Articles Clustering Report

Problem Definition:

The problem that is being addressed here is how to group similar articles in each category into the same group without the need of a human reading them, this problem could be classified as an unsupervised problem and also could be classified as an NLP problem because the data that's being used is text data

Solving the problem:

Because the data that's being used in this problem is unstructured text data, the Pre-processing step was the most important and the biggest .step

In that step I lowered all the letters, deleted all the digits and the stop words from the text, I also lemmatized all the words in the text, then after that, the text was fed into a Tf-idf Vectorizer to vectorize the text and to calculate the overall texts weightage of the words, and finally, the output of the Tf-idf vectorizer will be fed to a dimensionality reduction model called Truncated Singular Value Decomposition

The output of that step was then fed to the modelling algorithms: K-means, Affinity Propagation and Agglomerative Clustering, to choose the best of them for clustering the data based on their silhouette score. I have calculated the silhouette score for the three algorithms and came to the conclusion that the best algorithm to cluster the three categories is Affinity Propagation

.The clustered articles then were written into a JSON file called output