Project Documentation CS-410 / Fall 2021
JunyangWang

An overview of the function of the code

Code can be used to do an aspect based sentiment analysis. As seen in the code, we first tokenize all reviews. Then extract bigrams NN-ADJ pairs to form a word cloud and visualize the features that stand out the most. We also extract unigrams that are NN as the aspects to use for sentiment analysis. Note that in order to train our classifier, we use the overall column from the airline dataset and then label our aspects with a pos, neu, neg sentiment. We then visualize the aspects and the associated sentiments using a bar plot. we use the sentiment classifier trained before to classify the sentiments of the aspects extracted.

Software Implementation and Usage

1. download GitHub repository
2. change cmd directory to code folder
3. pip install -r requirements.txt
4. Start Jupyter-notebook from shell using command: jupyter notebook
4. open *project_code.ipynb* and preprocessing_code.ipynb in the code folder
5. Change the file path to where the preprocessed files are.

GitHub file structure

```
.
├── code                                        # project code files
|   ├── preprocessing_code.ipynb                 # project preprocessing code
|   ├── project_code.ipynb                       # project code
|   ├── requirements.txt                         # lib requirement
|   ├── sentiment_analyzer.joblib                # sentiment classifier dump
|   └── vectorizer.pickle                        # vectorizer dump
├── data                                        # data files
|   ├── dataset                                 # dataset files
|   |      ├── airline.csv                      # dataset2
|   |      └── capstone_airline_reviews3.xlsx    # dataset1
|   ├── airline_reviews_preprocessed.csv         # preprocessed dataset1, for training sentiment classifier
|   └── airline_reviews_preprocessed_dataset2.csv   # preprocessed dataset1, for evaluate sentiment classifier
├── Progress Report.pdf
├── Project Proposal.pdf
├── Project Documentation.pdf
└── README.md
```

Note for project classifier testing

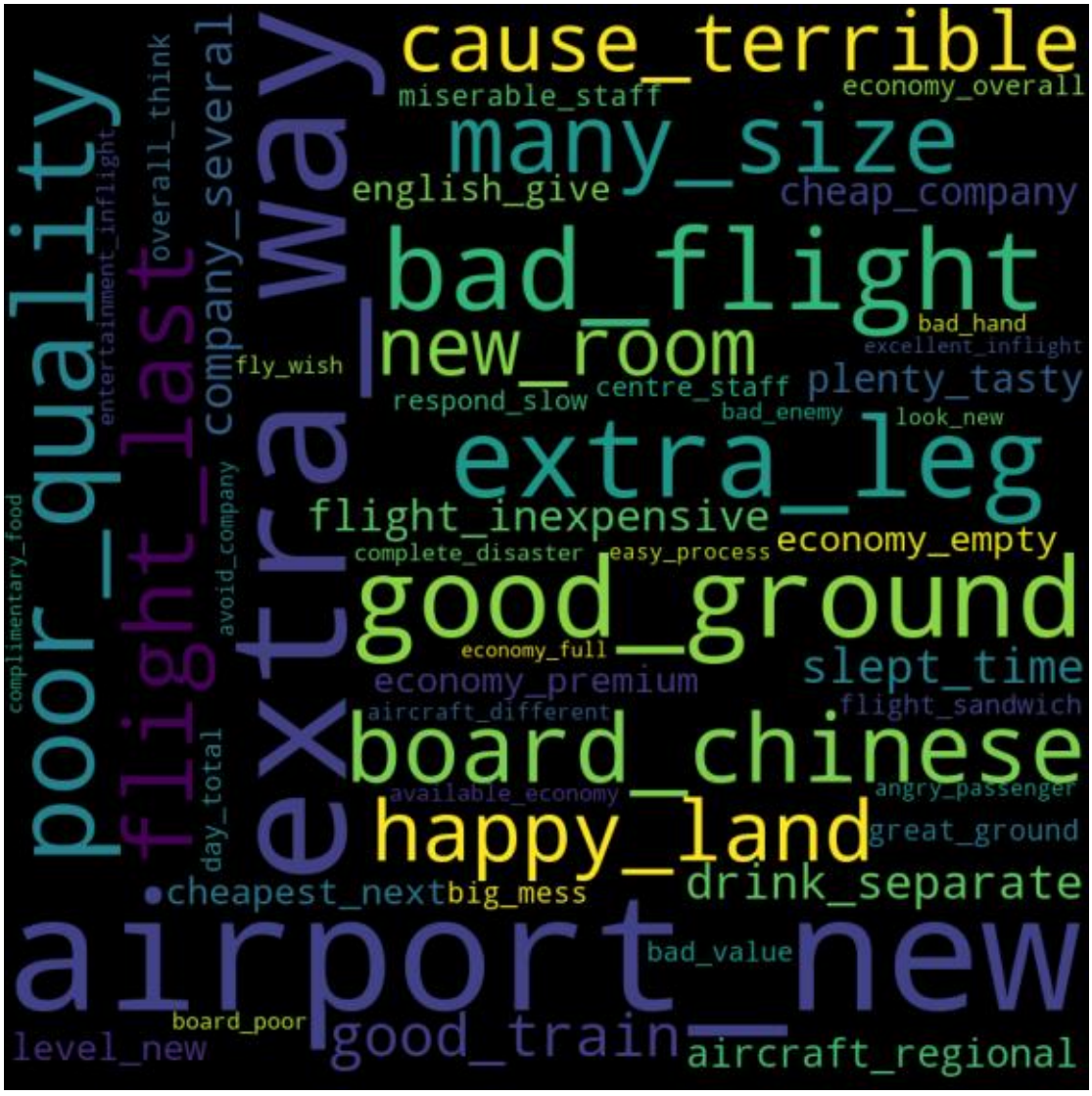
There is a joblib file that you can use to test the sentiment classifier. The classifier has been trained on capstone_airline_reviews3 dataset. Check Step

5 under Airline Review Analysis in preprocessing_code.ipynb, or test the sentiment classifier in Step 4.b in project_code.ipynb. I have already shown how to use it in the video presentation.

Entire Code including the preprocessing and training sentiment classifier can be found in **preprocessing_code.ipynb.**

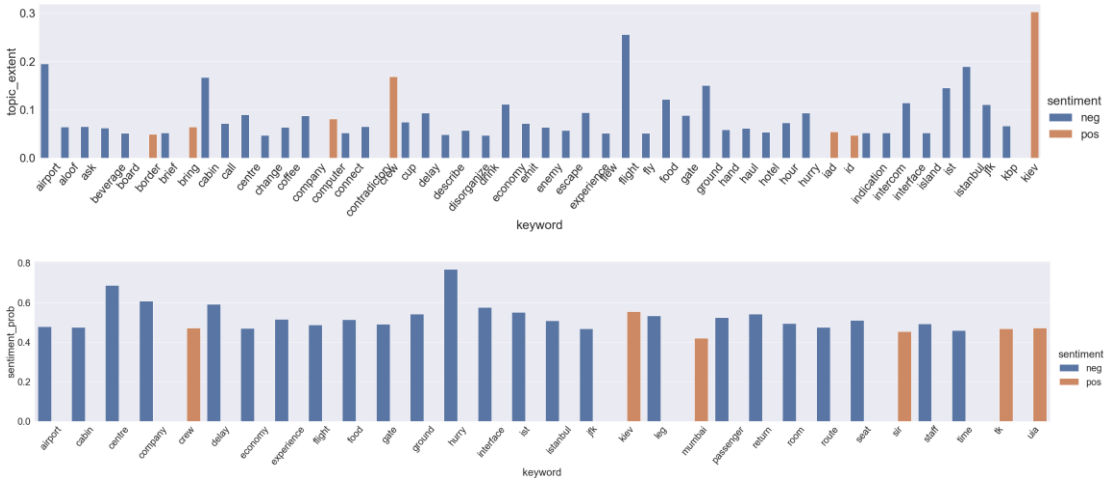
Modified Code for Testing can be found in **project_code.ipynb**

Final Results Understanding Plots and Graphs
[Airline Review Bigram WordCloud]



Use WordCloud to visualize bi-grams. NN-ADJ pairs are extracted from reviews and TF-IDF is used to retrieve top n bigrams. There will also be a bar-plot associating the sentiment with every unigram NN keyword extracted with the probability of the sentiment.

[Airline Review Aspect Sentiment Graph]



In the Wordcloud we can observe that people tend to talk about the aspects of cabin, flight and airport. As for the bar plot, we can see that airline have “experience”, “food” and “cabin” aspects that have been associated with negative sentiment. Also the highest positive sentiment is observed among aspects like “crew” and “border”.

What is completed and what could be better?

I have successfully been able to analyze the aspects that drives people to choose different airline. However, using LDA for bi-grams does not work well but specialised algorithms for Bi-gram topic extraction can be used in future. LDA for unigrams also did not group the categories very well, but top weighted words could have been considered. Therefore, I used TF-IDF to find the key aspects and only used nouns to do so.

Video presentation link:
https://drive.google.com/file/d/1B9AbRe3bDtQ6OBJ8PLIjWzteyrNSP8Y0/view?usp=sharing