TEXT CLASSIFICATION

1. FEATURE SELECTION

A major challenge in text classification consists on the selection of features due to the rich nature of natural languages. The English language for example contains more than a million words; which represents a very large feature space, and which results computationally infeasible.

*Stop words and Punctuation*

Stop words and punctuation characters were removed by using the regular expressions Python module *re* and the natural language toolkit *nltk*. Manual manipulation was also implemented to avoid dedicated stop words such as *\_\_eos\_\_.*

*Mutual Information*

In this project we selected a feature selection method based on our need to reduce computational cost, and the specific intention to classify text. We implemented the Mutual Information (MI) feature selector, which measures how much information the presence/absence of a word contributes to making the correct classification decision of a conversation [3].

The concept of MI is defined in information theory for two discrete random variables as follows:

In the case of text classification, one of the random variables indicates if a document contains a given word, while the other variable one indicates if the conversation belongs to a given topic.

Chapter 13 in [3] does a good job at describing the algorithm to implement mutual information in order to extract features in a text classification problem. In this chapter the equation to calculate MI is represented in terms of counters for implementation purposes.

Where for example represents the number of conversations where word X is absent, but label with class Y. Please refer to [3] for more in detail description of this equation.

Since MI indicates how much information a word contains about a given class, we selected the words with the highest MI values for each class to build the features space of our classifier.

The following table shows the top 5 words, with descending priority, from each of the classifiers when using the MI algorithm to prioritize features.

|  |  |
| --- | --- |
| **Topic** | **Top Five Features** |
| *Author* | book, write, read, author, story |
| *Movies* | film, movie, scene, actor, director |
| *Music* | song, album, band, record, play |
| *Interview* | president, time, say, government, look |

REFERENCES

[3] [Christopher D. Manning](http://nlp.stanford.edu/~manning/), [Prabhakar Raghavan](http://theory.stanford.edu/~pragh/) and [Hinrich Schütze](http://www.cis.uni-muenchen.de/personen/professoren/schuetze/), Introduction to Information Retrieval, Cambridge University Press. 2008.