Patricia Neckowicz
Sahil Seekond
COSC 074

# Project Proposal:
# Text Classification using Neural Networks

**Problem**

Our goal is to develop a system that uses the content of a text (such as a news article) to identify its category from a predefined list. For example, given a passage that includes the words “ball” and “bat”, we want to identify the text as belonging in the “sports - baseball” category. This is not a task that can be easily hard coded, as words can appear in multiple categories and the frequency of a word, or proximity to certain other words in a category, might be more indicative than the presence of the word alone.

**Methods**

We will address this problem using a neural network approach and compare the results to other supervised learning algorithms.

For the document preprocessing step, we will perform the following:

1) Eliminate all common/uninformative words, such as “and”, “from”, etc.
2) Experiment with various ways of using word counts as feature vectors. For example, a document could be represented by a vector counting the number of times each of the top 2000 words in the training set appear.[7]
3) Reduce the dimensionality of these vectors (using the DF method, CF-DF, etc.[1]).

Once we have reduced features, we will train our neural network using a backpropagation technique[2], where we will label feature vectors with their corresponding category and adjust weights of the network accordingly. The output of the network will be a classification vector in which the nth position indicates the relative likelihood of the input belonging to the nth category. We will compare the performance of our neural network with a Naive Bayes[3] approach, a decision tree approach, and a supervised SVM approach, which may be more efficient considering the dimensionality of our data[6]. Our goal is to pinpoint the top-performing combination of supervised learning approach and dimensionality reduction technique.

**Data Sets**

We will incorporate the following dataset of approx. 20,000 newsgroup documents organized into 20 topic groups. The groups include different sports, medicine, technology, politics, etc.
**http://qwone.com/~jason/20Newsgroups/**

In addition, we will use a list of the most common words in the English language, most likely to remove the top 500-1000 from our text during preprocessing.
**http://www.wordfrequency.info/free.asp?s=y**

**Milestone**

By the milestone, we will have all of the data preprocessed according to our needs and have a functional neural network classifier. Moving forwards, we will implement the other supervised learning approaches for comparison.

**Sources**


[1] http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=765752
[2] http://neuralnetworksanddeeplearning.com/chap2.html
[3] https://web.stanford.edu/class/cs124/lec/naivebayes.pdf
[4] http://ronan.collobert.com/pub/matos/2008_nlp_icml.pdf
[5] http://courses.unt.edu/ruiz/Publications/asis-sigcr8.pdf
[6] http://link.springer.com/chapter/10.1007%2FBFb0026683
[7] http://web.mit.edu/6.863/www/fall2012/projects/writeups/newspaper-article-classifier.pdf