Statistical NLP Report

Young Sang Kwon

**1. Introduction**

“Statistical NLP” integrates NLP(natural language processing) techniques to suggest highly relevant articles to users. A unique blend of personalized recommendations and content diversity.

**2. User Interface (UI) Overview**

* Dataset Selection: Users start by choosing between two datasets: 'BBC News Articles' (bbc\_news.csv) and 'Wikipedia Sample' (wikipedia\_sample.json).
* Article Recommendations Display: A list of 102 article recommendations is displayed, each identifiable by a number and title.
* Detailed Article View: Users can view the full content of an article by inputting its corresponding number.
* Continuous Interaction: The program continually offers fresh recommendations based on user selections, enhancing the discovery experience.
* Exit Option: Exiting the program is made simple by entering 'q' at any interaction point.

**3. Test method**

1) Copy “statisticalNLP.py” and data folder to the appropriate directory.

(data folder contains "bbc\_news.csv" and "wikipedia\_sample.json")

2) Run the program from that directory. (Type “python StatisticsNLP.py” at the prompt and press Enter)

3) Use the number keys to select the desired dataset.

(If you select incorrectly, the program will terminate)

4) When selected, a list of 100 items is randomly displayed.

5) Select the article you want

(If you select a wrong number, the list will reappear, press 'q' to exit the program)

6) The selected article content is displayed.

7) Click ‘Enter’ and you'll get 100 recommendations of similar articles to keep things relevant, plus an additional 2 diverse articles to introduce diversity and expand user exposure.

**4. Main Function**

* **load\_articles**: Loads articles from the selected dataset, supporting both CSV and JSON formats, thus accommodating different data sources.
* **vectorize\_documents**: Leverages TfidfVectorizer to transform articles into numerical vectors, enabling the application of similarity measures.
* **calculate\_similarity**: Employs cosine similarity to determine the closeness between article vectors, which is crucial for recommendation logic.
* **generate\_recommendations**: A pivotal function that generates a recommendation list based on the user's last viewed article. It intelligently provides 100 recommendations of similar articles to maintain relevance and includes 2 additional diverse articles to introduce variety and broaden user exposure.
* **display\_recommendations and display\_article:** These functions are integral to the UI, showcasing both the list of recommended articles and the full text of a selected article.