Data Mining Project 2020-21

Identify Frequent Topics in Covid-19 Related Tweets

Daniele Passabì
Data Science - Year I
daniele.passabi@studenti.unitn.it

ABSTRACT

TODO: complete after finishing everything

The goal of the work is to identify consistent topics in time, starting from Covid-19 related tweets.

Different techniques will be presented, such as A, B and C, useful for finding sets of frequent items over time.

1 INTRODUCTION & MOTIVATION

The aim of this project is to identify popular themes over time, starting from Tweets texts. The first step of the analysis consists in identifying popular topics on a specific day, but the focus of this work lies in finding themes that recur over time.

The problem that is faced falls within the large family of problems aimed at finding frequent items. More specifically, the data used belongs into the category of *market-basket* model. In this kind of problems, there is on the one hand a set of *items* and on the other *baskets* containing them. A set of items present in many baskets is considered *frequent*.

In the specific case of this project, dealing with social media texts, words are identified as items and tweets as baskets. By cleaning the text from noise (punctuation, extremely common words, ...) it is possible to identify common themes present in tweets. Once this is done, it's easy to investigate which topics are only momentarily popular and which are popular for longer periods of time.

Solving the problem is very relevant because it not only allows us to investigate what are the causes that make a topic popular, but also to predict the reaction of the public in the future. Furthermore, the solution algorithm can be extended to solve many other similar problems where the initial premises are the same.

The proposed solution exploits a Python optimized version of the *APriori algorithm*, the most used in solving these problems. It is important to focus on the efficiency of the algorithm, as approaching the problem in a naïve way or trying to use brute-force could be very expensive in terms of time and resources.

2 RELATED WORK

TODO: complete after finishing everything

Max 1 page. Briefly describe for the methods you will use, what they do and what is their role.

E.g., you describe what clustering does and what techniques exist for clustering.

3 PROBLEM STATEMENT

Input

In order to better understand the problem, it follows a small sample of the cleaned dataset used in the analysis.

date	text
2020-07-25	[if, i, smelled, scent, hand, sanitizers, ...]
2020-08-29	[thanks, iamohmai, nominating, who, ...]
...	...

Column **date** shows the day the tweet was posted, in the year-month-day format. It was decided to use only the day on which the tweet was published, ignoring the precise hour and minutes.

In the **text** column it is stored the text of the tweet, after being cleaned by removing extremely common words in natural language (stopwords), punctuation, numbers and symbols. The text was also transformed into a list of words, or terms. Therefore, a tweet can be defined as a text composed of several terms.

Output

The interest is placed in finding groups of frequently repeated words within the tweets of several days, in order to find frequent topics over time.

Before proceeding, it is wise to decide when to consider a set of words *frequent* on the tweets of one day. Since there is a different number of tweets available for each day of the dataset, it is not feasible to use an absolute measure.

For this reason it was decided to consider as *frequent* the groups of words that pass the following test:

$$\frac{\text{frequency of group of words}}{\text{total number of daily tweets}} \geq \text{threshold}$$

After several attempts and experiments with the threshold, its value was set at 0.015.

After running the APriori algorithm on each day of the dataset and combining the results achieved, a dataset is obtained with:

- the groups of frequent words
- the days they were used
- their frequency as a percentage of the tweets of the day

An example follows.

group_of_words	dates	frequencies
(pandemic)	[2020-7-24, 2020-7-25, ...]	[0.03, 0.05, ...]
(covid, trump)	[2020-7-24, 2020-8-16]	[0.04, 0.07]
(covid19, india)	[2020-7-25, 2020-7-26, ...]	[0.03, 0.04, ...]
...	...	...

4 SOLUTION

TODO: complete after finishing everything

The actual solution in details. Note that there is no need for code or specific software component tools description here. Also, you do not explain things already known by the theory, e.g., do not start elaborating on what is clustering and how useful it is.

5 IMPLEMENTATION

TODO: complete after finishing everything

Description of what tools you have used to implement the solution you described above.

6 DATASET

TODO: complete after finishing everything

Description of the dataset, and all the possible preprocessing that you performed to it from the original form to the one you need in order to run your program.

7 EXPERIMENTAL EVALUATION

TODO: complete after finishing everything

Perform the necessary steps to illustrate that the method is good – or is not good. You can do this through a user evaluation and also through comparison with some base line method. It is up to you to select the base line method. Then you can compare the results and comment on what you observe. You should also care not only about the quality but also about the scalability, i.e., time, related to the size of the data. In this section, you should also have a subsection called Dataset in which you describe how you created the test dataset.

Just a test of a reference [1]

REFERENCES

[1] Poker-Edge.Com. 2006. Stats and Analysis. Retrieved June 7, 2006 from http://www.poker-edge.com/stats.php