Project – Natural Language Processing
Problem Statement

This project is divided into two parts. In the first part you are required to do sentiment analysis on customer review data of a product. And the second part is based on text analytics on data containing complaints received in a banking firm.

Part-I (15 points)

Think of yourself as the product manager for Google Home. Your team has come up with a new product Google Home Mini which has been launched in the market. The product is sold via Flipkart website where customers buy your product and give their review about the product on the web site.

As a product manager, your role is to identify the reviews from the website and run a sentiment analysis tool to understand what the sentiments of the customer are. Based on their sentiments, your first task is to identify what users think of the current product. Are they happy? Discontent?

The second task would be to come up with the task document which lists what your team needs to focus on for making the product better.

Real World

Real world NLP problems are way complex and require data scientists deep NLP expertise to solve the problem justifiably. However, the above project, though a shorter version of how data scientists solve NLP projects in the real world, will give you a huge insight into what are all the steps involved in solving a classical sentiment analysis problem.

By learning these steps hands-on, you would be better equipped as a leader to lead a team of data scientists whose job is to come up with a sentiment analysis algorithm.

Data Collection using Web Scraping (Optional):-

- Use the web-scraping tool (Parsehub) to scrape the comments out of your product website.
- Link to download ParseHub - https://www.parsehub.com/
- Go through the provided video to understand how the web scraping tool works in extracting the comments from the website.
- Below is the product URL from the Flipkart website. (Kindly click on “read more” using “Select Mode” of ParseHub if the entire comment is not getting selected.)

https://www.flipkart.com/google-home-mini/p/itmf3xz9exnjbzzm

- Using the web-scraping tool (Parsehub) extract the first 20 reviews from the customers about the product. Upload this .csv file as a submission.

Graded Steps:-

1. You are given the extracted csv having 20 reviews of the above product. Analyse the first 10 reviews independently using Amazon Comprehend and answer the below questions.
    a. Report the sentiment and its confidence percentage for each comment. (5 points)
    b. Report the overall summary of the 10 comments. The summary should contain the percentage of negative, positive, mixed, and neutral comments. (2 points)
    c. Comment whether the majority of comments are positive or negative? (1 points)
2. Analyse all the negative & mixed comments from the scraped reviews and answer the below questions.
    a. What are the common key phrases in these comments? Report at least 10 key words. (1 points)
    b. Report part of speech of this first 10 words of key phrases of the previous comment.(1 points)
    c. List down the common problems or areas of improvement in the product, by analysing the key phrases of negative comments. (3 points)

d. List down suggestions to improve the product which will result in better customer satisfaction. (2 points)

Part-II (15 points)- AZURE ML

Complaint Categorization Baseline Model

Fast and efficient handling of complaints on consumer forums is vital to the e-commerce industry today. The dataset contains consumer complaints on financial products.

Dataset Description:-

- Consumer complaint narrative:- complaint of the consumer
- Product:- Product in reference to
- len:- length of the comment

1. Upload the provided dataset in AzureML and convert the “product” feature into categorical data.
2. Visualize the “len” column and report min and max values of the length. Comment on the distribution of the “len” feature. (1 points)
3. How many unique products are there? Which product has the highest number of complaints? (1 points)
4. It is said that the average length of comments related to the “debt collection” product is comparatively lower than other products. Is this statement true? Provide necessary evidence to prove/disprove this claim. (2 points)
5. Check for null values and drop them if there are any. (1 point)

Text Analytics using Azure ML:-

6. Preprocess text data. Remove punctuations, stop words, lemmatization, etc. (2 points)
7. Extract N-gram features from the above preprocessed text and report the final number

of columns. Report the size of the resulting vocabulary. (3 points)

Product class prediction using multi-class neural network:-

8. Select variables of interest for the model building process. (1 point)
9. Build a neural network model to predict the product name. (2 points)
10. Evaluate the model and report the accuracy. (1 point)
11. Write your observations and findings. (1 point)