## Grouping / Classification of documents based on text analysis

I am trying to use the text analysis pipeline to analyze the texts from various articles and try to come up with a grouping of the articles based on similarity of topics of the articles.

The data set identified for this purpose is from the kaggle website.

**Dataset**:News\_Category\_Dataset\_v2.json

**url**: https://www.kaggle.com/rmisra/news-category-dataset

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from [HuffPost](https://www.huffingtonpost.com/). The model trained on this dataset could be used to identify tags for untracked news articles or to identify the type of language used in different news articles.

**Acknowledgements**

This dataset was collected from [HuffPost](https://www.huffingtonpost.com/).

So as per the above description the dataset is cleaned and so there is not much of work in getting it cleaned. The questions I am trying to answer from this data are:

* Can you categorize news articles based on their headlines and short descriptions?
* Do news articles from different categories have different writing styles?
* A classifier trained on this dataset could be used on a free text to identify the type of language being used.

So in the end I am trying to come up with a model that could categorize new articles if fed to the model.