Assignment 3- Advanced Natural language processing

Purpose: This class is a continuation of the first three lessons on text analytics, text classification and introduction to natural language processing. We will be combing the first three lessons in this assignment.

This assignment provides you with an opportunity to demonstrate the achievement of the following course learning outcomes:

- Understand how to extract information from text
- Understand how to analyze sentence structure (context-free grammar)
- Build grammars
- Perform classification with text data

Key Information

- Type: *Individual*
- Weight: 5%
- Delivery: Course website upload
- Due Date: End of lab session

Expectations

You are expected to complete this assignment individually.

Respect for academic integrity is crucial to your success. Make sure you understand what constitutes acts of academic dishonesty in the page: What is Academic Dishonesty?

Instructions

1. The goal of this assignment is to apply NLP techniques to a set of IMDB movie reviews to get a prediction of whether or not a movie will be “Positive ” or “Negative”
2. Download the necessary files (provided by the instructor) to your local file system
3. Your datasets should contain a list of positive tweets and a list for negative tweets.
4. Now use Python’s built-in re package to remove punctuation and numbers, by using the re.sub function, this is a built-in Python function.
5. Now, convert all the reviews to lower case, and perform a tokenization, to split them into individual words.

6. Now use the nltk function to remove stop words (words in the English language that don't have much meaning individually such as: "a", "is", "the", etc...) by making use of **from nltk.corpus import stopwords**, and then stopwords.words("english")

7. Now combine all of the aforementioned steps into one function so that you can apply this function to each of the 25,000 IMDB reviews.

8. Now, use the scikit-learn function CountVectorizer to create bag-of-words features.

9. Use the bag-of-words created in the previous step, print a count of each word in the vocabulary.

10. Use a machine learning algorithm (SVM, Naive Bayes) of your choice to now create a supervised learning model to predict the sentiment of a given movie review using the test datasets.

Rubric

To achieve full marks on this assignment, you must have answered all questions above correctly with code submitted that has no errors.

Natural Language Processing
Course Code: BDA 102