Identify the Categories of Texts

**The Goal**

In the electronic era that we are currently living in, there are gigabytes or terabytes of text data being generated every day. These data are generated from corporations, people, and machines. There are millions of documents being store in corporations, millions of tweets from twitter, and millions of data being generated from automatic sensors. These data can all provide useful information, but they are disorganized and hard to analyze. The most important step in analyzing text data is to categorize them into groups. In order for analysts to analysis these data, they need to organize the data before they can determine which type of data are needed for analysis. For this project, my main goal is to identify the texts and determine which of these three categories they belong to: Fiction, Non-Fiction, and Article.

**Key Metric**

The method that I will be using to distinguish the texts is *Tf-idf score*. Tf-idf scores can be viewed as translating from ‘human-readable’ language to ‘computer-usable’ numeric form. It transforms the words into numbers based on the frequency of the words appearances. There are two types of frequencies: document frequency and collection frequency, which represent the number of appearances in each text and total appearance throughout all the samples respectively.

All the content of the texts will also be divided into paragraphs in preparation to calculate Tf-idf scores. The three other variables that I will also be using to support Tf-idf scores are: length of total text, length of each paragraphs, and ‘K-Mean labels’ generated from Clustering Analysis.

Sample Size: 100

30 Fiction, 30 Non-Fiction, and 40 Articles

Training Set: randomly pick 75 texts

Test Set: remaining 25 texts

**Clustering**

|  |  |
| --- | --- |
| C:\Users\Mike\AnacondaProjects\Unit 4\Capstone\KMean.png | C:\Users\Mike\AnacondaProjects\Unit 4\Capstone\Mean_Shift.png |
| C:\Users\Mike\AnacondaProjects\Unit 4\Capstone\Spectral.png | C:\Users\Mike\AnacondaProjects\Unit 4\Capstone\Affinity.png |

These 4 clustering are generated from the training set.

I have decided to use the labels from K-Mean Clustering as a feature to differentiate the texts. The reason is that Mean-Shift has 2 clusters and Affinity Propagation has 7 clusters. For this project, the number of categories is set to 3; therefore I want to pick a clustering technique that closely represents that number. The reason why K-Mean is picked over Spectral Clustering is because the clusters are more closely grouped together and easier to identify the groups.

**Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest | Logistic Regression | Gradient Boost |
| Training | 0.997 | 0.901 | 0.990 |
| Test | 0.586 | 0.572 | 0.578 |

|  |  |  |  |
| --- | --- | --- | --- |
| Cross-Validate | Random Forest | Logistic Regression | Gradient Boost |
| 1 | 0.684 | 0.733 | 0.458 |
| 2 | 0.789 | 0.851 | 0.719 |
| 3 | 0.716 | 0.772 | 0.725 |
| 4 | 0.788 | 0.850 | 0.839 |
| 5 | 0.708 | 0.738 | 0.901 |
| Mean | 0.731 | 0.789 | 0.728 |
| Standard Deviation | 0.037 | 0.052 | 0.152 |

Note: Cross-Validation is using all 100 samples

Judging from the results from cross-validation, Logistic Regression appears to be the better model to be used for this particular case. The key factor that makes Logistic Regression to be the best model is its **mean**. This value represents the accuracy in which the model can accurately categorize the texts into the correct category.

**Note for Future:**

* The sample size is small for this project => work with larger dataset
* Articles and Books are easier to distinguish because of the number to texts => include more varieties of samples such as: brochures, medical documents, book chapters, poems, and etc.
* In cross-validation, test with more trials => increase accuracy
* Can try different models
* Use statistical test for more solid supports on the claim.