The processor’s story

This document presents the track of our final implementation for the processor part.

* The first attempt was to apply without any pre-processing step Latent Dirichlet Allocation (LDA) and Non- Negative Matrix Factorization (NMF) algorithm. We saw that the topics generated were very hard to be labeled because there was no semantic relationship between the topics’ words.
* Second attempt involved clustering the words segnificant for topics using word2vec and applying LDA on the devided reviews (for food businesses and for hotels). The number of words for each topic gave as input for LDA was the maximum number of words considering the clusters for each category of business. Using the LDA’s topics, we computed a confussion matrix (each element in the matrix is the intersection between the two topics that indexes the element). But the problem was that the intersection was very high, meaning that the topics were not disjunct at all. That is because the smaller the number of words per topic in the LDA algorithm, the better. So it had to be another approach.
* Third attempt was to take into account a smaller number of words per topics (a quarter of the number used in previous version). The topics were disjunct this time but there were some topics hard to label.
* This lead to another approach. We manually created a topic dictionary with some words that we thought that they’re relevant for the topic. We enriched it using word2vec and wordnet (used to find words that are semantic related). Then, we parsed the reviews for 10 businesses chosen for case study, we took the context for each word that was involved in a topic and we did a sentiment analysis of the context that the word appears. We tried to find manually some pros and cons for businesses to confront the output of the method above. The result of this seemed to work well.