CS 410 Text Information Systems 2021 Fall - Project Proposal
Causal Topic Modeling

Team Information

- Team Name: MP
- Team Members: Masami Peak (NetID: masamip2, Email: masamip2@illinois.edu) - Captain

Topic Information

Description:

This project is building a causal topic model for identifying hidden key topics in the MLB (Major League Baseball) articles correlated to the annual MVP (Most Valuable Player) winner.

Tasks:

1. **Obtaining Datasets**: Preparing a python script to perform web crawling on explicitly defined URLs, such as Reuters, MLB, Wall Street Journals, NY Times and ESPN, to fetch MLB articles published in the month between April and October of the year between 2011 and 2021.
2. **Data Cleaning and Preprocessing**: Cleaning the dataset to fix any data issues and transforming it to well-formed dataset for topic modeling.
3. **Dictionary and Corpus Creation**: Creating a dictionary and a corpus (TermID, Frequency) for a topic model.
4. **Topic Model Building and Evaluation**: Building a topic model LDA (Latent Dirichlet Allocation) and evaluate the quality of the model.
5. **Topic Model Visualization and Analysis**: Visualizing and analyzing the model for the project report.

Interest/Motivation:

At the end of the MLB season, each MVP in the 2 leagues is determined by the voters in the Baseball Writers' Association of America and the MVP need not come from a division winner or other playoff qualifier. Due of this, topic modeling on the MLB articles published before the MVP announcement can be used to discover the trend topics correlated to the MVP for the year. Additionally, this technique can be applied for analyzing any other sports awards and the Academy Awards (the Oscar).

Approach:

MLB articles as documents can be retrieved from MLB related webpages. The most popular topic model LDA extracts the topics discussed in the documents. Visualizing top 30 relevant terms to each topic distributed over documents interprets the topic.

Tools/Systems/Datasets:

On Python environment in a local machine, a Python script performs web crawling for obtaining MLB article dataset with id, headline, summary, created and source columns. Then, the rest of the listed tasks are completed in Jupyter Notebook while the project report is prepared.

Expected Outcome:

This project is an experiment, but it is interesting to find out if there are hidden trend topics from one year to the next.

Evaluation:

In addition to human judgement, topic model can be measured by computing coherence score which indicates the interpretability between the topics and the relevant words inferred by the model.

Development Environment Information

Programming Language: Python using Jupyter Notebook
Main Modules: bs4 for BeautifulSoup, selenium for webdriver, numpy, pandas, nltk, gensim, pyLDAvis, matplotlib for pyplot

Workload Information (Total: 30 hours)

1. **Obtaining Datasets**: 4 hours
2. **Data Cleaning and Preprocessing**: 4 hours
3. **Dictionary and Corpus Creation**: 3 hours
4. **Topic Model Building and Evaluation**: 4 hours
5. **Topic Model Visualization and Analysis**: 15 hours