Can Visualizations be successfully recommended for a text document, even when the dataset is not directly associated with the article?

Abstract
In this project, we are investigating the efficacy of automatically recommending data visualizations to accompany a news story, even when there is no explicit data provided. To do this, we are creating a data pipeline that combines natural language and machine learning operations, including document summarization, keyword extraction, dataset search and retrieval, and visualization recommendation. Figure 1 shows an example of how an article is meant to run through the pipeline. A series of user studies is being conducted to fine tune the parameters of the pipeline, to evaluate how “best” to recommend data visualizations for text documents.

Research Goals
- Evaluate and validate the proposed pipeline to extract keywords from articles, search and identify relevant datasets, and recommend visualizations
- Assess the feasibility of this pipeline (Note the spring work involves keyword and dataset retrieval only)

Current Research Progress
- Note the spring work involves keyword extraction and dataset retrieval only
- NLP Algorithm Implementation: Have developed a module for extracting keywords from 3 NLP (Natural Language Processing) algorithms (NER, Gensim, Rake)
- Google Dataset Search: Have developed a module to run Google Dataset Searches on articles using 2 approaches. One method is using the top keyword result while the other method is called the blended approach where we used the top k keywords as a single query
- Dataset Collection: Used FiveThirtyEight’s database of articles that have predetermined datasets that are associated to each of the articles. The articles range in topics from Sports to General News to Politics in order to ensure that the system works on the broadest range of articles
- Statistical Analysis: Used the results of the 6 different types of searches (NER, Gensim and Rake with the 2 different search approaches) in order to determine the amount of times the correct FiveThirtyEight Dataset

Statistical Analysis
Chart 1 shows the results from the statistical analysis. Out of the 171 Articles in the dataset less than half of the time the correct dataset was found using the queries. Out of the 6 different NLP and query methods, it was found that the best results were produced by a Blended Gensim Approach with the correct articles mean search ranking being within the Top 3 searches.

Next Steps
Currently, we have generated text summaries (using the SMMRY library) for the 171 FiveThirtyEight articles in our dataset, and scraped metadata for the datasets retrieved via Google Dataset Search. These items will be used in subsequent user studies to further tailor the pipeline (User Studies 2 and 3).

Future Work
- User Study 2: Identify the optimal NLP algorithm (RAKE, GenSim, NER) that finds "relevant" articles, even if not technically "correct."
- User Study 3: Generate visualizations using NLP approaches/datasets from studies #2 using an existing visualization recommendation techniques and evaluate its usefulness.

References
- “An Overview of Graph-Based Keyword Extraction Methods and Approaches” : 2015 https://b.hrak.ir/ir/244867
- Data2Vis: Automatic Generation of Data Visualizations Using Sequence-To-Sequence Recurrent Neural Networks https://arxiv.org/abs/1907.05324

Chart 1: Results of the Statistical Analysis