|  |
| --- |
|  |
| **Newspaper Article Classification Contest** |
| Proposal |

**Terry LiTao Yu**

*tianweil@andrew.cmu.edu taoyu@andrew.cmu.edu*

**Abstract**

This is the proposal for the Newspaper Article Classification Contest project. The content includes how we retrieve the features from articles, which classifier we plan to use and how to train the classifier.

**1 Project members**

Terry Li (Andrew ID: tianweil) and Tao Yu (Andrew ID: taoyu)

**2 Plan A**

**Feature Extraction**

We plan to try the “Bag of words” first. We plan to parse the training articles and construct a dictionary for the words (except the stop words). Then use the word count in each article for the feature.

Feature reduction: rather than using the whole vocabulary (except stop words), we will reduce the vocabulary to the t highest frequency words.

1. Calculate the histogram of all features.  
2. Select the most frequent features using a threshold t  
3. Try to use different values for t, and choose one which gives the most accurate result.

Singular value decomposition (latent semantics) approach which (as far as I understand) needs me to construct that matrix in the first place to calculate the SVD matrix. I found some research of people doing this and was wondering if there was a trick not needing to construct the full matrix I didn't get until now.

**Classifier Selection**

We plan to try the Naïve Bayes Classifier first.

LR--------------------

**Training Classifier**

For the training articles, we calculate the percentage of count for each word in the dictionary for each class, which is the MLE probability for the words happening in the given class.

Then for the new article, we multiply the MLE probability for each word in the article for each class and get the class that has the highest result.

Cross validation ---------

Overfitting ---------

**3 Plan B**

**Feature Extraction**

We plan to try the “Bag of words” first. We plan to parse the training articles and construct a dictionary for the words (except the stop words). Then use the word count in each article for the feature.

Feature reduction: rather than using the whole vocabulary (except stop words), we will reduce the vocabulary to the t highest frequency words.

1. Calculate the histogram of all features.  
2. Select the most frequent features using a threshold t  
3. Try to use different values for t, and choose one which gives the most accurate result.

Singular value decomposition (latent semantics) approach which (as far as I understand) needs me to construct that matrix in the first place to calculate the SVD matrix. I found some research of people doing this and was wondering if there was a trick not needing to construct the full matrix I didn't get until now.

**Classifier Selection**

We plan to try the Naïve Bayes Classifier first.

LR--------------------

**Training Classifier**

For the training articles, we calculate the percentage of count for each word in the dictionary for each class, which is the MLE probability for the words happening in the given class.

Then for the new article, we multiply the MLE probability for each word in the article for each class and get the class that has the highest result.

Cross validation ---------

Overfitting ---------

**References**

[1]