Lan Mei

BAX 452 HW Assignment #9 [HW#9 NLP_APIs.ipynb]

Overview

In this project, I used natural language processing(NLP) tools from Amazon Web Services(AWS) and Google Cloud Platform(GCP) to analyze the sentiment based on a dataset of 100K Coursera's Course Reviews.The original dataset: [100K Coursera's Course Reviews Dataset]

Steps of Analysis

- **Data Ingest**

  Original dataset is like this, which is clean to use:

Id	Review	Label
0	good and interesting	5
1	This class is very helpful to me. Currently, I...	5
2	like!Prof and TAs are helpful and the discussi...	5

- **AWS Configuration**

  In this process, python packages “aws” and “boto3” are used. After the configuration of credentials and config, I used the “comprehend” service under boto3.client, which is used for natural language processing. The syntax is:

  comprehend.detect_sentiment(Text=row, LanguageCode='en')

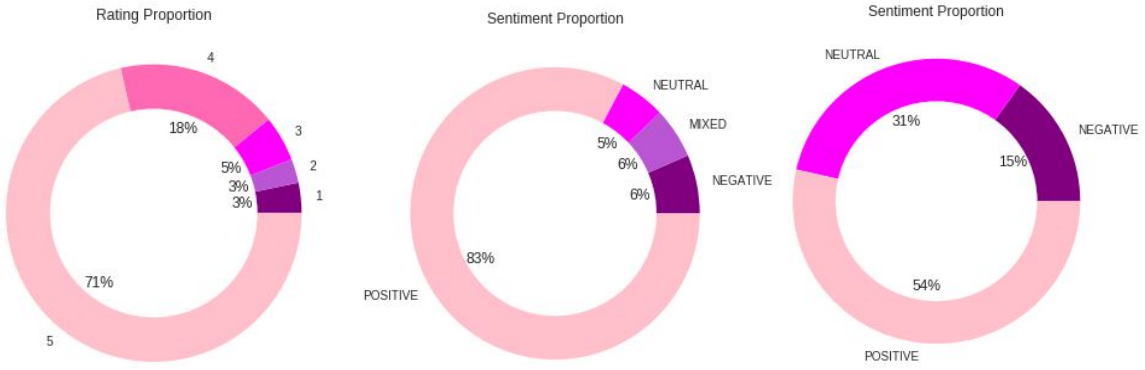
- **GCP Configuration**

  In this process, language, enums and types from google.cloud are used. A json file with the credentials is used to connect to the server. The syntax used to analyze is:

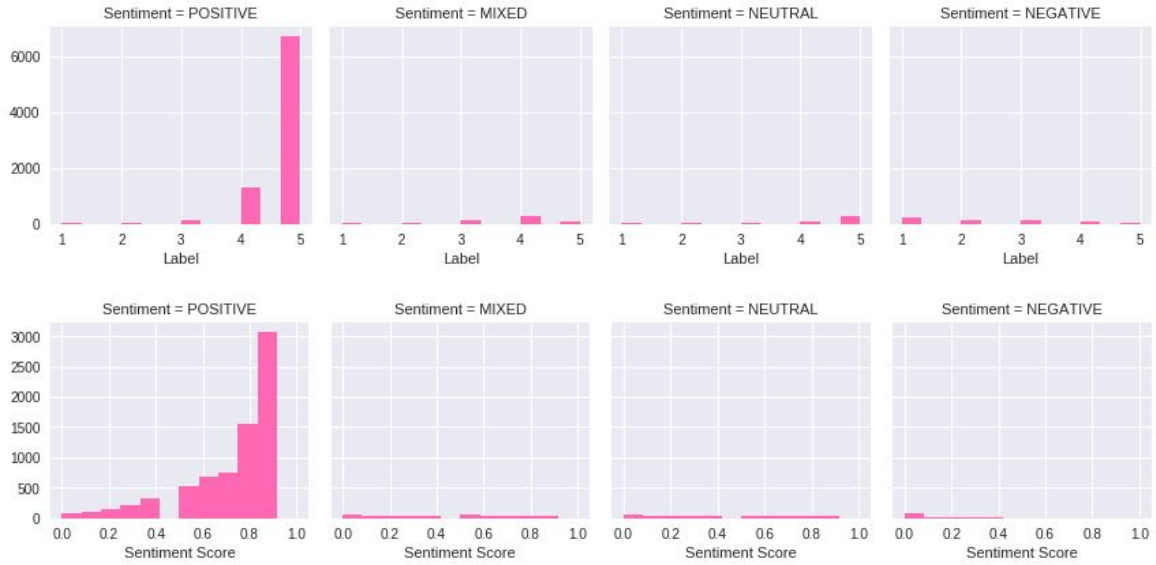
  client.analyze_sentiment(document=document).document_sentiment.score

  Sentiment score ranges from 0 to 1, the larger the score, the more positive sentiment.

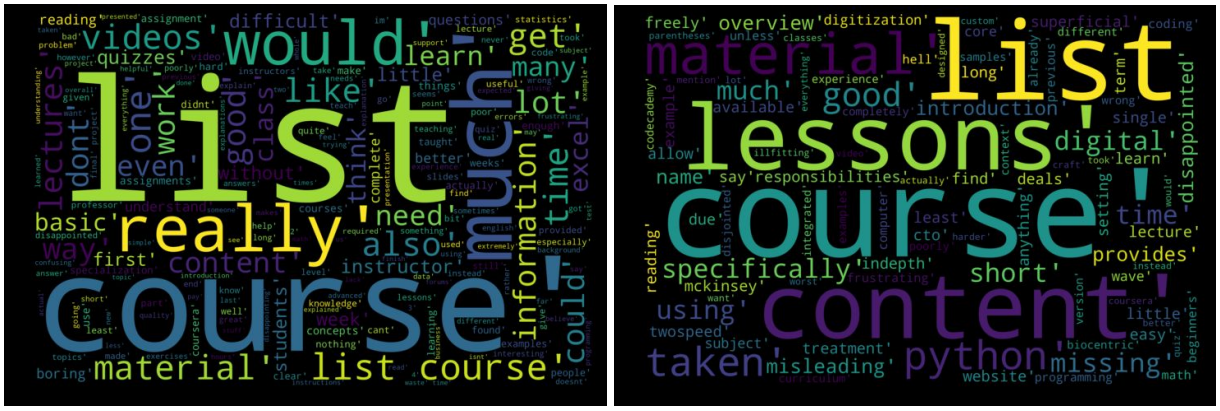
- **EDA based on the result of AWS and GCP**



  Comparing the proportion of rating, sentiment(AWS) and sentiment(GCP, categoriezed with 0.3 and 0.7 as the threshold of positive and negative), we can find that the result of AWS is basically more similar, with “positive” sentiment in accordance with rating “5” and “4”. Therefore, maybe the threshold of 0.7 is a little high for the result of GCP.



This is to analyze the ratings and sentiment score(GCP) distribution based on the sentiment category(AWS). Generally, we can infer from the chart that the sentiment(AWS) is accurately categorized, with more high scores in “POSITIVE” and more low scores in “NEGATIVE”.



Comparing the keywords of negative sentiment from AWS(left) and GCP(right), we can find that students on Coursera will give lower rating to courses related to “difficult”, “boring”, “python” and “misleading”.

Conclusion

- The NLP tools in AWS and GCP are convenient and accurate, and the tool of GCP offered customized definition on the threshold of sentiment.
- Coursera should provide necessary assistance on the courses that are more difficult or related to python. Moreover, Coursera should look into the courses with reviews of “boring” and “misleading”.