NLP Assignment 1: Sentiment Analysis

Summary

In this first assignment, you will implement your own sentiment analyser to predict the polarity (positive/negative) of the sentiment that is expressed by individual consumers in their online reviews of products they bought. Reviews consist of a short evaluative text and a product score (1 to 5) that are assumed to express the same underlying sentiment. Your sentiment analyser will learn to predict the reviewer’s score from the written text, so that afterwards, the system can also detect the sentiment in reviews without explicit scores. You have considerable liberty for making your own choices in the different parts of the assignment. First, you can choose which data to train and test your system on by selecting reviews from 1 specific product category out 24 different ones. Next, you can make personal and creative choices in the design of your system’s architecture: You can choose which machine learning algorithm you use for training a probabilistic sentiment analyser, and which textual features this algorithm should learn from. Depending on the algorithm, your sentiment analyser will assign a sentiment class or predict a sentiment score for unseen reviews. You will evaluate the results of your sentiment analyser on a held-out test set, using standard evaluation measures like precision and recall, and macro/micro averaged F1.

As background material to this assignment, you can consult the slides of Class 3 on Canvas and read Chapters 4, 5 and 19 of Jurafky and Martin, (3rd Edition, draft) *Speech and Language Processing*.

Implementing a probabilistic sentiment analyser

Dataset

For his research, Dr. Julian McAuley at the University of California (San Diego) has collected product reviews from Amazon, covering 24 different product categories. He has publicly released samples of the datasets on his website [http://jmcauley.ucsd.edu/data/amazon/ (Links to an external site.)Links to an external site.](http://jmcauley.ucsd.edu/data/amazon/). The datasets come in the format of JSON-dumps.

* Choose a product category that you like to work on and download the corresponding dataset.
* If you think the dataset is too large to be processed on your computer, you can take a sample (min. 10,000 reviews)

Pre-processing

* **Scores** (field “overall”): Depending on whether you choose a regression or classification approach for your sentiment analyser (see next step), you can either keep the original 5-point scale scores or discretize them into 3 sentiment categories (positive/neutral/negative), for which you can choose interval boundaries yourself, but motivate these in your document.
* **Text** (field “reviewText”) Preprocess the text data using python libraries of your choice (NLTK, spaCy,..). The pre-processing should contain at least tokenization and you should handle the problem of *negation* (see section 4.4 in Jurafky and Martin 3rd Ed.)
* Split your dataset in a training set and test set, and, if needed, a development set. (Optionally you can also make splits for 10-fold cross validation)

Training your sentiment analyser

* Choose a supervised machine learning algorithm you would like to use for training your sentiment analyser (motivate your choice in your document). You can use the Naïve Bayes classifier we discussed in class or any other type of regression and/or classification algorithm you think is suitable. You don’t have to implement the algorithm yourself: use any publicly available Python library you like.
* Use the training and development splits from your dataset to train and fine-tune at least 2 different models to predict the review scores and/or sentiment categories from text.
  + At least one model should use n-grams as features (you can choose the order of n-grams yourself and optionally fine-tune what works best using the dev split of your dataset)
  + At least one model should use features derived from a sentiment lexicon. You can choose any of the publicly available sentiment lexicons discussed in class and J&M Ch.19.

Evaluation of your sentiment analyser

* Use the test split from your dataset to evaluate each of your models with the appropriate evaluation measures (see Jurafsky & Martin 3rd Ed. section 4.7, or any other evaluation measure you think is appropriate)
* (Optionally you can evaluate using 10-fold cross validation (see J&M Ch. 4.8))
* Compare the performance of your models
* Pick out 10 misclassifications (negative classified as positive and vice-versa) for further manual error analysis. Discuss in your report why you think your sentiment analyser misclassified these reviews.

Handing in your assignment

The format of your assignment should be a Jupyter Notebook. This is a dynamic document that mixes Markdown and code blocks. Use the Markdown blocks to document your code and motivate the decisions you make! Make sure to describe the dataset that you have used, the pre-processing steps to clean up the data and your choices regarding the machine learning algorithm and features. You can see the Jupyter Notebook as a code file and report bundled into one.

If you have never used Python or Jupyter Notebooks before, we recommend installing everything through [Anaconda (Links to an external site.)Links to an external site.](https://www.anaconda.com/distribution/).