Text Mining

Text classification with R

The following packages are required for this lab: `text2vec`, `Matrix`, `stopwords` and `NNML`. In order to install them with the `install.packages` command, you need to update to the latest version of R (as of now, 3.5.1).

Exercices

Part 1 - Supervised text classification

1. Load the corpus from "reviews.csv" into a data frame.
2. Split the corpus into a training set by randomly sampling 60% of the exemples and a test set.
3. Extract the raw vocabulary from the training set.
4. Vectorize both the training and test sets using this vocabulary.
5. Implement a function, `mle_mnb<-function(X,Y,k)`, to train a multinomial naive Bayse binary classifier; implement the decision rule in terms of the log-ratio of the conditional probabilities.
6. Compute the confusion matrix using the training and test sets and report the overall accuracy, for k (i.e. the Laplace smoothing constant) $\in [1; 5]$.
7. Do the same again, after pruning the vocabulary.

Part 2 - Unsupervised text classification

1. Vectorize the whole corpus using a pruned vocabulary.
2. Apply tf-idf weighting.
3. Compute the decomposition with `nnmf`, for $k = 50$.
4. Print the top words for each topic.
5. Print the top review for each topic.

Documentation

- Random Samples: https://stat.ethz.ch/R-manual/R-devel/library/base/html/sample.html
- Tf-idf weighting: http://text2vec.org/vectorization.html#tf-idf
- Row and column sums: https://stat.ethz.ch/R-manual/R-devel/library/base/html/colSums.html
- Fast non-negative matrix factorization: https://rdrr.io/cran/NNLM/