**Title - News Makers Network**

**Team (names) - Glen Colletti, Peter Stewart, Shubkirti Prasad**

**Project Description**

* **Objectives 10pts: State as clearly as possible what you want to do. What problem do you solve etc?**
  + We intend to allow a user to quickly analyze a named person in the news, see what other people or entities they are connected with and the general sentiment of news about them. We will do this by accomplishing the following:
    - Identify named entities
    - Sort entities into people, places, things
    - Assess article sentiment
    - Graph network of articles and people with sentiment links
    - Provide tools for users to investigate relationships and articles
* **Usefulness 10pts: State as clearly as possible why your chosen application is useful. Make sure to answer the following questions: Are there any similar or equivalent applications out here?  If so, what are they and how is yours different? Which user group/stakeholders is your application targeting?**
  + Our application is useful because it enables a user to explore a named person’s representation in the news. This might be useful when researching politicians to vote for or someone who is under public scrutiny. The user will be able to see if news coverage trends positive or negative and what other people or entities they are related to.
  + Similar applications such as <https://ground.news/> provide tools for users to evaluate media bias and find articles about subjects under-represented by their preferred media outlets.
    - The focus of this application is the political perspective of the media organizations publishing and not on networking related entities.
  + Our application is targeting users who do not follow the news on a daily basis and may feel lost when a particular person or persons become the subject of a prolonged news cycle.
    - The subreddit r/outoftheloop frequently features posts to the effect of “Who is [person] and why does everyone feel [good, bad, otherwise] about them.
    - Example

A screenshot of a computer

Description automatically generated

* **Data 30pts:**
  + **Describe dataset origin (who collected, when, and for what purpose). Do not use data without any information provided. Do not use text-preprocessed clean data. What is data format: csv, json …**
    - The dataset was created by Hadas Unger, a kaggle contributor, by using a web crawler on CNN’s websites.
      * A github link was provided to the code used for the web crawler
      * https://github.com/hadasu/CNN\_web\_crawler
    - The data was collected by Hadas to practice data visualization and text classification and later provided to kaggle.com.
  + **Provide initial data description.**
    - **Besides a text field, what other information do you have (for example, review data -> location, stars, users; twitter data -> user, likes;...)**
      * Other than article text we have author, date of publication, category and sub-category of news (example business / investing or news / world), URL, headline, description, keywords, and alternative headline.
    - **do you have labeled data?**
      * Somewhat, we have news categories and keywords, we do not have specific NLP, or ML purpose labels.
    - **#records, #fields (columns if available), #NA values**
      * There are 9308 articles, with 11 features, including the text and a redundant index column.
      * An initial screening of the data did not find any NA values or other missing data entries.
    - **What type of cleaning does it require?**
      * Stopwords
      * Root / Lemmatization
      * NER

* **Functionalities 20pts:  Describe tentatively what tasks your application will perform. There are two types of functions you would need to offer:**
  + **NLP Functions: specific to your NLP tasks**
    - Preprocess data
      * We will preprocess the raw text of each article by removing stops words and either rooting or lemmatizing each word before vectorizing the text.
    - Named Entity Recognition and classification into people, places, and things.
      * + The preprocessed data will be analyzed for Named Entities using a variety of pre-trained models or pre-built dictionaries. Given that our dataset covers a wide variety of general interest news articles we do not feel that specific model tuning or dictionaries will be necessary.

The user will have some tools to compare the resulting entity lists and choose which one they like best

* + - Article clustering
      * Articles will be clustered based on co-mentions of named entities. The articles will then be analyzed for sentiment. We may also investigate document similarity as a criteria for clustering.
  + **User interaction: For example, allowing users to select/filter/search**
    - Present graph (graph in computer science context) to user of named entities in the news article corpus centered on a user selected entity.
      * Will be dynamically calculated and displayed
        + We are considering Tableau to implement this feature.
      * Each node will include articles which co-mention the user specified named entity and other named entities
      * Each node will summarize if articles in the node are mostly positive, somewhat positive, etc.
    - Allow user to select which model is used to generate entity list (Bert, NLTK, etc.)
      * Option will be presented as entity lists, processing will be done ahead of time.
  + **Note - your application should be focused on NLP methods**
* **Communication and Sharing 10pts: Set up your preferred communication methods (zoom, teams, discord ...) and provide a github repository link for the project:**
  + **create a Read.me file with the initial project description**
  + **upload dataset**
  + We have setup a github project, established a weekly meeting on zoom, and a whatsapp group chat
    - Github link
      * https://github.iu.edu/shubpras/Fall-2023--NLP-Group-3

**4. Personal Contribution Statement (10pts)**

* **team project: What was your contribution during the Planning stage (Part 1)**
* **individual project: Reflect on how you will manage time/tasks to complete milestones**