POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

POLO TERRITORIALE DI COMO

Master of Computer Science and Engineering

Data Analytics of Milano Design Week 2016 Collected Using the Official Mobile App

Supervisor: Prof. Emanuele Della Valle

Co-supervisor: Christian Marazzi

Master Graduation Thesis by:
Blertan Libusha id. 836096

Academic Year 2015/16

POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

POLO TERRITORIALE DI COMO

Corso di Laurea Magistrale in Ingegneria Informatica

Data Analytics of Milano Design Week 2016 Collected Using the Official Mobile App

Relatore: Prof. Emanuele Della Valle

Correlatore: Christian Marazzi

Tesi di laurea di::
Blertan Libusha matr. 836096

Anno Accademico 2015/16

Acknowledgement

I cannot describe everything that happened since I started my Master Degree. Consequently, I will only limit the gratitude to some of the many people who helped me to fulfil my studies, which I would like to express.

A special thanks goes to my supervisor Professor Emanuele Della Valle, for his assistance and precious support offered to me throughout my work and for the contribution in finalizing my thesis, the product of much consultation, advice and support from him.

Thank you, professor!

I would also like to thank my co-supervisor Christian Marazzi with all my deepest gratitude.

I thank my friends for the support they have given me at every moment.

I would like finally to express my deepest gratitude to my family, which I owe a lot to from the beginning and the successful finalization of this trip, where without their support I would not have been able to finish this road, which was hard as well as beautiful.

Thank you!

Abstract

Milano Design Week (MDW) is one of the biggest events in Milan, which attracts a lot of people and is very hard to manage and analyse, as it involves many different participants from sponsors to visitors.
The main goal of this work is to analyse the MDW in terms of behaviour of the visitors extracted by "digital footprints" that they leave every day (for example through their smartphones) and the impact of the events on social networks to understand, which events are most successful. These analyses can be done in several other ways, but they can result in getting to be expensive both in terms of cost and time consuming: social media and digital footprints can offer a better, cheaper, and faster alternative without a loss of quality in the results.

This work presents an analysis performed starting from data gathered from Politecnico di Milano and Fluxedo, defining a set of functions, techniques, and a code to compare them with information about the events of MDW. The object of the analysis is Milano Design Week 2016, with more than 300.000 visitors from 165 countries and over 1.000 events spread all over the city. The proposal techniques use SPARQL and R programming language to obtain some meaningful results that can be useful to organizers to improve the organization for next years.

Sommario

La Milano Design Week è uno dei principali eventi che si svolgono in città, in grado di attrarre un grande numero di persone e che presenta molte difficoltà sia nell'organizzazione sia nella valutazione dei risultati in quanto coinvolge una grande varietà di partecipanti, sponsor e visitatori.
L'obiettivo di questo lavoro è analizzare la Milano Design Week in termini di comportamenti dei visitatori, estratti dalle "impronte elettroniche" che le persone lasciano ogni giorno (ad esempio attraverso i loro smartphone), e di effetti generati sui social network, allo scopo di determinare quali eventi hanno avuto più successo. Queste analisi possono essere svolte in molti altri modi ma comporterebbero maggiori costi sia in termini economici che di tempo necessario. I social network e le "impronte elettroniche" possono offrire un'alternativa migliore, più economica e più veloce senza comportare una perdita di qualità nei risultati.

Questo lavoro presenta le analisi svolte a partire dei dati raccolti dal Politecnico di Milano e da Fluxedo, definendo un insieme di funzioni e tecniche per comparare questi dati con le informazioni relative agli eventi della Milano Design Week. L'oggetto dell'analisi è la Milano Design Week 2016, che ha attratto più di 300.000 visitatori da 165 paesi e con oltre 1.000 eventi sparsi in tutta la città. Le tecniche proposte fanno uso di SPARQL e R per estrarre alcune interessanti informazioni che possono essere utili agli organizzatori per migliorare l'organizzazione degli eventi nei prossimi anni.

List of Figures

Contents

Chapter 1

1 Introduction

This thesis tackles the problem of improving the Milano Design Week event for the coming year. Using the adequate tools, I try to discover and find an efficient way to display meaningful visualization that can be useful for the organiser, managers, and sponsors of Milano Design Week for improving the event for the coming year. My work can be considered as an extension of the research project by Politecnico and Fluxedo that deals with the problem of semantic analysis of social media postings during a Milano Design Week (MDW).

1.1 Problem statement

The analyses take place during the week of Milano Design Week (MDW) (7) the largest fair regarding interior design and furnishing, which takes place in several zones of Milan. Milano Design Week 2016 attracts more than 300,000 visitors from as many as 165 countries. During MDW, more than 2,400 exhibitors were featured among the Saloni and the Fuorisalone and over 1,000 events spread all over the city, and the MDW was open for every creative mind from all over the world and is ready to enliven the city with one-of-a-kind events. MDW was very colourful from the number of events such are:

- Salone Internazionale del Mobile, i.e. International Furnishing Accessories
- International Furnishing Accessories Exhibition, divided into two core themes such as "Classic Furnishing Accessories" and "Design Furnishing Accessories"
- EuroCucina (International Biennial Kitchen Exhibition) and FTK (Technology for the Kitchen)
- International Bathroom Exhibition etc.

For such a large event the potential reward for getting a feedback from visitors is very important. Everyone who takes part in the event starting from organizers, sponsors, and managers need both quantitative and qualitative analysis about the events and how they went. So, it was very important to gather descriptive analyses of MDW to give an idea to the management of the MDW.

Gathering data of MDW matters to improve the performance of the event for the coming year to perform better than the previous year. This can be achieved by trying to find the results, which can show something through visualization and answering a question like what I can do better looking at the results of the current year of the event. In this way, the organiser can take steps trying to perform better in the event for the coming year.

For example, during the visualization we can see how people for Tortona events are posting from Porta Venezia. So, the organiser can make advertisements of the Tortona events in Porta Venezia and it can attract more visitors since it was a successful event. However, in the end I can assume by making analyses of MDW it can increase the performance of the event in coming year, trying to decrease the weakness of the events that weren't so successful through the analyses I did.

So, Social media can be used as a cheap and good source of comments, feedbacks, and criticisms for these kinds of big events but it needs to be analysed in the correct pattern to be useful for the management.

So, Fluxedo and PoliMi gathered data of social posts of events from SocialOmeters, Fuorisalone Official database and Fuorisalone official application.

SoxialOmeters is a tool developed by Fluxedo that analyse posts of the public social streams, according to hashtags and/or geographical bounding box, so Fluxedo and PoliMi gathered posts related with MDW and were saving in RDF format and the data gathering from the official app when users gave authorization through the app with social login.

Sources as I mention above are from SocialOmeters, Fuorisalone Official data base and Fuorisalone official application.

- Fuorisalone Official data base (events/locations/itineraries)
- Fuorisalone Official App (GPS positions[1] of the App users, Events inserted in the agenda on the app, private social post (Facebook) of App users[2])
- Social Listener (Keyword-based public social post (Twitter/Instagram), Semantic analysis).

The size of the data in numbers, which were taken to be analysed are:

- 47334 public post social
- 19308 app users
- 167971 GPS observations in Milan during the day of the events
- 4216 social logins (21,8 % of app users)
- 90249 events insert in the personal agenda by all app users
- 55979 private social post analysed

[1] When the Application was running

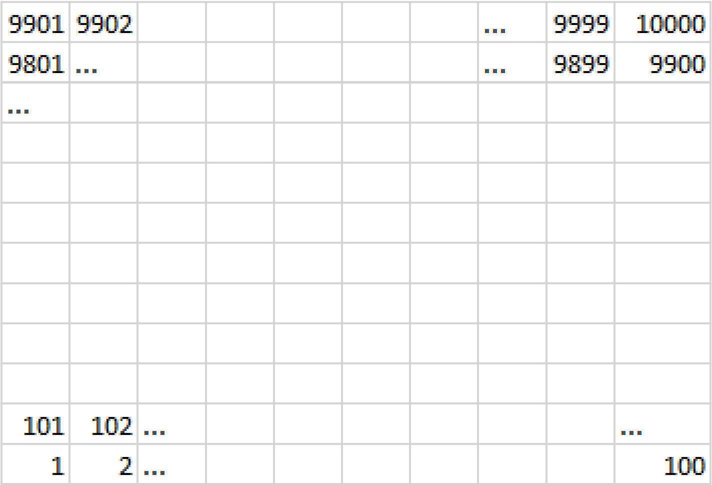
[2] To use some App features the users had ti perform a social login

1.2 Proposed solutions

Fluxedo and PoliMi performed several analyses regarding the MDW. They analysed the social coverage of MDW, by analysing how many public posts talk about MDW and Fuorisalone, how many social posts refer to each of the itineraries, which moods emerge from social posts related to each itinerary.

They also did some analyses fusing the data. All data are georeferenced and aggregated by (29) citypixel (100 x 100 mt squares). Milan was divided into citypixels and a list of cells had been associated to each district (Tortona, Brera, Lambrate, etc.).

Figure 1 shows the way the cells are numbered.



9901	9902						...	9999	10000
9801	...						...	9899	9900
...									
101	102	...							...
1	2	...							100

Figure 1. Example of numbering of the cells into that how an urban area can be divided

Merging multiple data sources, they were possible to infer information such:

- – Which events attract more visitors?
- – Which areas have the greatest presence of visitors?
- – Do people talk on the social networks about the events they are interested in?
- – Do people use social networks while visiting the events?

They as well checked the provenance. They analysed where the foreign visitors and Italian visitors come. Based on the provenance of the visitors they analysed, which venues were most visited by the people from different places.

So, at the beginning of my thesis I did at the Fluxedo company, I received help from the professor and assistant in various forms to understand the data, which they already gathered and the concept of how to deal with those data to find some solution to improve the MDW for the coming year.

As I said Fluxedo and PoliMi were working on the MDW data when I joined them and started to understand more deeply the semantic web and R programming language, which was required to visualize the results, that I was in charge to solve.

I was requested to make some meaningful results to answer the question as: summing up the number of posts with same content but being written in different way without changing the content (ex. #MilanoDesignWeek and #MilanDesignWeek or #MDW etc.). Another task was displaying the number of the posts after summing up for every day but in several intervals of time. That can show during what time the visitors are most frequent and which locations are mostly posted.

Though, I was trying to find if the posts of one event of one district are being posted in another district and if the number is large then the organiser can make advertising of the event in that district. So, advertising the most talkative district from a different area they were talking about that.

The last task was extracting most unofficial posts to increase the research and improve analysis results. These new insights are important to the organiser of the MDW. This information is important in the way that the organiser can organise the events in a better way, focusing and giving more attention to those events, which were most visited. Those insights can give an overview from the provenance of the users and can offer to them some offers through the official app to attract them to come again in the next year.

1.3 Outline of the thesis

- **Chapter 2** introduces the background concepts of the presented work, such as the Semantic Web, SPARQL Query Language, Georeferencing, Social Media, and R programming language.
- **Chapter 3** describes how the problem was solved. The ontology of the data at my disposal. It will explain in detail the importance of data I had from MDW events and the particularities of our case-study.
- **Chapter 4** describes the implementation experience that covers all the development activities, lessons learnt, which I carried out throughout the project development and reports the results of the experimental evaluation through sample data sets and some codes from Fuseki and R.
- **Chapter 5** draws the conclusion of the work and proposes future extensions of the present work.

Chapter 2

2 State of the art

2.1 Social Media

Social media technologies are computer mediated, which we can share and create ideas, information, and a lot of other interests of that can express in different forms through networks and virtual communities. The variety that was created over social media services currently available presents challenges of definition. Social media we can definition in some most common features.

1.Social media are interactive Web 2.0 Internet-based applications.

2. Construct a public or semi-public profile, such as text posts or comments, photos or videos.

3.Users create profiles for their service-specific to articulate a list of other users with whom they share a connection.

4.Social media simplify the development of online social networks by allowing the users view and traverse their list of connections and those made by others within the system.

The terminology and nature of these networks can be different base on the site how they diversity.

Internet and the World Wide Web they were trying always to make much easier the social interaction, the appearance and fast diffusion of Web 2.0 functionalities in the beginning of new millennium allowed an evolutionary jump ahead in the social component of web use. This makes to decrease the costs and online data storage made it possible to offer Internet to the users to have access to an array of user-centric when they can fill up with user-generated content, through this diverse with set of opportunities for connecting these together to form virtual social networks.

2.2 Differences between Social media and other forms of communications

Social media is different from what we have seen in electronic media such as TV, radio or paper-based. Social media architecture is based on web-based technologies, desktop computer and mobile technologies (Smartphones and Tablets), where it creates extremely interactive platforms along individual, organizations and communities can share, discuss, modify their content or content posted online. They are different in several ways. Social media interacts in dialogic system as many sources to many receivers. So, this is different from what we have seen in the traditional media, they operate in a monologic[3] system as one source to many receivers. There is no real interaction between participants since the communication is only one-directional and has no interest or concern for the listener's attitudes. The most popular social media services websites are Facebook, WhatsApp, Tumblr, Instagram, Twitter, Pinterest, LinkedIn, Google+, YouTube, Snapchat etc. These social media websites have more than 100,000,000 registered users. The next table will show the most used social network sites worldwide as of January 2017, active users.

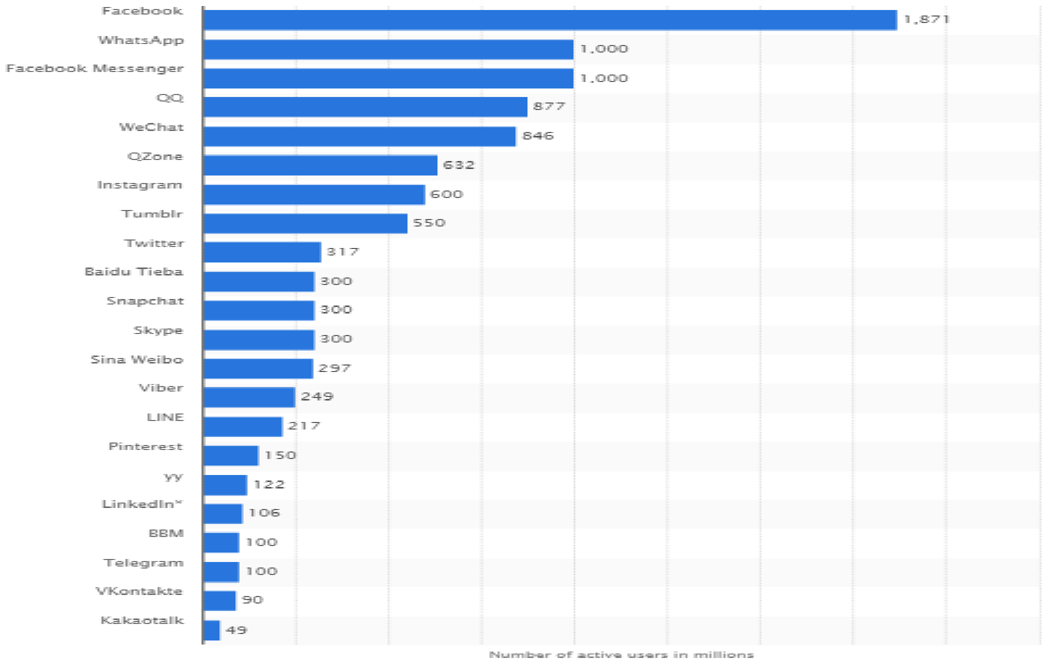


Figure 2. Leading Social Media Services Worldwide by active user accounts, Millions

[3] Monologic - monologic system can be described as an occasion where one person speaks, and the other listens.

2.3 Why Social media are unique and their timeline

The social networks are unique because those allow us to meet strangers individually, also allow the users to post and make their social networks visible. Social network sites (SNSs) sometimes are not there just to look at meeting new people, but they are there to communicate with the people who are already part of their social network.

The world of social media started in 1997, with six degrees.com being the first major social network.

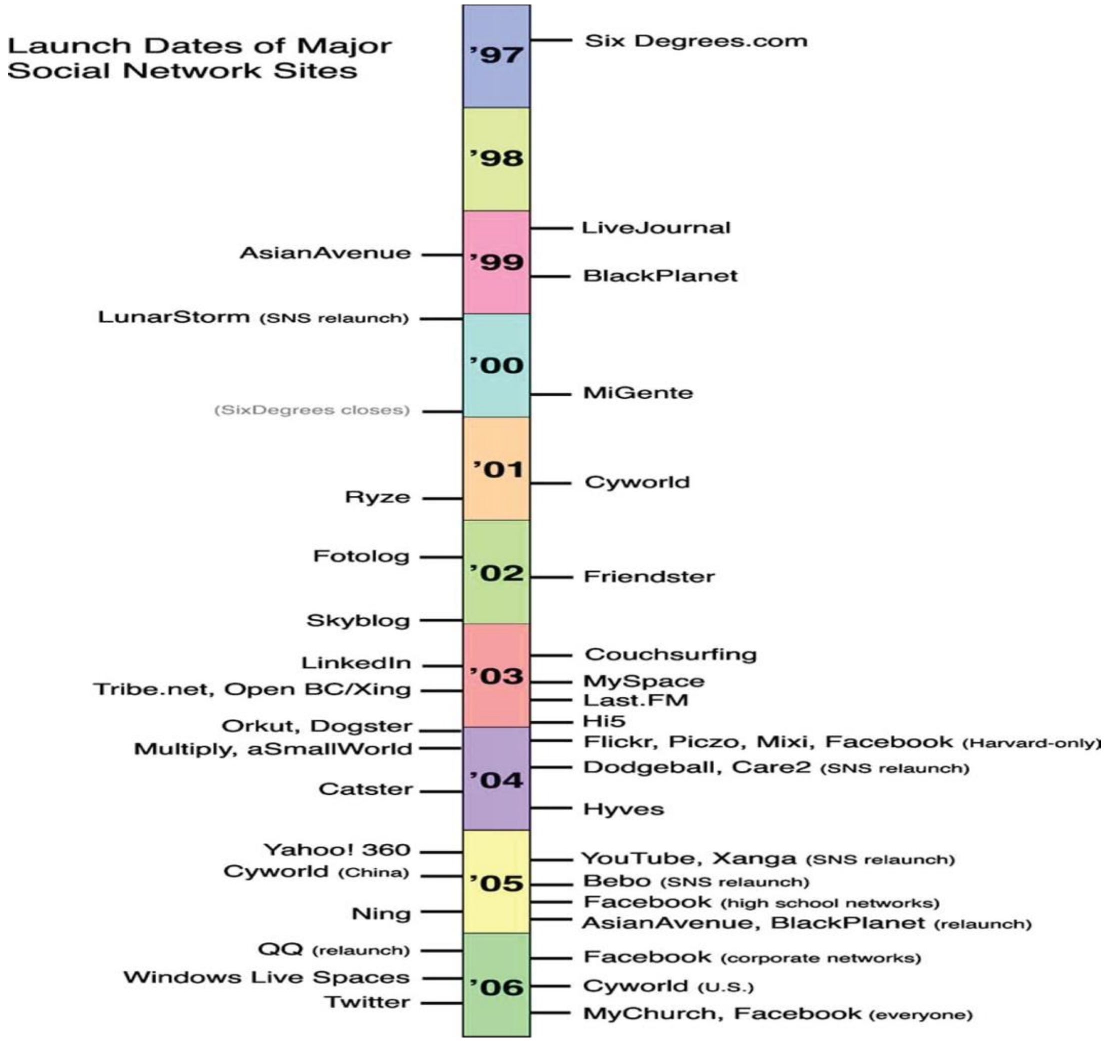


Figure 3. History of Social Media Networks

2.4 Geo reference data

Georeferencing means that you can turn to a ground system of geographic coordinates the coordinate system of a map or aerial photo image. The software of Georeferencing it has effect to display ground coordinates such as latitude/longitude and measure areas and ground distances. So, Georeferencing referred to location in physical space. The term 'geographical object' it is related with everything that can be associated with geographical location such as bridges, roads, buildings etc. Geographical location signifies a spatial reference. Then the Georeferencing we can use to a variety of objects or structures, which they are linked with geographical location. Geographical locations are presented using coordinate reference system mostly.

Geographical locations can be represented in several ways of spatial dimensions:

- 0-dimensional (points),
- 1-dimensional (lines),
- 2-dimensional (areas)
- 3-dimensional (bodies).

Ex. To 0-dimensional point locations can be referenced points of interest(POIs), to line locations can be referenced road segments. Buildings can be represented as 3-dimensional models since they are except to not move and also as 2- dimensional referenced to area locations.

Geographical locations are distinguished between informal and formal referring to location said by Hill (2006). Informal georeferencing are linked to geographical objects such are place names. Formal georeferencing tell us the exact location referenced in technology and science.

To reference to geographical objects can used three types of information:

1) Geometrical information that specify properties of an object (layout, shaped etc.).

2) Topological information, which specify properties and are protect under some circumstances deformations of object.

3) Semantic information, which can be linked with geographical location.

So, combining this method can identify the geographical object uniquely. The process, which identify the geographical objects and link them to geographical location is called matching.

2.5 Semantic Web definition

The Semantic Web, as an interdisciplinary research field, emerged out of the longing to improve the World Wide Web in such a way that interoperability and integration of multi-authored, multi-thematic and multi-perspective information and services could be realized seamlessly and on-the-fly.

Semantic (Web) Technologies are under generous examination in many disciplines where information reuse and integration on the Web promises significant added value, e.g., in the life sciences, in geographic information science etc. In the meantime, Semantic Technologies are likewise being grabbed to upgrade arrangements in application regions, which are not principally focusing on the World Wide Web but rather need to get to comparative difficulties, such as enterprise information integration, intelligence data analysis, and expert systems.

The Semantic Web ranges from foundational disciplines to application areas. In terms of size and impact of its scientific community, it has emerged as a major field within Computer Science during the last decade.

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [Tim Berners-Lee et al. 2001] (1)

2.6 Semantic Web Stack

In a classical view, we can build a tower or stack from technologies and concepts that are used in Semantic Web. The most famous stack of Semantic Web, offered by the World Wide Web Consortium (W3C) members, has seven levels or layers. Layers are categorized as: Hypertext Web technologies, Standardized Semantic Web technologies and Unrealized Semantic Web technologies.

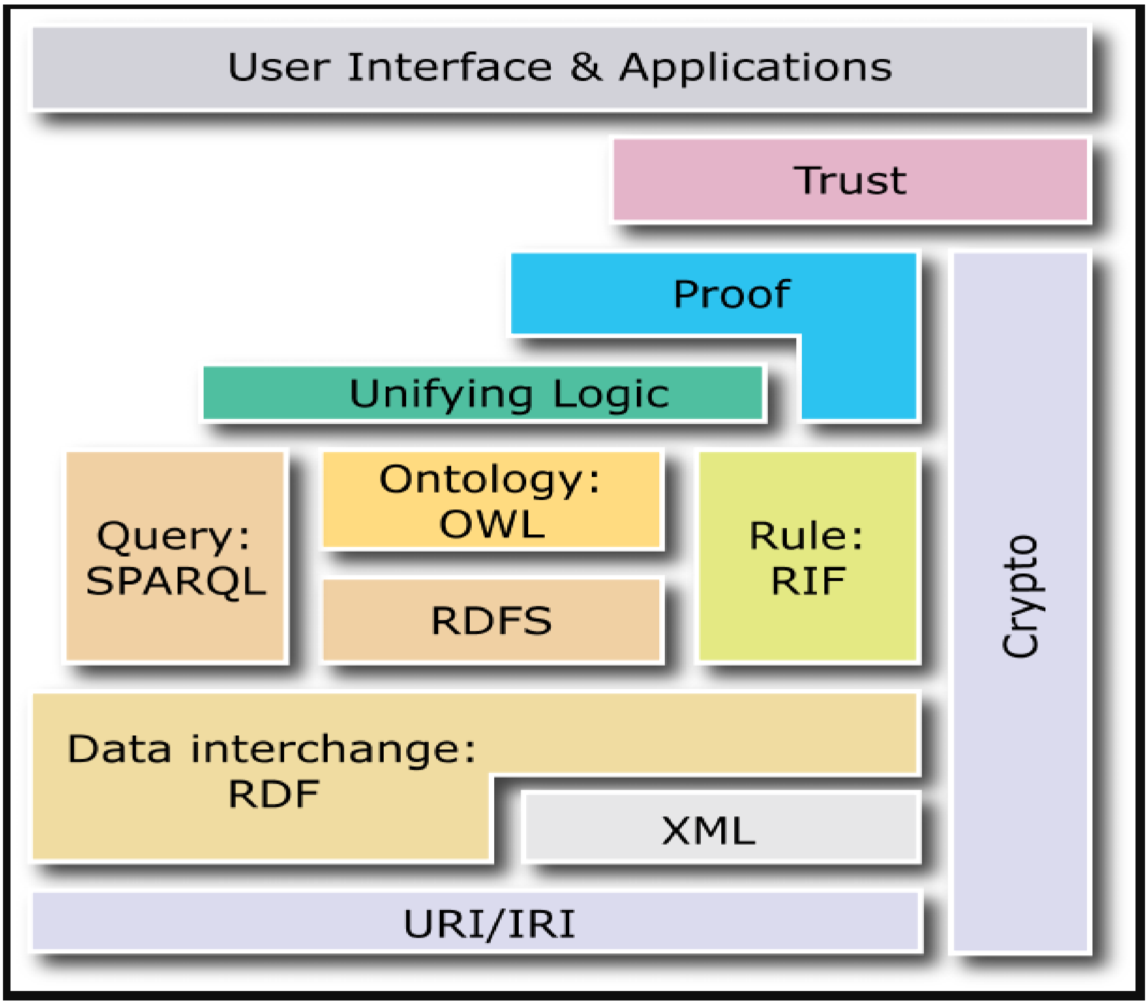


Figure 4. Semantic Web Stack

2.6.1 Hypertext Web technologies

Hypertext Web technologies layer is the bottom layer and include technologies that are notable from hypertext web and that without change secure basis for the semantic web.

- Unicode is a standard of encoding international character sets and it permits that every human language can be used (written and read) on the web utilizing one standardized form.
- Uniform Resource Identifier (URI) allows us to uniquely identify resources (e.g., documents) and it is a range of a standardized form that. A subclass of URI is Uniform Resource Locator (URL) that has mechanism access and a location of a document – like: http://www.example.org/.
- Extensible Markup Language (XML) layer with XML namespace and XML schema definitions ensures that there is a typical syntax used in the semantic web. XML is a general language for documents containing structured information. With XML namespaces, we can use different mark-up vocabularies in one XML document. Then XML schema it serves for stating schema of a exact set of XML documents.

2.6.2 Standardized Semantic Web technologies

The Standardized Semantic Web technologies layers include the technologies that are standardized by W3C to allow building semantic web applications.

- RDF Schema and Web Ontology Language (OWL): ontology definition languages for defining the vocabulary of RDF graphs
- Resource Description Framework RDF: a data model to define the structured data
- RDF/XML, Turtle, N-Triples: a variability of data interchange formats
- SPARQL: a query language for RDF data
- (31) RIF is a rule interchange format. A set of dialects to enable rule exchange among different rule systems.

2.6.3 Unrealized Semantic Web technologies

Unrealized Semantic Web technologies include technologies that are not standardized yet or include only ideas that need to be implemented in the way to be able to realize Semantic Web.

- It is normal that all the semantics and principles will be executed at the layers beneath Proof and the outcome will be utilized to demonstrate reasoning. Formal proof together with trusted inputs for the proof will imply that the outcomes can be trusted
- Cryptography - For dependable information sources, cryptography means are to be used, for example: digital signatures for verification of the origin of the sources
- User interface is the final layer that will allow humans to use semantic web applications.

2.7 Semantic Web Technologies

There is an arrangement of technologies and frameworks that makes the fulfilment of the Web of Data conceivable and are expected to give a formal depiction of ideas, terms and relationship inside a given knowledge area. They are:
Currently, the most important ontology languages for the Web are the following:

- RDF Schema and Web Ontology Language (OWL): ontology definition languages for defining the vocabulary of RDF graphs
- Resource Description Framework RDF: a data model to define the structured data
- RDF/XML, Turtle, N-Triples, N3: a variability of data interchange formats
- SPARQL: a query language for RDF data

2.7.1 The Resource Description Framework (RDF)

The Semantic Web is built on the W3C's Resource Description Framework (RDF) that is a World Wide Web Consortium (W3C) proposed standard for defining the architecture needed for supporting the encoding, exchange and reuse of structured Web metadata.
RDF states mechanisms that support regular conventions of semantics, syntax and structure, enabling metadata interoperability. RDF stands for “Resource Description Framework”), a property name, and a property value:

subject (resource identifier)	predicate (property name)	object (property value)
richard	homeTel	(229) 276-5135
cindy	Email	cindym@gmail.com

Resource Description Framework it's a data model, which the basic unit of information is known as a triple.

A triple consists of a subject, a predicate, and an object. Subject is resource identifier what object the triple is describing. Predicate defines data in the object we are giving a value and Object is the value.

The triple's subject and predicate must be URIs to remove any uncertainty from the information specified by a given triple.

RDF isn't a data format, but a data model with a choice of syntaxes for storing data files. In this model, you have three-part statements known as triples. Every triple states a fact. RDF does not define the fixed semantics, but it allows every community to do in their own metadata as needed. RDF uses XML as a usual syntax for exchanging and processing metadata.

A RDF data can be used also in other forms not only as statement, as well as a graph, in that assets are nodes, property types are coordinated marked circular segments and atomic value are cited strings. Every bolt or edge is a RDF articulation. The name toward the begin of the bolt is the announcement's subject, the name toward the finish of the bolt is the announcement's object, and the name that names the bolt is the predicate. The relationship goes in one bearing thus the triple can be perused in just a single way. A RDF-graph communicates the very same data as a RDF piece of information composed as triples, but the graph shape makes it simpler for human creatures to see the structure in information.

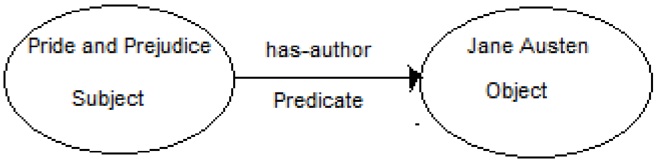


Figure 5. The RDF- graph representation of the example statement

RDF graphs (22) can be collected and organized into RDF datasets, that consist of one default graph and zero or more named graphs. Putting data in different named graphs can help to divide data in order to keep their contents separate and to ease query operations. Each named graph is identified by a unique URI called graph name.

It's important to use unique identifiers for resources for allowing the unambiguous association of properties; for example Jane Austen may be the author of ‘Pride and Prejudice’, but also the member of a sports club. The use of unambiguous identification of resources allows the correct use of the same resource for both the properties (the author relationship and the membership one).

2.7.2 Data Interchange formats

A RDF model is a graph and a technique to exchange RDF graphs and writing down this data into a textual file is needed. They are some formats the allows this conversion. They are RDF/XML, Turtle, N-Triples, N3. The difference between then it is not in context but in the form how are them written.

RDF/XML in the beginning was the only format to write the RDF models. But, it was some restrictions, which was compulsory by XML.

After, to pass these restrictions, Turtle (the Terse RDF Triple Language) (23) format was defined by Dave Beckett as a subset of Notation3 (N3) language, a syntax defined by Tim Berners-Lee and Dan Connolly's. Turtle can only serialize valid RDF graphs, differently from N3 that has a higher expressive power and is used to write "N3 Logic".

Turtle it is easier to be readable and can be easier edited manually than the RDF/XML formats. Moreover, a Turtle syntax it used by SPARQL, the query language for RDF, for the triple patterns and for RDF triples in the CONSTRUCT clause.

Simple triples are a sequence of subject, predicate, object terms, divided by whitespaces and terminated by '.' after each triple.

Here is an example of Turtle syntax:

```
@prefix ns: <http://ex.com/mynamespace/>
ns:object1 ns:property1 ns:object2 .
ns:object3 a ns:type1;
ns:property2 ns:object4 .
```

note that the ";" lets you specify different predicates that apply to a subject.

(24) Simplified version of Turtle it is called N-Triples. N-Triples it removes almost all of the short hand. N-Triples has more words than Turtle, but N-Triples can be suitable when millions of triples have to be handled. The above piece of code would look like this written in N-Triples:

```
<http://www.w3.org/2001/sw/RDFCore/ntriples/>
<http://purl.org/dc/elements/1.1/creator> "Dave Beckett" .
<http://www.w3.org/2001/sw/RDFCore/ntriples/>
<http://purl.org/dc/elements/1.1/creator> "Art Barstow" .
<http://www.w3.org/2001/sw/RDFCore/ntriples/>
<http://purl.org/dc/elements/1.1/publisher>
<http://www.w3.org/>
```

Which represents the following RDF/XML:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://www.w3.org/2001/sw/RDFCore/ntriples/">
    <dc:creator>Art Barstow</dc:creator>
    <dc:creator>Dave Beckett</dc:creator>
    <dc:publisher rdf:resource="http://www.w3.org/"/>
  </rdf:Description>
</rdf:RDF>
```

Peoples currently thinks Turtle and N-Triples are replacements for the obsolete RDF/XML format. However, if you write few hundred triples by hand is better to use Turtle, and if you have to publish large RDF data sets like DBpedia then it is better to use N-Triples.

2.7.3 RDF Schema: Basic Ideas

RDF is a language, which allows the users to describe the sources using their own vocabularies. RDF doesn’t make any supposition about any particular application domain, but neither doesn’t the semantics of any domain. So, is it based to the user to do so in RDF Schema (RDFS).

It is required from resource description communities to say some things about certain kinds of resources.

To describe bibliographic resources, for example, descriptive attributes including ‘author’, ‘title’, and ‘subject’ are common. It exists also digital certification, attributes such as ‘authorization’ and ‘checksum’, which they are often required. Properties (attributes) and their corresponding semantics are defined in the context of RDF as an

RDF schema. A schema does not define only the properties of the resource (Title, Author, Subject, etc.) but also things of resources being described define (books, Web pages, people, etc.).

This does not specify just descriptive elements such as ‘author’ but it specifies the mechanisms need to define those elements, to define the classes the may be used with possible combination of classes and relationship in restricted way, and trying to find violations of this restrictions.

RDF Schema mechanism affords a basic type system for use in RDF models. RDF schema defines properties and resources such as Class and subClassOf where they are used in specifying application- specific schemas.

An example of RDFS vocabulary presenting taxonomy of classes and properties and usage of variety and domain of properties:

```
@prefix :     <http://www.example.org/sample.rdfs#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
:Dog      rdfs:subClassOf :Animal.
:Person   rdfs:subClassOf :Animal.

:hasChild rdfs:range :Animal;
        rdfs:domain :Animal.
:hasSon   rdfs:subPropertyOf :hasChild.

:Max      a :Dog.
:Abel     a :Person.
:Adam     a :Person;
        :hasSon :Abel.
```

2.7.4 Querying the Semantic Web - SPARQL

The name is a recursive acronym for SPARQL Protocol and RDF Query Language, which is described by a set of specifications from the W3C. SPARQL (25) is a query language designed to query RDF databases and, as a query language, it only queries the information held in the models, without any kind of inference. SPARQL takes the description of what the application wants, in the form of a query, and returns that information, in
the form of a set of bindings or an RDF graph.

A SPARQL query typically says "I want these pieces of information from the subset of the data that meets these conditions." Queries in SPARQL are sent from a client to a service known as SPARQL endpoint using HTTP protocol. URI, is a SPARQL endpoint, which you send the request. The SPARQL Request can be differs as query request or as an update request.

Different kind of queries are defined in SPARQL. They are:

- SELECT (26) Returns all, or a subset of, the variables bound in a query pattern match.
- WHERE clause says "pull this data out of the dataset," and the SELECT part names, which parts of that pulled data you actually want to see.
- ASK – asks query processor to return something if it has any matches the given graph, if 'yes' the processor returns true if 'no' it returns false.
- DESCRIBE – queries return a single result RDF graph holding information related to the nodes matched in the graph pattern.
- CONSTRUCT - Returns an RDF graph constructed by substituting variables in a set of triple templates.

An example using SELECT and WHERE:

```
SELECT ?craigEmail
WHERE {
?x ab:email ?craigEmail
}
```

Illustrated in Figure 5, a SPARQL query's WHERE clause says "pull this data out of the dataset," and the SELECT part names, which parts of that pulled data you want to see.

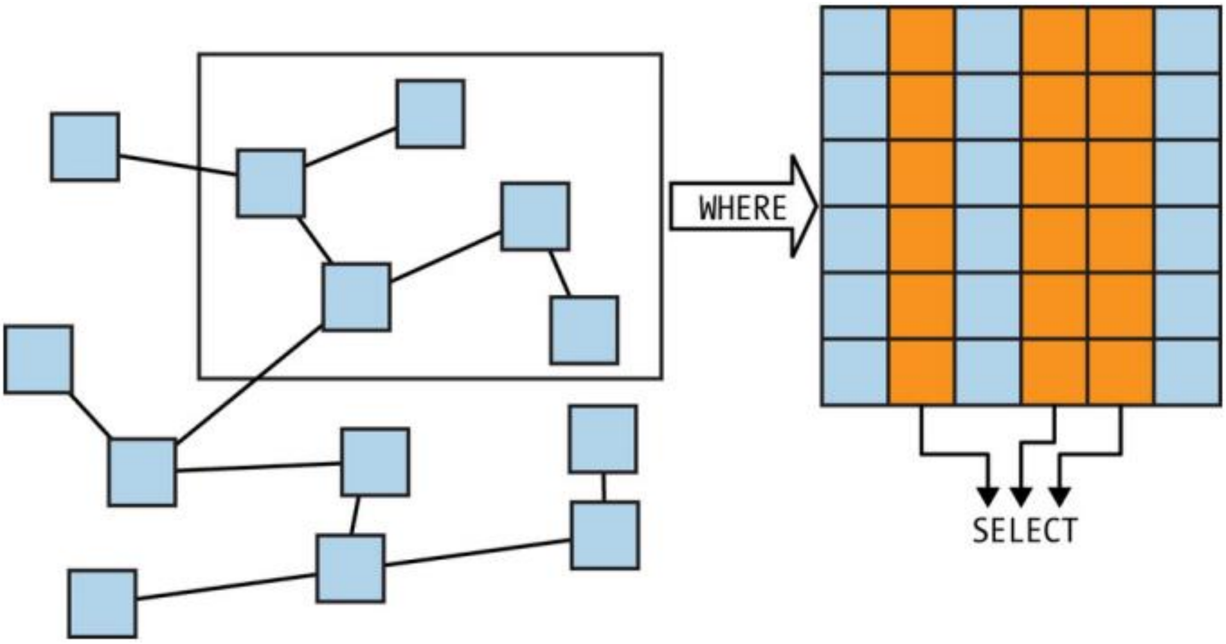


Figure 6. WHERE specifies data to pull out; SELECT picks, which data to display

2.8 Apache Jena

Apache Jena is an open source Semantic Web framework for Java. Jena affords an API when allow us to extract data from and write to RDF graphs. The main Java class is the Model one that have many methods for writing RDF-based programs and applications. The serialisation of RDF graphs Jena support the relational databases or in RDF/XML, Turtle or N3 formats. ARQ is a good solution for embedding SPARQL functionalities, if you are using the Jena RDF framework. ARQ is query engine for Jena that supports the SPARQL RDF Query language.

Fuseki is part of the Apache Jena project. Fuseki is a SPARQL server and consist an HTTP interface to RDF data. (4) Fuseki it supports SPARQL for updating and querying using simple HTTP requests and get responses in various formats (JSON, XML and CSV for instance).

2.9 What Is "R"

R is one of the most powerful scripting language for analysis and statistical data manipulation. This language it is compatible and inspired by a statistical language S, which is developed by AT&T, where S stand for statistics.
The model of the R looks like Figure 7. (6) The model of the tools how it looks in a data science project.

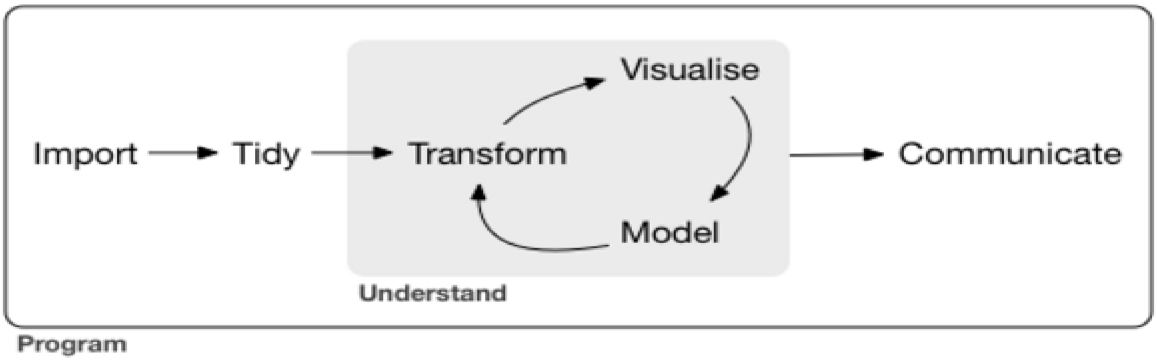


Figure 7. How looks the model of the "R"

2.9.1 How it works "R"

R is very powerful language. So, working with R firstly needs the data to get imported then to tidy them it is a good idea for better efficiency. So, the data that have already saved in a file, web API or database we load then into a data frame in R. So, working on R we need to get the data or we can't start working before getting the data. Tidying your data implies putting away it in a predictable form that matches the semantics of the dataset with the way it is stored. In short, when your data is tidy, every column is a variable, and each line is a perception. Tidy data is significant because of the fact that the reliable structure gives you a chance to concentrate your battle on inquiries concerning the data, not battling to get the data into the correct form for various functions.

Looking up to the Figure 7 we see after we have tidy the data, then we start to transform the data. Transforming mean to arrange the data on the way you can use for your interest and you can group in the way like people from one country, people of their birth year etc. But, R it allows us during the transformation creating a new variable with their functions, like computing the gross salary plus taxes and having a new column of their net salary.

Tidying and transforming make the form that’s natural to work sometimes they feels like they are on bout. So, together they are called wrangling.

After we have finished with tidy data then we deal with two main engines of knowledge generation: visualisation and modelling. These have corresponding qualities and weaknesses so any real investigation will emphasize between them ordinarily.

Visualization is made through person’s activity. Visualization it can shows things like wasn’t expected if visualization is doing in good way. Also during that you can raise a question like are those data what we needed if the visualization is made in good way but the visualization if given us different results, So, then are questions are those data what we needed or we need to collect different data. Visualization not needed to be scale particularly well because the visualization is going to be interpreted from people.

Models are harmonizing tools to visualisation. If you have done your questions adequately, then you can use a model to answer them. Models are a basically mathematical or computational tool, so they scale well. Models are made to make assumptions, so by nature it is known a model cannot question it is own assumption, so for that we cannot be surprise from fundamentally model.

Communication is a critical part of any data analysis project and it is the last step of data science. So, depend how well are visualization and model have been made you will understand when you will communicate the results to others.

Programming it is all about these tools. Programming is a cross-cutting when you can use in your project in every part.

Chapter 3

3 Problem setting

In this phase my work is to present the motivations behind the research work as well as the issues I faced. I analysed both general and specific aspects of the problem focusing on assuming the results of Milano Design Week (MDW). For Milano Design Week 2016 I applied Data Analytics approach to public and (authorized) private social networks posts and to the data retrieved by the official app. Starting from the large number of posts from the bottom and finding the features based on the predictive results and data behaviour grew the performance of final extraction. My role was trying to answer some of the questions and being able to find some solutions to helping the organiser to organise the event in a better way for next year.

- Trying to display the official social posts of MDW hourly to verify if something new emerging with respect to Fluxedo analysis.
- Verifying if there are some city areas where some events are mostly discussed, that can be useful for advertising solutions.
- Extracting the most used unofficial posts to increase the research and improve the results of the analysis.

3.1 Extraction of the results of MDW from Social Media

Large quantity of information is gathered from the application of MDW and posts from Social networks (Twitter, Instagram and Facebook). Extracting this amount of data, can increase the number of data scientist and machine intelligence to discover the data for different analytical approaches. Nowadays users can make a status in short posts distributed by mobile, web etc. Nowadays people post or share something, which for them is interesting. My problem was during the MDW. These kinds of events attract very large crowds and are particularly hard to analyse, as they involve different entities: city districts, managers, organizers, sponsors, visitors etc.

Gathering a huge amount of data from posts, where people were posting was a real challenge for Fluxedo and PoliMi to deal with problem. We could obtain the posts geolocation of the users who were using the app of MDW. Meanwhile, knowing more about public response from, which events they are participating, where they are from, for what they are more interested in, etc. can help the organizer to have an idea for the near event and save time for thinking what to do. In this case, they can be more successful.

For these things analysing data of social networks for particular events can be useful for the organization. Analysing the social networks using the geolocation of the user where they were using the official app of the event, quantity of the users that were visiting the event and quality of the post like with what was related the event can be analysed. So, an organizer could get an indication of popularity of its events by analysing the feedback gathered from those data.

3.2 Extraction of the number of visitors based on the area

All places during the MDW are not equal. Some districts of Milan, are more historical and they are linked to the design scene more than others, and this brings attention of more visitors to visit those venues and most popular events are concentrated in that part. From my analysis, the biggest design districts are Brera, Tortona and in consideration numbers Lambrate/Ventura, organisers for these districts even provide specific guides to the MDW for the public and tools such as official app to manage and promote their events, giving help in matters such as marketing and communication, or giving listings of rentable locations to interested organisers.
I analysed the MDW 2016 (April 12th-19th) in several aspects from the data I was handling and from different sources gathered by Fluxedo and PoliMi. Example number of posts from the cell districts as it is shown in Figure 8.

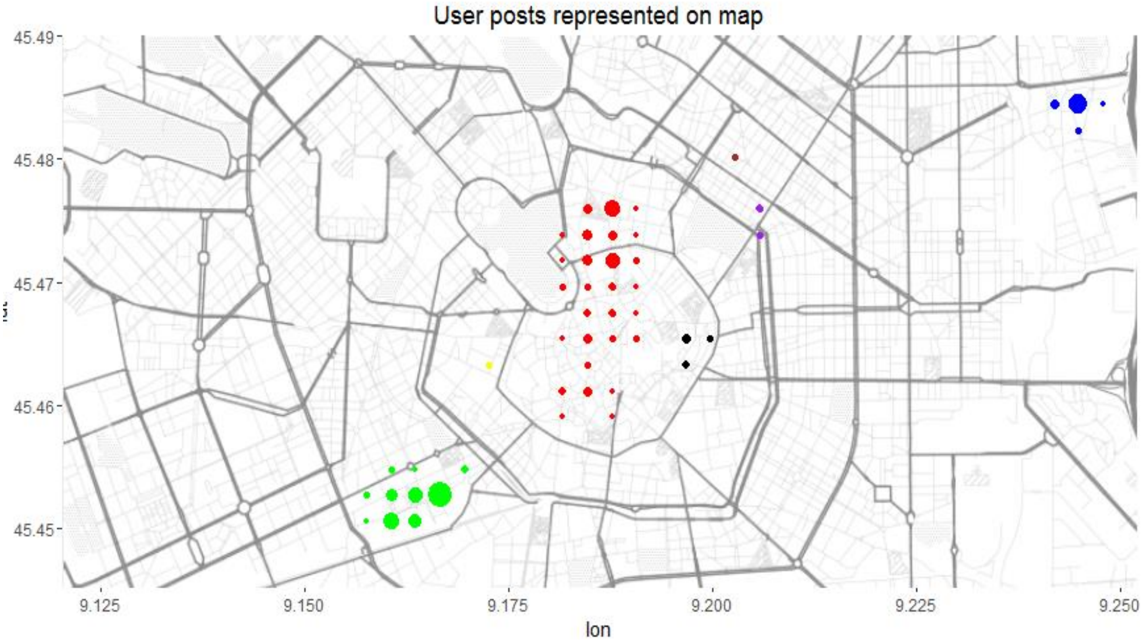


Figure 8. Users posts represented on map

- Red dots - Brera
- Gren dots - Tortona
- Blue dots – Lambrate
- Purple dots – Porta Venezia
- Black dots – Durini
- Brown dots – San Gregorio
- Yellow dots – Sant Ambrogio

3.3 Hourly distribution analysis of social posts

The problem was to find hourly distribution of social posts. I deal in solving this problem based on the data of social media I had and that were saved as JSON files from Fluxedo and PoliMi. I have used Fuseki to extract those data, then transferring in "R" in order to be able to work with those data. Fuseki is a SPARQL server and it allows to create queries using simple HTTP request and get responses in various formats in our case as JSON file. So, having access on those data through Fuseki, I could create queries and extract the most official posts used during the MDW. After, creating meaningful queries and importing in "R", then I tidied the data in order to manipulate and use in a proper way and with “R” being able to present the required results. Firstly, I have worked on collecting the posts of Milano Design Week from given hashtag: “MDW”, “MDW2016”, “Fuorisalone”, “Fuorisalone2016”. Those four-main official social posts I added main posts after I discovered by the content. What I did was summing up the number of same posts with same content but being written in different way as (ex. #MilanoDesignWeek and #MilanDesignWeek or #MDW etc.) and I added those hashtags with a name as #MDW. I did this after I had extracted the most used post with same meaning. I did this also for Fuorisalone, Fuorisalone2016 and MDW2016 as for #MDW.

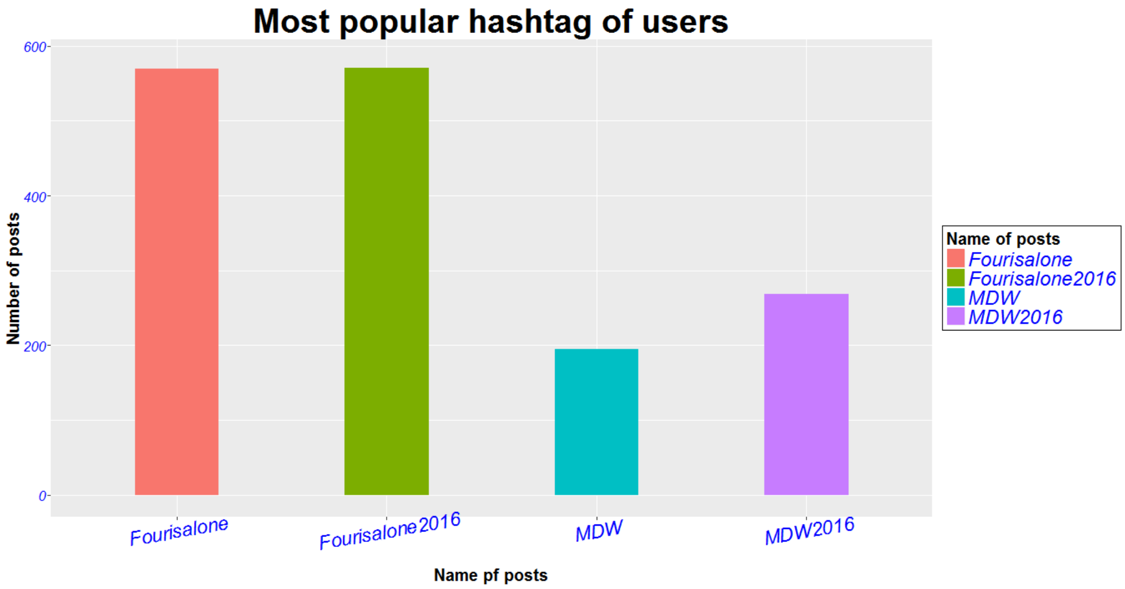
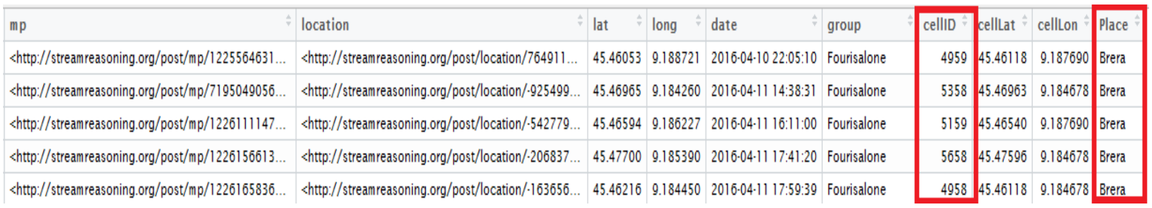


Figure 9. Milano Design Week comparison of most popular hashtag

The problem was to see the distribution of the most popular posts hourly, and where those posts were spread around the city. As I mentioned above, firstly I tidied the data to be able to work more precise. I did that in the way trying to find the post from, which specific area it is posting. Firstly, when I imported with Fuseki, I queried in the way the social posts must need to have the geo-position (latitude and longitude) in order to be useful posts to find the area where it was posted. Since the city of Milan is divided by pixels 100x100, then with the given function I applied to every post to have his identifier cellID based on their geo-position. Based on their cellID I can group in, which area the posts were posted.

An example of how I tidy the data in "R". Making a data frame of Brera posts, grouping based on the cellID that corresponds with geo-position of Brera.



mp	location	lat	long	date	group	cellID	cellLat	cellLon	Place
<http://streamreasoning.org/post/mp/1225564631...	<http://streamreasoning.org/post/location/764911...	45.46053	9.188721	2016-04-10 22:05:10	Fourisalone	4959	45.46118	9.187690	Brera
<http://streamreasoning.org/post/mp/7195049056...	<http://streamreasoning.org/post/location/-925499...	45.46965	9.184260	2016-04-11 14:38:31	Fourisalone	5358	45.46963	9.184678	Brera
<http://streamreasoning.org/post/mp/122611147...	<http://streamreasoning.org/post/location/-542779...	45.46594	9.186227	2016-04-11 16:11:00	Fourisalone	5159	45.46540	9.187690	Brera
<http://streamreasoning.org/post/mp/122615661З...	<http://streamreasoning.org/post/location/-206837...	45.47700	9.185390	2016-04-11 17:41:20	Fourisalone	5658	45.47596	9.184678	Brera
<http://streamreasoning.org/post/mp/122616583б...	<http://streamreasoning.org/post/location/-163656...	45.46216	9.184450	2016-04-11 17:59:39	Fourisalone	4958	45.46118	9.184678	Brera

Figure 10. Data frame of Brera posts grouped based on the cellID

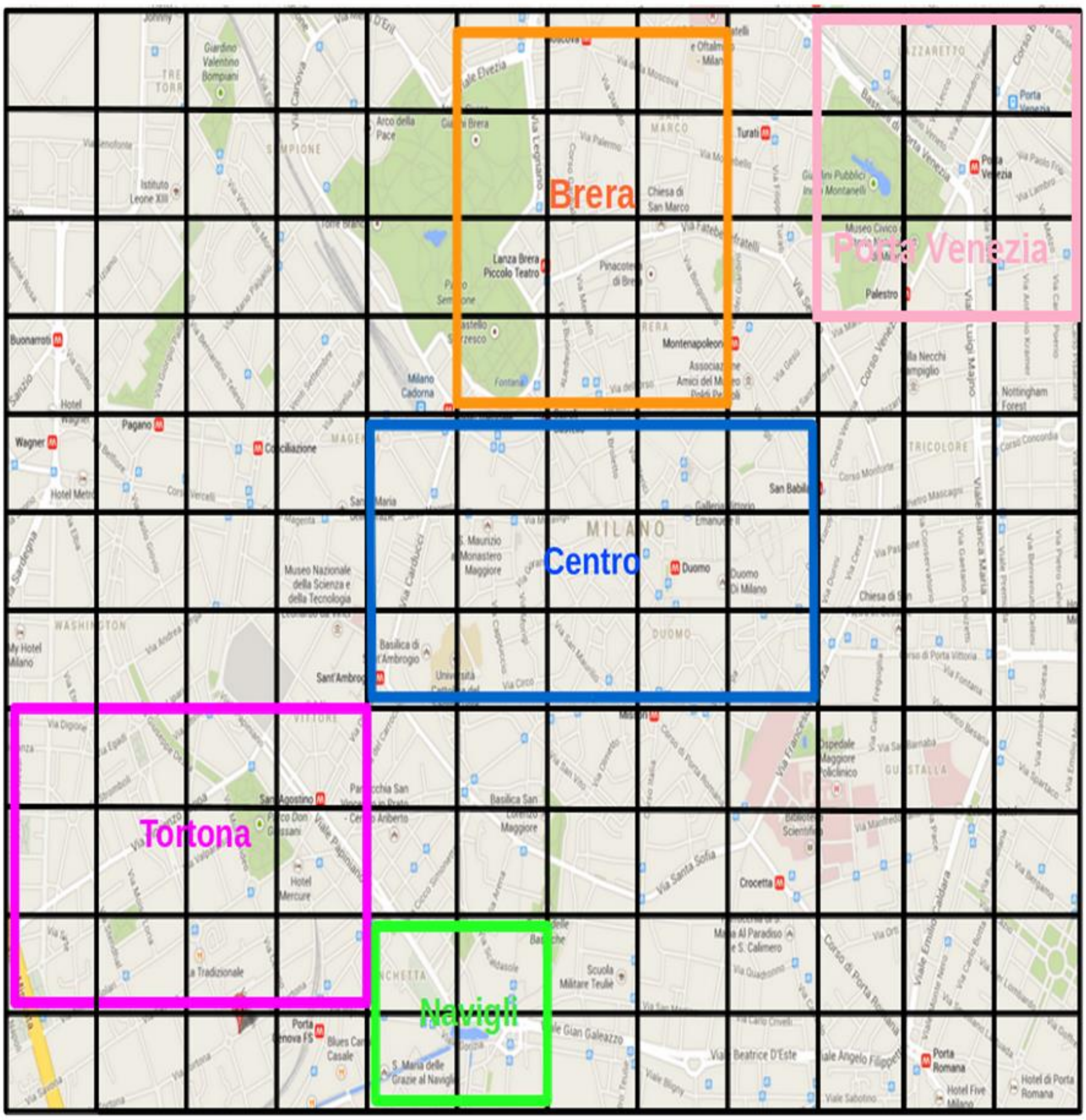


Figure 11. The map of Milano divided by areas

Based on the Figure 11, the posts were grouped according to their location. An example of implementation of the graphs that shows that depending on the day and hour trend in the number of posts were different. Also, some areas as Brera, Tortona are the most visited during the event and then Lambrate and the following areas but fewer numbers of posts. Number of the posts can be an indication to assume, which area was the most visited during the MDW event. Also, I have divided the time span by every two-hour checking in that interval of the day the areas were most visited based on the posts that were posted at that time.

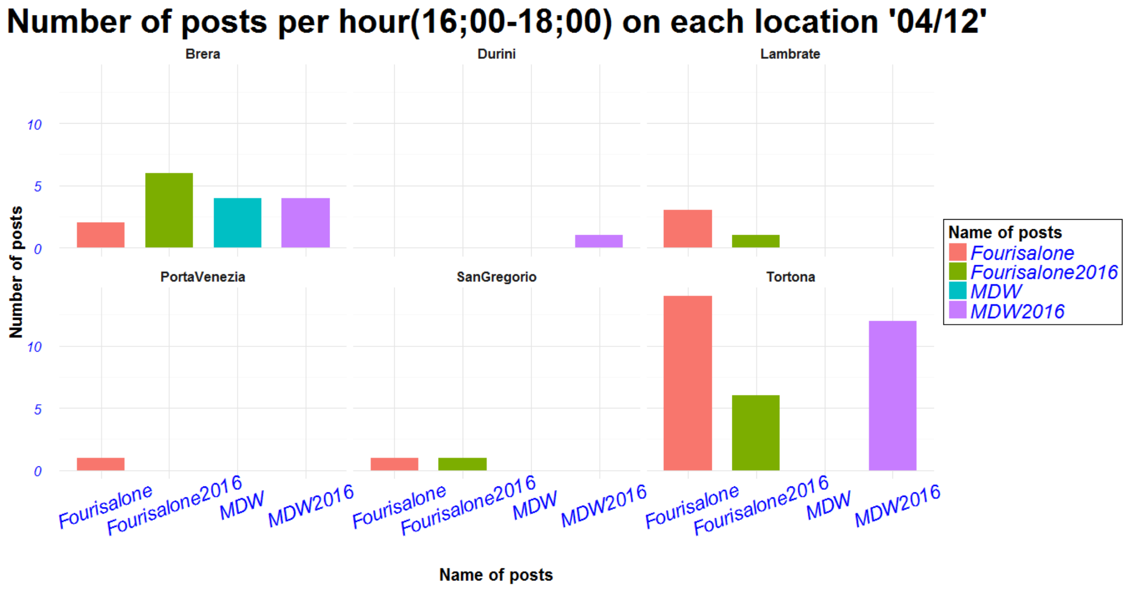


Figure 12. Number of posts per hour (16:00-18:00) on each location '04/12'

Based on the Figure 12, we can assume from 16:00-18:00 by 12th of April the zones of Tortona and Brera are visited by visitors in every hour and every day while the rest of the areas are not in every hour. But it seems that the area of Lambrate is less visited then Brera and Tortona but more visited compared with the other areas.

3.4 Geographical distribution analysis of social posts

The problem in this task was trying to find geographical distribution of social posts and the meaning of the post if it is related with the specific area of the event. In this task, I have created several data frames, like data frame with social posts with content of Tortona area and with a specific code I implemented to get the map of Milan in order to present on the map the position of the users and posts. The area of Milan was divided into citypixels and a list of cells has been associated to each district (Tortona, Brera, Lambrate, etc). In this request, firstly I was gathering the posts with the content of each area. For example, the posts with content of Tortona, which they got most used during event are like: (#tortona, #tortonadesignweek, #zonatortona, #tortonadistrict, #tortonadesigndistrict). The positions of the users using the app was already collected by Fluxedo and PoliMi and set as "Positions"[4] and I just transferred in "R" to be able to manipulate with data and have the data, which I want to use for further steps. For this specific task, I was focused only on the users they were using the app in Milan. Based on the coordinates of Milan city I subset from "Positions" only the users that were in Milan and I grouped as "Positions Milan". Then, I classified based on their cellID in that part of the area they were using the app. After, I classified the data frames, I compared the posts with content of the area from where it posted and the users from the specific area checking also where the event was held and which other part of the city was discussing. For example, in Figure 13 you will see except of Tortona, people were posting for Tortona events also from different part of the city. Red dots represent the users and the blue dots the posts with Tortona content.

[4] "Positions" you can find in the section 4.2 how I transfer into "R" and the "Positions" got from the log of the Fuorisalone official app.

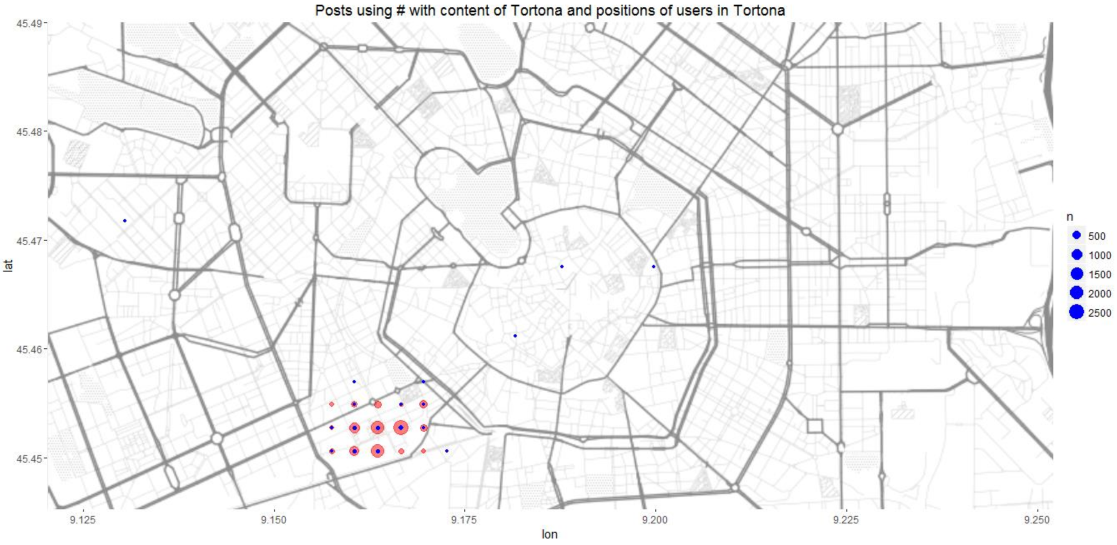


Figure 13. Posts using # with content of Tortona and positions of users in Tortona

The problem was trying to figure out for the given area of the event except its current location and which other part of the city it is mentioned. This is important for managers, organiser, and sponsor to advertise the area in that part of the city where was mentioned but not taking into consideration its neighbourhood areas. This can increase the number of the visitors seeing the advertisement example in Porta Venezia for the Tortona event. So, generally the curiosity of the visitor can increase by going there to visit for the sole purpose that the advertisement was in a different part of the city. So, the main idea of this request was advertising the events from different areas and counting the number of posts for those events even if the event was held in a different area.

3.5 Extraction of popular unofficial hashtags

In this problem, I was trying to extract the most popular unofficial hashtags. I did this using Fuseki. I created the queries to extract the social posts that was most used in numbers and they weren't official posts of the MDW and then importing in "R". After importing I was just summing up the posts and displaying by graph the most unofficial posts used during the MDW.

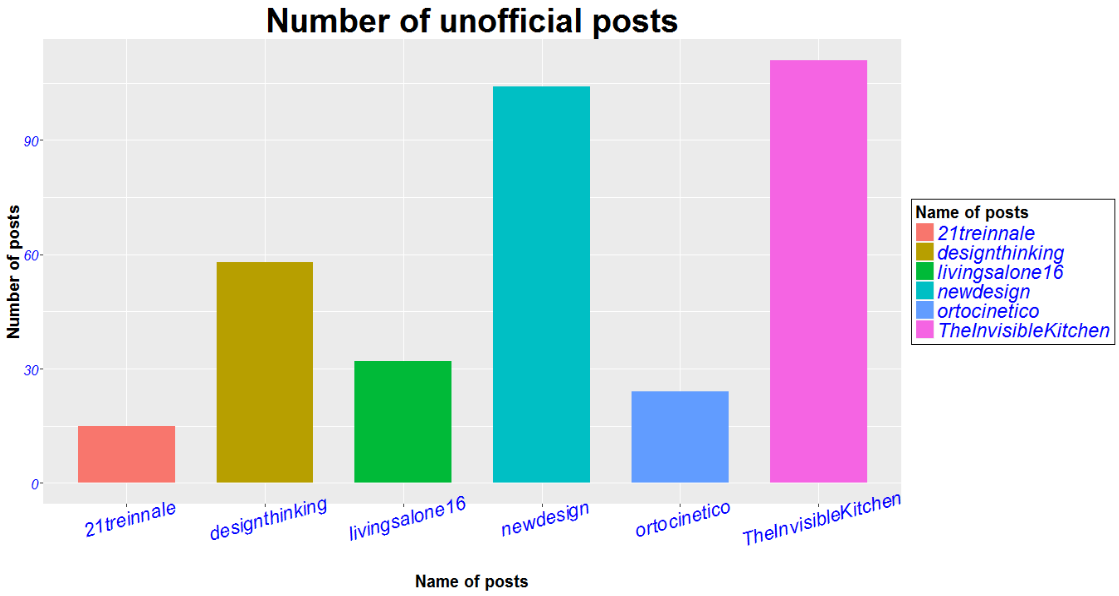


Figure 14. Number of unofficial posts

So, from the Figure 14 we can see that the event that was related with the kitchen or some posts that were telling the new design era were mostly used as unofficial posts. Extracting most unofficial posts to increase the research and improve analysis results. Also, an indication to the organiser knowing, which unofficial posts were most used and for what kind of event. For example, #TheInvesibleKitchen is related to the new kitchen design. The analysis of knowing the most unofficial posts can be useful for the success (or failure) of the current events, because that can help you to enlarge the set of analysed discussions and give you more reliable results for taking future decisions. So, this analysis can create an idea to the organiser and sponsor for designing improvements in the organization of future events and conferences for the hashtag that was mostly used.

Chapter 4

4 Implementation experience

In this chapter I will go more deeply into details with the practical aspects of the work and present the results of the work step by step following the technical process. The thesis has been done with "R" programing language and Apache Jena Fuseki. "R" supports different packages that makes it statistically more powerful language and makes it easier for me to work with "R".
Some of the main libraries in "R" I have worked with are:

- SPARQL - allows us to import data into "R"
- ggmap - is a collection of functions to visualize spatial data and static maps from online sources
- ggplot2 - is based on the grammar of the graphics and tries to take the good part of the base instead of the bad one
- tidyr - for data tidying that works well with 'dplyr' data pipelines etc.

I have used "R studio" to get the result in my thesis. (29) RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

4.1 System Architecture

I decided to analyse the data of MDW 2016 around "R", a programming language with a huge variety of tools that you can use to build graphs and the strongest language in the statistical field.
"R" gives you liberty in terms of how to solve the problems, because it has a lot of different libraries that you can use to build graphs. I was trying to use the most appropriate tools and techniques in regard to the results and data that I was using in my thesis.
The important issue was to find meaningful results that can help the organiser to take better decisions for the upcoming events. During the thesis, I faced problems with posts

from different languages. As I said, the data from social networks was saved as RDF format. So, I have used SPARQL to access the data and queries that I want in the end.

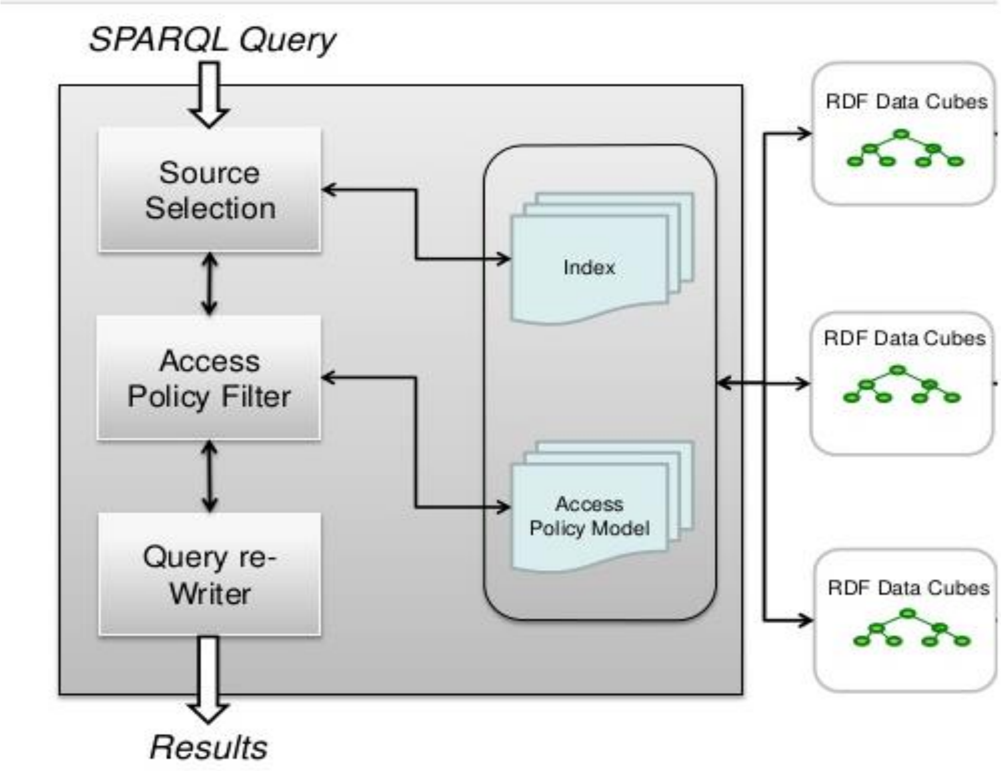


Figure 15. System Architecture

Querying the results, I have used SPARQL server called Apache Jena Fuseki. It can run as an operating system service, as a Java web application (WAR file), and as a standalone server. It provides security (using Apache Shiro) and has a user interface for server monitoring and administration.

4.2 Data

The data I have used in the thesis is a collection of data from MDW 2016. There are several data that I have collected from MDW as: recorded data from Socialometer and the data from the official app of the event. Several information was represented about MDW dataset, MDW started from 12th of April till 19th of April but the data has been collected one month before the MDW started and 3 weeks after. An example of importing the data into “R” represented as .csv files.

```
agenda <- read.csv("C:/Users/Blertan/Desktop/Master thesis/agenda.csv")
```

```
positions <- read.csv("C:/Users/Blertan/Desktop/Master thesis/positions.csv")
event_location<-read.csv("C:/Users/Blertan/Desktop/Master
thesis/event_location.csv")
events <- read.csv("C:/Users/Blertan/Desktop/Master thesis/events.csv")
users <- read.csv("C:/Users/Blertan/Desktop/Master thesis/users.csv")
```

Looking from the datasets that were collected from Fluxedo and PoliMi, I was focussed mostly to use the posts with text, and post hashtags contained in the message and geographical coordinates: latitude and longitude, which I could use to see the venue of the city where the posts were published.

An important measurement of my analysis was the time, were I faced a problem in events session due to StartDate and EndDate of different events that were same in the same district, which led to assumption problem, due to the fact that it was hard to determine from which event the posts were coming, so in this case I was taking as a reference the venue at a particular time. I will need to assume the geospatial information was pretty accurate and rich in identifying a venue of the city.

4.3 Technical Snapshot

In this section, some of the technical snapshot encountered during the implementation of the thesis.

4.4 Hourly distribution analysis of social posts

Analysing the number of the posts during the events daily but having them divided in several intervals of time during the day. So, I can say the steady trend is not always the same as it can impact the number of visitors or holidays.

Firstly, I installed the Apache Jena Fuseki that is a SPARQL server. Then using R studio and installing the specific libraries, I used the package of SPARQL that allows me to connect with Fuseki. In Fuseki I have stored the data that was extracted before from Fluxedo and PoliMi and saved those as JSON files. I was programing with R in order to get what I want from the data to solve the problem that I was working on. Connecting the "R" with Fuseki I was using the specific code:

```
endpoint <- http://localhost:3030/SocialOmeter/query
```

Then starting to query:

```
query <- "prefix j.5:   <http://rdfs.org/sioc/ns#>  ....
```

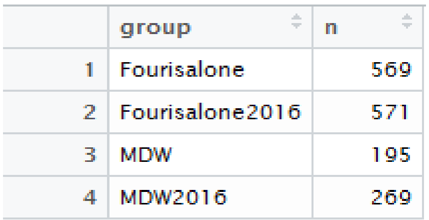
Since Fuseki is a SPARQL server, I extracted the data using proper queries with this particular the specific code:

```
SELECT ?mp ?location ?lat ?long ?date
WHERE
{  ?mp j.0:created ?date. ...
}"
res <- SPARQL(endpoint, query)
```

With the specific queries, firstly I imported the posts that had date and geo-positions. Having the date and the geo-positions, I could create useful graphs showing the users geo-position and date. After I imported all the posts, I started to search and group by content each post, and clearing what is not related with the official post of MDW. Since the posts could be written in different ways, I collected them based on the content of the post (ex. #Milano Design Week 2016 and #MilanDesignWeek2016district or #MDW2016 etc.) and I added those hashtags with one data frame name as #MDW2016. I did the same thing also for Fuorisalone, Fuorisalone2016 and MDW since they were official hashtags of the Milano Design Week 2016. The following code shows you how the data is grouped separately.

```
Posts_ Milano_Design_Week_2016$group <- " Milano Design Week 2016"
Posts_ MilanDesignWeek2016district $group <- " MilanDesignWeek2016district"
Posts_ MDW2016$group <- " MDW2016 "...
Hashtag_MDW2016Content<-rbind(Posts_Milano_Design_Week_2016,
Posts_MDW2016, posts_MilanDesignWeek2016district)
```

Same structure of code is used for Fuorisalone, Fuorisalone2016 and MDW. I did them for the four main groups, as counted posts in one group as in the Figure 16.



	group	n
1	Fourisalone	569
2	Fourisalone2016	571
3	MDW	195
4	MDW2016	269

Figure 16. Counted official hashtags

Then I graphically represented the official posts used during the event from the data that was gathered. Graph is a way of representing the gathered data, which makes it easier for random people to understand.

The following code shows how I represented it graphically:

```
PopularHashtag_plot <- ggplot(PopularHashtag,aes(x=group,y=n,fill=group))+
geom_bar(stat="identity", position=position_dodge())+  ggtitle("Most popular
hashtag of users")+ylim(0,580)+ylab("Number of posts")+xlab("Hashtag names").
```

After having an idea of the overall posts that were mostly used, have started to divide the posts based on the area that they were posted. As I said in the beginning of this problem, all of the posts have a geo-position (latitude and longitude). Then, with a function I was giving every single post the cellID considering their latitude and longitude. The given code represents the function that was built to give their cellID to each post after the function was applied.

```
get_cell_ID <- function(lat, lon) {
 cell_ID <- -1
 verticalIndex <- as.integer((lat - minLat) / cellHeight)
 horizontalIndex <- as.integer((lon - minLon) / cellWidth)
 cell_ID <- as.integer(((verticalIndex * columnsNumber) + horizontalIndex) + 1)
 if(cell_ID >= (columnsNumber * rowsNumber)) {
  cellID <- -1
 }
 return(cell_ID)
}
get_lon_cell <- function(cellID) { return(minLon + (cellID - 1) %% columnsNumber
* cellWidth + (cellWidth / 2)) }
get_lat_cell <- function(cellID) { return(minLat +  as.integer((cellID - 1) /
columnsNumber) * cellHeight + (cellHeight / 2)) }
minLat <- 45.356686
minLon <- 9.01149
cellHeight <- 0.00211101
cellWidth <- 0.00301197
columnsNumber <- 100
rowsNumber <- 100
```

And this following line of code shows how all the posts get their cellID.

```
total_post_byname$cellID<-
get_cell_ID(total_post_byname$lat,total_post_byname$long)
```

After applying the function to the posts, I was grouping them in data frames according to that area where the post belongs, subsets on the areas based on their latitude and longitude are expanded. I grouped all the posts by the areas that they were posted, then I tidied the data by date and hour to represent graphically as in Figure 17 to have an idea, which area was most visited assuming based on the posts that were posted.

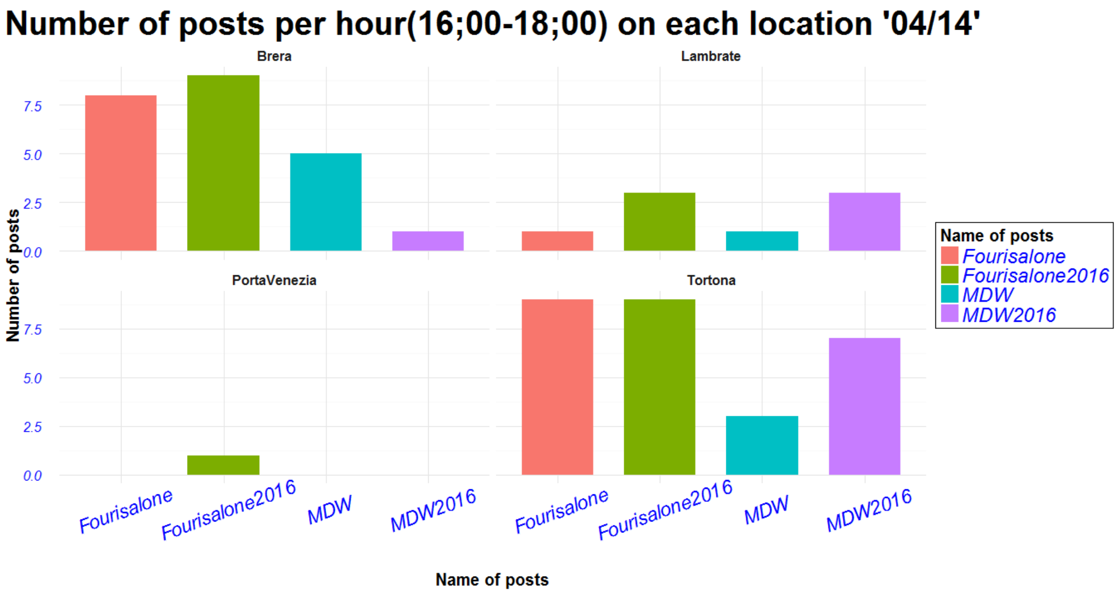


Figure 17. Number of posts per hour (16:00-18:00) on each location '04/14'

4.5 Geographical distribution analysis of social posts

I was trying to gather the whole posts whose content had the hashtags of the most used city areas like Brera, Tortona and Lambrate/Ventura. Ex. #tortonadesignweek, #zonatortona, ..., those hashtags were collected together as Tortona Content. Same for Brera and Lambrate/Ventura. After that, I was checking, which part of the city the hashtags were spreading in the main areas mentioned above and assuming, for those events being mentioned also in different areas. This could be valuable information for the organiser to know and advertise the main area events in the places where visitors were posting about those areas.

After connecting R with Fuseki, I started creating queries to extract the needed information. Firstly, I started extracting the hashtags from the content of Tortona, Brera and Lambrate/Ventura. The following lines of code will show an example of the Lambrate/Ventura area, and how I query the contents of the hashtags related with the main areas.

```
query <- "prefix j.5:   <http://rdfs.org/sioc/ns#>
…
SELECT ?mp ?location ?lat ?long ?date
WHERE
{
?mp j.0:created ?date.
?mp j.7:location ?location.
?location j.7:lat ?lat.
?location j.7:long ?long.
?mp j.2:topic  t:venturadesigndistrict. (“lambrate", "Lambrate Ventura", "Ventura",
"Ventura Lambrate",  "Ventura Lambrate 2016",  "Ventura Design District")
}"
```

Based on the query, I was trying to extract every possible hashtag whose content was related with the exact area of the event. Then, I gathered all the social posts based on their content, I grouped them in main data frames as TortonaContent, BreraContent and Lambrate/VenturaContent. I did this only for those three main areas because looking from the previous task those are the areas that can be advertised in different part of the city since they are most visited. The goal was to show to the organiser, managers, and sponsors where they can advertise the events except of the areas where it is held. Then trying to make assumptions based on where the posts of different areas were posting in different areas of the city. For example, posts for the Tortona events were posted even from the centre and Porta Venezia so it could be a good idea for next year to advertise the Tortona events from those areas. So, presenting this in visualized way, I have used “R” and several libraries to present in a better way how posts are spread around the city. The most important libraries to work in this task properly are ggamap, maproj, maptools, and other libraries that allow me to tidy the data. Using the libraries, I have created the map of Milan that includes the three main areas Brera, Tortona and Lambrate/Ventura. The following code shows how I deal to create the map in given geo-positions and background of the map trying to make it readable for the audience.

```
map2 <- ggmap(get_map(location = c(9.120357,45.445243,9.252000,45.490000),
maptype = "toner-lite"))
```

The next problem were the users and posts to present on the map. So, I was comparing the number of the users from Lambrate/Ventura with the posts that got the content of Lambrate/Ventura. Users that were already grouped in a single area depending on their app activity whether they are Brera visitors, Tortona visitors etc. The comparison I did was to check the number of the users and posts of the given area. The following code shows how I compared the users of Lambrate/Ventura with posts that had content of the Lambrate/Ventura.

```
Posts_vs_Positions_LambrateVenturaContent <- map2 + geom_point(data =
Positions_Lambrate_CountedbyCellID, aes (x = cellLon, y = cellLat, size = n), color =
"red", alpha = 0.5) +
  geom_point(data = Hashtag_LambrateVenturaContent_CountedbyCellID, aes (x =
cellLon, y = cellLat, size = n), color = "blue", alpha = 1)+ggtitle("Posts using # with
content of Lambrate/Ventura and positions of users in Lambrate/Ventura")
```

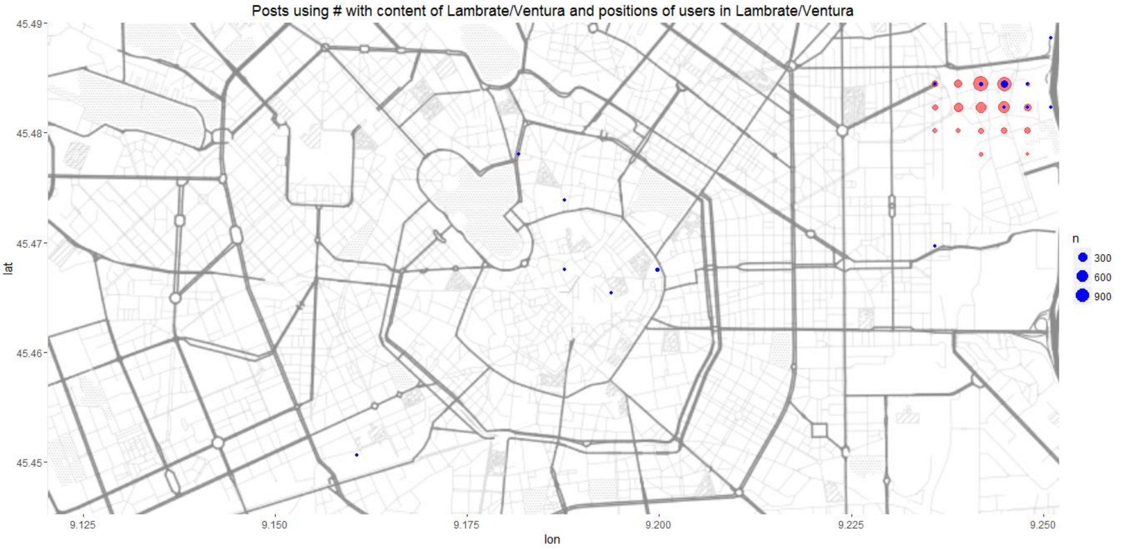

Posts using # with content of Lambrate/Ventura and positions of users in Lambrate/Ventura
45.49
45.48
45.47
45.46
45.45
lat
9.125
9.150
9.175
9.200
9.225
9.250
lon
n
300
600
900

Figure 18. Posts using # with content of Lambrate and positions of users in Lamb

4.6 Extraction of popular unofficial hashtags

Analysing the number of the unofficial posts during the events. So, I can say the steady trend of unofficial posts is not always the same, and this content can impact the theme of the event.

Analysing the number of the unofficial posts firstly it uses the same procedure as the previous tasks to be connected with Fuseki in order to create queries to get the desired data. I classified the whole social posts through Fuseki. Then I was classifying the unofficial posts for the given dates of the events, trying to find the most useful posts during the MDW. Some of the most unofficial tags that were being used are: #theinvesiblekitchen, #21trienale, #designthinking, #newdesign, etc. The following code shows how I created the queries to get the most used post, and in this case, I will show #TheInvesibleKitchen that has the date and checking if it corresponds with the event held from 12th-19th of April.

```
SELECT ?mp  ?date
WHERE
{
?mp j.0:created ?date.
?mp j.2:topic  t:TheInvisibleKitchen.
 }"
```

Showing the number of the posts during the events, firstly I grouped based on their hashtag

```
Posts_mostunofficial_used<-
rbind(Posts_TheInvisibleKitchenHashtag,Posts_newdesignHashtag,Posts_designth
inkingHashtag,Posts_ortocineticoHashtag,Posts_21triennaleHashtag,Posts_livingsa
lone16Hashtag)
```

Then I represented in graph using “R” with the given lines of code.

```
Posts_mostunofficial_used_bar<-
ggplot(Posts_mostunofficial_used,aes(x=factor(group),fill=group))+
geom_bar(stat="count", width=0.7)+   theme_minimal()+ylab("Number of
posts")+xlab("Hashtag names")+ggtitle("Number of unofficial posts")
```

Chapter 5

5 Conclusion

Summing up, I am going to have an overview about the problems I faced, how I contributed to solve them, and the possible future developments.

5.1 Problems faces and solutions adopted

The problem I tried to solve in this thesis work was to find an efficient way to display meaningful visualization that can be useful for the organiser, managers, and sponsors of Milano Design Week for improving the event for the coming year. I focused my attention on Milano Design Week and social media generated by users: I tried to find the relationship between posts and locations of the Milano Design Week 2016.

I analysed the contents and hashtags of social media to discover the area or events mentioned in the social posts. I defined a set of suitable code for comparing these data with each other for finding the best answer between the relationships of events, posts, areas, and users.

During my work, I have faced some problems that can have impact on the precision. Another issue was the difficulty to extract the location of some posts since there was no geo-position provided by the user. So, the posts that didn't have geo-position or date had to be deleted from my data frames. I did that so I could have the precise results displayed by visualisation.

After clearing the data, I could find only the posts that were related with Milano Design week, and to perform a search based on their location whether that's from the whole city or a single district as Tortona, Brera, etc.

The organiser of the Milano Design Week could focus on how the posts are spread around the city, and where is the possibility to advertise the event to increase the number of the visitors.

5.2 Future works

Considering the current project as a starting point I can define a future work to be focused on extracting the pattern of visits. The future problem is to be able to follow the path of the visitors' activity in similar ways. This can predict the behaviour of visitors and improve the organization. This can increase the potential of the organisers and sponsors to make group offers that can attract groups of people to visit the places as previous visitors following most similar paths extracted from this analyses.

Another future work could be making promotional offers for users that had already installed the official app of MDW, before the event started (example two weeks in advance). The organiser can have an idea from where they are and what they are more interested in based on what events they booked as favourites on the app.
Those works might be challenging but really interesting and useful for the event.

Bibliography

[1] http://marenas.sitios.ing.uc.cl/talks/pods11.pdf

[2] OReilly, Learning SPARQL, 2nd Edition July 2013, p.16

[3] J. Perez, M. Arenas and C. Gutierrez. Semantics and Complexity of SPARQL. ACM Transactions on Database Systems, 34(3), Article 16 (45 pages), 2009

[4] https://semanticwebrecipes.wordpress.com/2014/07/17/fuseki-it-really-is-that-easy/

[5] The Art of R Programming Norman Matloff September 1, 2009, p.1

[6] R for Data Science, Garrett Grolemund & Hadley Wickham, 2016

[7] http://www.wheremilan.com/events/milano-design-week-2016-salone-del-mobile/

[8] Jonathan A. Obar, Steve Wildman (2015). "Social media definition and the governance challenge: An introduction to the special issue, Working paper, p.3-4

[9] Andreas M. Kaplan, Michael Haenlein (2010). "Users of the world, unite! The challenges and opportunities of social media", p.60-61

[10] Nicole B. Ellison, (2007). "Social Network Sites: Definition, History, and Scholarship". Journal of computer-mediated communication. p.210–230

[11] https://en.wikipedia.org/wiki/Social_media

[12] Andreas Hackeloeer, Klaas Klasing, Jukka M. Krisp & Liqiu Meng (2014) Georeferencing: a review of methods and applications, Annals of GIS, 61-69

[13] Clough, P., Sanderson, M. and Joho, H. (2004) Extraction of semantic annotations from textual web pages. Technical report, University of Sheffield.

[14] https://en.wikipedia.org/wiki/Georeferencing

[15] Pascal Hitzler (Kno.e.sis Center, Wright State University, Dayton, OH, USA), Krzysztof Janowicz (University of California, Santa Barbara, CA, USA), Semantic Web, Paper Work

[16] Grigoris Antoniou, Frank van Harmelen (2008). A Semantic Web Primer, second edition. P. 8-154

[17] Grigoris Antoniou, Paul Groth, Frank van Harmelen and Rinke Hoekstra (2012). A Semantic Web Primer, third edition. P. 44-47

[18] https://www.obitko.com/tutorials/ontologies-semantic-web/rdf-schema-rdfs.html

[19] http://dig.csail.mit.edu/2010/Courses/6.898/resources/sparql-tutorial.pdf

[20] https://www.hausmanmarketingletter.com/16-differences-between-social-media-and-traditional-media/

[21] https://www.obitko.com/tutorials/ontologies-semantic-web/semantic-web-architecture.html

[22] Graham Klyne and Jeremy J. Carroll. Resource description framework (RDF): Concepts and abstract syntax. World Wide Web Consortium, Recommendation REC-rdf-concepts-20040210, February 2004.

[23] E. Prud'hommeaux D. Beckett, T. Berners-Lee and G. Carothers. Turtle: Terse rdf triple language. 07 2012.

[24] https://www.w3.org/2001/sw/RDFCore/ntriples/

[25] Andy Seaborne and Eric Prud'hommeaux. SPARQL query language for RDF. W3C recommendation, W3C, January 2008.http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

[26] https://www.w3.org/TR/rdf-sparql-query/#describe

[27] B. McBride. Jena: a semantic web toolkit. IEEE Internet Computing, 6(6):55-59, 2002.

[28] https://en.wikipedia.org/wiki/Jena_(framework)

[29] Marco Balduini, Emanuele Della Valle:
FraPPE: A Vocabulary to Represent Heterogeneous Spatio-temporal Data to Support Visual Analytics. International Semantic Web Conference (2) 2015: 321-328

[30] https://www.rstudio.com/products/RStudio/

[31] http://teaching-wiki.sti2.at/uploads/7/74/08_SW-RIF.pdf