[HW1] Sentiment Analysis (Classification task)

Available dataset

- Movie Review Dataset (Small)
    - Dataset download (http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz)
    - Size
        - 1,000 positive and 1,000 negative movie reviews from IMDB
    - Reference site: http://www.cs.cornell.edu/people/pabo/movie-review-data/

- Movie Review Dataset (Large)
    - Dataset download (http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz)
    - Size
        - 25,000 training examples, 25,000 test examples
        - 80 MB compressed (Approximately 480 MB uncompressed)
    - Reference site: http://ai.stanford.edu/~amaas/data/sentiment

Tools You Can Use

Here is a list of some Python implementations of algorithms that you may find useful for your assignments:

- Scikit Learn
- NLTK

Tasks

Reading the dataset

Download and unpack the file provided above.

Training the Naive Bayes classifier

Write a Python function that uses a training set of documents to estimate the probabilities in the Naive Bayes model. Return some data structure containing the probabilities. It could look something like this:

```
def train_nb(training_documents):

    ...

    (return the data you need to classify new instances)
```

Classifying new documents

Then write a Python function that classifies a new document. The inputs are 1) the probabilities returned by the first function; 2) the document to classify, which is a list of tokens.

```
def classify_nb(classifier_data, document):

    ...

    (return the prediction of the classifier)
```

Evaluating the classifier

Test your NB classifieir representations for each category (Pos, Neg) and report Precision, Recall, and F1 for each category using scikit-learn

References (Books)

- Opinion Mining and Sentiment Analysis by Bo Pang and Lillian Lee. (Download)
- Introduction to Information Retrieval by Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze (Download)
- Foundations of Statistical Natural Language Processing by Christopher Manning and Hinrich Schuetze