Text Classification Summary Report

Tools or Frameworks Used

- Web Scraping: A combination of BeautifulSoup and Selenium is used for scraping article titles from Medium.com. BeautifulSoup is utilized for parsing the HTML structure, while Selenium is employed to interact with the dynamic webpages of Medium.
    - An additional tool, chromedriver, is downloaded and used during scraping. It is mainly for Selenium to control the Chrome browser and perform the scrolling automation.
- Named Entity Recognition (NER): Industrial tool spaCy and NLTK are both used here.
- Sentiment Analysis: TextBlob is used here.
- Topic Modeling: Gensim and scikit-learn are both used to perform LDA and NMF here.

Results Obtained

- Web Scraping: 100 article titles without topic specification on medium.com are acquired.
- Named Entity Recognition (NER): SpaCy recognized about 10+ entities with different frequencies, with ORG and CARDINAL being more prevalent, while NLTK recognized only 4 kinds of entities, PERSON, and ORGANIZATION dominating the entity frequency.
- Sentiment Analysis: Polarity and Subjectivity of titles are gathered through the analysis; the result shows a quite prevalent spread across polarity while about half of the articles are neutral.
- Topic Modeling: 5 clusters of the topic are acquired with both LDA and NMF

Analysis/Reflection on the Results

- Efficiency of Web Scraping: The use of BeautifulSoup accompanied by Selenium suggests effective scraping of dynamic content, which is crucial for gathering a robust dataset. Medium.com tends to change its structure after a driver opens its webpage a few times, making it harder to scrape down the content desired. Hence, restarting the kernel occasionally can solve this problem effectively.
- Named Entity Recognition: Not all input data got entity successfully detected by either method, but SpaCy is definitely doing a better job here than NLTK since it can detect more diverse and detailed entities. Although it may be due to the length or the richness of the input text, some of the recognized entities are not accurate, spaCy is providing us with a better outlook of the different categories being discussed among all the titles.
- Sentiment Analysis: The final analysis was not what I was expecting, it is more prevalent than I thought in terms of polarity and subjectivity. I was initially thinking that most of them should be close to neutral, but it turns out there are more article titles with stronger sentences. I guess this is the way authors trying to attract people to read their articles and to speak louder about what they care about.
- Topic Modeling: We can see some groups of topics having eye-distinguishable differences, this may relate to the result from sentiment analysis where the words used for different polarity are different and thus categorized into different topics. However, it may also be due to the length and the richness of the input text; the categorized topics of similar sentiment are not easy to tell apart. I think the result can be related to the result of NER,

when some of the entities are wrongly recognized, it is easy to see the topic modeling being ambiguous.

Visualizations

To enhance the interpretability of these results, several visualizations could be helpful:

- Entity Distribution: A bar chart is shown for results from spaCy and NLTK.
- Sentiment Distribution: A scatter plot is chosen to represent the overall sentiment for all 100 articles. The x-axis represents polarity, ranging from -1 to 1, -1 represents super negative 1 represents super positive, and the y-axis represents subjectivity, ranging from 0 to 1 with 0 being total objective and 1 being total subjective.

Notebook Repo:

https://github.com/Yi-HsuehYang/94812_Assignment1.git