Il faut sauver les datas, Ryan !

Boosz Paul paul.boosz@gmail.com
Estrade Victor victor.estrade@u-psud.fr
Gensollen Thibaut thibaut.gensollen@gmail.com
Rais Hadjer rais.hadjer@gmail.com
Sakly Sami sami.sakly@u-psud.fr

January 26, 2016

The objective of the challenge is to determine the polarity of an opinion from raw text. Since it's a challenge for starter you will only focus on classifying opinion to positive or negative. You can go further in detailing sentiments like happiness, sadness, satisfaction

1 Background

The area of text analytics includes techniques are using multiple areas, such as linguistics, A.I., or statistics. It has a wide range of applications from information retrieval to marketing. Sentiment Analysis and Opinion Mining are one of these application. For the last one, it utilizes a subset of text analytics techniques with the goal to categorize opinions/reviews within a pre-specified range of "non-favorable to favorable" rankings.

In this project you will tackle the problem of Opinion Mining in movie reviews with a basic set of techniques used in text classification.

Many sentiment-analysis methods for the classification of reviews use training and test-data based on star ratings provided by reviewers. However, when reading reviews it appears that the reviewers' ratings do not always give an accurate measure of the sentiment of the review.

You will be asked to perform an annotation study in order to categorize reviews as positive or negative, thanks to the vocabulary used in the reviews.

2 Material and Methods

2.1 Dataset Description

The train test is composed of 25 000 reviews taken from the IMDB website. These reviews include 12 500 positive and 12 500 negative reviews. Each example is composed of 3 fields :

- `id`
- `text` : raw text
- `label` : 0 (negative), 1 (positive)

All reviews come from the IMDB website. The positive reviews are reviews that gave the movie a score of 7/10 or higher, the negative reviews are reviews that obtained a score of 4/10 or lower.

You will be evaluated on a test set of 25 000 unlabeled reviews.

There may be reviews of the same movie in the train and in the test set

2.2 Preprocessing

We proposed the following basic preprocessing :

- We remove the HTML tags
- We represent the text in the bag of words model
- We use as features a TF-IDF matrix

Improving the preprocessing will be a part of your work

2.3 Evaluation

Given that we have the same number of examples in both classes, we will evaluate with the accuracy of the predictor, ie the percentage of right predictions.

Given the class equirepartition, random guess will give a score of 50%. You will hopefully get a higher score than this.

2.4 Baseline Method

Our baseline method is a Naive Bayes classifier

3 Preliminary Results

3.1 Starting kit

The starting kit is a jupyter notebook containing a way of processing the data. The raw text is converted to a matrix of Tf-Idf feature representation before being passed to a random forest. This model gives an **accuracy** around **82%**.

3.2 Improvement Ideas

- You can improve the preprocessing (removing stopwords, stemming)
- The bag of words model can be improved to take into account the structure of the text (ngrams, graph representation, word vectors,...)
- You can engineer new features from the text (length, punctuation,...)
- You can test more complex predictors

3.3 Other tools

Data and our documents are available on our github page : `https://github.com/tgensol/projet`

Sources are available here : `http://ai.stanford.edu/~amaas/data/sentiment/`

References

[1] Kuat Yessenov and Sasa Misailovi *Sentiment Analysis of Movie Review Comments* 2009.

[2] Bo Pang and Lillian Lee and Shivakumar Vaithyanathan *Thumbs Up? Sentiment Classification Using Machine Learning Techniques* 2002

[3] Bo Pang and Lillian Lee *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales* 2005