CS 410 Text Information Systems 2021 Fall - Project Progress
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.

Completed Tasks:

1. **Obtaining Datasets**: Prepared 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**: Cleaned the dataset to fix any data issues and transforming it to well-formed dataset for topic modeling.
3. **Dictionary and Corpus Creation**: Created a dictionary and a corpus (TermID, Frequency) for a topic model.
4. **Topic Model Building and Evaluation**: Built a topic model LDA (Latent Dirichlet Allocation) and evaluated the quality of the model.

Pending Tasks:

1. **Topic Model Visualization and Analysis**: Visualizing and analyzing the model for the project report.
2. **Software Usage Tutorial Presentation**: Documenting and presenting software usage/implementation.

Challenges:

1. **Web Crawling**: Some sites (e.g. mlb.com) had 2 popups that interrupt web crawling. The time those popups appear did not seem consistent. I needed to set the time by which the popups should have shown up (e.g. 30 seconds) to handle the interruption. Also, some advertisements seemed interfering web crawling, so I needed to run the script multiple times to retrieve the articles on the site.
2. **Data Cleaning and Preprocessing**: Some manual inspections were needed to build a better vocabulary for a topic model. For example, a part of a sentence ‘Buffalo, N.Y., the’ would become ‘buffalo n y the’ after lowercasing and replacing any special character with a whitespace were applied. To achieve the ideal result: ‘buffalo ny the’, more preprocessing was required to apply.
3. **Stop-words Removal, Lemmatization and Stemming**: The order of applying stop-words removal, lemmatization (grouping together the inflected forms of a word) and stemming (reducing the derived forms of a word) gave slightly different effects on the vocabulary creation. For example. if a word ‘player’ is one of the stop-words, then stop-words removal is applied first and lemmatization is applied second, the word ‘player’ can be still found in text because lemmatization has transformed ‘players’ (plural) to ‘player’ (singular). I needed to try different orders of applying the document processes and different types of lemmatization to see the best result.
4. **Model’s Hyperparameter Tuning**: Configuring the topic model’s hyperparameters was done by building the model several times while the performance of the outcome was observed. The main topic “Baseball” is already known and analyzing the outcome of the optimized LDA model for finding the meaning of the hidden topics is somewhat challenging.