Project Plan

Author: Bagus Maulana

Project Title: Using NLP and sentiment analysis to analyse media representation of people with disabilities.

Supervisor's Name: Catherine Holloway, Nicholas Firth

Aims and Objectives:

Aims: To utilise computational natural-language processing, sentiment analysis, and machine learning methods to understand how individual news articles represent people with disabilities, and analyse a large collection of news articles at scale to reach meaningful conclusions towards specified themes (e.g. how different news sources represent people with disabilities)

Objectives:

1. Utilise public APIs and Python web scrapers to collect news articles
2. Utilise natural-language processing to understand which articles are relevant to a specific topic (i.e. a specific disability)
3. Utilise natural-language processing and sentiment analysis to gauge each individual article's sentiment towards a specific topic (i.e. a specific disability)
4. Utilise data-analytics to reach conclusions over specified research questions (e.g. how different news sources represent people with disabilities), based on data from Objective 2 and Objective 3.
5. Visualise results of findings from Objective 4.

Deliverables:

- Design specification:
  - List of research questions to be analysed
  - Input data required to analyse each theme
  - High-level overview of process to collect information from online sources, filter only relevant articles, extract input data
- Results of the experiment
  - Raw data: articles, topic (which disability), sentiment, source, datetime
  - Visual data based on several research questions:
    - How different data sources (e.g. different newspapers) represent people with disabilities
    - How events (e.g. campaigns, policy changes) affect media coverage of people with disabilities (quantity of articles and sentiment)
      - How long does the effect last, if any?
      - Compare positive events (e.g. campaigns) vs negative events (e.g. budget cuts)
    - Other research questions to be discovered during the 'work through iterations' stage
  - Statistical analysis / discussion of statistical significance of results (if possible)
- Project plan
- Interim, final reports
- Python scripts, used for:

    - Data collection (scraping news articles about a certain topic from various web sources, filtering for relevant articles), processing (e.g. tokenisation)
    - Natural language processing (determining key nouns, adjectives to these nouns, etc. from a text article, given a topic)
        - Scripts to generate features for data analytics (e.g. n-gram counts)
    - Sentiment analysis (determining how a given article represents a specified topic (people with a certain disability))
    - Time-series data analytics, visualisation
    - Code to be functional and properly commented/documented.
        - Does not need to reimplement everything, mainly concerned about how existing Python APIs (e.g. scikit-learn, spacy) are used for this project

Work Plan

- Project start to early November (4 weeks):
    - Literature search and review
    - Start defining requirements (disabilities and events to search for relevant articles, research questions to analyse).
    - Create project plan
- Early to end of November (4 weeks):
    - Refine your requirements
    - Start the initial iteration(s) for data collection (gather articles with a certain search query, filter to only relevant articles).
- Early December to early March (3 months): Work through the iterations.
    - Data collection: Gather relevant articles to specific disabilities, relevant events, etc.
    - Natural language processing: Extract relevant information from relevant articles
        - Extract references to specific disabilities and/or people with disabilities from the article text (nouns/phrases)
        - Extract verbs and adjectives 'attached' to these nouns/phrases related to certain disabilities and/or people with disabilities
    - Sentiment analysis: Model how positively/negatively does the article text refer to specific disabilities and/or people with disabilities (based on 'attached' verbs and adjectives, and the rest of the sentence)
    - Data analytics and visualisation: Process and visualise time-series, sentiment analysis, and other data extracted from article text based on several research questions (e.g. how events impact quantity and sentiment of articles related to a disability)
- Mid to end of January (3 weeks): Work on interim report.
    - Based on available information
        - Introduction, background research, requirements, and data processing / research method should be completed by this time
        - Results, evaluation, and conclusion to be appended in final report
- Mid-March to end of April (6 weeks):
    - Write conclusion based on data and experiments
    - Work on final report