DataByte

The Official Machine Learning and Data Science Club of NITT

FIRST-YEAR INDUCTION

TASK-2

NATURAL LANGUAGE PROCESSING

1. Sentiment Analysis:

Develop a sentiment analysis model to classify Twitter tweets as positive, negative or neutral based on the sentiment expressed in the text.

- The model should be capable of hearing any sentiment-bearing words, contextual hints bearing positive or negative sentiments, and indicate the overall sentiment present in a given tweet.
- Measure the performance of the aforementioned model with standard evaluation metrics such as accuracy, precision, confusion matrix or any appropriate set of parameters.
- This technique is used by e-commerce companies to gauge and understand customers' sentiments on various products to provide feedback to manufacturers.

You are free to use the following dataset compiled from a Twitter API: **Dataset:** https://www.kaggle.com/datasets/kazanova/sentiment140

BROWNIE POINTS: Or feel free to use web scraping to obtain the required data to be added onto your dataset and make it on any topic i.e Elon Musk, Kerala Story, Farmer's Protest.

Recommended Software: VSCode, Jupyter Notebook

2. Recommender System:

Develop a movie recommender system using the K Nearest Neighbors (KNN) algorithm and its implementation in the field of Natural Language Processing (hint: you can use Porter Stemming, Lemmatization). You can use the dataset below to build and train your model.

- You should implement your model from scratch using only numpy (for building the model).

- Clean and preprocess the textual data, such as movie titles, product descriptions, or user reviews. This may involve tasks like tokenization, lowercasing, removing punctuation, stop words, and special characters.

- Reduce words to their base or root form to handle variations of words. Lemmatization and stemming help in reducing the vocabulary size and finding related items.

- The evaluation and performance metrics for recommender systems are different from those used in classification or regression tasks, as the focus is on measuring the quality of recommendations, such as accuracy, diversity, and coverage, rather than predicting a specific target variable.

Dataset:
https://files.grouplens.org/datasets/movielens/ml-latest-small.zip

->**BROWNIE POINTS:**
->Instead of using the given dataset, scrape movie data from a website to build your movie recommender system. For this purpose, use web scraping techniques to collect movie titles and other relevant information required for the recommender system.
->build a web application using Flask, a Python web framework, to deploy and demonstrate our movie recommender system. The web application will allow users to enter a movie title, and the system will recommend similar movies based on the input.
Recommended Software: VSCode, Jupyter Notebook

Important Note: We will check your submitted codes for plagiarism and if found, you'd be immediately removed from the induction process without any notice.

Resources:

1. https://www.geeksforgeeks.org/introduction-to-natural-language-processing/
2. https://youtu.be/X2vAabgKiuM
3. Resources for web scraping in twitter:
   1.https://docs.tweepy.org/en/stable/
   2.https://medium.com/@skillcate/python-tweepy-a-complete-guide-f0ff5ba54fce

https://realpython.com/beautiful-soup-web-scraper-python/

4. Basics of Flask: https://www.youtube.com/watch?v=Z1RJmh_OqeA