Natural Language Processing – Project Proposal

Siva Karthik Gade (sxg137530)

Title: Movie Review Sentiment Analysis

Data Set: Large Movie Review Dataset. This data set contains 50000 distinct movie reviews tagged as either positive or negative. (Download Link)

Ref - http://www.aclweb.org/anthology/P11-1015 (ACL 2011)

Description:

- This project is aimed at performing sentiment analysis on a large set of movie reviews.
- Learn different data models in supervised manner using the training data.
- Use these data models to perform sentiment analysis on the test set.
- Compare performance of all the data models.
- Build an ensemble classifier, using all the learnt data models, which perform better than all the individual classifiers.

Feature Engineering:

Planning to use the below feature engineering methodologies and compare the performance of the respective data models:

Features:

- Bag of words (Unigram)
- Bag of words (Bigrams)
- Consider only a subset of tokens (unigrams/bigrams) as features, using a feature selection heuristic.
- Use some semantic approach to engineer features.

Feature Values:

- TF-IDF
- TF
- Term present/not present

Natural Language Processing – Project Proposal

Siva Karthik Gade (sxg137530)

Data Model / Classifier:

Planning to learn at least four different classifiers (Naive Bayes, Support Vector Machines, Maximum Entropy Classifier, Random Forests). Compare the performance of all the classifiers.

Testing:

Test the classifiers using set aside test dataset. Planning to use the following metrics:

- Accuracy
- Precision
- Recall
- F-Measure

Software:

Planning to use out of box feature engineering and classifier support available in WEKA and publicly available NLP libraries. Develop rest of the unavailable algorithms, wrapper code and ensemble classifier.

Expected Deliverables:

- Project implementation source code.
- Readme and Result Report for different preprocessing and classification approaches. Performance comparison matrix among them. Working of the ensemble classifier built to perform sentiment analysis of the movie reviews.