CSC591/791: Project Proposal (ALL_12)

Anusha Manur
amanur@ncsu.edu

Kartik Shah
kcshah@ncsu.edu

Shailaja Mallick
smallics@ncsu.edu

Project Title

Applying Machine Learning Techniques for Text Classification

Data Set

We will be using Twenty Newsgroup data set which is publicly available from CMU website. This data set contains 1000 text articles posted to each of 20 online newsgroups, for a total of 20,000 articles. The newsgroups are hierarchically organized with articles belonging to categories like sports, science, etc. Each record in the data set is a text file.

Project Idea

With the massive volume of data available online such as news feed, email, medical records, automatically classifying text documents is a problem. Such text classification usually requires large amount of labeled data to train any model which affects the accuracy. Hence, to improve the accuracy, unlabeled documents can be used to augment the labeled ones.

Although prior work suggests various text classification algorithms such as Naive Bayes, we in this project will be using concepts from Natural Language Processing such as stemming, text tokenization and lemmatizing for data pre-processing. We will be also incorporating topic modeling by vectorising and classifying the labels. For handling unlabeled data to augment labeled documents, we will using Semi-Supervised Learning techniques. We will validate our model by reporting the performance via accuracy and error rate.

Software Needed

Python and it's machine learning libraries

Papers to Read

We will be reading the seminal papers by McCallum et al., such as Text classification from labeled and unlabeled documents using EM and A comparison of event models for Naive Bayes text classification.

Work Division

The project work will be divided equally among all the three members. As the project is based on predicting labels for unlabeled document, each of us will be working on one model to analyse the data and compare the results. Data pre-processing, feature selection, model training, testing and prediction of labels will be shared tasks. The project report will also be split equally.