Université de Manouba

Institut Supérieur des Arts Multimdia

Achievement of the requirement of the degree of Diploma in Computer Science and Multimedia Engineering

Specialty

Computer Science and Multimedia

Option

Software engineering

Analysis of twitter data on sustainable development goals

Ines Aouadi

Pedagogic advisors :

Sehl Melouli professor, LAVAL UNIVERSITY
Tarek Hamrouni professor, High Institute of Arts Multimedia

Research Unit: Centre de Recherche sur les Communautés Intelligentes

Contents

List of Figures

List of Tables

List of Algorithms

Introduction

Chapter 1

General context

1.1 Introduction

During this introductory chapter, we will talk about the Problematic and we will try to define the challenges and constraints of our project.Later, we will explain our objectives in order to accomplish our mission. Finally, we will end this section with a presentation of the framework of our internship.

1.2 Problematic

The concept of smart city is evolving as a new approach to mitigate and remedy current urban problems and make urban development more sustainable (Alawadhi et al., 2012).in this case, we must have citizen with operational feedback.
We focus on twitter microblog to extract information.This platform is accessible through websites or cellphone applications allowing users to instantly post relevant information about what they are seeing, hearing and experiencing around them. In a SDG case, such platform provide valuable information shared voluntary to inform or alert the society .
Information retrieval from microblogs is hindered by many challenges such as: streaming data analysis,the variety of information format and language processing, large datasets, and extracting relevant and fresh information from a huge amount of outdated and redundant data.

1.3 Objectives

The objective of the proposed research is to provide insights on the online interactions and revealed perceptions of stakeholders about the SDG related to cities. We will provide cities with new informational tools, supporting their strategy in order to achieve the SDG11 (sustainable cities and communities) goal and its targets.
These informational tools will enable to bridge the gap between the vision of the diverse stakeholders, from United Nations, government and non-government agencies, to the civil society and the private sector. Specifically, the proposal will analyze the data posted on the Twitter platform to provide agencies and citizens with operational feedback on how SDGs are perceived and implemented.

1.4 framework of our internship

The internship is on the context of obtaining of the degree of Diploma in Computer Science and Multimedia Engineering.
The internship was proposed by CeRCI Unit. It have an overall duration of four months.

- Phase of documentation and bibliographic study which have a duration of one month.
- Phase of development and experimentation which was concertized by a practical internship in the laboratory CeRCI at Laval University with two months of work.
- Phase of The synthesis and evaluation of the obtained results which was elaborated in one month.

1.4.1 laboratory CeRCI

Laboratory CeRCI,Intelligent Communities Research Center,was created in XXXX in Laval University.

Figure 1.1: Logo Laval University

Organizational chart

Laboratory CeRCI was founded by five professors: Sehl Mellouli, Monia Rekik, Adnne Hajji, Jacqueline Corbett et Karim Ben Boubaker. The research activities is centered to three axes: citizen, governance and technology.

1.5 Conclusion

During this chapter, we have presented the project, the framework of internship and the work obligations. In the next chapter,we will present and compare the commonly used approach of information retrieval.

Chapter 2

State of the art

2.1 Introduction

This section aims at introducing micro blog platform ,SDG and information retrieval in order to provide the reader with basic notions and necessary technical background to understand in more details next chapters.

2.2 Sustainable development goals

The Sustainable Development Goals (SDGs), otherwise known as the Global Goals, are a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity.[1] On September 25th 2015,the 194 countries of the UN General Assembly adopted the 2030 Development Agenda titled Transforming our world: the 2030 Agenda for Sustainable Development.Countries adopted those goals which are a set of 17:

1. No Poverty
2. Zero Hunger
3. Good Health and Well-Being

[1] https://sustainabledevelopment.un.org/

4. Quality Education
5. Gender Equality
6. Clean Water and Sanitation
7. Affordable and Clean Energy
8. Decent work and Economic Growth
9. Industry, Innovation and Infrastructure
10. Reduced Inequality
11. Sustainable Cities and Communities
12. Responsible Consumption and Production
13. Climate Action
14. Life Below Water
15. Life On Land
16. Peace, Justice and Strong Institutions
17. Partnerships for the goals

The SDGs work in the spirit of partnership and pragmatism to make the right choices now to improve life, in a sustainable way, for future generations. They provide clear guidelines and targets for all countries to adopt in accordance with their own priorities and the environmental challenges of the world at large.To make a positive change for both people and planet,They tackle the root causes of poverty and unite us together.

2.3 Microblogs Platform

Nowadays, everyone can easily share and access to content on the web within few second.According to internet live stats website, 90% of information shared daily on the web is essentially provided from microblogging platforms[2] .

[2]http://www.internetlivestats.com/

Data acquisition and extraction from microblogging platforms is being an important research field. We focus on analyzing extraction and acquisition techniques adapted to the microblogging platform Twitter cause this is platform is one of the most popular microblog platforms according a large set of rich information shared and updates publicly regarding various topics and events.
With around 500 million of daily shared tweets,Twitter microblogging platform is ranked in the second position.Such shared information are generally not provided by search engines websites as they are usually not yet indexed and thus available only through the microblogging platform search tools.
The wealth of information shared in Twitter is attracting an increasing attention of researchers in many fields especially knowledge discovery and data mining.Exploring such information qualitatively and quantitatively could lead to understand and propose new powerful models learning the particularities of human behavior and interests.

2.3.1 Twitter Microblog

Twitter was created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass and launched in July 2006. Twitter allows users to publish messages, known as tweets, expressed with no more than 140 characters via SMS or web or/and mobile applications.
Nowadays, Twitter has gained a huge popularity. In our daily life, we used twitter to comment on any news and discuss trending topics. It has integrated richer characteristics by enabling users to publish various data content formats such as texts, images, links and videos. Twitter combines elements of social network sites and blogs , but with a few notable differences.

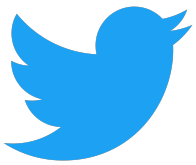

Figure 2.1: Logo Twitter

specificity of Tweets

Usernames: In order to direct their messages users often include twitter usernames in their tweets. A de facto standard is to include @ symbol before the

username (e.g.@towardshumanity).
Usages of links: Users very often include links in their tweets.(only 140 characters)
Stop words: There are a lot of stop words or filler words such as a, is, the used in a tweet which does not indicate any sentiment or meaning.
Repeated letters: Tweets contain very casual language. For example, hello with an helloooo,helllo

2.4 Data Collection Technique

2.4.1 Direct Data Access

Data Access via Research Data Collections: Test collections are generally composed of a set of documents related to one or various topics suited to specific information needs.
Data Access via US Congress Library: Since April 2010, the Library of Congress has announced its intention to archive public historic tweets for conservation and research.
Data Access via Data Grants: Twitter has introduced,in February 2014, its Data Grants project accepting applications from any member of research institutions to access to the needed historical and public information required in their research studies.

2.4.2 Data Access via Ad-hoc Applications

Data Access via Public Twitter APIs
Data Access via Crawling Techniques

2.5 Natural Language Processing

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text. NLP researchers aim to collect knowledge on how human beings understand

and use language so that fitting tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the preferred tasks.

2.5.1 Steps in NLP

NLP includes five general Steps:

- Lexical Analysis: it means dividing the whole text into paragraphs, sentences and words. It is analyzing the structure of words.
- Syntactic Analysis: It implicates analysis of words in the sentence for grammar and arrangement of words in a manner that shows the relationship among the words.
- Semantic Analysis: It defines the exact meaning or the dictionary meaning from the text.
- Disclosure Integration: The meaning of any sentence depends upon the meaning of the sentence just before and after it.
- Pragmatic Analysis: It involves deriving those aspects of language which require real world knowledge.

2.6 Information Retrieval

Information retrieval is the science of searching for information in a document, searching for document themselves, searching for metadata that describe data and for databases such as text, image or sound.

2.6.1 IR based on Information Content Classification

classification strategy consists of analyzing tweets content for situational information retrieval. Many classification dimensions have been explored to separate between relevant and irrelevant situational information:
location

time
Credibility
Information provided

2.6.2 IR based on Information Providers Classification

classification strategy consists of identifying the prominent microblog users who are susceptible to provide the targeted relevant information
User's connectivity graph
User's interaction graph
User's activities

2.7 Approaches analysis Tweets

Many models were proposed to analyze tweets such are:

- TF-idf(term preponderate)
- LDA(probabilistic)
- Vector space model or term vector model
- LSA(latent semantic analysis)(presentation term and document in the same space : improvement of Vector space model)

2.8 Conclusion

NLP is a powerful tool if understood and properly used.In the next chapter, we will discuss our contribution in order to extract meanings and links between the tweets collected

Conclusion

References

Alawadhi, S., Aldama-Nalda, A., Chourabi, H., Gil-Garcia, J. R., Leung, S., Mellouli, S., ... Walker, S. (2012). Building understanding of smart city initiatives. In H. J. Scholl, M. Janssen, M. A. Wimmer, C. E. Moe, & L. S. Flak (Eds.), . Berlin, Heidelberg: Springer Berlin Heidelberg.

Eiben, A., & Smith., J. (2007). *Introduction to evolutionary computing*. Natural Computing.

Jensen, R. (2005). *Combining rough and fuzzy sets for feature selection*. Doctoral Dissertation, School of Informatics University of Edinburgh.

Matzinger, P. (2001). The danger model in it's historical context. *Scandinavian Journal of Immunology*, *54*, 4-9.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, *8*, 338-353.

Zimmermann, H. (1996). *Fuzzy set theory and its applications*. Kluwer Academic.