Advanced Human-Machine Interaction
Interaction Data Analysis
TD04: Text Mining

INSA Rouen Normandie - Normandie University

Practical session : Positive / Negative classification of messages

Data collection and processing

1. Download and analyze the dataset at `http://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences`. It contains a set of sentences posted by users and labeled as 0 (negative) or 1 (positive).
2. Merge the three files and mix the sentences.
3. If necessary, format all lines to be easily processed

Baseline (no pre-processing, bag of words, simple presence of word, no cross validation)

1. Divide the dataset into 2 sets : a learning set (2/3) and a test set (1/3).
2. Construct the learning matrix (some toolboxes propose their own function, others need you to do it yourself) : each line is a sentence, each column a word, a 1 means that the word is in the sentence, 0 it is not.
3. Learn a model (ex : SVM with linear kernel -if you don't want to wait for a too long time-, Naive Bayes, ...) on the learning set.
4. Test the model on the test set and compute the F-measure

Baseline (no pre-processing, bag of words, simple presence of word, cross validation)

1. Divide the dataset into K sets.
2. Do a k-fold cross validation (learn on sets 1...n-1 and test on set n ; learn on sets 1...n-2, n and test on set n-1 ; ... ; learn on sets 2..., n and test on set 1.
3. compute the mean F-measure

Advanced models (pre-processing, bag of words, cross validation)

1. Evaluate the efficiency of various pre-processing : stop-words filtering, stemming, lemmatization...
2. Evaluate the efficiency of various bag of word representations (word counting, tf.idf)
3. Compare various ML models (SVM, Naive Bayes, Random Forests, ...)

Neural network approaches (pre-processing, word2vec, cross validation)

1. Evaluate the efficiency of word2vec representation and classic ML algorithms (SVM, Naive Bayes, ...).
2. Evaluate various neural network models (CNN, LSTM, etc.).
3. Evaluate novel NN representations such as BERT.