sunrise	morning	53708906207	0,00927
sunrise	sunrise	52666974753	0,00902
sunrise	sunrise	33280666825	0,00901
sunrise	sunrise	51395537292	0,00728
sunrise	sunrise	51575815627	0,00720
sunrise	clouds	53839589174	0,00670
sunrise	sunrise	81465824	0,00652
sunrise	sunrise	51603056713	0,00628
sunrise	sunrise	51531014665	0,00555
sunrise	ocean	54068491637	0,00496

Figure 47 Top 15 Photos by views ratio



topic	tag	photo	FavsRatio
sunrise	morning	51809773505	0,02334
sunrise	morning	53708906207	0,00597
sunrise	sunrise	52666974753	0,00589
sunrise	sunrise	52661673862	0,00550
sunrise	ocean	53808101863	0,00524
sunrise	ocean	53988991978	0,00481
sunrise	clouds	53920446423	0,00478
sunrise	sunrise	51395537292	0,00417
sunrise	sunrise	51603056713	0,00415
sunrise	sunrise	51469736905	0,00395
sunrise	sunrise	51575815627	0,00383
sunrise	sunrise	51425076463	0,00375
sunrise	morning	53677148246	0,00372
sunrise	sunrise	51531014665	0,00357
sunrise	sunrise	51426520258	0,00342

Figure 48 Top 15 Photos by favorites ratio

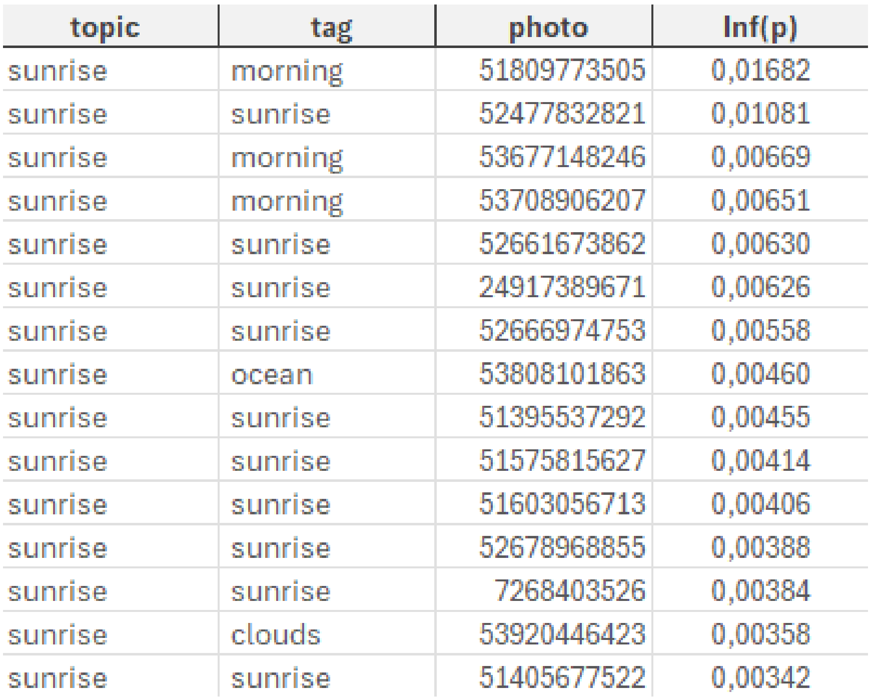


topic	tag	photo	CommentsRatio
sunrise	morning	51809773505	0,01271
sunrise	sunrise	51405677522	0,00807
sunrise	morning	53751786948	0,00704
sunrise	sunrise	51477835852	0,00671
sunrise	ocean	53808101863	0,00657
sunrise	sunrise	52678968855	0,00627
sunrise	sky	53971117183	0,00582
sunrise	clouds	53865421758	0,00577
sunrise	clouds	53838693452	0,00574
sunrise	ocean	53834979234	0,00563
sunrise	sunrise	7268403526	0,00516
sunrise	clouds	53923131170	0,00471
sunrise	morning	53949843531	0,00435
sunrise	morning	53708906207	0,00429
sunrise	morning	53847491711	0,00429

Figure 49 Top 15 Photos by comments ratio

Some key points of the influence factors we use:

- By presenting these metrics together, we gain a comprehensive understanding of the most influential sunrise-related photos, demonstrating how views, favorites, and comments each contribute to the overall impact of these photos within the community.
- Tags like *#morning, #clouds*, and *#ocean* frequently appear in photos with high comments ratios, showing that these elements are likely to generate user engagement and conversations.



topic	tag	photo	Inf(p)
sunrise	morning	51809773505	0,01682
sunrise	sunrise	52477832821	0,01081
sunrise	morning	53677148246	0,00669
sunrise	morning	53708906207	0,00651
sunrise	sunrise	52661673862	0,00630
sunrise	sunrise	24917389671	0,00626
sunrise	sunrise	52666974753	0,00558
sunrise	ocean	53808101863	0,00460
sunrise	sunrise	51395537292	0,00455
sunrise	sunrise	51575815627	0,00414
sunrise	sunrise	51603056713	0,00406
sunrise	sunrise	52678968855	0,00388
sunrise	sunrise	7268403526	0,00384
sunrise	clouds	53920446423	0,00358
sunrise	sunrise	51405677522	0,00342

Figure 50 Most influential photos regarding #sunrise topic

Key common photos are:

- Photo *51809773505* emerges as the most influential across all three metrics (high views, highest favorites, and highest comments), showing a strong balance of visibility and user interaction.
- Photo *52477832821* has the highest views ratio, showing that it attracts many viewers. Although it doesn't rank highly for favorites and comments, its significant visibility still gives it the second-highest influence score.
- Photo *53708906207* has second highest favorites ratio and third highest comments ratio, with a moderate views ratio, indicating strong community engagement. This shows that while the photo may not be as widely viewed as others, it generates significant appreciation and discussion, making it important for deeper engagement within the community.

Photos like *51809773505* and *53708906207* are highly influential because they rank strongly in multiple influence metrics, demonstrating both visibility and engagement. Photographs that rank highly in more than one factor are more likely to have a broad and balanced impact within the community. They not only draw attention but also foster engagement and conversation, making them key drivers of user interaction.

8. CONCLUSION

Our solution has successfully identified key influential users and photos based on a combination of factors, including views, favorites, comments, and influence metrics like user interest, involvement and endorsement. The solution effectively identifies the most impactful users and photos, providing a robust understanding of the community trends surrounding the *#sunrise* topic.

Our current dataset is limited in scope, as it only covers content tagged with *#sunrise* which, while effective for focused analysis, restricts the diversity of insights. Additionally, the dataset's current homogeneity, being entirely composed of sunrise-tagged photos, limits its variability and generalizability. Increasing the data volume by including related topics (e.g., *#sunset, #morning, #clouds*) would provide a richer dataset for comparison. Broader data coverage would help in uncovering additional relationships between tags and identifying influential users and photos in overlapping or adjacent contexts.

We have used equal weights across the different metrics - *interest, involvement, endorsement,* and *preference* - to calculate the influence scores for users and photos. Adjusting these weights based on specific objectives, such as placing more emphasis on community interaction (endorsement or comments), could yield more targeted results.

Furthermore, while we have already implemented basic tag and photo filtering, more sophisticated techniques could improve results by filtering based on engagement thresholds like minimum comments or favorites to identify truly influential content. Incorporating temporal factors to capture how influence evolves over time could provide a more dynamic view, identifying trends and highlighting viral content.

Finally, segmenting users based on behavior patterns (e.g., amateur/professional photographers, content creators, casual viewers) and implementing content-based analysis using machine learning would help deepen our understanding of what drives engagement beyond metadata, such as photo features or themes.

REFERENCES

[1] Isabel Anger and Christian Kittl (2011) Measuring influence on Twitter, i-KNOW '11: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, Article 31, 1 - 4, https://doi.org/10.1145/2024288.2024326

[2] E. Koutrouli, G.Kanellopoulos, and A. Tsalgatidou (2016) Reputation Mechanisms in on-line Social Networks: The case of an Influence Estimation System in Twitter, SEEDA-CECNSM '16: Proceedings of the SouthEast European Design Automation, Computer Engineering, Computer Networks and Social Media Conference, 98 – 105, https://doi.org/10.1145/2984393.2984400

[3] Charalambos Tsekeris & Ioannis Katerelos (2012) Web 2.0, complex networks and social dynamics, Contemporary Social Science, 7:3, 233-246, DOI: 10.1080/21582041.2012.721896

[4] ‘Number of monthly active Facebook users worldwide as of 1st quarter 2021’, https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/, Published by H. Tankovska, May 21, 2021

[5] ‘Social Media Fact Sheet’, April 7, 2021, https://www.pewresearch.org/internet/fact-sheet/social-media/

[6] https://www.simplilearn.com/real-impact-social-media-article

[7] https://www.nytimes.com/2016/11/17/technology/social-medias-globe-shaking-power.html?_r=0

[8] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[9] Hilbert, M. (2015). Global information Explosion. Part of the University of California course: ‘Digital Technology and Social Change’, Open Online Course at the University of California, freely available at: https://canvas.instructure.com/courses/949415

[10] https://www.strategyand.pwc.com/gx/en/insights/2019/creating-value-from-data.html

[11] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

[12] Data Collection and Analysis by Dr. Roger Sapsford, Victor Jupp ISBN 0-7619-5046-X

[13] https://en.wikipedia.org/wiki/Representational_state_transfer

[14] ‘What is a REST API?’, https://www.redhat.com/en/topics/api/what-is-a-rest-api

[15] https://developer.mozilla.org/en-US/docs/Web/HTTP

[16] https://www.ibm.com/topics/rest-apis

[17] https://www.tibco.com/reference-center/what-is-data-transformation

[18] https://en.wikipedia.org/wiki/Data_cleansing

[19] https://techterms.com/definition/newsgroup

[20] Burke R. Hybrid recommender systems: survey and experiments. User Model User-adapted Interact 2002;12(4):331–70.

[21] Bobadilla J, Ortega F, Hernando A, Gutie´rrez A. Recommender systems survey. Knowl-Based Syst 2013;46:109–32.

[22] D.H. Wang, Y.C. Liang, D.Xu, X.Y. Feng, R.C. Guan(2018), "A content-based recommender system for computer science publications", Knowledge-Based Systems, 157: 1-9

[23] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29(2–3):131–63.

[24] Duda RO, Hart PE, Stork DG. Pattern classification. John Wiley & Sons; 2012.

[25] Bishop CM. Pattern recognition and machine learning, vol. 4, no. 4. Springer, New York; 2006.

[26] https://en.wikipedia.org/wiki/File:Collaborative_filtering.gif

[27] Xiaoyuan Su, Taghi M. Khoshgoftaar, A survey of collaborative filtering techniques, Advances in Artificial Intelligence archive, 2009.

[28] "Collaborative recommendations using item-to-item similarity mappings".

[29] Sarwar, Badrul; Karypis, George; Konstan, Joseph; Riedl, John (2001). Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on the World Wide Web. ACM. pp. 285–295. CiteSeerX 10.1.1.167.7612. doi:10.1145/371920.372071. ISBN 978-1-58113-348-6. S2CID 8047550.

[30] Suriati, S., Dwiastuti Meisyarah, Tulus T., (2017) Weighted hybrid technique for recommender system, Journal of Physics: Conference Series, Volume 930, Issue 1, article id. 012050 DOI 10.1088/1742-6596/930/1/012050

[31] Article: 7 Types of Hybrid Recommendation System, Jeffery chiang, Analytics Vidhya, https://medium.com/analytics-vidhya/7-types-of-hybrid-recommendation-system-3e4f78266ad8

[32] https://en.wikipedia.org/wiki/Social_networking_service

[33] Obar, Jonathan A.; Wildman, Steve (October 2015). "Social media definition and the governance challenge: An introduction to the special issue". Telecommunications Policy. 39 (9): 745–750. doi:10.1016/j.telpol.2015.07.014

[34] Amichai-Hamburger, Yair; Hayat, Tsahi (2017). "Social Networking". The International Encyclopedia of Media Effects. pp. 1–12. doi:10.1002/9781118783764.wbieme0170

[35] J. D. Biersdorfer, Sept. 14, 2016, See How Popular Your Flickr Photos Are, https://www.nytimes.com/2016/09/15/technology/personaltech/see-how-popular-your-flickr-photos-are.html

[36] “What is Flickr?”, by Mark Guertin, September 4, 2015

[37] "The man behind Flickr on making the service 'awesome again'", The Verge, March 20, 2013

[38] “Fun Fact: Flickr and Slack Started as “A Game that Never Ends”, by Kaden Ng, September 21, 2017

[39] “How many public photos are uploaded to Flickr every day, month, year?”, by Frank Michel,
March 20, 2012 (updated Oct. 2019)

[40] https://www.flickr.com/services/, The App Garden, Flickr.com

[41] https://www.interviewbit.com/blog/spring-boot-architecture/

[42] https://spring.io/projects/spring-boot

[43] https://angular.dev/

[44] https://material.angular.io/

[45] https://www.mysql.com/

[46] https://en.wikipedia.org/wiki/Long_tail