**EXTENDED OUTLINE OF FINAL PROJECT**

**TITLE**

**Classification of documents by using Naive Bayes**

**ABSTRACT**

This project is to identify the class of the document, based on the documents used for training. To be clear, few documents and the corresponding class of document will be given.  Now, by using Multinomial Naive Bayes algorithm given documents are analyzed. Navie Bayes uses the conditional probability of Bayes theorem.  When test document is given multinomial algorithm analyzes every word in the text document in perspective to all classes and assigns a score. Finally, the class with more score is the class of the testing document.

**RELATED DOMAIN OF STUDY**

Bayes Theorem helps us to find the probability of an event to occur, based on conditions related to the event in advance. For example, if a disease is related to geographical location, then using Bayes theorem, a person’s geographical location can be used to more accurately assess the probability that they have the disease, compared to the assessment of the probability of disease made without knowledge of the person’s location.

Multinomial Naive Bayes uses Bayes Theorem to find the probability of the document to be in particular class or category. And this is based on analysis of the previous document set used for training the algorithm.

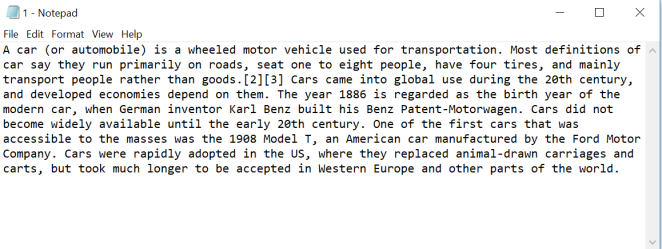
**ALGORITHMS**

* The Multinomial Navie Bayes algorithm is used to classify the given test document. And this algorithm works based on the conditional probability.
* Training data has to be given to training algorithm. Training algorithm calculates the conditional probability of the word in a particular class.
* This algorithm can be used to train offline by using the local training dataset.
* The result of training algorithm has to be the conditional probability of every word to be in a document. And the final application algorithm specifies the class of the given test document.

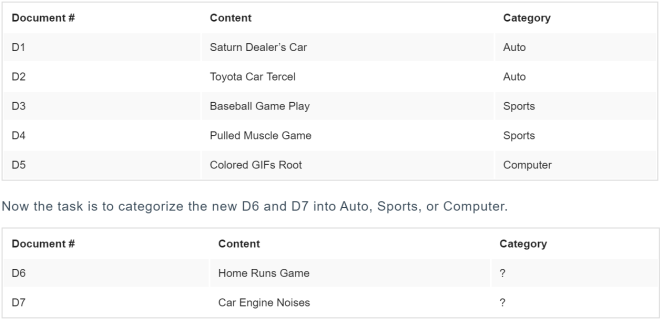
**DATA SOURCES**

Data Sources for my project would be, documents with some text in it. I need to have different sets of documents one with categorized documents for training the algorithm and other with uncategorized documents for testing.

For example, have a look at my sample text file in my dataset,



For reference, my dataset looks like this:



**GRAPHICS**

I would like to show results in the form of Bar graphs by using Matplotlib in python.

**CURRENT CHALLENGES**

Finding the existing dataset was the challenge. But I was able to create a dataset for my project by collecting text document by surfing.

**REFERENCES**

[NLP Stanford](https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html)

[Wikipedia](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)

[Pillar Global](https://www.3pillarglobal.com/insights/document-classification-using-multinomial-naive-bayes-classifier)

[Sebastian Raschka](http://sebastianraschka.com/Articles/2014_naive_bayes_1.html)

[SCIKIT-Learn](http://scikit-learn.org/stable/modules/naive_bayes.html)

[SciKit Library](http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html)

[Basics of Naive Bayes Algorithm](https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/)