WQD7001 - Alternative Assessment 1

Name: Kar Hong Sam

Matric No: S2191926

Question 1:

Explain why data cleaning is important in text analytics.

Answer:

The data cleaning process is important in text analytics because the raw data is unstructured text. This preprocessing method ensures that the unstructured text is converted into a machine-understandable, readable text structure. The process is necessary before feeding the data into algorithms as input, while also ensuring the data's quality.

Question 2:

Recent studies have been focusing on the increasing relevance of the social networks in the public image of politicians and political parties. Several of them have been focusing on the opinion on candidates to presidency and on how this opinion changes. You are planning to develop an application accessing a set of tweets containing the name of a specific politician and for each tweet determine if it is about the politician and if it is positive or negative. The application input will be the name of a politician, and the output will be the tweets referred to the politician, indicating for each of them, if the politician polarity is positive or negative.

a. Discuss about the application described in the scenario.

   Answer:

   The application described in the scenario is sentiment analysis that is able to analyze the opinion from tweets and classify it as either positive or negative. This allows the public to know about the politicians' performances based on the sentimental results.

b. Discuss any challenges that you might face to find the name of a specific politician.

   Answer:

   One of the difficulties that may arise is the presence of two or more politicians with identical names. This can lead to ambiguity when searching for tweets about a specific politician, resulting in inaccurate search results. Furthermore, dealing with unique names that could be similar to role-playing words or slang might also lead to misleading search results. Other issues may arise as a result of misspellings, nicknames, or differences in the depiction of the politician's name across different tweets.

c. Tweets also present several difficulties when applying natural language tools. Describe any issues or problems (at least 3) related to processing of tweets using relevant examples.

Answer:

One of the issues will be detecting sarcasm or ironic tweets, the NLP tools hardly can tell whether the tweet is positive or negative due to the key phrase that looks positive but it's not. The next issue will be the ambiguity of text, that can be interpreted in different ways like 'I saw a sick bird' that perceived positive. At the same time, this will make the NLP tools misinterpret the original meaning from 'sick' word. Lastly, the other issue will be the new slang or acronyms like 'ICYMI' or 'In Case You Missed It' that are not available from the existing corpus.

d. Discuss how to address the issues described in Question 2(c).

Answer:

Possible solutions to address the issues described will be:

1. Train the NLP model with certain cues that appear to be sarcastic phrases like 'yeah, right', 'whatever', 'I'm fine' and so on. In return, the retrained model can recognize the sarcastic phrases.

2. Part-of-Speech tagging can help to curb the pain points of ambiguity. This technique helps disambiguate the meaning of words by analyzing the grammar structure and the relationship within the tweets.

3. Regularly feeding new text data as input and retraining the NLP model so it can be up to date on the latest trends, especially on social media platforms.