Project Week 6

A Song Contest

Building an NLP Classification Project

The goal of this challenge is to present your knowledge about web scraping and NLP. We are using a public song lyrics website. We pick 3-4 artists and train a model to predict who performed an unseen song.

TASK 1 - Data Gathering - Web Scraping

- Define 3 or 4 artists you want to collect lyrics from. Each artist should at least have 50 songs
- Scrape the links to the sub pages with the lyrics from the artist’s main page of lyrics.com (list of URLs)
- Write a loop to download the songs and save them (it’s up to you if you want to store them separately in .txt files in a subfolder or put them all into a dataframe and store them as a csv).
- Repeat for the other 2-3 artists

TASK 2 - Text Cleaning & Preparation

- Write a loop to reload each of your stored lyrics and perform the following steps:
  - Text Cleaning (get rid of punctuation and unwanted characters)
  - Removing Stopwords
  - Lemmatization or Stemming (your choice)

TASK 3 - Model Training and Tuning

- Create a dataframe with the cleaned lyrics in each row (make sure that there is the same amount of lyrics for each artist!)
- Do a train-test-validation split (use stratification to make sure that the proportions are right)
- Build a pipeline with an Vectorizer (either WordCounter or TF-IDF)
- Train a classification model of your choice
- Pick a suitable performance metric and print the results for the training data set as well as the validation data set
- Repeat this process for at least two other algorithms of your choice - one of them being Naive Bayes
- Optimize using your validation set
- If your satisfied, test it with your test set

Bonus: feel free to add GridSearch, Cross Validation (don’t forget to shuffle), other algorithms...

TASK 4 - Present your findings

- On Friday, you’ll present your problem as well as your solution
- You can do this with slides or your jupyter notebook
- Tell us about challenges you had to overcome and what you learned from it

BONUS

- You can feed the model any other lyric or text and it will tell you which of your artists was the most likely one to have it as a song text
- Add a sentiment analysis of your lyrics