NLP PROJECT

(2024)

2023/2024

MODERN ACADEMY
Class 1

Team Members of class 1:

Rana emad – 12300020 – c1
Maryam emad – 12300542 - c1
Shahenda emad - 12300430 – c1

Project Idea:

Document Similarity (Quora question pair similarity)

Data Set Name:

https://www.kaggle.com/datasets/quora/question-pairs-dataset
we used 500 dataset only

Tools used in project :

- Pandas: For data manipulation and analysis.
- Scikit-learn: For model selection (train_test_split, GridSearchCV), feature extraction (TfidfVectorizer), similarity metrics (cosine_similarity), and evaluation metrics.
- NumPy: For numerical operations.
- NLTK: For natural language processing tasks like stopword removal and lemmatization.
- TF-IDF: For text feature extraction.
- Cosine Similarity: For measuring similarity between text vectors.
- Machine Learning Models:
    - Support Vector Classifier (SVC)
    - Random Forest Classifier
    - Gradient Boosting Classifier
    - XGBoost Classifier
- Accuracy Score: For evaluating model performance.

Summary:

this code processes text data from a csv file to determine if pairs of questions are duplicates. It import data from a csv file, clean and lemmatize text, removing stopwords , convert text to TF-IDF features, calculate cosine similarity between question pairs ,train several machine learning models on the similarity data and test the models and print their accuracy.