Data Mining Project's Report

Aurora Maria Tumminello
University of Trento
1st year, Data Science, Trento, Italy
aurora.tumminello@studenti.unitn.it

Figure 1: Twitter Logo

ABSTRACT

This report provides a description of the work related to the data mining project of the academic year 2020/2021. An introduction and motivation will be provided at the start, accompanied by some references to related work from a theoretical perspective. Furthermore, the problem statement and a detailed explanation of the solution proposed will be presented. In conclusion, the report will contain some experimental evaluation.

1 INTRODUCTION AND MOTIVATION

The project presented focuses on Twitter data in a specific time period and it aims to identify popular topics contained into tweets. A topic can be considered as a set of terms that regularly can be found together in tweets and that it is popular at a certain time.

Before talking deeply about the project, some motivations are provided. First of all, we will focus on **Twitter data**, because they provide a clear overview of events and news in the political, cultural and sociological worldwide sphere. Secondly, **tweets** are short texts sent at specific times by users all over the world, thus people have to be more directed. In order to do so, they commonly use hashtags, i.e. words preceded by the `#` character, to spread the word and communicate directly the message with few words. It is interesting to notice that redundancy and ambiguity of natural language are diminished thanks **hashtags**: they are easy to read, to follow and they communicate directly an idea. Moreover, they become word of mouth and they are frequently used to communicate something that maybe has multiple meanings or way to be told. In addition, Twitter data, contrary to other social networks, are open and downloadable through its API, despite there is a download limit.

The usefulness of this project resides in understanding what is happening in the world, taking into account time and place (when location information is available). Furthermore, to reach this goal, the project aims seeing how words are correlated by looking at how many times they appear together and if they can be considered as a frequent itemset, i.e. as a topic.

A relevant point to take into consideration is the **timing** factor, because it bestows a contextual background to topics discovered. If data contextualization was not acknowledged, we would lose all the necessary background for accurate data analysis and interpretation. Therefore, the project presented will look at consistent sets of terms that are frequently used throughout time, i.e. popular topics.

2 RELATED WORK

This section will be completed in the next days.

3 PROBLEM STATEMENT

The input of the project presented is a dataset with the date, text corpus of tweets and hashtags used, therefore of dimension $N \times 3$, where N represents the number of rows. The `date` column should be a string in the format `YYYY-MM-dd hh:mm:ss`, while `text` contains tweet body, entire or truncated. If the corpus is truncated, there will be suspension points followed by the link to the referred tweet. The column `hashtags` refer to the hashtags that can be found in the text body, which are maintained separated from the tweet itself because some of them might be omitted for textual truncation. The input file should have extension `csv`.

Notice that the solution displayed in this report accepts all dataset that respects those assumptions. Therefore, it can be applied to the **COVID dataset** used as suggested by the project description, as well as any other Twitter dataset. In particular, the COVID dataset used is available at https://www.kaggle.com/gpreda/covid19-tweets and refers to the time period from 2020-07-24 to 2020-08-30. The original dataset also contained information about the user who published the tweet, such as username, description, number of followers and location. Related to the latter information, places where tweets' authors come from are omitted, since there are many missing values. Moreover, location is inserted by users and therefore there is no absolute way to indicate, for instance, New York (e.g. 'New York, New York', 'New York, NY, USA','Brooklyn, New York','New York, LA, Miami, Houston'). Thus, this project will focus more on the timing instead of the place where topics come from.

The solution proposed will return as output a set of significant frequent itemsets discovered in the input dataset. The topics contained in the output will be ordered increasingly by the popularity of that topic, i.e. the more that topic is cited in the dataset, the highest position will cover in the results.

4 SOLUTION

5 IMPLEMENTATION

6 DATASET

7 EXPERIMENTAL EVALUATION

8 CITATIONS AND BIBLIOGRAPHIES

REFERENCES