

READ ME

Nikolas Lippmann Pareschi – nikolaslippmann@gmail.com

Version of Python: 2.7.13 Version of Spyder: 3.23

Description: In this program we calculate calculate the tf-idf for the movie_review corpus and the top ranking words for that corpus.

Installation: Please install on your pc the latest version of Anaconda with Python 2.7 and the latest version of Spyder. The program was run in Windows 10 but it should work also if you use a Mac PC or a Linux distribution.

One can download the Anaconda / Python 2.7 / Spyder here:

<https://www.anaconda.com/download/>

Packages Used:

The following packages should be installed by the anaconda environment or via through pip install package in anaconda prompt: Pandas, Numpy, Matplotlib, random, tnltk and Sklearn.

When you run the code from the file, make sure you follow the readme instructions. In the readme file the user is guided regarding where to put the movie files. Please download the files from:

<https://www.cs.cornell.edu/people/pabo/movie-review-data/>

Movie Review Data

This page is a distribution site for movie-review data for use in sentiment-analysis experiments. Available are collections of movie-review documents labeled with respect to their overall *sentiment polarity* (positive or negative) or *subjective rating* (e.g., "two and a half stars") and sentences labeled with respect to their *subjectivity status* (subjective or objective) or *polarity*. These data sets were introduced in the following papers:

- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, [Thumbs up? Sentiment Classification using Machine Learning Techniques](#), *Proceedings of EMNLP 2002*.
- Bo Pang and Lillian Lee, [A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts](#), *Proceedings of ACL 2004*.
- Bo Pang and Lillian Lee, [Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales](#), *Proceedings of ACL 2005*.

Until April 2012 (but no longer), we maintained a [list for of other papers using our data](#) the purposes of facilitating comparison of results.


Please cite the version number of the dataset you used in any publications, in order to facilitate comparison of results. Thank you.

Sentiment polarity datasets

- [polarity dataset v2.0](#) (3.0Mb) (includes [README v2.0](#)): 1000 positive and 1000 negative processed reviews. Introduced in Pang/Lee ACL 2004. Released June 2004.
- [Pool of 27886 unprocessed html files](#) (81.1Mb) from which the polarity dataset v2.0 was derived. (This file is identical to movie.zip from data release v1.0.)
- [sentence polarity dataset v1.0](#) (includes [sentence polarity dataset README v1.0](#)): 5331 positive and 5331 negative processed sentences / snippets. Introduced in Pang/Lee ACL 2005. Released July 2005.

And place them in your working directory:

```
10 from nltk.corpus import movie_reviews
11 from nltk.corpus import stopwords
12 from nltk.tokenize import word_tokenize
13 import nltk
14 import random
15 import sklearn
16 from sklearn.datasets import load_files
17 from sklearn.feature_extraction.text import CountVectorizer
18 from sklearn.naive_bayes import MultinomialNB
19 from sklearn.model_selection import train_test_split
20 import nltk.classify.util
21 from nltk.classify import NaiveBayesClassifier
22 from nltk.corpus import movie_reviews
23 from nltk.corpus import stopwords
24 from nltk.tokenize import word_tokenize
25
26 moviedirectory = r'C:\Users\nikol\txt_sentoken'
27
28
29 movie_train = load_files(moviedirectory, shuffle=True)
```



Also unpack the positive files and negative files to the txt_sentoken folder. This is mandatory.