**Problem description**

* **Tweets**: short texts (sequence of words) sent at specific times by specific individuals. The collection of terms that a tweet contains is collectively describing a message. Appearance of a term in many tweets makes the term popular.
* **Topic**: a set of terms. A popular topic is a set of terms that appear together in many tweets.

Identifying a popular topic can be done by finding highly frequent combinations of terms.

Tweets have a **time** in which they are published, and by taking this into consideration, one can see that the popularity of terms is changing over time. Some term that is popular today may not have been popular 4 days ago but may have been popular the week before that.

The **goal** of this work is to **identify consistent topics in time**, i.e., set of terms that become frequent together throughout time.

For instance, the terms “hurricane” is becoming often popular, but is not continuously like this. It becomes popular when some hurricane hits a country, e.g., when it was hitting the US south eastern coasts. The same applies for the term “nuclear”. It becomes clear when there is interest due to some accord or some disarming efforts. However, there is no correlation between nuclear and hurricane popularity. On the other hand, the terms “hurricane” and “emergency” become popular together, meaning that whenever *hurricane* is mentioned frequently in tweets, so does the word *emergency*. This leads us to decide that a popular consistent topic is the {“hurricane”, “emergency”}.

Intuitively, we are looking for frequently correlated terms in time.

**Dataset**

As a dataset, you can use the twitter dataset on COVID from Kaggle.

Link: <https://www.kaggle.com/gpreda/covid19-tweets>

You probably need to do some pre-processing of the text of each tweet to turn it into a sequence of terms

* **Terms**: they can be words or hashtags

**Report structure**

The report should contain the following sections:

1. **Introduction** & **Motivation**
2. **Related Work**

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.

1. **Problem Statement**

See the formal model as we defined it in the Rec Systems Lecture for an example

1. **Solution**

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.

1. **Implementation**

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

1. **Dataset**

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.

1. **Experimental Evaluation**

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.

**Report format**

The report should be written in latex using the following template.

Link: <https://github.com/EDBT2021/Template>

The first page should contain apart from the name of the author, the year of studies (e.g. 1st or 2nd), the program (CS, EIT DS, DS, etc.).

**Delivery components**

The project is delivered in different phases.

1. Report Part A (Introduction, (maybe motivation), and problem statement sections)
2. Dataset
3. Final Report
   1. Report (Previously submitted report updated + Solution Section + Experiments Section)
   2. Code, Updated Dataset, Results

**Marks**

1. Report Part A 10% (3 marks)
2. Dataset 10% (3 marks)
3. Final Report & Code: January 31st 80% (24 marks)

As per the course syllabus, delivery of the project in the June/July period entails a 10% penalty. Delivering in Aug/Sep a 20% penalty.

**Delivery schedule**

1. Report Part A December 20th
2. Dataset January 7th
3. Final Report & Code: January 31st January 31st

**Delivery method**

Create a directory on **google drive** with your <firstname, lastname, matricola> in this order and separated with spaces. Share the specific drive with the professor [velgias@unitn.it](mailto:velgias@unitn.it) before the delivery deadline of the first phase. The directory should contain 3 subdirectories:

1. **doc** containing the report
2. **src** containing the source code of the project
3. **data** containing the dataset
4. **bin** the binary file of the execution of the program alongside the results it produces through

the various runs you have performed

**Methods to use**

**Jaccard similarity (not really in this case)**

* Definition: |S intersect T|/|S union T|
* Aim: find textually similar documents in large corpus. We can directly compare the words in the tweets.
* Note: the similarity is at character level, we are not talking necessarily about “similar meaning”

**APriori**

Definition