Project 2: Market-basket analysis Ukraine Conflict

Tommaso Locatelli
student no. 965066
tommaso.locatelli1@studenti.unimi.it

This project implement the Toivonen's algorithm to find frequent itemsets over tweets, considering tweets as baskets and words as items. To scale up to larger datasets all the steps are implemented in Spark in order to eventually distribute the computation.

Data

Tweets were collect through a dataset available on Kaggle. The dataset was downloaded through the official API, after a token key had been obtained. The data are ingest in the spark contest in a way that allow to distribute the reading of the data. The italian tweets are the only ones taken into account.

Data cleaning

Tweets were pre-processed to provide valid results, in particular:

- tweets are splitted in tokens, that are words or punctuation marks
- links and punctuation are filtered out
- Italian stopwords are removed
- each word is stemmed

Algorithm

At first, each step of the Toivonen algorithm is implemented.
Regarding the sample selection, instead of just picking the first portion of elements, an actual random sample is taken. This imply that the actual sample size is subject to some variance and the threshold to be used in the sample need to be correct.
To obtain the frequent itemsets in the sample, an aPriori generator implemented in spark is imported from another project.
Then the pass over the sample is implemented, at first, by considering alone: singletons, couples, triplets candidates and after generalized in a single function that find elements of the negative border, one cardinality at time, relying on the frequent itemsets of the preceding cardinality. Its functioning is validated also with the book example.
Finally, the pass over the entire dataset is implemented by a specific function which return the frequent itemsets or a string that specify the number of negative border sets that happen to be frequent if the algorithm failed.
When steps were clear and well defined, a function that implement the algorithm is defined by putting in sequence the singles steps.
Some external functions have been also defined to let the pipeline be easily readable.

Scalability

A Spark-Hadoop environment was set to scale up with data size. Data are first ingest, allowing distribution, and then loaded in an rdd. The entire implementation uses extensively different methods of the pyspark.RDD class, letting computation be parallelized over many nodes, thus let archive a good degree of scalability.

Results

The algorithm is tested over 1898 tweets in two experiments with: minimum frequency support equal to 170, sample size fraction 0.05; showing how lowering the alfa value (from 0.95 to 0.7) can bring from a failing algorithm scenario to the actual finding of frequent itemsets. For instance there are six frequent itemsets: 'ucraina', 'russia', 'zelenski', 'nato', 'putin', 'russia' and 'putin'. For each itemsets the support is also reported.

Declaration

"I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study."