Data Knyts Project Proposal

Anirban Banerjee (aba177), Ayush Raina (ara95), Tanmay Jain (tja34), Siddhartha Haldar (sha285)

Dataset(s)

Our project will be centered around the New York Times Articles & Comments (2020) dataset from Kaggle. It contains the following information (non-exhaustive):
News Articles (11 features, 16K rows): Category, Headline, Abstract, Keywords, Article Length, Length, Publication Date, Number of Comments
Comments (23 features, 5M rows): Username, Location, Text, Recommendations, Editor's Choice, Reply Count, Is a Reply

Some of our goals also involve using similar datasets containing New York Times article data such as New York Times Articles 1920-2020 and NYTimes Article Lead Paragraphs 1851-2017.

Proposed Goals

Our main goals will involve analysing user comments with regard to articles. This can be split into the following broad categories:

1) Basic popularity finder: ranking articles based on basic comment data
2) Basic recommendation system: recommend related articles to users based on their basic comment data
3) Predictive analysis: Use popularity analysis to find out which articles and categories will result in the most user engagement. This can be used on the other NYT article datasets (which do not have associated comment data) to predict their reception.

We will also be creating visualizations based on comments, such as their geographical location.

Stretch Goals

Iterating on our previous goals, we can create:

1. Advanced popularity finder: Analyse comment reactions, comment sentiments, comment engagement
2. Advanced recommendation system: recommend related articles to users based on advanced comment data (sentiment, frequency, geographical location, etc))

Technologies To Be Used

We will utilize PySpark, along with other Python libraries as needed (Numpy, Pandas, SciKit-Learn, PyTorch/TensorFlow for NLP, etc). We will likely use Amazon S3 and EMR to help in processing our dataset. We can add more technologies as our requirements grow.