Assignment 1 “Natural Language Processing”

Module 1 “Mine your Own Business in Blogs, Reviews and Tweets”

Learning from Big Data

Rotterdam School of Management

Student name: …

Background

Every day, thousands of people are voluntarily writing about things we are interested in such as movies, games, events, international experiences, safety, public health, and holidays. Unfortunately, we cannot manually read and interpret such huge amount of text by ourselves. We need appropriate tools for that. In this assignment, you will expand your toolbox and fine-tune such skills with supervised and unsupervised machine learning methods. You will receive a dataset of movie reviews and will be asked to come up with a relevant research question that can be solved with this dataset and the machine learning methods seen in class. The report includes the implementation of your solution and the assessment of the performance of your model.

Input data

The datasets for this test are available on the Github of module 1: https://github.com/guiliberali/Learning-from-Big-Data-Module-1 , in the folder \data.

There are seven datasets:

	Files	Content
1	**reviews_tiny.csv** **reviews_short.csv**	1k and 10k movie reviews. Each observation is a movie review. Each observation in the data includes the textual review, a numerical rating from 1 to 10 (i.e., the number of stars), the movie title, the reviewer and the date the review was written. The observation includes data from the movie being reviewed: the movie release date, the box office in the first week (as that is the strongest predictor of movie success), the studio that produced the movie, the number of teathers that the movie was released and the MPAA rating. The review also includes the number of readers who found the review useful, and the number of readers who rated the review as useful or not useful. There are reviews that non one rated as useful or not useful.
2	**Sentiment_stopwords.csv**	List of stopwords used in sentiment analysis. Optional.
3	**generic_stop_short.txt**	Generic, short list of stopwords (not specific to sentiment analysis). Optional.
4	**acting_33k.txt**	Training text for the topic acting. This file contains 33 thousand words.
5	**storyline_33k.txt**	Training text for the topic storyline. This file contains 33 thousand words.
6	**visual_33k.txt**	Training text for the topic visual and special effects. This file contains 33 thousand words.
7	**Judges.csv**	1,981 sentences labeled by a panel of human judges as being about acting, storyline or visual/special effects. This file is to be used to evaluate your model.

Task

Your task is to predict the content of the user-generated data at hand. You have two options:

1. Use the movie review data you are given

2. Use your own choice of tweets, blogs or reviews
   You are able to choose the domain that most interest you, as long as it is legal. For example, you can choose blogs, posts, tweets, or reviews in any industry such as hotels, laptops, massage, rock concerts, songs, chocolate, makeup, your family line of business, your favorite sport, your internship, etc. I do not recommend using data leaks of personal emails (so no wikileaks, please ☺). In this second option, you need to produce the files #1 and #4 to #7.

Your report will be evaluated based on a simple average of the accuracy and precision of your predictions, and the quality in terms of depth and breadth of the report. Please refer to Canvas for the rubric.

Your submission must contain the following deliverables:

1. Your report 'Assignment_1 _ERNANUMBER.pdf` (see report template on Canvas for details)
2. Your predictions for every review in the data
3. Your evaluation metrics indicating the quality of the predictions
4. Your word likelihoods

Please submit your report (pdf) to Canvas before the deadline. In addition, please upload your predictions, likelihoods, and your code in the section "extra material". Instead of uploading your code you can also add a link to your GitHub repository in your report.

All the best and have fun ☺!

Prof. Gui Liberali