COMS 4772 Advanced machine learning project proposal

Automatic summarization of article using regularised SVM and Lasso regression

Jiamin Huang jh3768 Shuyang Zhao sz2631 Lingjun Zhao lz2495

Objective:

For a large content of text file, we want to retrieve the most important information and summarize the content in a few lines of words. In other words, we want to reduce a large amount of data into a useful small amount of data by extracting the most important content. The data comes from the web in a specific field (e.g. political news, reports).

Data representation:

We may present the article with the following broadly used methods:

- Term frequency-inverse document frequency (tf-idf) weighting
  Which puts more emphasis on rare words and less emphasis on common words
- Graph based method
- Clustering based method

Machine Learning Methods:

We will gather the training data from the web and learn the data using the following techniques:

1. Naive Bayes
2. SVM with linear kernel
3. LASSO regression
4. Recurrent Neural Network

Detailed and Future Plan:

We will first use Term frequency based method to start with the data and use two of the Machine Learning methods listed above to test our learning approaches. If the result is positive and time is sufficient, we will improve this project further by trying different ways to feature data and train them in order to make comparison. In addition to an article summarizing, we may use our existing works to see if it is feasible to also get a good solution for text summarization on multiple articles.

Programming Language and Computing Platform:

We would use Python in the entire project. As training data is extremely large, training and testing process would be conducted on Google Cloud Platform with GPU support.

References:

http://pakacademicsearch.com/pdf-files/sci/74/121-129.pdf
http://thescipub.com/PDF/jcssp.2016.178.190.pdf
http://cs229.stanford.edu/proj2013/ZhouMashuq-WebContentExtractionThroughMachineLearning.pdf