MSc in Data Science

Natural Language Analytics

Academic Year: 2017-2018

Exercise 2: Sentiment Analysis

Delivery Date: **29/05/2018**

This exercise concerns the classification of tweets in categories, according to their polarity (positive, negative, neutral). We are going to use a subset of the dataset available for SemEval-2017, Task 4, "Sentiment Analysis in Twitter": http://alt.qcri.org/semeval2017/task4/index.php?id=results

1. All training data can be found here.
2. The test data can be found here.
3. The gold labels, submissions and scores for all teams can be found here.
4. The task paper can be found here.

The needed datasets have been downloaded using the SemEval tools, and downloaded files can be found at GitHub:

https://github.com/MSc-in-Data-Science/class_material/tree/master/semester_2/Natural_Language_Analytics/Exercise_2-TwitterSentimentAnalysis/data

You must not download the datasets by yourself, you must use only the datasets available in the GitHub link shown above. All tweets that were not available during downloading (containing the message "Not Available"), must be removed before any further analysis and processing.

Question A: (60%)

- Use the training, development and test examples, in order to create and evaluate a supervised classifier for tweets. You can use any machine learning algorithm and feature set. External resources (like lexicons, embeddings, etc.) can be used, if such resources seem fit for this classification task.
- You must briefly justify (i.e. no more than a paragraph) all the choices made (machine learning algorithm, selected features, etc.).

Question B: (40%)

Apply an existing application for sentiment analysis on twitter data, in order to classify and evaluate only the test examples. Compare its performance with the performance of the supervised classifier obtained in question A.

Helpful resources:

The following packages provide some **indicative**, ready-to-be-used sentiment analysis packages:

1. Pattern (Python): https://www.clips.uantwerpen.be/pattern
2. Stanford CoreNLP (Java): https://stanfordnlp.github.io/CoreNLP/index.html
3. twitteR (R): https://github.com/geoffjentry/twitteR

   a. https://dataaspirant.com/2018/03/22/twitter-sentiment-analysis-using-r/
   b. https://www.kaggle.com/rtatman/tutorial-sentiment-analysis-in-r

4. TextBlob (Python): https://textblob.readthedocs.io/en/dev/
   a. https://dev.to/rodolfoferro/sentiment-analysis-on-trumpss-tweets-using-python-
   b. https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/
   c. https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/
5. NLTK (Python) – Trainable: https://www.nltk.org/
   a. https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis