Web and Text Analytics

Course Project 2: Sentiment Analysis

Background

You will be provided with movie reviews (text) annotated with their respective sentiments (positive, negative), which have been assigned by users.

Your aim will be to train a machine learning algorithm on these reviews, and to use the trained model to predict the sentiment of new reviews.

Data

The movie reviews can be downloaded at http://www.cs.cornell.edu/people/pabo/movie-review-data/ (see "polarity dataset v2.0").

(After downloading and unzipping), you will notice that the actual reviews are organized into 2 subfolders, containing movies with a positive and negative sentiment respectively.

Requirements

Cleaning & Preprocessing (5 marks)

You are required to clean the data as follows:

- Remove stopwords
- Remove punctuation symbols (these are not useful as we will adopt a bag-of-words models)
- Convert to lower-case

You may need to perform other cleaning and pre-processing activities.

Feature Selection (10 marks)

- Compute the relevancy of words using tf-idf
- Select only those words whose relevancy score > a user-defined threshold. To select a threshold:
    - View the words and their respective tf-idf scores (sorted in ascending/descending order of the score)
    - Find out the threshold via inspection.

Training & Predictions (10 marks)

- Train a Naïve-Bayes classifier and a RandomForest classifier on the reviews
- Generate the confusion matrix. Remember to adopt cross-validation.
- Which methods give the best performance?
- Apply both of your classifiers on new movie reviews. You can collect them from any suitable source, such as RottenTomatoes.
- Comment on the performance of your classifiers on these new reviews?

Extra (10 marks)

- In addition to tf-idf, apply either the chi-squared or mutual information tests to further select the most promising features.
- Retrain your models with this new feature set and estimate its performance on new reviews as described above.
- Instead of using words as features
    - Represent each word by its respective part-of-speech tags. For e.g. nouns will be replaced by NN and adjectives by JJ.
    - You can skip the tf-idf step
    - Compute either the chi-square or mutual information tests and select only the most promising POS-tags
    - Retrain your models using this new feature set (POS tags) and estimate its performance on new reviews as described earlier. Comment on its performance.

Deliverables & Deadline

- Short report (2-3 pages), describing the entire framework:
    - Tools used for the various steps above
    - Results:
        - Intermediate results, e.g. tf-idf scores of 10 words, chi-square scores of 10 words
        - Final results: confusion matrices and performance scores
- Source code
- Submissions to be made via lol@. Zip the report and the source codes together (or give the github link in your report).

Deadline: Thurs. 29th Nov. , midnight.

Teams will present their work in my office during the lecture hours of the 30th Nov. Details will follow.