Foundations of Data Science

Course Project
Fall 2024

Phase 0

January 2025

Overview

In this project, the end-goal is to develop an AI assistant that can help a user find relevant papers. You will be given a dataset of scholarly articles up to the year 2017. You will try to infer useful information about the articles and use this information to enhance your AI assistant's inference. More data will be scraped from internet resources. The data you have crawled will be used as testing data throughout the course of the project.

This document contains instructions on **scraping** the required data from internet resources. Make sure to start working on Phase 0 of the project as soon as possible, as gathering all the data requires a *substantial* amount of time.

1 Crawling

You will be working with the DBLP dataset throughout the course of the project. The dataset provided for the scholarly papers has a cutoff date of 2017. This dataset will be used to infer useful information about the articles and use this information to enhance your AI assistant's inference.

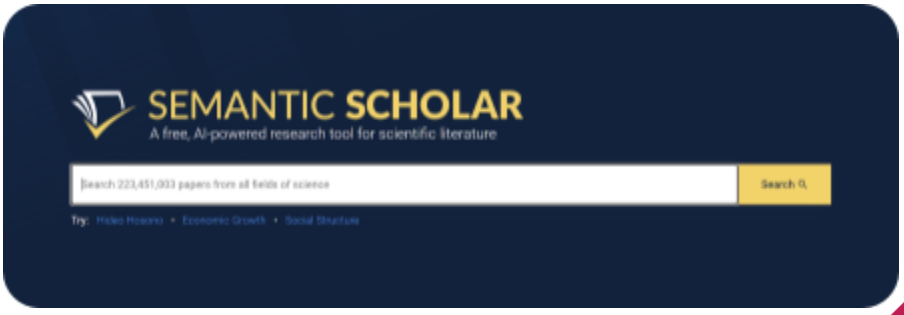
However, for testing and deploying the assistant, we are interested in newer papers related to 5 specific topics, listed below. These topics are your queries when scraping the data through the search engine.

- Foundation Models
- Generative Models
- LLM
- VLM
- Diffusion Models

1.1 Introduction

Semantic Scholar is a free, AI-powered research tool for scientific literature, developed at the Allen Institute for AI. It's designed to help researchers discover and understand academic papers across a wide array of disciplines. By leveraging advanced algorithms and machine learning techniques, Semantic Scholar can provide relevant search results, extract key information from papers, and suggest related works, making it a valuable resource for anyone in the academic and scientific community

Starting from its homepage, you can search for either of the specified topics. Upon searching for a query, you will find various options for filtering the search results and sorting them by any preferred metric. Feel free to move around the Semantic Scholar and try out its functionalities!



1.2 To Do

- ★ Can you figure out any proper way to search for the topics discussed above, sort the papers by relevance and also filter the results to only the papers after 2017? One way would be to send GET requests with proper arguments (Hint: Take a look at the URL in your browser after you apply a certain filter to the results!).
- ★ Propose a *thorough* crawling pipeline for searching these topics out, retrieving the results for papers after 2017, and storing the **Title, Abstract, Authors** and **Citations** of the paper. Other fields of the articles are not required, but you are allowed to store other useful information about the papers too in case you might find them useful.
- ★ For each of the 5 topics given in the previous page, scrape *at least* 2000 records. Store the results in a proper database so that they can be properly maintained and utilized later.

1.3 Uploading Material

You will upload your crawled results to the courseware. A small report on your methodology for crawling this dataset is also required.