Automated Text classification

Niharika Kunaparaju(nk1851), Narasimman Sairam(ns3184), Kavitha Vishwanathan(kv668)

Proposal:

Text classification is a widely used machine learning technique in many applications. Some of the examples are like email spam detection, automated categorization of documents, etc. Text Categorization is the task of assigning documents to predefined categories. Using text categorization, we can perform Information Retrieval that allows users to browse more easily the set of texts of their own interests, by navigating in category hierarchies. This can further be extended to hierarchical text classification where the documents are classified based on a hierarchy and place at the leaf nodes of the hierarchical tree.

Example:

After training our model, given an unknown document our system should be able to classify the document into a flat-structured/hierarchical category. A document on Apollo shuttle should categorize as science -> space (hopefully what we are trying to achieve).

Possible Corpus:

1. Wikipedia
2. Reuters Corpus
3. Newspaper Articles

We have not decided on the exact corpus that will suit our purposes and hence we will explore more on these options and use the one that fits best.

Tentative experiments to perform:

Some of the key steps that could be performed to make the prediction better is having a good feature extractor (TF-IDF, POS tagging, exploiting context). With the resulting multi-classification vector representation, we could build a model, which will assign one or more predefined classes or categories to a document, making it easier to manage and catalogue.