Final Project Proposal: The Text Categorization Project

Jin Yong Shin
Jae Goan Park

Team Members

- Jin Yong Shin
- Jae Goan Park
- TBD

Project Description

Problem

We have decided to explore the “Text Categorization Project” for our final project for Deep Learning. The goal of the project as mentioned in the project guidelines is to achieve the lowest possible ‘classification error rate’ when determining the topic (or source) of a text corpus. Many options are available for us to utilize in achieving

Dataset

We have decided to experiment with the dataset of news headlines from 4 categories. This experiment will primarily be in the form of supervised learning on textual data. The default features given by the Kaggle dataset is obviously not enough to perform high performing classification, therefore we will need to determine which features to further segment and quantify to feed into neural nets to train the data. In the Kaggle dataset, we are given the Title, URL, Publisher, and the timestamp of the respective news articles. We will primarily be using the corpus of titles and perhaps the publisher names to create good features and fabricate an efficient and high performing model.

Approach

The four categories we are given are business, entertainment, science and technology, and finally health.

- Feature Selection
    - Tokenization
        - Tokenize title corpus’
    - Stemming
        - Use nltk python library to stem title corpus
    - Normalize multinomial feature set
        - Find normalized frequency of every unique word
    - Feature Engineering
        - Create set of all unique words in the tokenized and stemmed title corpus dataset and for each example create feature vector (defaultdict for python) and assign values ‘0’ or ‘1’ to words to indicate whether or not they are present in the set of unique words
        - For publisher names, in order to distinguish them from title corpus dictionary key, value set, we will prepend the string “p-“ to categorize all

the publisher names as publishers and append to the dictionary of unique title words

Modelling

We will take two major approaches this problem: General Machine Learning Algorithms and Deep Learning (both separately and combined)

- Machine Learning
  - Naïve Bayes (sklearn package)
  - Support Vector Machine (sklearn package)
- Deep Learning
  - Multi-layer feedforward networks
    - Using term weighting method with neural network approach
  - Recurrent Neural Networks
- Hybrid Method (adventure!! if we have time)
  - Use neural networks with naïve bayes
  - LS-SVM

Work Distribution

Unfortunately, since we still do not have a third team member (spreadsheet indicates there are 4 groups of 2 leftover), we will temporarily assign project duties evenly between the two team members.

Jae: Extract data and perform feature engineering

Jin: Create models and evaluate performance

Both of us will mostly do little bit of both when taking the hybrid method