CA4023

Assignment Two

Semester Two, 2021

Build a sentiment analysis model which attempts to classify a movie review as either positive or negative.

A baseline system which uses Naive Bayes is described in the *sentiment-naive-bayes.ipynb* notebook. This system is trained on the movie review polarity data described in A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. The dataset itself is available at http://www.cs.cornell.edu/People/pabo/movie-review-data (section "Sentiment polarity datasets"). It contains 1000 positive and 1000 negative reviews, each tokenised, sentence-split (one sentence per line) and lowercased. The model is evaluated using average 10-fold cross-validation accuracy.

Your goal is to improve the accuracy of the baseline model by carrying out a series of experiments (documented in a Jupyter notebook), which attempt to improve the accuracy by

1) Experimenting with **different learning algorithms**, e.g. logistic regression, decision trees, support vector machines, etc.
2) Experimenting with **different feature sets**, such as:
    a) Handling negation
    b) looking at bigrams and trigrams rather than just words
    c) using sentiment lexicons
    d) performing linguistic analysis of the input (e.g. part-of-speech tagging)

You should use the same training and test set-up as the baseline system.

Marking Criteria

Marks will be awarded for

1. Scope of experiments (4 marks)
2. Description of experiments (4 marks)
3. Average 10-fold cross-validation accuracy (4 marks)
4. Accuracy on 'hidden' test set (also movie reviews) (2 marks)
5. Clear, readable code (1 mark)

Assignment Submission (via gitlab)

1. Jupyter notebook

Assignment Deadline

Mon 22nd March, midnight