Homework Assignment-2

Start Assignment

Due Friday by 11:59pm **Points** 40 **Submitting** a text entry box or a file upload
Available Oct 12 at 5pm - Oct 15 at 11:59pm 3 days

Part 1 - 20

(10) Load the ISEAR corpus using Python. Find maximum, minimum, and average length of sentences for each emotion type. Output this to a CSV file so that each entry are tab separated.

Emotion Name Max-length Min-length Avg-length [Use \t as delimiter]

Also separate rows with a space (i.e. rows to be double spaced)

(10) For each type of emotion, calculate the vocabulary size and analyze the distribution of tokens. Experiment using NLTK's tokenizer and tokenizer from Spacy and document the difference in a text file.

Original data: **https://www.unige.ch/cisa/research/materials-and-online-research/research-material/**
(https://www.unige.ch/cisa/research/materials-and-online-research/research-material/)

Data in CSV format: https://github.com/PoorvaRane/Emotion-Detector/blob/master/ISEAR.csv

Special Notes:

Ignore rows with more columns than expected.

Round the average length to 2 decimal places.

Sort tokens by frequency.

Limit the number of the most frequent tokens to 50.

Remove tokens containing non-alphabetic characters.

Remove stop words.

Part -2 - 8

Import Sem-eval dataset (https://www.kaggle.com/azzouza2018/semevaldatadets) using Python.

Data file to use for assignment: semeval-2017-train.csv

(4) - Print the sentences which contains URLs and hashtags

(1) - Print all neutral sentences (annotated with 0)

(3) - Find the distribution of positive, negative and neutral sentences

Part-3 - 7

Load the XML file 'movies-new.xml' using Python DOM parsing and do the following:

(4) Find all the movie title, year, rating, and description and print the output

(3) Output them in a JSON file 'movies.json' as follows:

{title: "a title", year:1982, rating: PG, description: "a description}

Part-4 - 5

Now load the XML file 'movies-new.xml' using Element-Tree and do the following:

- Find the movie title, and year of all Action movies and print them