Week 10: Deliverables

Group Name: Awwab Ahmed
Name: Awwab Ahmed
Email: awwab.k.ahmed@gmail.com
Country: USA
College/Company: University of South Florida
Specialization: NLP

Problem description:

Computer Science is a growing field and online computer courses are sold by many different vendors and companies. Because of this companies which sell courses have great competition and small differences in the courses or their structure can yield large returns in profit.

My team's job is to analyze reviews of different top CS courses through NLP techniques such as sentiment analysis, summarization, and NER to present insights into my company that can help them design the ideal CS course that will yield the most profit.

Data cleansing and transformation:

1) Getting rid of all lowercase:
    a) This was done using the lower() method.
    b) This is needed as whether a character is lowercase or uppercase has no effect on whether it is hate speech or not.
    c) Therefore, normalizing this leads to better model results.
2) Getting rid of characters that are not alphanumeric:
    a) This was done using python's isalnum() method.
    b) Hate speech only comes in the form of words and not characters therefore they have no effect on the classification of a tweet.
3) Lemmatization:
    a) This is an NLP technique that uses the ideology of stemming to better model results.
4) Tokenization:
    a) This is done during preprocessing in model development.

EDA:

As this is a natural language processing project, the EDA comes in different forms compared to regular data science projects.

This project will explore chars per sentence, words per sentence, and obviously, sentiment analysis.

Each tweet was about 100 characters.
This was about 10-15 words per tweet.

Next step is to carry out sentiment analysis and build model.