Project Proposal Title: Evolution of NYT Headlines over the Last Century

Group Members:
Chris Kopacz (School: chk0833@g.harvard.edu Work: Christopher.Kopacz@dodiis.mil)

Background and Motivation:
The world (its cultures, peoples, successes, failures, etc.) is reflected through the reporting of major media companies. These companies take on as their profession observing and subsequently disseminating what is going on around us in a digestible and meaningful way. Through the use of natural language processing (NLP) and additional machine learning (ML) modeling techniques, the headlines of New York Times (NYT) articles over the last 100 years will be examined to see how the distribution of news reporting has evolved. What was important to readers over the years and how is it different to what is important to readers now?

Data:
There are two principle, publicly available datasets that are of interest to this project.

1) “Three Decades of New York Times Headlines”
https://www.kaggle.com/datasets/johnbandy/new-york-times-headlines
This is a robust data set that contains not only article headlines themselves (plus a few other columns), but also contains which “news desk” the article came from. News desks will, potentially, be used as the categories into which article headlines are grouped. This dataset covers the years 1990-2020 and will likely be used as the training set for an NLP+ML model that takes the article headlines as input (features) and uses the “news desk” as the classification output (labels).

2)”New York Times Articles 1920-2020”
https://www.kaggle.com/datasets/tumanovalexander/nyt-articles-data
This is a more concise dataset whose only relevant columns are publication year and headline, but it covers a much larger timeframe: 1920-2020. After training a model using the data from the first data source (Which includes the “news desk” column), listed above, the model will subsequently be used to categorize the article headlines in this second data set. Notably, this data set does not include the “news desk” column, so any classification made on these headlines will be representative of the strength/fitness of the trained ML model.

Lastly, NYT also provides a publicly available developer API for fetching news data and the articles it has published.
https://developer.nytimes.com/docs/articlesearch-product/1/overview
This API lists all possible values for “news desk” and additionally can be used (should the need arise) to fetch further data on NYT reporting.

Scope:
The first step in this process will be processing the NYT article headlines into a format useful to a ML model (i.e., transform the headlines into something that a computer can “read”). This will be accomplished by natural language processing. Specifically, the Sentence Transformers python package, plus the all-MiniLM-L6-v2 pretrained sentence transformer model (publicly available on huggingface: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2) will transform the NYT article headlines into numerical vectors (useable by computers) while retaining the semantic meaning of the original text. From there, ML will be implemented (either on a simpler scale such as random forests or decision trees – or a more complex scale involving a tensorflow/keras neural network) to classify the NYT headlines. Once classification is complete, exploratory data analysis will be conducted on the results to assess how society’s focus/interests have adapted over the last 100 years.