Week 10 Assignment - MSc AIDA UoM

Professor Yannis Refanidis

Academic Year: 2021-2022

For this assignment, you will need to use Python's NLTK library. You can consult the following book,`https://www.nltk.org/book/`, where you will find many relevant examples. The assignment has two questions:

1. Using Python's NLTK library, load 10 books of your choice from the Project Gutenberg corpus. Create a dictionary of all the words that appear in the books (use a tokenizer of your choice; NLTK offers several).

   Divide the texts into sentences (sentence tokenizer) and into tokens. Calculate the frequencies of unigrams, bigrams and trigrams. For bigrams, you will need to define a token for the beginning of a sentence, as well as a token for the end of a sentence. For trigrams, you will need to define two tokens for the beginning of a sentence and one token for the end of a sentence (no second token is needed for the end of the sentence).

   Generate and report in your assignment 10 sentences using: a) Bigrams b) Trigrams

2. From Python's NLTK library, load the movie_reviews dataset, which includes 2,000 movie reviews, of which 1,000 are positive (pos) and the other 1,000 are negative (neg).

   Train a classifier (e.g., Naïve Bayes classifier, neural network, etc.) to correctly classify the positive and negative reviews. Choose which features you will use as input (e.g., individual words, ngrams, etc.).