Nile University
جامعة النيل

COVID-19 Research Paper Clustering and Recommendation System

Advanced Big Data Analytics
Spring 2023

Project Documentation

Prof. Ahmed Awad
Eng. Heidy Hazem

Introduction

The COVID-19 Research Paper Clustering and Recommendation System is a project aimed at assisting researchers and healthcare professionals in navigating the vast amount of COVID-19 literature. The project leverages the COVID-19 Open Research Dataset (CORD-19), which contains over 57,000 scholarly articles related to COVID-19, SARS-CoV-2, and related coronaviruses. By clustering similar research papers and providing recommendations based on those clusters, the system aims to simplify the search for relevant publications and facilitate the discovery of new knowledge.

Setup

To set up and run the COVID-19 Research Paper Clustering and Recommendation System, the Databricks platform is utilized. Databricks provides a scalable and collaborative environment for data processing and analysis, making it an ideal choice for this project. The following steps outline the setup process:

- Create a Databricks account or access an existing account.
- Create a new workspace or select an existing workspace.
- Import the project files, including the research paper dataset and the necessary libraries.
- Set up a Databricks cluster with appropriate configuration and specifications.
- Launch a notebook and connect it to the cluster.
- Execute the notebook cells to run the project.

Project Requirements

1. Read the dataset using Spark:
    - The research paper dataset is provided in the form of a metadata CSV file. This file contains information about the papers, such as title, authors, abstract, and publication date. The dataset is read into a Spark DataFrame for further processing.
    - We found that the schema of the file is as follows:
        - sha: String type, nullable
        - source_x: String type, nullable
        - title: String type, nullable
        - doi: String type, nullable
        - pmcid: String type, nullable
        - pubmed_id: String type, nullable
        - license: String type, nullable

        - abstract: String type, nullable
        - publish_time: String type, nullable
        - authors: String type, nullable
        - journal: String type, nullable
        - Microsoft Academic Paper ID: String type, nullable
        - WHO #Covidence: String type, nullable
        - has_full_text: String type, nullable
    - The size of the dataset is 30057
    - Then we joined the metadata with the JSON files of each paper to get the title, authors, and abstract of each paper.
    - Then we put the title, authors, and abstract in all rows where the actual is empty.
    - We extracted the year from the publish_time column into the publish_year column.

2. Exploratory Data Analysis:
    - Perform an initial analysis of the dataset to gain insights into its structure and content.
    - Identify the relevant columns for clustering and recommendation.
    - Clean the data by replacing empty values and dropping unnecessary columns.
3. Preparation and Cleaning of the Data:
    - Clean the dataset by handling missing values, duplicates, and irrelevant columns.
    - Filter the dataset to include only English-language papers for analysis.
    - Perform necessary preprocessing steps, such as removing stop words and punctuation and converting text to lowercase.
4. Preprocessing:
    - Tokenize the text data into individual words.
    - Preprocess the text data by tokenizing, removing stop words, and converting to lowercase using the PySpark ML feature transformers.
    - Apply Word2Vec embedding to represent the papers as dense vectors.
    - Perform dimensionality reduction using techniques like PCA (Principal Component Analysis) to reduce computational complexity.
5. Clustering:
    - Utilize clustering algorithms such as K-means, Gaussian Mixture, or Bisecting K-means to group the research papers into clusters based on their embeddings.
    - Evaluate the clustering algorithms and select the most suitable one based on accuracy metrics.
6. Recommender System:
    - Calculate the cosine similarity between each paper in the clustered data.
    - Implement a recommender system that takes a paper title as input and recommends the most similar papers based on its cluster.

- Preprocess the input paper title using the same preprocessing steps applied to the research papers.
- Transform the title into a vector using Word2Vec.
- Calculate the cosine similarity between the input title vector and the clustered data vectors.
- Retrieve the N most similar papers from the same cluster as the input paper.

Conclusion

The COVID-19 Research Paper Clustering and Recommendation System aims to address the challenge of navigating a vast amount of COVID-19 literature by providing an automated approach to cluster similar research papers and recommend relevant publications. By leveraging techniques such as data preprocessing, vectorization, clustering, and a recommender system, the project facilitates easier access to relevant information and fosters the discovery of new insights in the field of COVID-19 research.