PapersPlus

An implementation of paper retrieval google extension base on NLP

Zeyu Liao
zeyu9@illinois.edu
University of Illinois

Liziqiu Yang (leader)
Liziqiu2@illinois.edu
University of Illinois

1 Project Overview

With the continuous advancement of technology, our methods of information retrieval and processing have become increasingly intelligent and efficient. In this context, we are developing a Google Chrome extension based on Natural Language Processing (NLP) technology (under Theme 1: Intelligent Browsing). The primary aim of this extension is to assist users in better understanding specific academic papers. Whether a user inputs a sentence or asks a question, the extension will swiftly and accurately locate relevant paragraphs within the paper, providing users with a more convenient and precise academic information retrieval service. We will leverage the power of Python and JavaScript technologies, integrating efficient text processing and frontend interaction functionalities to achieve this goal.

2 Project Background

In today's digital age, academic research is flourishing, producing a vast number of new scholarly papers every day. This explosion of information provides rich resources for both the academic community and the general public. However, it also brings forth a new challenge: the complexity of academic articles often makes it difficult for the average reader to grasp their content accurately. Addressing this challenge, we are developing an intelligent Google Chrome extension aimed at providing users with a more convenient and intelligent way to access academic information. Leveraging Natural Language Processing (NLP) technology and the advantages of artificial intelligence, this extension will analyze scholarly papers and transform their content into language that is easy to understand. Through this approach, we aim to help users comprehend complex concepts, specialized terminology, and profound theories in the academic field with greater ease.

In this project, we develop a highly intelligent system capable of precise matching in response to user queries. Whether users are searching for research on a specific topic or trying to understand the latest developments in a particular professional field, our extension will serve as their most reliable assistant. Deeply integrated with Google, users will only need to input concise questions in the search bar. Our extension presents them in simplified, understandable language. This precise matching and intelligent interpretation will significantly enhance the efficiency of ordinary users in acquiring academic information, enabling them to navigate the sea of knowledge with confidence.

In summary, this project is not just a technological innovation; it represents a revolution in service. We aim to connect the academic world with the general public through intelligent means, breaking down the barriers that limit knowledge to expert domains. Through our efforts, we will open a gateway to the realm of knowledge for ordinary users, ensuring that the fruits of academic research benefit every corner of society.

3 Project Objectives

First, our project focuses on implementing an efficient NLP algorithm. This algorithm is not just a simple language processing tool, it is an intelligent academic knowledge extraction system. It can accurately parse academic papers, understand the semantics and logical relationships in them, and then generate a relevant keyword index. This index will be the key to the user's query, through which the user can accurately find the academic information they need. To achieve this goal, we use the latest deep learning technology, combined with the theory of natural language processing and information retrieval, to design and train a highly intelligent academic information extraction model.

Secondly, we develop a user-friendly interface for Google Plugin. The design of the interface focuses on user experience and strives to be concise and user-friendly. Users can enter questions or keywords in the search box of the plugin without tedious operation to trigger the academic information query function of the plugin.

Most importantly, we will realize the efficient interaction function between the front end and the back end. The front end will be responsible for receiving user inputs and accurately passing the user's query request to the back end for processing. The back end will receive the query request from the front end and utilize advanced NLP algorithms to parse the user's needs and quickly locate the relevant academic paper paragraphs. Then, the backend will present the processing results to the user in an intuitive way, which may be in the form of concise and clear text descriptions, charts, or other visualization forms. In this process, we will make full use of asynchronous processing and caching technologies to improve the response speed of the system and ensure that users can obtain satisfactory search results in the shortest possible time. At the same time, we will design a flexible back-end architecture to support high concurrency and large-scale data processing to meet the needs of users at different times and under different scenarios.

4 Project Plan

The topic we chose is intelligent browsing. We plan to calculate word frequency and TF-IDF, perform stemming, and use Gensim with TextRank to generate the summary. It would be a Chrome extension. In the demonstration, we will work through our extension on Chrome. Our expectation is when we select a paragraph of text, the extension will generate a summary of the text. The coding languages we will use are JavaScript and Python (with Flask to combine them). The total workload of our topic is about 40 hours. The table below is our initial plan for the project.

task	
Implement preprocessing and word frequency statistics	~ 5
Implement TF-IDF and other algorithms	~ 6
Implement Gensim and test summary generation for local files	~ 6
Design front-end interface and explore Google extensions	~ 5
Create extension based on the aforementioned Python algorithm	~ 12
Summarize and prepare for demo	~ 6

Table 1: project task estimated time.

5 Project result

Our project meets the basic requirements described above. When users enter questions/keywords, our extension will return explanations based on five different ranking functions. These functions are BERT, ChatGPT-3.5, ChatGPT-4.0, Doc2Vec, and TF-IDF. Users can select these functions by their preferences. We use FLASK, python, and JavaScript for the project.

For running the code, there are two parts, the back end and the front end. Please download FLASK first. For the back end, please run app.py and the front end is popup.html. But we did not web host. To validate our implementation, we include a demo video of our project in our GitHub repository.

6 Summary

Through this project, our goal is not only to provide users with an intelligent and efficient academic paper interpretation service but also to open the door to academic knowledge for a wider audience. We aim to make knowledge accessible to everyone, removing the privilege of understanding complex academic concepts solely for a select few experts. In this age of information explosion, the depth and breadth of academic research are constantly increasing, making it challenging for the average user to keep up with the pace of information. Our project will bridge this gap, freeing users from the complexities of the academic domain. Through intelligent interpretation and precise localization, we will help users better comprehend intricate academic concepts and profound knowledge, providing them with a broader academic perspective and inspiring their academic interests and creativity.

Simultaneously, this project represents the fusion of technology and innovation. We will incorporate cutting-edge natural language processing technology, combining artificial intelligence and big data analytics to create a powerful academic interpretation engine. This is not only a challenge to our technical capabilities but also a test of our team's collective intelligence. We anticipate encountering various technological and innovative challenges during the project because challenges drive progress. Leveraging the wisdom of our team, we will overcome these obstacles and propel the project toward success. Through this project, we will also provide new insights and methods for the future processing of academic information. We will accumulate valuable experience and contribute to the development of the academic and technological fields.