**TEXT CLASSIFICATION USING NLP**

**PURPOSE**

This purpose of the project is to build a ML Model that categorizes the document using NLP.

**INTENDED AUDIENCE AND READING SUGGESTIONS**

This project is a prototype and it is restricted within the group. This has been implemented under the guidance of our trainer. This project is useful for the librarians and other researchers who are in need of categorizing the document.

**PROJECT SCOPE**

The purpose of the project is to create a convenient and easy-to-use application for librarians, researchers. The model is based on Machine Learning and Natural Language Processing. We will have a model that will be used widely to categorize any type of document.

**EXTERNAL REQUIREMENTS**

**Processor**: Intel i3 or higher/ AMD Radeon

**Operating System:** Windows 7/8/10 or Ubuntu

**RAM:** 1GB or Higher

**STORAGE:** 100 MB or Higher

**SOFTWARE INTERFACES**

**Tools:** Python IDE, PyCharm 2018 (1.4)

**LIBRARIES:** Pandas, numpy, Matplotlib, sklearn, nltk

**TEAM SIZE:** 2

**TEAM MEMBERS:** ANOOP AND REENA

**ASSUMPTION DEPENDENCIES**

Let us assume that this is a completed ML Model and it is used in the following application:

* The document handed will be categorized between library books, media articles, or is it a web page will be processed by certain key words and grouped.
* Calculation of frequency (most frequent words) and reduce the weightage of more common words.
* Generating the confusion matrix and performance report and plotting the points for the same.

Assuming the dataset is already extracted and pre-processed beforehand and the output of the document is stored as a vector matrix from the dictionary.

**DATE:** 24/07/2018 **COMPANY**: Infidata Technologies