2. Classification of Tweets using Frequent Pattern Mining

You are given a dataset of tweets and asked to organise them according to their topics. The expected outcome is a set of classes, where each class has a label that describes the topic of the tweets belonging to it. An example a topic label is "US elections".

To assess the effectiveness of the resulting classification, you should make use of the hashtags. Each tweet is associated to a hashtag. Thus, all tweets of the same topic should have the same hashtag (in case the data is not noisy). Based on that, the quality of a class decreases when the number of hashtags associated to its contained tweets increases. The steps to follow for tweets classification are:

1. Define the attributes that represent the tweets (hashtags must be excluded).

2. Reduce the number of attributes by applying principled strategies. For example, removing stopwords, and taking only top-k frequent attributes.

3. Write the data in an ARFF format where rows represent tweets and columns represent the attributes you have chosen.

4. Use Weka to extract frequent patterns from the tweets and define class labels

5. Assess the effectiveness of the resulting classes

Deliverable: Write a report of 1-2 pages about your findings explaining (1) how did you construct the attributes, (2) the set of class labels and the strategy you followed to extract them, (3) the assessment of the quality of the classes. In addition, provide a file with all the classified tweets.

Deadline: 16/11/2017 at 2pm

1http://www.inf.unibz.it/mkacimi/teaching.shtml
2http://www.textfixer.com/resources/common-english-words.txt