Sentiment Analysis of IMDb movie reviews

Paritosh Gaiwak
Pragam Gandhi
Sanjana Kacholia
Tushar Pahuja

1 Project idea

Sentiment analysis is one of the popular sub-fields of Natural Language Processing which helps in building systems to identify and extract opinions from the text. With the help of sentiment analysis systems, unstructured information (in the context of machine learning) in the form of sentences can be classified into a structured format. For example, the public opinion of a product can be estimated by the reviews it gets on online shopping portals.

For this project, we will design a sentiment analysis system based on movie reviews given in the IMDB movie reviews dataset. The dataset contains a total of 50000 reviews, with 25000 reviews each in the training and test sets. The positive and negative reviews are equally distributed in training and test sets.

2 Software needed for the project

- Programming Language: Python 3.x
- Libraries: NLTK, numpy, Pandas, scikit-learn
- Software Tools: Spyder, Jupyter Notebook, Anaconda

3 Data set

Large Movie Review Dataset [http://ai.stanford.edu/ amaas/data/sentiment/]

4 Midterm milestone

For this project, data preprocessing mainly includes conversion of text documents to a structured format. Various data manipulation steps like lexicon generation and stemming are performed and the data, in the desired format, would then be used for exploratory data analysis (EDA). Based on the result of the EDA, feature selection and experiments will be performed using basic Machine learning algorithms (e.g. Naive Bayes). Results from this analysis can be used to build final models to check for improvement in accuracy.

5 Relevant research papers

- Baselines and bigrams: Simple, good sentiment and topic classification [link]
- Describes both symbolic and machine learning techniques for understanding sentiments from the text [link]
- Overview of sentiment analysis, particularly opinion mining with techniques and tools that could be used [link]