ANKARA YILDIRIM BEYAZIT UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INTERACTIVE SOCIAL MEDIA VISUALIZATION

M.Sc. Thesis by

Elif ŞANLIALP

Department of Computer Engineering

September, 2018

ANKARA

INTERACTIVE SOCIAL MEDIA VISUALIZATION

A Thesis Submitted to

The Graduate School of Natural and Applied Sciences of

Ankara Yıldırım Beyazıt University

In Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Engineering, Department of Computer Engineering

by

Elif ŞANLIALP

September, 2018

ANKARA

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"INTERACTIVE SOCIAL MEDIA VISUALIZATION"** completed by **ELİF ŞANLIALP** under the supervision of **ASSIST. PROF. DR. M. ABDULLAH BÜLBÜL** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. M. Abdullah BÜLBÜL

Supervisor

Prof. Dr. Tolga Kurtuluş ÇAPIN

Jury Member

Assist. Prof. Dr. Hilal KAYA

Jury Member

Prof. Dr. Ergün ERASLAN

Director

Graduate School of Natural and Applied Sciences

ETHICAL DECLARATION

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,
- All information, documents and assessments are presented in accordance with scientific ethics and morals,
- All the materials that have been utilized are fully cited and referenced,
- No change has been made on the utilized materials,
- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date: 2018, 13 September **Sign:**

Name&Surname: Elif ŞANLIALP

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my supervisor, Assist. Prof. Dr. M. Abdullah BÜLBÜL for his tremendous support and motivation during my study. His immense knowledge and precious recommendations constituted the milestones of this study. His guidance assisted me all the time of my research and while writing this thesis.

I also would like thank Prof. Dr. Tolga Kurtuluş ÇAPIN and Assist. Prof. Dr. Hilal KAYA for their valuable contributions and constructive criticisms during my thesis defense examination.

I am grateful to my husband İbrahim ŞANLIALP for his endless support. He always motivated me during my hard times and always trusted me. I am very fortunate to have such a person.

Finally, I must express my profound appreciations to my family for providing me emotional support, loving care and continuous encouragement throughout my years of study and through the period of writing this thesis. This accomplishment would not have been possible without them. Thank You.

2018, 13 September **Elif ŞANLIALP**

INTERACTIVE SOCIAL MEDIA VISUALIZATION

ABSTRACT

The usage of social media is increasing day by day. People use the social media platforms to communicate with their friends or other users and to share the things are interested in such as photos, texts, videos on their profiles. Also, many people share their thoughts on where they are and what they do by attaching location information in social media platforms mostly. The shares that have location information are called geo-tagged posts in social networks. The posts are using in lots of researches for social media analysis in recent years. The number of such studies is increasing with the popularity of social media platforms using location.

When these studies are examined, we have realized the lack of visualization of interactive mapping researches. Due to, we propose an interactive tourist map that displays the selected geo-tagged images which are gathered from Flickr using latitude and longitude information of images on web based map. The main purpose of this study is to detect and show the popular locations and events according to geo-tagged images of users in Flickr on the map.

Keywords: Social media analysis, geo-tagged images, Flickr, interactive mapping, visualization, location information.

SOSYAL MEDYANIN ETKİLEŞİMLİ OLARAK GÖRSELLEŞTİRİLMESİ

ÖZ

Sosyal medya kullanımı günden güne artmaktadır. İnsanlar sosyal medya platformlarını arkadaşları ya da diğer kullanıcılar ile iletişim kurmak için, fotoğraf, video ve metin gibi ilgi duydukları gönderileri kendi profillerinde paylaşmak için kullanırlar. Ayrıca, pek çok kişi, sosyal medya platformlarında çoğunlukla konum bilgilerini ekleyerek bulundukları konumları ve o konumlarda yaptıklarıyla ilgili düşüncelerini paylaşmaktadırlar. Konum bilgisinin eklendiği bu paylaşımlar, sosyal ağlarda coğrafi etiketli gönderiler olarak adlandırılmaktadır. Bu gönderiler son yıllarda sosyal medya analizi için birçok araştırmada kullanılmaktadır. Bu araştırmaların sayısı sosyal medyanın popülaritesinin artmasıyla birlikte artmaktadır.

Bu çalışmalar incelendiğinde, interaktif haritalamanın görselleştirilmesi hususunda eksikliklerin olduğunu fark ettik. Bu nedenle, fotoğrafların enlem ve boylam bilgilerini kullanarak Flickr'dan toplanan görüntülerin web tabanlı hartada gösterilmesini amaçlayan bir interaktif turist haritası oluşturmayı öneriyoruz. Bu çalışmanın ana amacı, Flickr kullanıcılarının paylaştıkları coğrafi etiketli görüntüleri kullanarak popüler lokasyon ve olayların tespit edilip interaktif harita üzerinde gösterilmesidir.

Anahtar Kelimeler: Sosyal medya analizi, coğrafi etiketli görüntüler, Flickr, interaktif haritalama, görselleştirme, konum bilgisi.

CONTENTS

NOMENCLATURE

Acronyms

3D	3 Dimensional
ANN	Approximate Nearest Neighbour
API	Application Programming Interface
BRIEF	Binary Robust Independent Elementary Features
CV	Computer Vision
DBSCAN	Density-Based Spatial Clustering of Application with Noise
DM	Direct Message
FAST	Features from Accelerated Segment Test
GIS	Geographic Information Systems
GPS	Global Positioning System
HTML	HyperText Markup Language
ID	Identification Number
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
MVC	Model – View – Control
ORB	Oriented FAST and Rotated BRIEF
RSS	Rich Site Summary
RT	Retweet
SIFT	Scale Invariant Feature Transform
SQL	Structured Query Language
SURF	Speed Up Robust Feature
TCP / IP	Transmission Control Protocol / Internet Protocol
TT	Trending Topic
TF-IDF	Term Frequency – Inverse Document Frecuency
UI	User Interface
URL	Uniform Resource Locator

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the widespread use of the Internet over the world, social media platforms have emerged as a result of people having to share and communicate each other over the Internet. Each of these platforms offers a variety of communication and content sharing methods, each focusing on different user needs. Increasing the number of social media platforms makes it easier and faster for people to search for content they are looking for. Through these platforms, each user can share texts, images, videos, surveys, etc. with another user. In addition, social studies enable people to easily understand their thoughts, behaviours, habits etc. through their shares. For these reasons, the analyses made on the social media platforms both for users and users have started to gain importance every day.

Today, there are dozens of social media websites and applications serving different needs and purposes. When the most used social networks in the world are searched: Facebook[1], Twitter[2], Instagram[3], YouTube[4], Flickr[5] and Snapchat[6] stand out. From these social networks Facebook and Twitter are especially prominent with the mass of users they have created all over the world. As Facebook and Twitter have been developed before other social networks, the need for analysis and using the obtained data has arisen together with the increase in the mass of users and the increase of the data obtained from these users every passing day.

Along with developing technology, with the widespread use of mobile devices and the increase of internet usage, the importance of locally based services has increased. The most common of these services is location based social networks. These systems allow

[1] https://www.facebook.com
[2] https://twitter.com
[3] https://www.instagram.com
[4] https://www.youtube.com
[5] https://www.flickr.com
[6] https://www.snapchat.com

users to share location information, interact with each other on-line, make suggestions about a location. Many people share lots of posts by adding location information recently. Because of that, these posts with coordinates data are also appropriate to be used for researches. The location information is attached to a post by using Global Positioning System (GPS) technology, which is included in most of our daily devices for communication, and this information is an important property for studies which especially includes world map or geographic map applications on social media data. One of the important application areas is tourism, as millions of tourists travel from one place to another for fun or other reasons. In any case, they want to see the popular attractive places among others. Meanwhile, when they decide to visit where to go, they start to search for facilities near the destination such as shops, markets, restaurants, hotels, cafe, library and so on. If the place has a geotagged photo, it can gain popularity on the social media and attract more tourists. Thus, the geotagged photo is critical for choosing geographical preferences of tourists. Under these circumstances, we propose a new platform on the map that informs people about popular locations and events using social media data in our study. Our platform is designed to dynamically display popular places on the map utilizing what people share. Our platform differs from static maps such as tourist maps in that it constantly updates itself according to social media data. In this way, the most up-to-date information according to static maps is obtained by the users [1].

As shown in Figure 1.1, in our study firstly, we collect the data from Flickr application programming interface (API) using developed software on the Python software development platform. Photos shared on Flickr with GPS coordinate information are used as the social media data. After this section, the data is processed using clustering algorithms. Two of the most commonly known and used clustering algorithms are k-means algorithm and density – based spatial clustering of application with noise (DBSCAN).

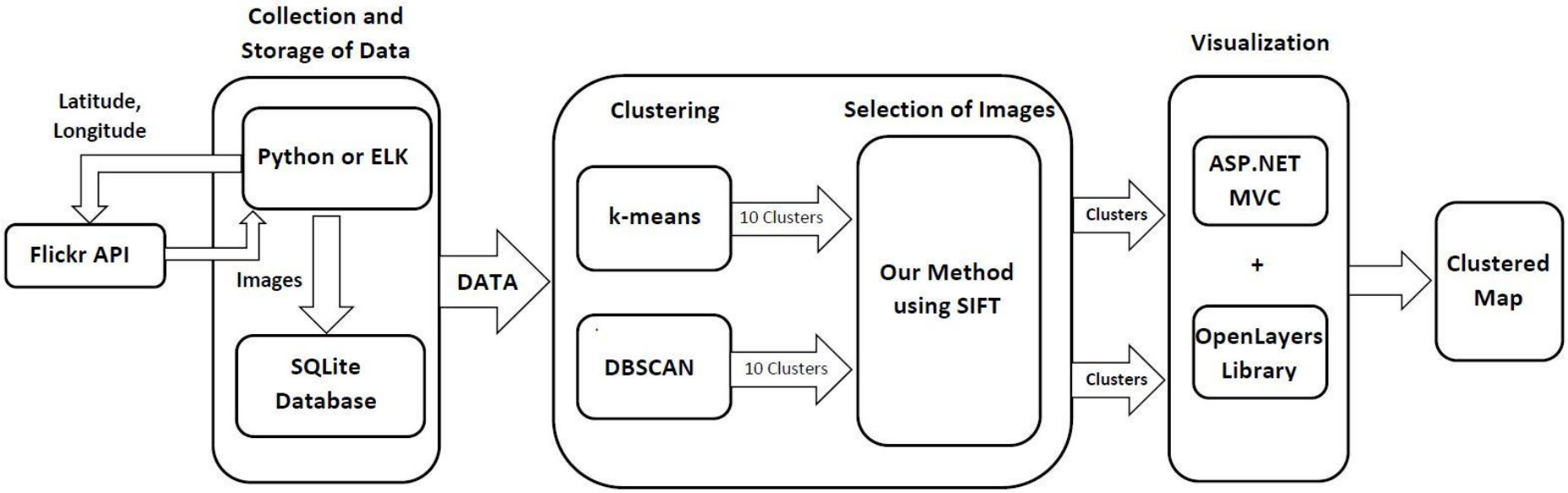

Figure 1.1 The overview of study

k-means and DBSCAN clustering algorithms are used in our study. The next step is to select the best photos among all of the shared photos in a specific location. An image feature based strategy is used for this step. To extract image features, we utilized Scale Invariant Feature Transform (SIFT) [2]. The last part of this study is the visualization of the processed data. To generate map, we use local web based project and implement one of the OpenLayers[7] Maps, which is a clustered map, to visualize the data we use JavaScript in the project.

The rest of the thesis is organized as five chapters. The first chapter of the study focuses a brief explanation of social media platforms and their usage areas, related works in the literature and the aim of this study. The second chapter of the study focuses how information is gathered for study and which methods are used to gather information. The third chapter of the work consists of how social media data is processed. The forth chapter focuses on generating maps using google static map API. The last chapter of the work mentions the results of study.

1.1 Social Media Platforms

The concept of social media has entered our life with Facebook founded by Mark Zuckerberg as a student of Harvard University in 2004. Facebook has been available for all over the world in 2006. This social network aims to make people communicate and share the things they are interested in like texts, photos, videos or websites with their friends [1]. Users can create groups, pages and also join them depending on their desires in Facebook People can create an account in this platform with an e-mail address and widen their social networks. In 2012, the number of the users of Facebook has exceeded 1 billion and as of December 2016, Facebook has 1 billion 860 million users [3]. In our country, the number of users has reached 42 million according to Facebook user statistics [4].

Besides Facebook, one of the mostly used social networks is Twitter. The thing posted by a Twitter user is called a *tweet* which has to be written with at most 280 characters.

[7] https://openlayers.org/

Users generally post tweets as a text but also they can attach photos, links and videos to their tweets. Besides these attachments, Twitter has an important term, Hashtag. Hashtag means a label on the world and starts with '#' character. Hashtag, was used on Twitter for the first time, and all social networks in the ongoing process began to use this feature. Users share their labels based on their interests and interact with target groups. Labels at the point are used all over the world in an active and continuous manner. Twitter has been developed by Biz Stone, Noah Glass, Jack Dorsey and Evan Williams in 2006 and now Twitter has 310 million active users. Users can follow or mention each others according to their interests [1]. Twitter has three important terms, Trending Topic (TT), which is a list consists of top 10 most talked topics in the world or a country, Retweet (RT), which is a tweet sharing another user's tweet, and Direct Message (DM) is a secret tweet between two users [5].

Instagram is another major social network which has founded Kevin Systrom and Mike Krieger in 2010. By the year of 2016, it has reached 500 million users. Users have an account and can share photos and videos on their profiles and follow other users like Twitter. Digital Filters for photos and videos and Hashtags are also used in Instagram [6].

Another social platform for photo and video sharing is Flickr which is created by Stewart Butterfield and Caterina Fake in 2004, older than Instagram and Twitter. Flickr is widely used by photo researchers and bloggers to host images that are buried in blogs and social media [7].

There are lots of social media platforms such as YouTube, Pinterest, Swarm, Snapchat, Tumblr, etc. But the most used platforms in researches are briefly explained above. The usage area of these platforms in academic field is mentioned below.

1.1.1 Usage Area of Social Media

Every social media platform has different attributes and allows users share different types of posts, such as Instagram and Flickr allows photos and videos, YouTube allows only videos, Twitter allows not only text limited 280 characters but also photos and

videos limited by one minute or instant locations can be shared at Swarm, using GPS coordinates, etc.

According to types of posts in social media platforms, academic studies in this area varies. For instance, data from Twitter is especially used for text mining to analyze people's reactions to specific topics, to predict elections results, to extract inferences from hashtags. On the other hand, the data gathering from Instagram or Flickr is mostly used for the area of image and video processing such as feature extracting from images, filtering and clustering the data, etc.

The usage of these topics associated with the thesis are covered in more details in the next Related Works section.

1.2 Review of Related Works

A wide variety of social media based studies have been extensively investigated in terms of gathering geotagged data, processing and visualizing the data.

The most comprehensive study is called *NewsStand: A New View on News* has been established by Teitler et al. NewsStand collects news stories from Google News[8], Yahoo! News[9], and Microsoft Live News, analyses the news data, and displays the results in a map interface. While gathering news, the application pursues Rich Site Summary (RSS) feeds from online sources and then gets articles as soon as news are published. Then geographic contents of the articles are examined and the location of the news are estimated. After that, NewsStand applies online clustering algorithm based on Term Frequency - Inverse Document Frequency (TF-IDF) on term centroid and time centroid to group news'. In the part of visualization, NewsStand shows the news in a web page using mapping API of Microsoft Virtual Earth [8].

Another important study *TwitterStand: News in Tweets* has been published by Sankaranarayanan et al.. The aim of this study is very similar to NewsStand on many

[8] https://news.google.com
[9] https://www.yahoo.com/news/

issues. In this study, tweets which contains breaking news are collected automatically, and applied online clustering like NewsStand according to the topics of tweets besides the geographic location of users. The interface is also similar to NewsStand but it has two panes; the left one has sorted clusters in descending order, the right one has the visualization part as a map that combines clusters with locations of tweets [9].

A similar study *PhotoStand: A Map Query Interface for a Database of News Photos* is conducted by Samet et al. to display photos of related news, which are stored in PostgreSQL database, on map according to their locations. PhotoStand geotags images by examining the article where they were found. This application does online clustering for the data of news photo like NewsStand and TwitterStand and also use the same interface. But it applies some process on photos such as feature extraction using TF-IDF image extraction based on keywords and near-duplicate image detection to eliminate similar or nearly same photos [10].

Another crucial study is *Social Street View: Blending Immersive Street Views with Geo-tagged Social Media* has been published by Du and Varshney. The aim of this study is to develop a new algorithm for visualizing the geo-tagged social media data by using immersive map which is panoramic 360^{o} and 3 Dimensional (3D) views of places from Google Street View. Instagram API is used to gather data as photos and videos and distributed MySQL database is used to store the social media data. To generate the interface, WebGL and WebVR techniques are used [11].

One of the 3D reconstruction based study named *Photo Tourism: Exploring Photo Collections in 3D* is done by Snavely et al. The system collects unstructured photos from Internet sites where geo-tagged photos are shared, processes them, and represents the scenes geometrically in 3 dimensional space. The techniques or algorithms used are the same with the previous study; SIFT is used for feature detection and matching for images and Structure from Motion (SfM) is used for construction 3D form of the scenes [12].

Agarwal et al. has established a significant study whose name is *Building Rome in a Day*. As the name implies the authors targeted 3D reconstruction on city-scale; and

worked on three data sets for Rome, Dubrovnik and Venice. The data sets are downloaded from Flickr by term based search, using the name of the city like ‘Rome’ or ‘Roma’ as the term. The system applies SIFT algorithm for feature extraction of photos, Approximate Nearest Neighbour (ANN) library for matching SIFT features as image matching and SfM algorithms for reconstructing cities [13].

Bulbul and Dahyot propose to automatically place people in geo-located virtual cities by harvesting and analyzing online data shared on social networks and websites. SIFT feature matching is used to analyze the data. Their method relies on Open Street Map[10] which is an online geographic information system to automatically generate 3D cities and populate these worlds using online data extracted from real people posting on Twitter and Instagram. To visualize the 3D city model, Unreal game engine is used [14].

Moreover, Bulbul and Dahyot proposed to use a geo-tagged virtual world for the visualization of people’s visual interest and their sentiment as collected from their social network activities. They firstly try to find out visually popular structures in the environment that attract the most visual attention and that deserves pictures to be taken and shared by analyzing pictures posted on social media. After this process, they automatically visualize these popular landmarks in a 3D environment by controlling illumination in game engine and by altering colours of the meshes to enhance sentiment visualization and enhance popularity using game engine technologies [15].

In another study, Zhang and Kosecka present finding GPS location of urban areas using images of the areas as inputs with SIFT feature matching technique [16].

Jaffe et al. proposed to facilitate a system which can automatically select representative and relevant photographs from particular spatial region, San Francisco. They gather images using tag SanFrancisco from Flickr, use Hungarian algorithm to cluster the data, tf-idf method to calculate scores of images and they summarize data-set according to users, tags and land marks; not photo features [17].

[10] https://www.openstreetmap.org/

Pavels et al.'s study named *Placing Flickr Photos on Map* aims to geographically label photos which are shared by Flickr users by using text annotations that the study uses as primary input and to place photos on the world map [18].

In order to generate an automatic map, Grabler et al. proposed to detect landmarks, paths, nodes, edges, and districts of San Francisco. They used 3D building geometry with texture, road geometry with type information, Traffic map and ground plane texture and webpages (Yahoo, Tripadvisor and Openlist) with landmark information to show buildings in 3D by using these properties [19].

1.3 Aim of the Study

Photo sharing and archiving services offer the possibility of "geotagging" photos to users. These services provide access to the geographical information of the online shared photo contents stored in the database using the application programming interface(API). The gathered data can be visualized according to photo similarities and photo contents shared within a certain area. In this study, it is aimed to visualize the geographical locations of photographs taken by social media users and to evaluate human activities in social media in different urban areas. Also, it is tried to develop a new interactive map by harvesting and analyzing online photos shared on social media.

The main objective of our study to show popular attractions, important events and places which are decided by using feature extraction on the geotagged images shared in places that are important or interesting for people on a current map. This map is designed to give information about what are the popular activities and where are the popular places to tourists in İstanbul or Dublin where the geotagged social media data gathering for.

CHAPTER 2

GATHERING DATA

In today's communication world, in social media, in video sharing sites, in health, security, information and other fields, with the progress of technology day by day, a large amount of data is being produced every day and the importance of these data increases day by day. The concept of Large Data has emerged by transforming these large quantities of data into meaningful and processable data. Large data is considered both an entity and a process. Large data as an entity usually encompasses information volumes that cannot be processed by conventional database and software techniques. Large data as a process represents a variety of data types that infrastructure and technology companies collect, store and analyze [20]. A significant proportion of these data are initially unstructured, scattered, and not very meaningful on their own. Therefore, there is a need to record, access, analyze and process this gigantic data in a performance-oriented manner.

Elasticsearch is one of the tools developed to deal with the problems mentioned in the big data world. When we have started our study, we used Elasticsearch for gathering data from Twitter. But we met some problems while gathering data. One of the problem is that the locations of tweets added photo do not come up in the data taken according to the coordinates, but instead the location information of users in profiles comes up. If we want to get the right results, we should collect tweets according to exact location information of tweets. Thus, we decided to change data collection technique and Python is chosen to use for gathering geo-tagged data from Flickr by using flickr.photos.search service in Flickr APIs. These methods used have described with the sequence below.

2.1 Elasticsearch – Logstash – Kibana

Elasticsearch is an open source full text search and analysis engine [21]. It is a scalable and built on lucene infrastructure using the java programming language. Additionally,

it allows to store, search and analyze large quantities of data in a fast and real-time manner so many people use it as a document database. It is used today for content searches, data analysis and queries on projects such as Mozilla, Foursquare, GitHub Facebook, LinkedIn[11].

To explore Elasticsearch, there are a few concepts and key features. Understanding these concepts and key features will help ease the learning process. Index, document, document type and mapping are concepts that are core to Elasticsearch [22]. The equivalents of these concepts in Structured Query Language (SQL) are shown in Table 2.1.

Table 2.1 Terminologies – Elasticsearch vs. SQL

Elasticsearch element	SQL element
Index	Database
Mapping	Schema
Document type	Table
Document	Row

2.1.1 Index

Elasticsearch stores all data in indexes. An index in Elasticsearch is a place like database (Table 2.1) in a relational database management system. Each record in Elasticsearch is a configured JavaScript Object Notation (JSON) document. Elasticsearch indexes are a collection of configured JSON documents. All fields of the document are indexed by default and can be used in a single query. Index can store different types of documents, search them, and update for them.

2.1.2 Document

Document is a main entity stored in Elasticsearch. Each document in Elasticsearch is a JSON object, analogous to a row in table in a relational database management system. A document consists of fields (row columns). Each field may occur several times and these types of field is called multivalued. Furthermore, each field has a type

[11] https://tr.linkedin.com/

such as text, number, date. Field types can also be complex. For example, a field can contain other subdocuments or arrays.

2.1.3 Document Type

One index can store many objects with different purposes in Elasticsearch. For example, a journal application can store comments and articles. These objects can easily differentiate by document type. Type is similar to table in a relational database management system.

2.1.4 Mapping

Mapping is the process of defining how a document must be transferred to the search engine. When creating the tables, they are created with mapping knowledge. It is similar to a schema definition in SQL databases [23]. Elasticsearch automatically generates the mapping according to the posting data.

Figure 2.1 Elasticsearch processing pipeline

Logstash is a server-side data processing pipeline that ingests data from a multitude of sources simultaneously, transforms it, and then sends it to users favourite stash[24]. It has open source technology. The real-time processing is especially powerful when coupled with Elasticsearch, Kibana, and also Beats.

Figure 2.2 Logstash processing pipeline

Logstash is a data flow engine that allows users to aggregate, extend, and merge multiple inputs from multiple common sources regardless of format or scheme. These variety of inputs data can be easily obtained web applications, user logs, data stores and different web services in continuous flow. It filters each event, defines named fields to build the structure and transforms it into a common formatter for easier analysis. Logstash has a different outputs that provide the route data that user wants. Moreover, the real-time processing is especially powerful when coupled with Elasticsearch, Kibana [24].

Kibana is a visualization tool. Kibana enables real-time analysis and visualisation of Elasticsearch streaming data. It has interactive dashboards and allows interactive data exploration. It also supports cross-filtering.

Figure 2.3 Interface of Kibana [25]

This tool has different chart types such as pie charts, bar charts, line and scatter plots, maps as shown in Figure 2.3. It is open source and no need to know programming or query language in most of the cases [25].

2.2 Python for Flickr

Python is a platform independent, high-level scripting programming language which has been developed by a Dutchman developer, Goido Van Rossum. When Python is run, source code is processed by an interpreter. It does not need to be compiled before running the program. Moreover, it is open source, it can be worked on commend window, does not need an Integrated Development Environment, called IDE. And this programming language supports object oriented programming, functional and structured programming methods. It also supports easy connection to almost any databases [26]. Therefore, this programming language works very fast. Because of these features of the Python programming language, in this study Python is chosen to use for gathering geo-tagged data from Flickr by using flickr.photos.search service in Flickr APIs [27].

To use Flickr APIs in an application, firstly user must have an account of Flickr and request two Flickr API keys; one of them is public key and the other is secret key. After then, these keys and the other input which is format as given parsed-json are used to satisfy the connection of Flickr using FlickrAPI() function. Then the latitude and longitude information of desired two locations, İstanbul, Turkey at the values of longitude 41.008238 and latitude 28.978359 and Dublin, Ireland at longitude 53.349805 and latitude -6.260310 in 5 kilometers diameter, are given as inputs to flickr.photos.search() function with some extra information as extras which are expected values. The values returned from search() function are parsed to obtain the desired information such as identification (id) of photo, user id, uniform resource locator (url) of image, text and exact location of the image for this study as shown in Figure 2.4.

```
flickr = FlickrAPI(FLICKR_PUBLIC, FLICKR_SECRET, format='parsed-json') #parsed-json
def getFlickrPhotos():

    extras = 'url_sq,url_t,url_s,url_q,url_m,url_n,url_z,url_c,url_l,url_o,geo'
    ist = flickr.photos.search(lat=41.008238, lon=28.978359, extras=extras) # for istanbul
    photos = ist['photos']
    photo = photos['photo']

    source = "Flickr"

    for p in photo:
        date = datetime.date.today()

        id = str(p['id']).encode('utf-8')
        user = str(p['owner']).encode('utf-8')
        url = str(p['url_m']).encode('utf-8')
        text = str(p['title']).encode('utf-8')
        lat = str(p['latitude']).encode('utf-8')
        lon = str(p['longitude']).encode('utf-8')
        location = str(lat + ',' + lon)
```

Figure 2.4 The part of Python code for collecting the İstanbul data

At the same time while collecting data, the images are also downloaded and saved the SQLite database together with the data properties.

2.3 SQLite Database

SQLite is an open source database product developed by C/C++ programming languages and provides a relational database management system. It has public-domain software package Thus, it is free for any purpose, private or commercial [28].

It has the following remarkable features: serverless, transactional SQL database engine, zero-configuration, self-contained.

Serverless means that it does not need a server to run. Relational database systems like MySQL require a separate server process to operate. The applications using client-server architecture want to access the database server. They use Transmission Control Protocal / Internet Protocol (TCP/IP) protocol to send and receive requests. But, SQLite works in a different way. It reads and writes directly ordinary disk files stored on disk.

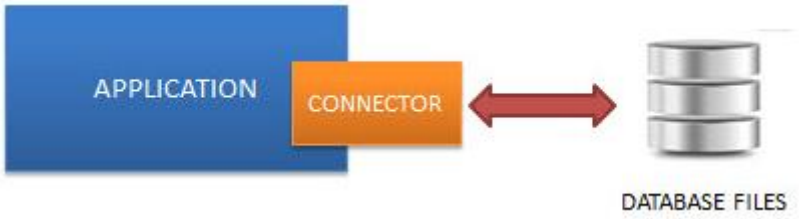

Figure 2.5 SQLite serverless architecture [54]

And also zero-configuration databases do not need to be installed before using; due to the serverless architecture for SQLite. No server means no server process that needs to be configured, started, stopped. Moreover, it does not need to use any configuration files.

Self-Contained can be explained as independent of servers. A database server system typically consists of multiple processes that are responsible for managing client connections, query processing, query optimization, file I / O and caches, and it has more than one directory trees and many files on the database server file system. When users want to access the database, database must be correct and stable. Because of this features, these database systems require resources and support from the host computer. However, SQLite has no separate server. Because all database engines are integrated into applications to access a database. If user need to back up or move the database, user can only copy the file on disk. It runs any operating systems. The entire SQLite library is encapsulated in a single source code files that are concatenated into a single large files of C-code named "sqlite3.c" and called "the amalgamation" [29]. Thus, SQLite requires little operating system support to read and to write some different type of storage from the operating system.

Transactional database means that all transactions in SQLite are fully ACID-compliant; all queries and changes are Atomic, Consistent, Isolated, and Durable and allows safe access from multiple threads or processes.

Owing to these features of SQLite database, in this study it is used to store the data gathering from Flickr in social media database. The data of İstanbul and the data of

Dublin are stored in different tables but they are very similar to each other and the ten clusters of each data and the scores of the best images of the clusters are stored in different scores tables which are connected on the other tables. These scores are calculated after the process of the data and saved to the scores tables of the database.

CHAPTER 3

PROCESSING DATA

Processing of social media data shows differences according to the type of data gathering from different social media platforms and the aim of studies. In this study, after gathering the images from Flickr, they are saved to SQLite database which is mentioned in the previous chapter. As first step of the processing the data, two types of clustering are applied. These types are K-means Clustering and Density-Based Spatial Clustering of Application with Noise which are disclosed below. And the second step of processing the data is feature extraction and selection of the best 5 images of each cluster.

3.1 Clustering

Clustering is an unsupervised classification of data items into groups called clusters and is used for data analysis basically. But one of the most important problems of unsupervised learning is clustering which has different definitions for the usage but it must have some rules:

1. Samples which are in the clusters should be as similar as possible.
2. Samples which are in the different clusters should be as different as possible.
3. For similarity and diversity, measurements should be clear and practical [30].

Clustering algorithms which has a wide range usage area such as image processing, information retrieval, computational biology, pattern recognition, mobile communication, medicine and economics have become very popular in recent years [31].

Clustering techniques vary in many types, some of them are mentioned below briefly:

1. **Hierarchical clustering:** It is based on the distance between instances. Algorithms of hierarchical clustering connect the objects according to their

distances. Because of that, this approach is also known connectivity-based clustering. BIRCH, CURE, ROCK, Chameleon are examples of this clustering.

2. **Partition-Based clustering:** Instances are divided into partitions by this type of algorithms which are more useful for large data sets, such as K-means, K-medoids, PAM, CLARA, CLARANS. This partitions are called clusters.
3. **Distribution-based clustering:** The algorithms of this type of clustering are used for artificial data sets which consist random objects as instances and Its clusters are called distributions. Complexity is the most important problem for the random objects. DBCLASD, GMM are examples of this model.
4. **Density-based clustering:** Algorithms of this type clustering separate clusters based on density. This approach can be used for spatial data that means two dimensional data, time and space, generally. DBSCAN is the best algorithm and OPTICS, Mean-shift are examples of density-based clustering.
5. **Grid-Based Clustering:** The aim of this clustering type is to divide the data set into grids. Grids are assigned a value initially, and instances are placed in the grid according to these values. Some examples of this clustering type are STING, CLIQUE.
6. **Model-Based Clustering:** It is based on a model that are created by datasets. Algorithms of this type clustering is closer to algorithms of clustering based on density. COBWEB, GMM, SOM, ART algorithms are examples of model-based clustering [32].

Table 3.1 Comparison of some of clustering techniques [32]

Clustering Technique	Shape of Cluster	Clustering Algorithm	Outlier Handling
Hierarchical	Arbitrary	BIRCH,CURE	Yes
Partition	Spherical	K-means, K-mode	No
Density	Arbitrary	DBSCAN	Yes
Grid	Arbitrary	CLIQUE, Wave Cluster	Yes

3.1.1 K-means Clustering Algorithm

One of the most known and simplest clustering algorithms is K-means algorithm that divides instances into K groups according to their features. K is the number of how many clusters that users want and it must be an integer number [33].

The target function of k-means algorithm is:

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{C_i} (\| x_i - v_j \|)^2 \quad \text{where,} \qquad \textbf{(3.1)}$$

$\| x_i - v_j \|$ is Euclidean distance between x_i and v_j.

C_i is the number of data points in i[th] clusters.

C is the number of centres of clusters.

K- means Algorithm has 5 step:

1. Determine 'C' centres of clusters randomly.
2. Compute the distance between each data points C_i and centres of clusters.
3. Appoint a data point to the centre of cluster which the distance between the data point and centre is more minimum than the other centres of clusters.
4. Calculate the new centres of clusters with the new data points by using:
   $V_i = (1/C_i) \sum_{j=1}^{C_i} x_i$ **(3.2)**
5. Compute again the distance between each data points and new centres of clusters.
6. If all data points did not change anymore, stop the algorithm, otherwise go to 3[th] step and continue.

K-means algorithm is an easy to understand, fast, and popular algorithm of unsupervised learning. While the data set includes well separated or distinct data points, it gives best outcome. The time complexity of k-means algorithm equals to 'O(tknd)', where n means data points, k means clusters, d means dimensions of each data points, and t means iterations. Generally, k, t, d << n [34].

Because of these advantages, it is used in this thesis used on longitude and latitude information of images in the Flickr data set by using OpenCV library in Python. To use k-means in OpenCV, cv2.kmeans() or cv2.kmeans2() functions must be called with 5 inputs; samples which should be np.float32 data type and each attribute must be in a single column, nclusters (K) that is the number clusters, criteria which is to stop the iterations, attemps which is to control how many iterations are done, and last, flags that is used for determine the initial cluster centres [35].

3.1.2 Density – Based Spatial Clustering of Application with Noise (DBSCAN)

DBSCAN algorithm relies on density-based clustering approach and gives better results when the data sets consist of spatial data, which has longitude and latitude coordinates [36]. DBSCAN uses distances to discover neighbour relationships by pre-allocating the number of radius and minimum dots to form clusters [37].

DBSCAN algorithm has 5 step:

Think about $X = \{x_1 , x_2 , x_3 , \dots , x_n \}$ is the data set; it requires 2 parameters: ε (epsilon) and the number of at least data points in clusters (minPts).

1. Start with a random and not visited data point as origin.
2. Find the neighbourhood of the origin point by using ε; every point within the ε distance are marked as neighbour.
3. In case of there are at least number of minPts around the origin points, to clustering points are marked as visited and labelled as noise.
4. If a data point includes in a cluster, 2^{nd} and 3^{th} step are applied to this point; for each data points this operation is repeated until all data points are placed in clusters.
5. A new data points which is not marked as visited in the data set can causes creation of a new cluster.
6. The algorithm continues till all data points has visited.

There are some advantages of DBSCAN algorithm, firstly the number of clusters is not needed to know. And also this algorithm can eliminate noise data while clustering [38].

In this study, DBSCAN algorithm is applied the data set collected from Flickr by using Python. Firstly, the longitude and latitude information of images are fetched from the database; after then they are converted to two dimensional array for DBSCAN function from scikit-learn implementation of DBSCAN library. The DBSCAN function has 4 compulsory parameters; first one is epsilon ε, second one is the number of minimum data points which is minPts mentioned before, third parameter is algorithm and last one is metric. In this study, used algorithm is 'ball tree' and metric is 'haversine'. Haversine metric expects radian parameters for epsilon and coordinates. Because of that, the units are converted to radians [39]. In addition, the time complexity of DBSCAN which is $O(n*\log_n)$ [30].

3.1.3 Comparison of k-means and DBSCAN Clustering Algorithms

The most commonly known and used clustering algorithm is k-means algorithm. To use k-means algorithm, the number of iterations and the number of clusters must be indicated firstly whereas the type of algorithm which is used by DBSCAN, the number of minimum members in clusters and the limitation, epsilon must be given as inputs before for DBSCAN clustering algorithm. DBSCAN algorithm decides the number of clusters using the inputs. DBSCAN algorithm is superior than k-means algorithm for spatial data because k-means minimizes variance, does not calculate the geodetic distance [39].

3.1.4 Results of Clustering

As a result of the clustering, the two aforementioned clustering algorithms, which are mentioned above, are applied on İstanbul and Dublin data. For both data, ten clusters are obtained with each algorithms. When k-means algorithm is applied, input k, the number of clusters, is given 10. Whereas, Epsilon, an input of DBSCAN algorithm for İstanbul data, is given 0,00006; for Dublin data epsilon equals to 0.00009; and min_samples, which is another input of DBSCAN algorithm, is given 50 means that

clusters have to have at least 50 samples of data. The clusters of İstanbul and Dublin data obtained from k-means and DBSCAN algorithms are shown in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.4.

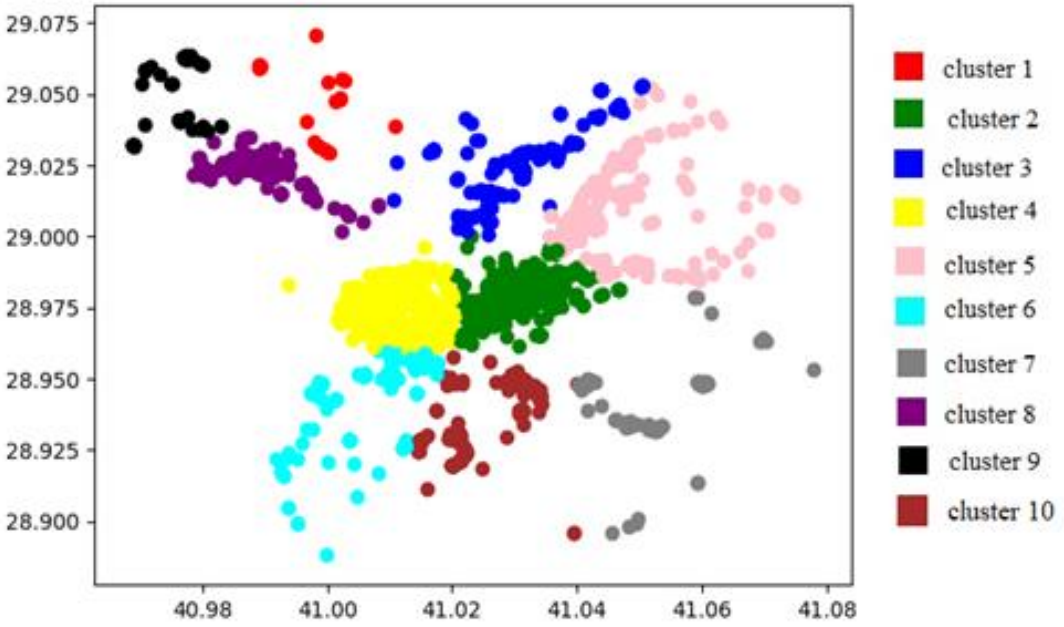

Figure 3.1 Clusters of k-means clustering algorithm for İstanbul Data

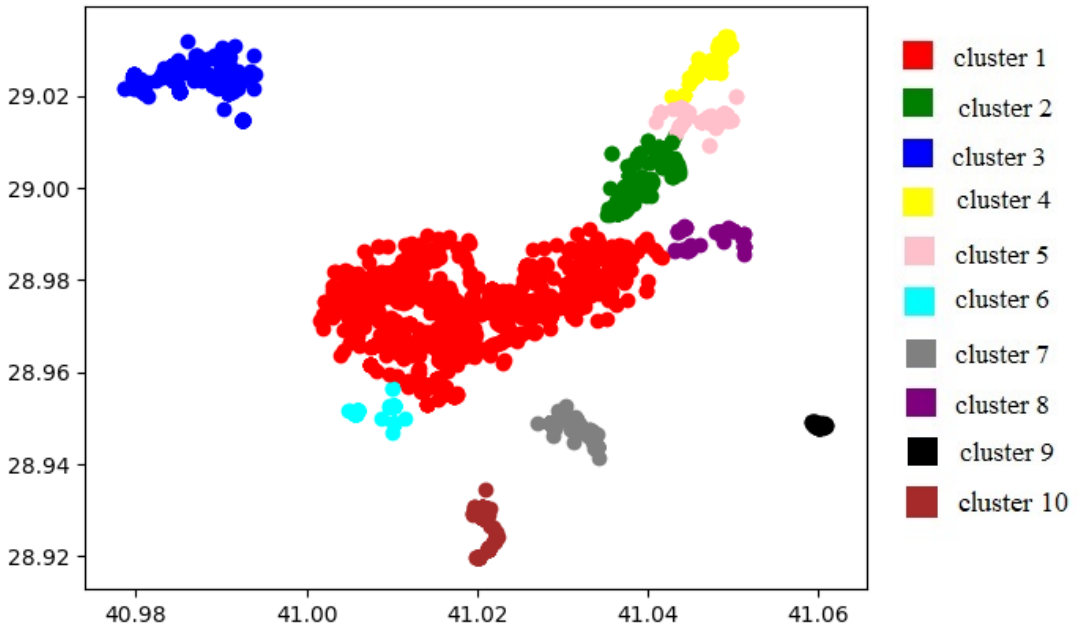

Figure 3.2 Clusters of DBSCAN clustering algorithm for İstanbul data

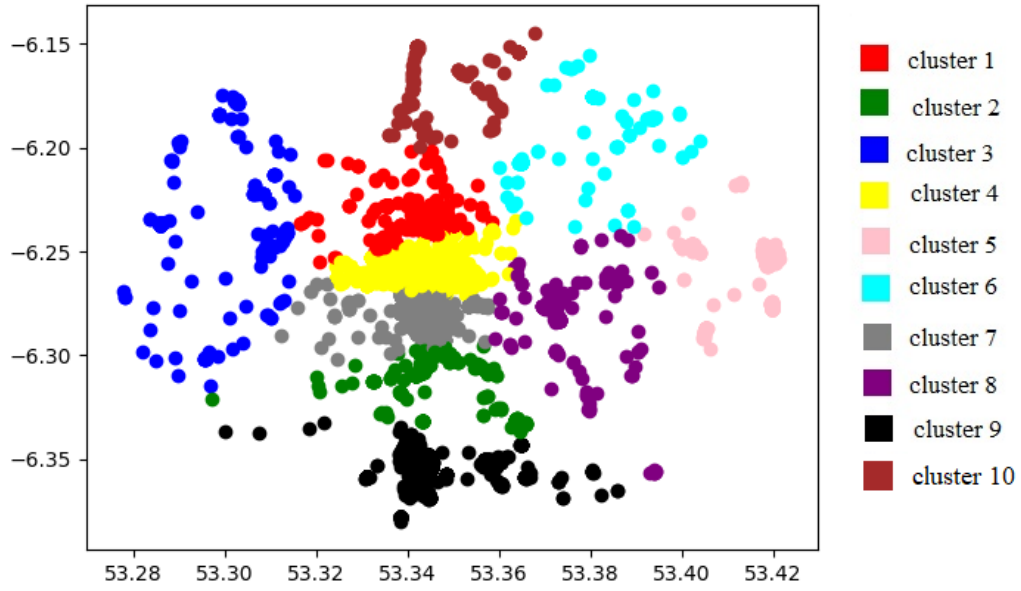

Figure 3.3 Clusters of k-means clustering algorithm of Dublin data

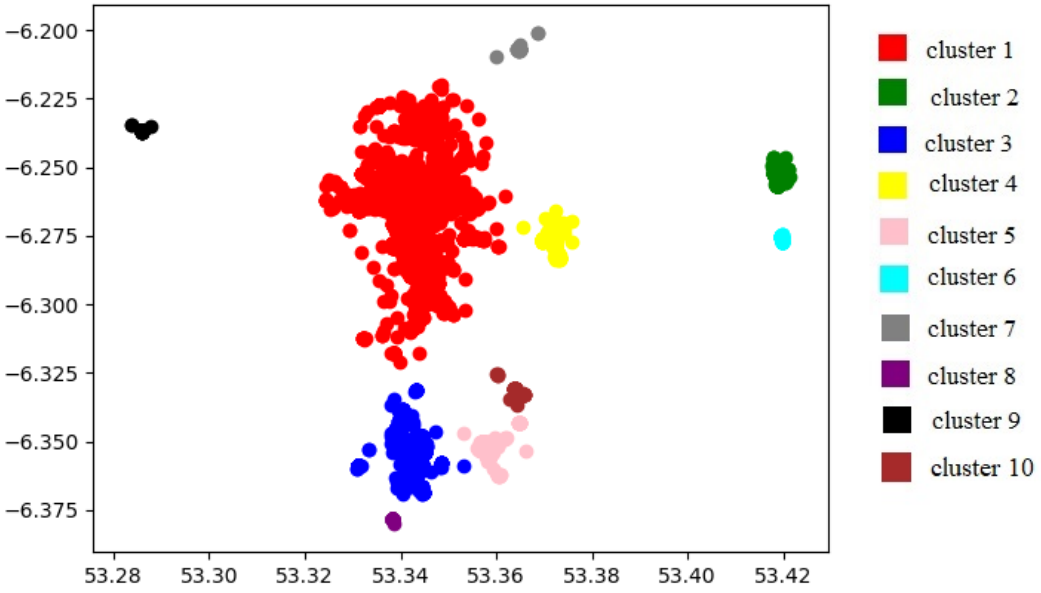

Figure 3.4 Clusters of DBSCAN clustering algorithm of Dublin data

As shown in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.4; x axis represents latitude, y axis represents longitude information and points represents latitude,

longitude information of each data in database. And also points of the same colour are in the same cluster.

Lastly, there are some data which cannot be clustered with DBSCAN algorithm. The reason of this that min_samples input of DBSCAN algorithm, given 50, limits the results. Also, the algorithm is based on spatial data but the data coordinates are very close from each other, all data in 5 kilometers.

3.2 Feature Extraction and Selection Best Images

Robust and fast image matching methods are very important subject for various applications in robotics, computer vision and image processing. There are some important methods for feature extraction and matching such as Oriented Fast and Rotated Brief (ORB), Scale Invariant Feature Transform (SIFT) and Speed Up Robust Feature (SURF). These methods are briefly mentioned below. And in this study, SIFT method is used as a part of our method for feature extraction from geo-tagged image data.

3.2.1 Scale Invariant Feature Transform (SIFT)

Scale invariant feature transform (SIFT) is one of the efficient method in feature detection and matching. Feature detection is a method for computing abstraction of the image information and determining whether there is an image feature at each point given type [40]. Scale invariant feature transform is a feature detector It connects feature points in image which is invariant to image transformations such as scale, rotation, noise [41]. In addition, SIFT provides important descriptors that can detect the correspondences between features in different images[42]. Moreover, it is very successful for estimating motion between images.

Generally, SIFT method is used for the mentioned purposes, whereas in this study, we developed a method which uses SIFT for the selection of the best 5 images of each clusters. While using the SIFT technique in our method, we use OpenCV library for Python programming language.

3.2.1.1 OpenCV Library

OpenCV is an abbreviation of Open Source Computer Vision Library which is an open source image processing and computer vision library built with C ++ and C languages and it is able to run on various platforms such as Mac OS X, Linux, Android and Windows[43]. OpenCV is designed with a strong focus on computational efficiency and real-time applications. It is published under the Berkeley Software Distribution (BSD) license developed by Intel and used for the first time by BSD, a Unix-like operating system. It is the family of free software licenses that allows this license. As such, OpenCV is free to all users, whether academic, commercial or individual.

With more than 500 functions in the OpenCV library, academic studies can be done and applications can be developed in fields such as medical imaging, product inspection, security, camera calibration user interface. Everyday every area is needed in daily life from industrial products to security products, from mobile cameras to stadium cameras. OpenCV is a library that allows users to do all these things.

OpenCV is an image processing library suitable for use in embedded operating systems because of the open source library. Furthermore, it can be compiled for the desired platforms through the open source library. OpenCV is a library that can be used for research and development in areas such as human-computer interaction, robotics, informatics and security [43].

3.2.2 Speed Up Robust Feature (SURF)

The Speed Up Robust Feature (SURF) is a keypoint detector. SURF is based on multi-scale space theory. It uses a BLOB detector which is based on the Hessian matrix to find the points of interest. Because Hessian matrix has good accuracy and performance [40].

SURF includes an orientation operator. This operator has two main phases to obtain distinctive features from images The first phase is that “key points” are extracted from distinctive locations from the images. These locations include edges, blobs, corner [44]. When key points are detected, a neighbourhood around the key point is selected

and divided into subregions [44]. These subregions are picked around every key point and distinctive feature descriptors are computed from each region.

3.2.3 Oriented FAST and Rotated BRIEF (ORB)

ORB is a fusion of the FAST key point detector and BRIEF descriptor with some modifications [41]. To briefly explain, FAST means Features from Accelerated Segment Test and it is a feature extraction method which can be used for extraction feature keypoints and tracking and mapping objects in computer vision. This method is faster than other well-known corner detection methods[45]. Also, FAST is used on real-time programs to find keypoints and match visual features [41]. Moreover, BRIEF means Binary Robust Independent Elementary Features which is feature descriptor using binary tests [46].

ORB is an alternative of SIFT or SURF. And also this method is faster than the other methods, and invariant with rotation. It resists to noise [47].

3.2.4 Applying Our Method and Results

To select best five images which represent their locations, our method is applied on the downloaded and clustered images groups. Firstly, for each images 100 features are extracted using SIFT and the score of each feature is calculated with $1/ |F(I_i)|$ where i represents images in clusters. The scores of the same or similar features are summed and the weighted features dictionary is generated. If a feature is not similar into features dictionary, the feature is added to features dictionary with its score. After the generation of the weighted features dictionary, each images scores are calculated by summing the scores of features in each images. The process is shown in Figure 3.5. At the end of the process, the images which has the highest score is selected as the best image. Then to select more than one image, the selected best image is removed from the images in its cluster and its scores of features are lowered by similarity level in weighted features dictionary. And then the scores of images are calculated again and the images with highest score is selected as second best image. The process is continued until selection of five best image in all clusters. The repeated calculation of scores of images is shown in Figure 3.6 [48].

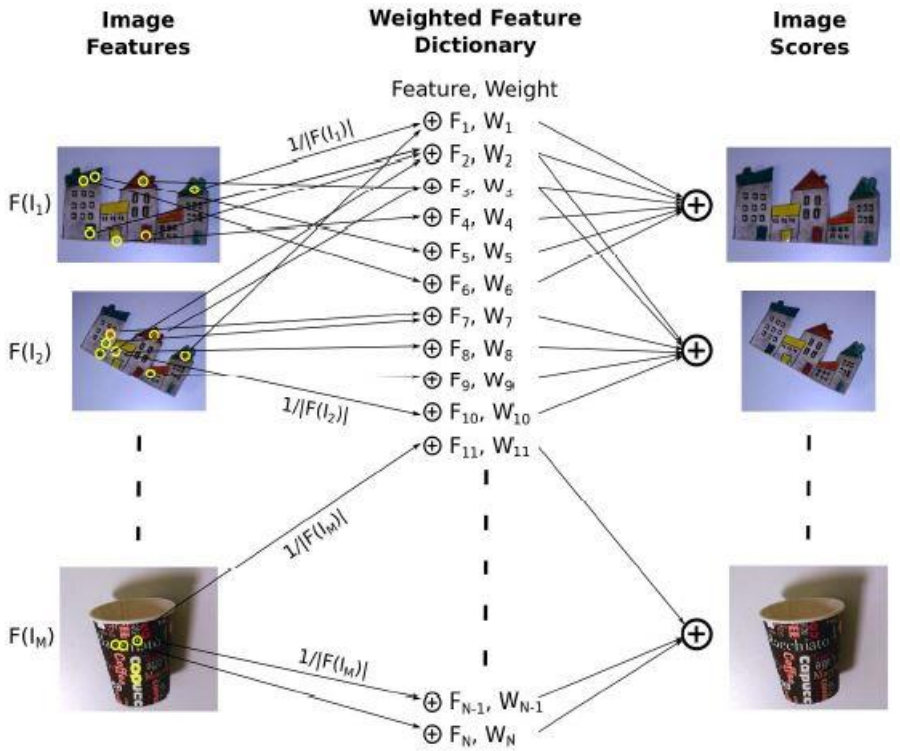

Figure 3.5 Computation of image scores [48]

1st round	1.85	1.9	3.53	3.95	3.86	**4.08**	1.2	1.3	1.69	1.57
2nd round	1.85	**1.9**	0.48	0.49	0.63	0	1.1	1.2	1.69	1.39
3rd round	0.34	0	0.48	0.49	0.63	0	1.1	1.2	**1.45**	1.39

Figure 3.6 Selection more than one images process [48]

CHAPTER 4

VISUALIZATION

The last part of this study is the visualization of the processed data. It is actually one of the main contribution of the thesis. To generate map, we use local web based project and implement one of the OpenLayers[12] Maps, which is clustered map, to visualize the data we use Javascript in the project.

4.1 Generating Map

4.1.1 ASP.NET MVC

ASP is the abbreviation of Active Server Page which is one of the most used server based web development platform that allows to develop dynamic web pages with .NET Framework [49]. Also, MVC is the abbreviation of Model-View-Controller which is architectural pattern that is shown in Figure 4.1. This pattern divides an application into three components, the model, the view, and the controller. The model defines data processes, the rules of a work and applies the logic of the data. The view shows the user interface (UI) of the application and the user interface is generated from the model data. And controller manage the user interactions, works with the model, and choses the view to render which shows user interface.

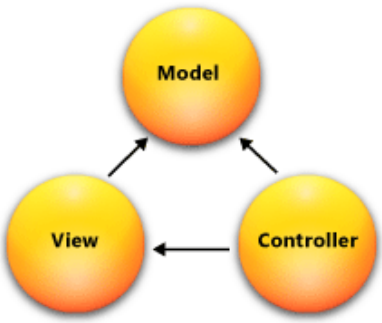

Figure 4.1 Design pattern of MVC [50]

[12] https://openlayers.org/

Thanks to ASP.NET MVC, it is possible to develop web applications that can be fast-running, testable, reusable parts using .Net framework languages and MVC pattern [50].

4.1.2 OpenLayers Library

Through the use of map technologies almost in every area today, the importance of map applications is increasing. OpenLayers has emerged at 2006 as an alternative of Google Maps and other web map applications such as MSN Virtual Earth API. This technology is a simple, open source JavaScript library that allows us to develop functional map applications to display map data on web browsers without any server side dependency [51]. It provides development of web based geographical applications. These geographical applications called Geographic Information Systems as GIS in nowadays. OpenLayers has two important terms, Map and Layer. OpenLayers stores object ans extent information on the map. The data on the map is displayed via the layer. A layer is a data source at the same time. To use OpenLayers, we need to add a script tag which contains the OpenLayers library. And OpenLayers places each map layer inside an html element to display map layers [52].

4.2 Display the Data

4.2.1 JavaScript

JavaScript is a script programming languages which is used for web programming and it is launched at 1995 by Netscape Communication Corporation. JavaScript is written in HTML codes with <script> … </script> tags. HTML is Hypertext Markup Language, not a programming language because a program that is written with HTML codes is not an executable program. Web browsers read, interpret, and visualize HTML code. Thanks to JavaScript, the HTML source codes can be changed, which makes it possible to prepare dynamic web pages. JavaScript allows user to interact on web pages such as press a button. When any interaction come up, JavaScript codes work and do necessary operations such as react to events, exhibit special effects, accept variable text, validate data, create cookies, detect a user's browser [53].

4.3 Results of Visualization

When we first started the project, all data in the database for İstanbul or Dublin selected is shown on the clustered map.

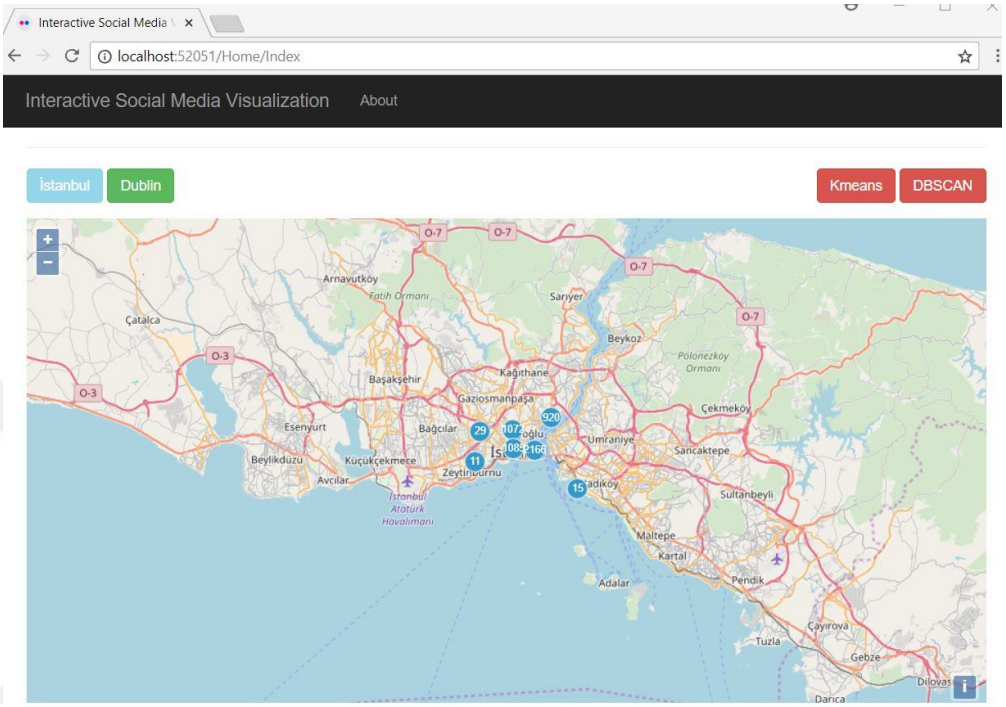

Figure 4.3 First display of all data of İstanbul

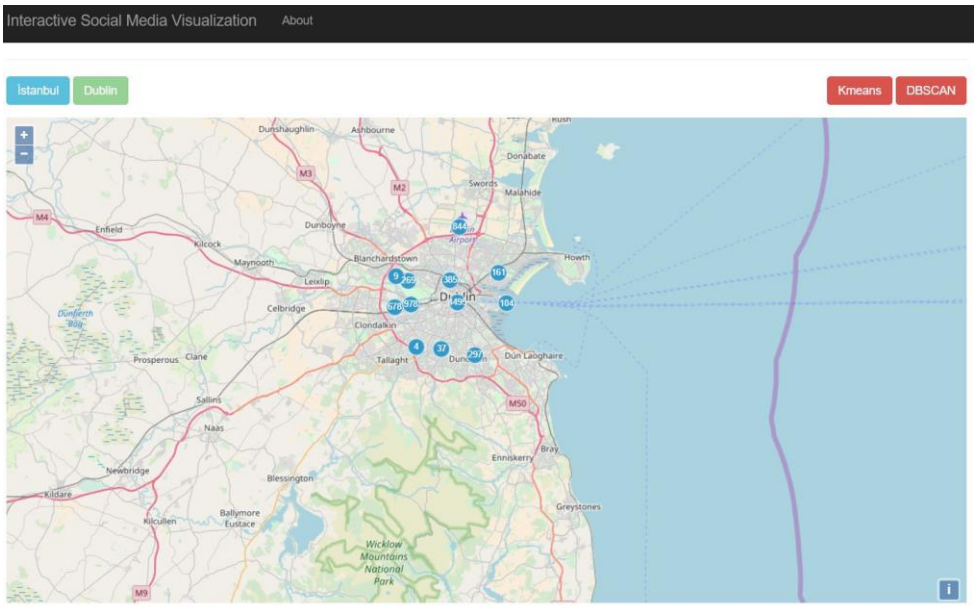

Figure 4.2 First display of all data of Dublin

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Results

While the application is started firstly, there are four buttons on the web page without a map. User must choose which city where the data is shared. One of the results obtained in the study is shown in Figure 5.1 for İstanbul**.**

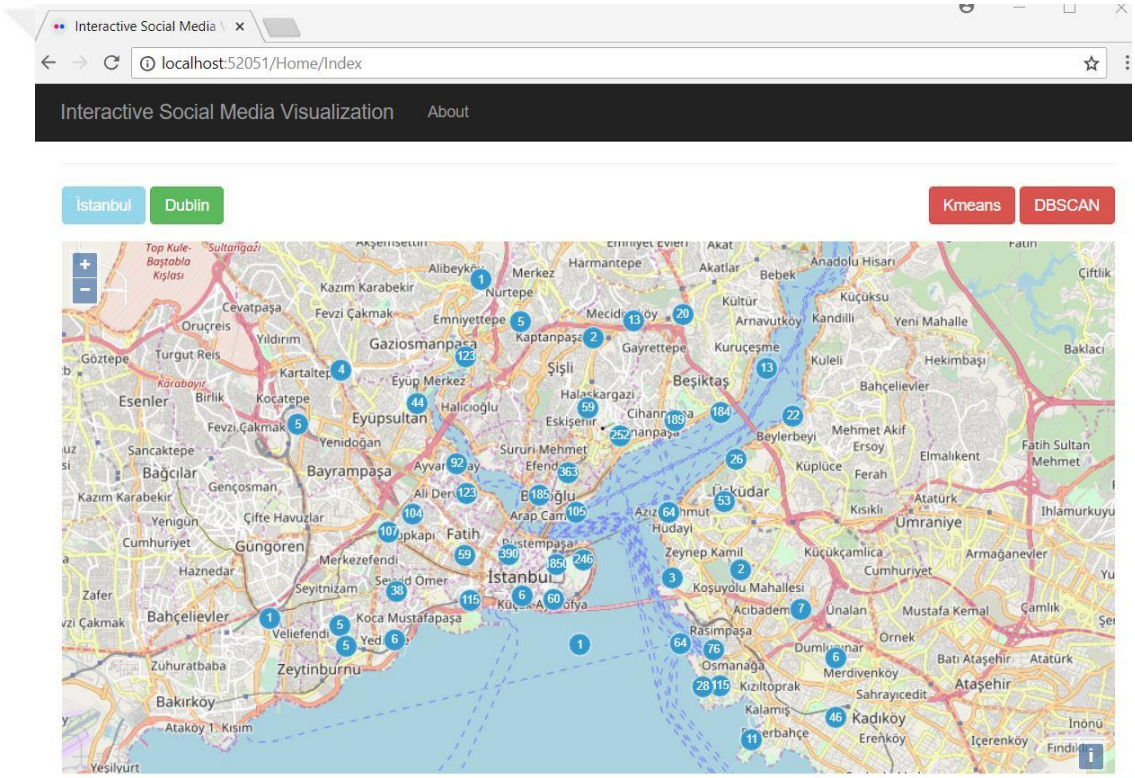

Figure 5.1 General display of İstanbul data

All data gathering from Flickr with coordinates of İstanbul are shown clustered circles of data that has closer coordinates on the map. The clustered circles are coloured blue and the number of data is written at the centres of circles. If user zoom in the map, these circles divided into smaller circles according to coordinates of data and zoom

level. And if the circles are clicked, the photos in the clicked circles are shown with a pop up window which is shown in Figure 5.2.

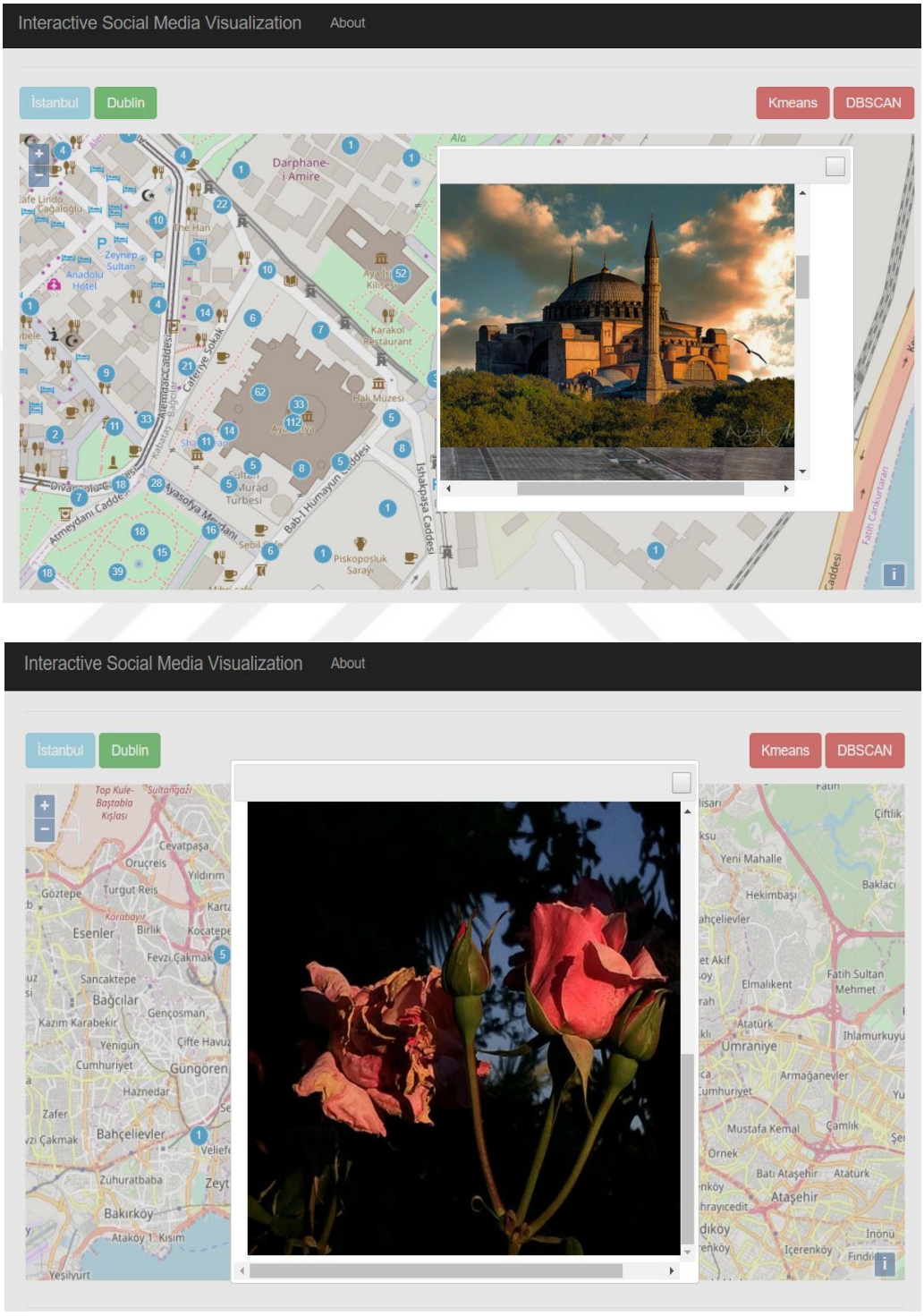

Figure 5.2 Displayed photos in clicked circles of İstanbul data

If user wants to see the selected images using k-means or DBSCAN clustering algorithms, one of the red buttons must be clicked. The result is shown in Figure 5.3 after the selection of the k-means algorithm for İstanbul, and one of the best images which are chosen using k-means clustering algorithm is shown in Figure 5.4.

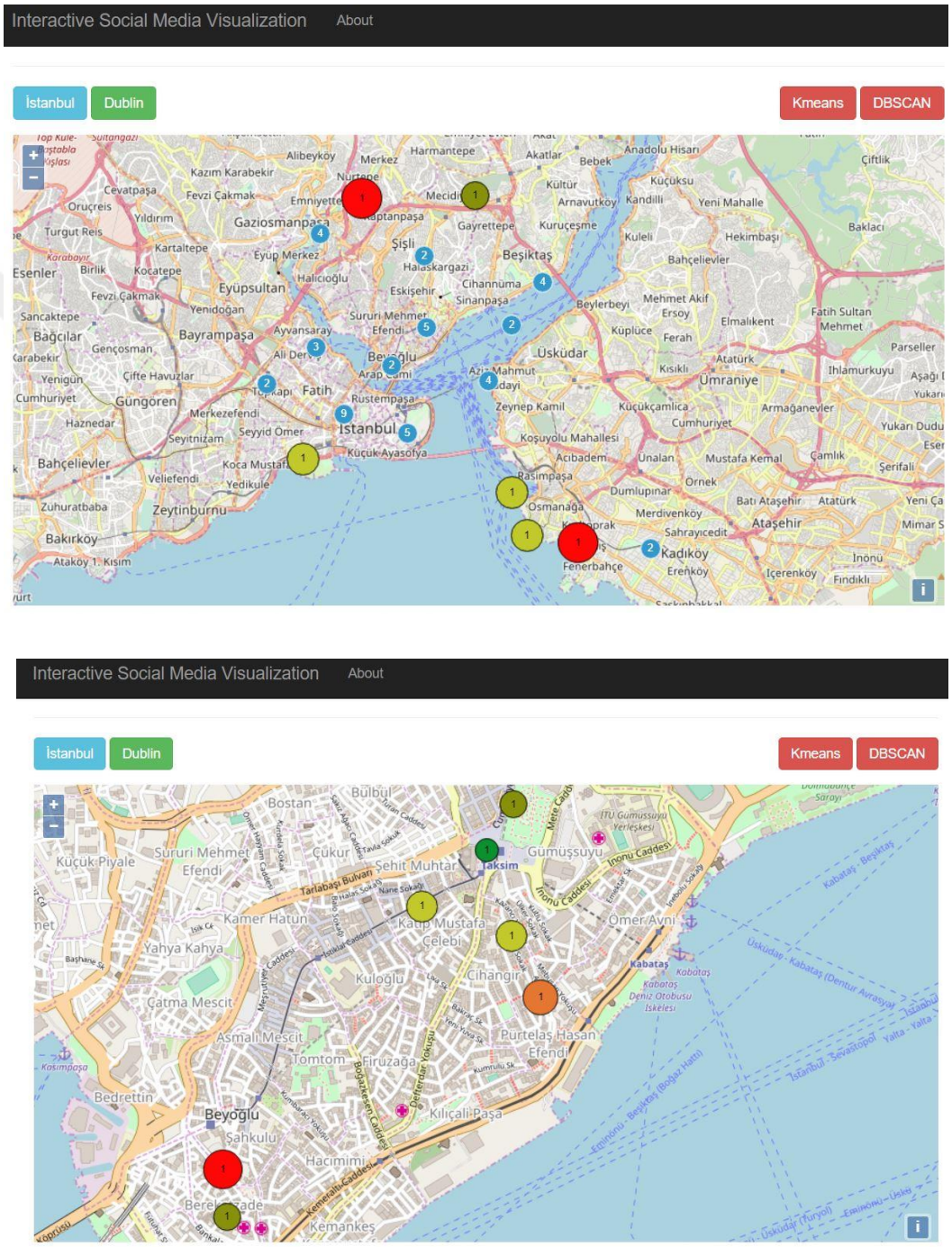

Figure 5.3 The selected images using k-means clustering algorithm for İstanbul

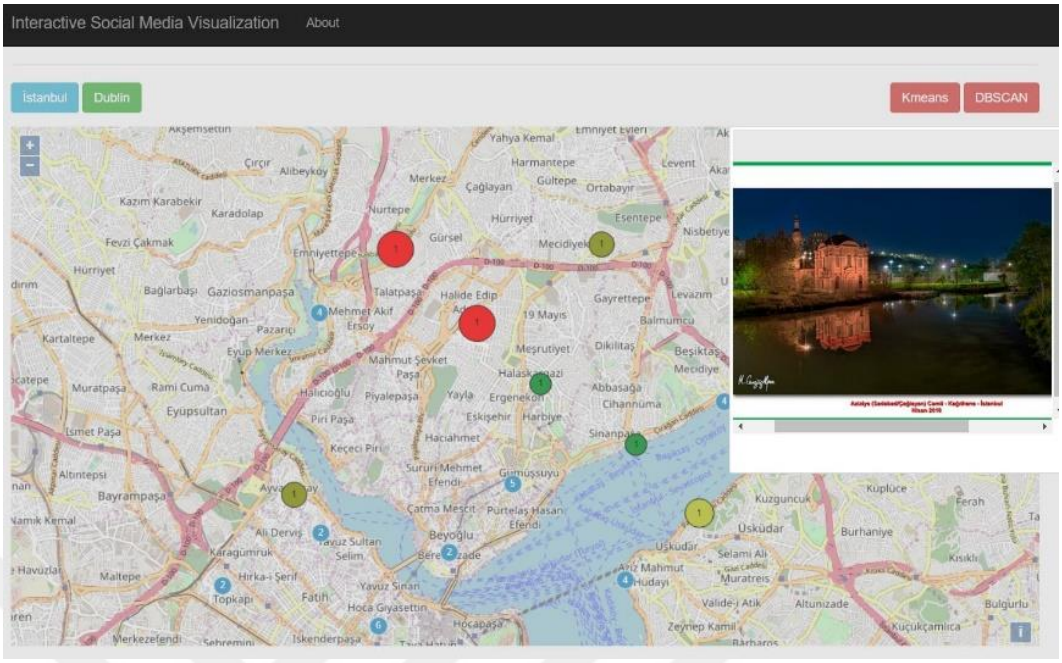

Figure 5.4 The view of one of the best images in k-means clusters

Tablo 5.1 Colour scale of selected images

COLORED CIRCLES	Ratio= Score of selected images of each cluster / Maximum score of selected images of the same cluster
RED	0.8 < Ratio <= 1.0
ORANGE	0.6 < Ratio <= 0.8
YELLOW	0.4 < Ratio <= 0.6
GREEN	0.2 < Ratio <= 0.4
DARK GREEN	0 < Ratio <= 0.2

The selected images are represented with different coloured and sized circles according to ratio as seen . The red circle means that the ratio of the image score with the maximum image score in the same cluster members is between 0.8 and 1.0, namely the red circles represents the best images in their clusters. The orange circles represent second images in each clusters which have the ratio between 0.6 and 0.8, the ratio is between 0.4 and 0.6 for the yellow circles, the ratio is between 0.2 and 0.4 for the green circles, and the last selected images in clusters are represented with the dark green circles where the ratio is between 0 and 0.2. Also the size of these coloured circles are decreases according to the ratio. The ratio of coloured circles is the same for the results of DBSCAN algorithm. The results of DBSCAN algorithm for İstanbul Data are shown in Figure 5.5, and one of the chosen images in DBSCAN clusters is shown in Figure 5.6.

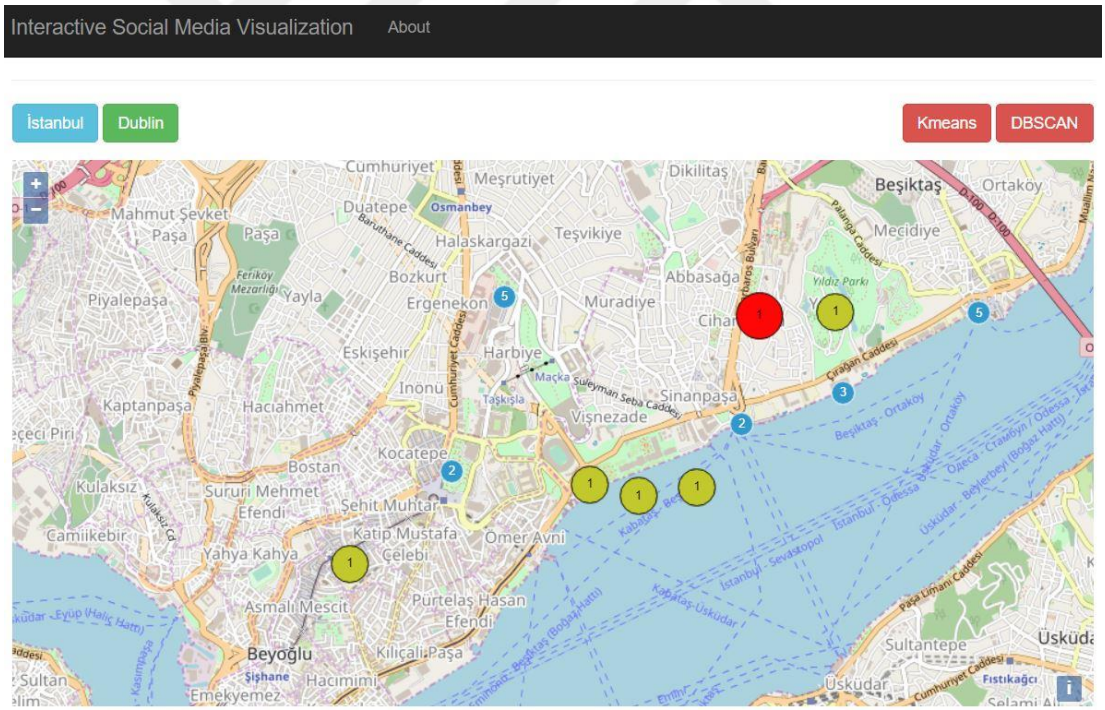

Figure 5.5 The selected results using DBSCAN clustering algorithm for İstanbul

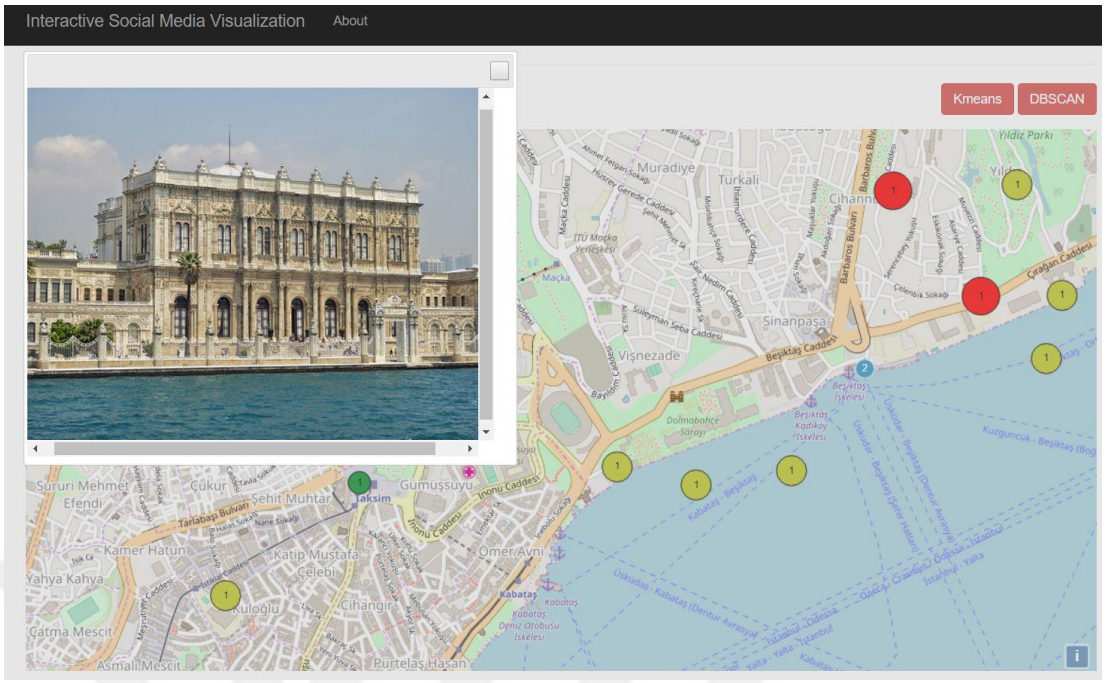

Figure 5.6 One of the photos chosen as third best using DBSCAN algorithm

The same approaches are applied on Dublin Data also. The results of the Dublin Data and the chosen mages are shown in Figure 5.7, Figure 5.8, Figure 5.9,and Figure 5.10.

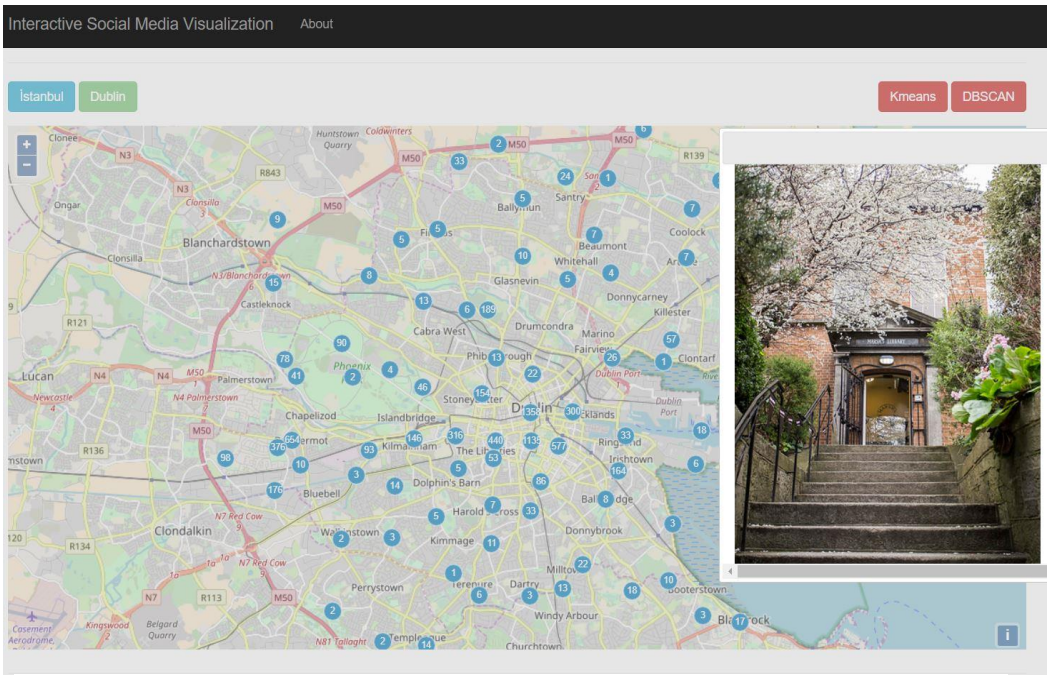

Figure 5.7 An image of all data gathered for Dublin

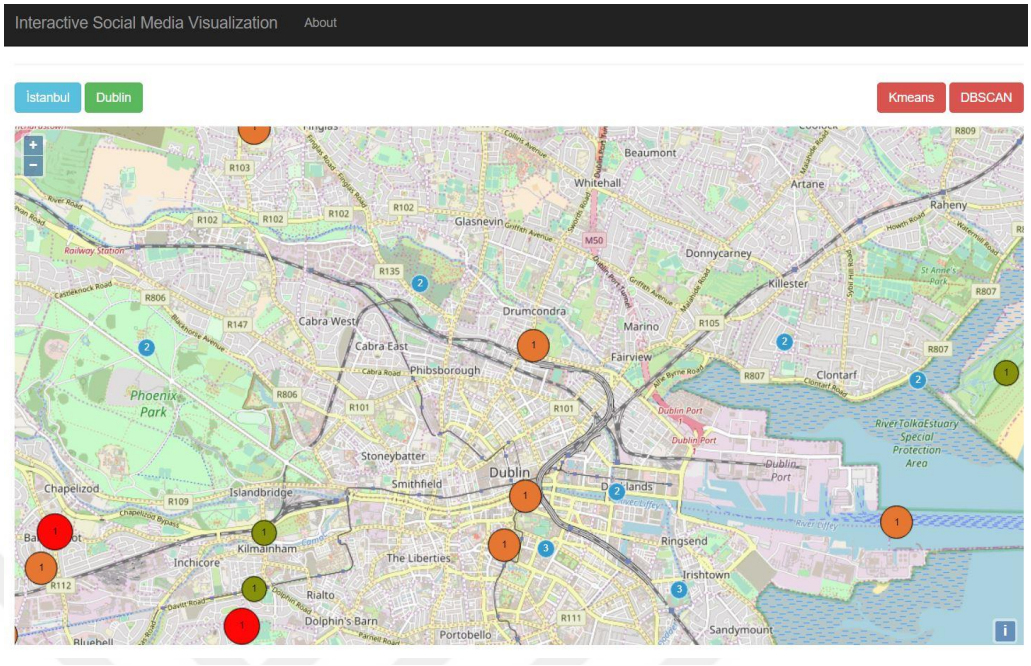

Figure 5.9 The results of k-means clustering algorithm

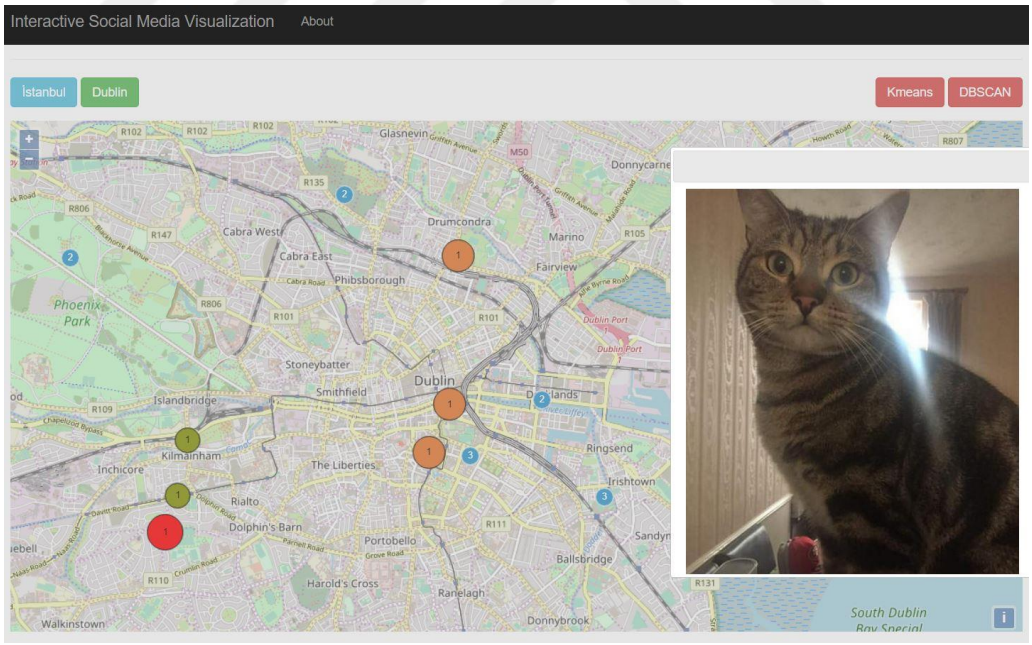

Figure 5.8 One of the selected images in k-means clusters for Dublin data

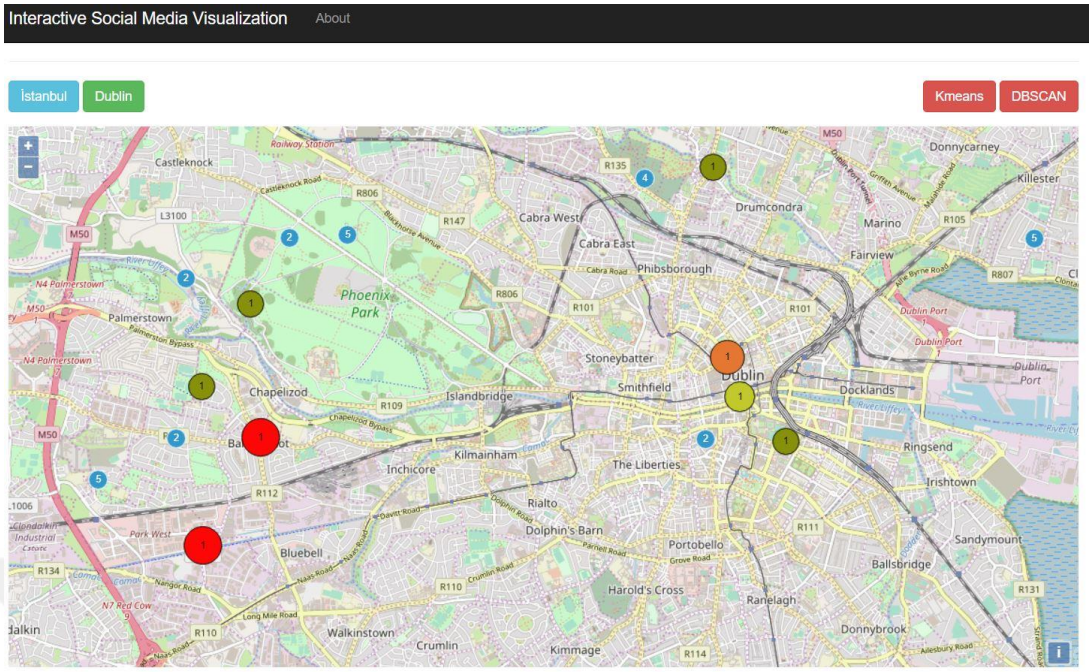

Figure 5.11 Displayed selected images in DBSCAN clusters for Dublin data

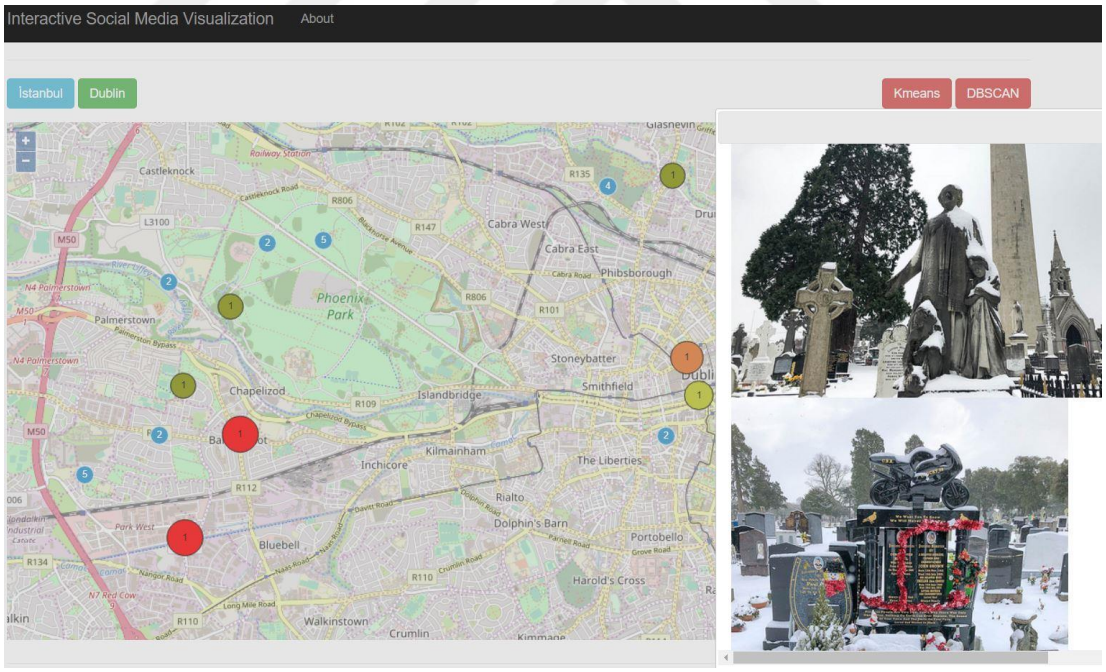

Figure 5.10 Displayed selected images in DBSCAN clusters for Dublin data

We have studied on 5298 data which are gathered in about four and a half months (between March 7 and July 20, 2018) for İstanbul and 8261 data which are gathered in

about three and a half months (between March 27 and July 20, 2018) for Dublin coordinates and when clustering the data, the clusters have the different numbers of data as seen Tablo 5.2 and

Tablo 5.3. The difference between the number of data in the clusters was a problem during the feature extraction and matching, especially when DBSCAN clustering algorithm is used. For instance, using DBSCAN algorithm for Dublin data, there are nearly 4000 data or third cluster has nearly 1250 data but seventh cluster has 80 data or tenth cluster in first cluster has 70 data.

Tablo 5.2 The number of each cluster for İstanbul data

Clusters	1	2	3	4	5	6	7	8	9	10
k-means	831	1561	245	438	278	923	149	414	386	73
DBSCAN	3128	199	246	170	128	177	154	119	123	161

Tablo 5.3 The number of each cluster for Dublin data

Clusters	1	2	3	4	5	6	7	8	9	10
k-means	3739	218	246	774	64	1356	430	330	308	796
DBSCAN	4900	680	1249	194	121	57	80	98	205	70

5.1.1 Think Results

The application has been tried by several users and comments have been taken from users to get better results about the study. First user Süleyman Furkan GÜL said that

this study is a very useful application for those who want to visit a city, they can look at the photos on the map and understand where they are worth visiting. Another user Fatime ÜZÜMCÜ said 'I can use the program when I am travelling a city that I do not know'. The other user Eda DOĞRU said that it is a useful work to use time more practically.

5.2 Discussions

In this study, our aim is to generate an interactive tourist map using our method on the geo-tagged data gathered from Flickr according to location information of photos. We use SIFT technique in order to extract 100 features from each photo in our method and select the best photos in each clusters using our method. As a result of our work, we saw that one of the photos with the same features in one location was selected as best. So, the more the images with the same visuals are shared, the better the system has chosen a photograph of that visual based on the calculated score.

As explained in Results section, the number of data of each cluster are different as shown in Tablo 5.2 and Tablo 5.3. If a cluster has more than 150 data, we picked up 150 images in the cluster randomly to apply the SIFT method and select best five images; because feature extraction and matching takes a long time. If a cluster has a very large data, the selected images may not actually be the best images because the better images may not possibly be selected. However, the clusters that has small data, all images are used for processing and the selection of best images gives main results for the clusters.

The time for selection of best five images of each cluster depends on the number of data in clusters. The processing time takes nearly between forty-five minutes and one hour for the clusters that has more than 150 data. Becuse the extracted features and feature dictionary is increasing with feature detection of every image and the comparison of the features takes longer time for the next image.

Additionally, while we are gathering data, we are saving the url information of the data to the database and we are downloading the images to the computer. And while processing the data, we use the downloaded images and select five images in the folder;

but in visualization part of the study we use the url information of the selected images. If the selected images are removed from Flickr, the images are not shown into pop up window. This is a deficit of this study. The instance of this situation is shown in Figure 5.12.

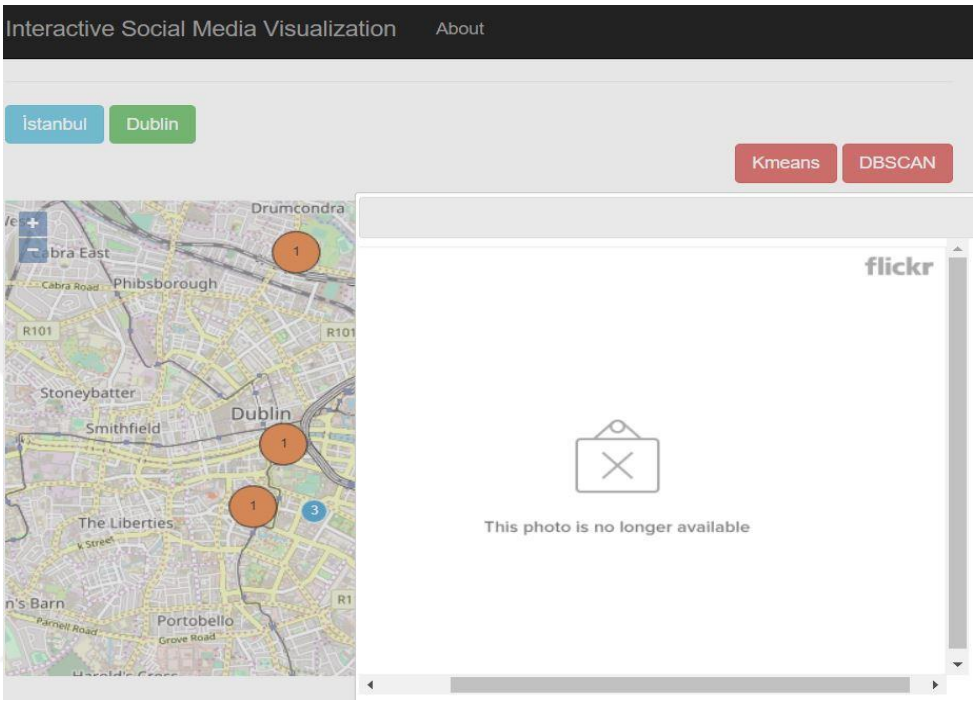

Figure 5.12 Image is not found with its url

The other issue which negatively affects the results is that the selection of the best images is influenced by the feature scores which is extracted from collected data from same users who share a lot of images on Flickr.

Theoretically, the system which is planned to develop can update with the new data and remove the expired images on map with the using time-based query. To achieve this process, the weighted feature dictionary must be updated with the new features of the new images, that are extracted using SIFT method and a dynamic database must be used.

5.3 Future Works

For the future work, we plan to include different social media platforms such as Twitter, Instagram to gather data. And we will apply ORB and SURF image matching methods to our clustering data to investigate image matching methods and to improve our study results. Furthermore, the visualization of the process data can be displayed on a 3D map. While gathering data, the number of data from same users can be limited to improve the selection best images using our method.

REFERENCES

[1] K. Linke, “Generation Facebook ? – the History of Social,” no. December 2011, pp. 1–10, 2015.

[2] D. G. Lowe, “Distinctive image features from scale invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, pp. 91–11020042, 2004.

[3] “How Facebook Was Founded - Business Insider.” [Online]. Available: https://www.businessinsider.com/how-facebook-was-founded-2010-3#we-can-talk-about-that-after-i-get-all-the-basic-functionality-up-tomorrow-night-1. [Accessed: 29-Aug-2018].

[4] “Boomsocial Facebook Ülke İstatistikleri.” [Online]. Available: https://www.boomsocial.com/Facebook/Ulkeler. [Accessed: 04-Jun-2018].

[5] “Twitter.” [Online]. Available: https://about.twitter.com/en_us/company.html. [Accessed: 29-Aug-2018].

[6] “What Is Instagram? - Business Insider.” [Online]. Available: https://www.businessinsider.com/instagram-2010-11. [Accessed: 29-Aug-2018].

[7] “Flickr.” [Online]. Available: https://www.flickr.com/jobs. [Accessed: 29-Aug-2018].

[8] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling, “NewsStand: A new view on news,” *Proc. 16th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, vol. 2008, no. November, p. 18, 2008.

[9] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, “TwitterStand: News in Tweets,” *Proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. - GIS ’09*, p. 42, 2009.

[10] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and J. Sankaranarayanan, "PhotoStand: A Map Query Interface for a Database of News Photos," *Proc. VLDB Endow.*, vol. 6, no. 12, pp. 1350–1353, 2013.

[11] R. Du and A. Varshney, "Social street view: blending immersive street views with geo-tagged social media," *Proc. 21st Int. Conf. Web3D Technol. - Web3D '16*, pp. 77–85, 2016.

[12] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo Tourism : Exploring Photo Collections in 3D."

[13] S. Agarwal, Y. Furukawa, and N. Snavely, "Building rome in a day," *Commun. ...*, pp. 105–112, 2011.

[14] A. Bulbul and R. Dahyot, "Populating virtual cities using social media," *Comput. Animat. Virtual Worlds*, vol. 28, no. 5, pp. 1–10, 2017.

[15] A. Bulbul and R. Dahyot, "Social media based 3D visual popularity," *Comput. Graph.*, vol. 63, pp. 28–36, 2017.

[16] W. Zhang and J. Kosecka, "Image Based Localization in Urban Environments," *Third Int. Symp. 3D Data Process. Vis. Transm.*, pp. 33–40, 2006.

[17] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," *Proc. 8th ACM Int. Work. Multimed. Inf. Retr. - MIR '06*, p. 89, 2006.

[18] P. Serdyukov, V. Murdock, and R. van Zwol, "Placing flickr photos on a map," *Proc. 32nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '09*, no. May, p. 484, 2009.

[19] F. Grabler, M. Agrawala, R. W. Sumner, and M. Pauly, "Automatic generation of tourist maps," *ACM SIGGRAPH 2008 Pap. - SIGGRAPH '08*, p. 1, 2008.

[20] V. Narayanan, "Using big-data analytics to manage data deluge and unlock real-time business insights," *J. Equip. Lease Financ.*, vol. 32, no. 2, pp. 1–7, 2014.

[21] “Basic Concepts | Elasticsearch Reference [6.3] | Elastic.” [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/current/_basic_concepts.html. [Accessed: 19-Jul-2018].

[22] T. Prakash, M. Kakkar, and K. Patel, “Geo-identification of web users through logs using ELK stack,” *Proc. 2016 6th Int. Conf. - Cloud Syst. Big Data Eng. Conflu. 2016*, pp. 606–610, 2016.

[23] O. Kononenko, O. Baysal, R. Holmes, and M. W. Godfrey, “Mining modern repositories with elasticsearch,” *Proc. 11th Work. Conf. Min. Softw. Repos. - MSR 2014*, pp. 328–331, 2014.

[24] “Logstash.” [Online]. Available: https://www.elastic.co/products/logstash. [Accessed: 20-Jul-2018].

[25] “Kibana.” [Online]. Available: https://www.elastic.co/products/kibana. [Accessed: 20-Jul-2018].

[26] “Python Nedir? | Python Türkiye.” [Online]. Available: https://www.python.tc/python-nedir/. [Accessed: 28-Jul-2018].

[27] “The App Garden.” [Online]. Available: https://www.flickr.com/services/api/. [Accessed: 08-Jun-2018].

[28] “About SQLite.” [Online]. Available: https://www.sqlite.org/about.html. [Accessed: 29-Jul-2018].

[29] “Features Of SQLite.” [Online]. Available: https://www.sqlite.org/features.html. [Accessed: 29-Jul-2018].

[30] D. Xu and Y. Tian, “A Comprehensive Survey of Clustering Algorithms,” *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015.

[31] “Data Clustering Algorithms.” [Online]. Available: https://sites.google.com/site/dataclusteringalgorithms/. [Accessed: 09-Jun-2018].

[32] S. S. Ghuman, “Clustering Techniques- A Review,” vol. 5, no. 5, pp. 524–530, 2016.

[33] a Nagpal, A. Jatain, and D. Gaur, “Review based on data clustering algorithms,” *Inf. Commun. Technol. (ICT), 2013 IEEE Conf.*, no. Ict, pp. 298–303, 2013.

[34] “k-means clustering algorithm - Data Clustering Algorithms.” [Online]. Available: https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm. [Accessed: 11-Jun-2018].

[35] “K-Means Clustering in OpenCV — OpenCV-Python Tutorials 1 documentation.” [Online]. Available: http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_ml/py_kmeans/py_kmeans_opencv/py_kmeans_opencv.html#kmeans-opencv. [Accessed: 11-Jun-2018].

[36] D. Birant and A. Kut, “ST-DBSCAN : An algorithm for clustering spatial – temporal data,” vol. 60, pp. 208–221, 2007.

[37] E. Güngör and A. Özmen, “Distance and density based clustering algorithm using Gaussian kernel,” *Expert Syst. Appl.*, vol. 69, pp. 10–20, 2017.

[38] “Density based clustering algorithm - Data Clustering Algorithms.” [Online]. Available: https://sites.google.com/site/dataclusteringalgorithms/density-based-clustering-algorithm. [Accessed: 12-Jun-2018].

[39] G. Boeing, “Clustering to Reduce Spatial Data Set Size,” pp. 1–7, 2018.

[40] E. Karami, S. Prasad, and M. Shehata, “Image Matching Using SIFT , SURF , BRIEF and ORB : Performance Comparison for Distorted Images Image Matching Using SIFT , SURF , BRIEF and ORB : Performance Comparison for Distorted Images,” no. February 2016, 2015.

[41] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2564–

2571, 2011.

[42] J. Xu, H. W. Chang, S. Yang, and M. Wang, “Fast feature-based video stabilization without accumulative global motion estimation,” *IEEE Trans. Consum. Electron.*, vol. 58, no. 3, pp. 993–999, 2012.

[43] “About - OpenCV library.” [Online]. Available: https://opencv.org/about.html. [Accessed: 30-Jul-2018].

[44] W. Lejmi, A. Ben Khalifa, and M. A. Mahjoub, “Fusion strategies for recognition of violence actions,” *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 2017–Octob, pp. 178–183, 2018.

[45] E. Rosten and T. Drummond, “Machine learning for high-speed corner detection,” pp. 1–14.

[46] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “BRIEF : Binary Robust Independent Elementary Features ⋆.”

[47] “What is ORB in computer vision?” [Online]. Available: https://www.quora.com/What-is-ORB-in-computer-vision. [Accessed: 31-Jul-2018].

[48] A. Bulbul and S. Ismail, “A Visual Demonstration based on Social Media Analysis of Refugees in Turkey,” 2018.

[49] “ASP .NET Nedir?” [Online]. Available: http://www.ismailgursoy.com.tr/asp-net-nedir/. [Accessed: 31-Jul-2018].

[50] “ASP.NET MVC Overview | Microsoft Docs.” [Online]. Available: https://docs.microsoft.com/en-us/previous-versions/aspnet/web-frameworks/dd381412(v=vs.108). [Accessed: 31-Jul-2018].

[51] “OpenLayers.” [Online]. Available: https://openlayers.org/. [Accessed: 01-Aug-2018].

[52] “OPENLAYERS’A GİRİŞ – CBS AKADEMİ.” [Online]. Available: https://cbsakademi.ibb.istanbul/openlayersa-giris/. [Accessed: 01-Aug-2018].

[53] “JavaScript Overview.” [Online]. Available: https://www.tutorialspoint.com/javascript/javascript_overview.htm. [Accessed: 01-Aug-2018].

[54] “What is SQLite?” [Online]. Available: http://www.sqlitetutorial.net/what-is-sqlite/. [Accessed: 29-Jul-2018].

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname : Elif ŞANLIALP

Date of Birth : June 19, 1991

Phone :

E-mail : egul@ybu.edu.tr

EDUCATION

High School : Köy Hizmetleri Anatolian High School / İSTANBUL (2005)
Burdur Anatolian High School / BURDUR (2006-2009)
(92.90/100)

Bachelor : Kocaeli University / KOCAELİ (2009-2013)
(3.19/4.0)

WORK EXPERIENCE

Research Assist. : Ankara Yıldırım Beyazıt University (2015-continued)
Karadeniz Technical University (2013-2015)

TOPICS OF INTEREST

- Information Retrieval
- Social Media Analysis