Market-basket analysis
Ukraine Conflict

Martin V. Ranieri
student no. 967158
martinvincente.ranieri@studenti.unimi.it

This project deployed a system that finds frequent word-sets over a collection of tweets about current ongoing Ukraine-Russia conflict.

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 dataset contains more than 1 million of tweets from different languages.

Pre-processing

The tweets were filtered by languages, only English tweets were taken into account. Tweets were pre-processed to provide valid results, in particular:

- special characters were removed
- words of each tweet were tokenized
- each word was stemmed
- stop words were removed

NLP tools were used to perform the steps above.

Algorithm

At first, each step of the classic Apriori were implemented 'till the end of the third pass. When steps were clear and well defined, the algorithm was generalized to work with an undefined number of pass.
At the end, a python generator function has been implemented. At the N-step, the generator yields all the frequent itemsets of size N, and their supports.
Some external functions (like sum or tuple ravelling) have been also defined to let the pipeline be easily readable.

Scalability

A Spark-Hadoop environment was set to scale up with data size. Pre-processed dataset was parallelized through a resilient distributed system.
Each Apriori step was coded through pipeline chunks, and the whole pipeline had been computed by the resilient distributed system on demand.
The pipeline was almost build with map or reduce steps. Map and reduce functions let computation be parallelized over many nodes, thus let archive a good degree of scalability.

Results

A sample size of 700 tweets were used to test the algorithm. A threshold of 4% was set, and Apriori managed to find around 200 frequent itemsets. Both confidence and lift were computed for all possible association rules.

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."