Assignment 2

Intelligent Systems - Labs

Description and overview

This is the second assignment you need to accomplish as part of the Intelligent Information Systems course. The material relevant for the following tasks is either provided as separate methods, or was considered during practical labs. The task accounts together for 100% of the second 25% of the (practical labs) grade.

The assignment must be submitted as a R Markdown file (file -> new file -> R Markdown), where you write both the code, as well as the theoretical answers (echo = T option must be enabled for each code block). Markdown files can easily be exported to PDF using ("Knit") button in R Studio; please submit also the PDF (along with the R markdown).

Mining insults

The task concerns the data present in the "insults.zip" folder, provided along with this assignment. The data has been pre-split into train and test sets (train.tsv, test.tsv). The first task of this assignment is as follows:

1. Cleaning
   - Remove punctuation and stopwords
   - Anonymize proper nouns
   - Remove unknown symbols
2. Exploration
   - Plot the frequency of words. What do you observe?
   - Perform a clustering for cluster number $k \in \{2, 4, 8, 16\}$ on the vectorized document space (select the vectorization of choice). (+ Extra 5% for using one neural (e.g., word2vec-based) and one non-neural text representation.)
   - Project representations to **two dimensions** (e.g., via PCA or t-SNE) and visualize the cluster assignments. What do you observe?
   - Color the document representations **according to class labels**. What do you observe?
   - Create a document representation comprised entirely of part-of-speech tags (one POS vector for each document)
3. Modeling
   - Discuss how balanced (or imbalanced) is the class label distribution.
   - Select at least 2 classification models and discuss their choice.
   - Select hyperparameter tuning method of choice and find the best representation-model configurations using the sufficiently split train.tsv (use e.g., cross validation)
   - Select relevant performance metrics (at least two)
   - Predict on test.tsv and report performance by using the selected metrics. Discuss the results (compare the two (or more) learners, why is the performance good/bad etc.)
   - Does adding POS tag-based representation to e.g., Bag-of-words improve performance?
4. Understanding
   - Perform feature ranking by using a filter method of choice (on a selected data representation from 3.)
   - Re-evaluate the models performance for top n features (according to the ranking).
   - Visualize model performance w.r.t. n by using the selected measure of performance.
   - Extract feature importances from a wrapper method (e.g., via a random forest)
   - Compare the two feature rankings (one by a wrapper and one by a filter) by using the Jaccard score at n: $\frac{|A_n \cap B_n|}{|A_n \cup B_n|}$, where A_n and B_n represent top n attributes from ordered sets A and B (rankings).

- Visualize the relation between n and $Jaccard(A_n, B_n)$. Interpret the results.