

Text Summarization of PubMed Articles using Gemini API

Introduction

The PubMed Summarization Project aims to simplify and expedite the review process of articles from PubMed by providing a tool that summarizes lengthy texts into concise versions. This project involves loading and preparing a dataset for summarization, and developing a web application where users can input PubMed articles to receive a summarized version using a state-of-the-art API-based model like Gemini by Google.

Data Exploration and Preparation

This all work has been done on Jupyter Notebook. Main libraries we used for data cleaning and preprocessing were pandas and nltk.

1. Loading the Dataset

We use the PubMed Summarization dataset available on Hugging Face. This dataset contains pairs of PubMed articles and their summaries, which will be instrumental in training and evaluating our summarization model.

For this you must have to install the dataset library of Python by using command: `pip install datasets`

2. Exploring the Dataset

This step helps us identify key components such as article text, summary, and any metadata.

3. Preprocessing the Dataset

Preprocessing is crucial for effective summarization. Typical preprocessing steps include Tokenization, Removing special characters and stop words, Lowercasing the text, Lemmatization.

After this process we stored our data in a csv file for further operations.

Web Application Development

Choosing the Framework

For this project, we use Flask, a lightweight and flexible web framework for Python. Flask allows for building scalable and customizable web applications with ease.

1. Set Up Flask:

Create a new directory, `app`, for the Flask application. Inside this directory, create the main Flask script, `app.py`, and a folder, `templates`, for HTML files.

2. Create `app.py`:

Set up the basic structure of the Flask app to accept user input or file upload.

3. Create HTML Templates:

In the templates directory, create index.html for the home page and summary.html for displaying the results.

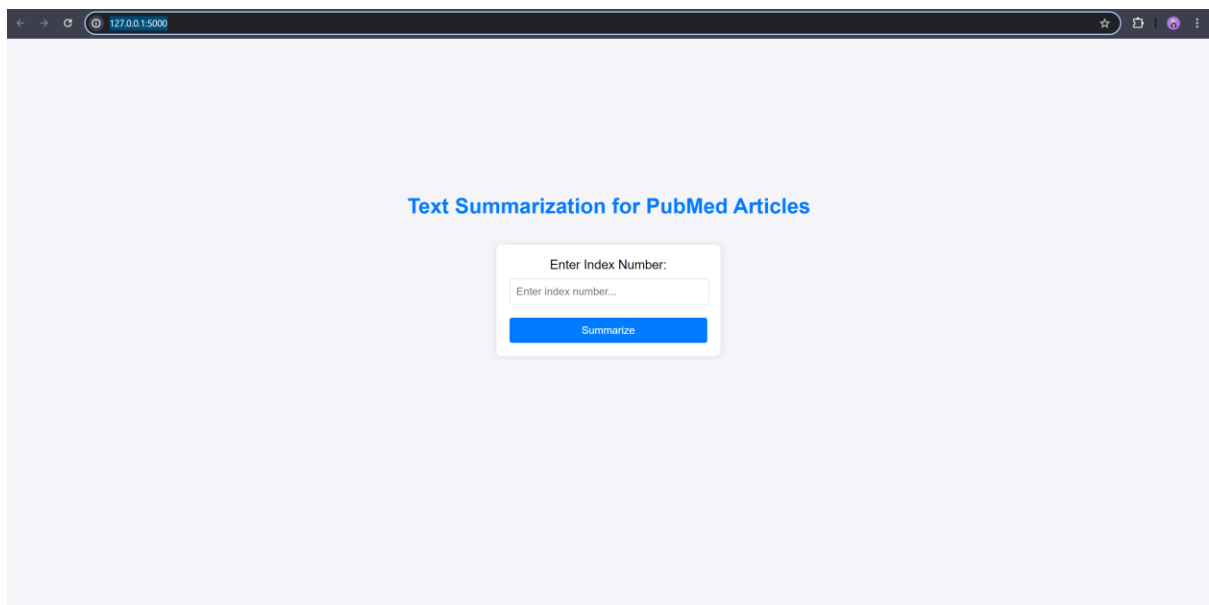
You just need to type this command on terminal to run the application: `python app.py`

Integrating Text Summarization

We used the API-based model (Gemini by Google) for summarization. Assuming you have API access and key. You can get it from Google AI studio.

Interacting with the Application

Get index of the article as an input from user. Click the "Summarize" button to see the summarized version of the article.



The screenshot shows a web browser window with the address bar displaying `127.0.0.1:5000`. The page has a light purple background and is titled "Text Summarization for PubMed Articles" in blue text. In the center, there is a white box containing the text "Enter Index Number:" above a text input field with the placeholder "Enter index number...". Below the input field is a blue button labeled "Summarize".

Result:

← → ↻ 127.0.0.1:5000/summarize ☆ 📄 🌙 ⋮

Text Summarization Result

Original Article

this study provides class iii evidence that more patients are seizure - free and have stopped aed treatment in the long term after resective epilepsy surgery than nonoperated epilepsy patients .

in sweden , all epilepsy surgery procedures are reported to the snesur , which was initiated in 1990 .

an internal control system rejects certain impossible combinations of data and regular external quality controls are performed by an independent controller .

since 2005 , the follow - up has been extended from 2 years to 5 , 10 , and 15 years postoperatively .

snesur contains baseline information on patient 's epilepsy history , preoperative seizure types and syndromes , mean monthly seizure frequency during the year preceding the presurgical investigation , aeds , preoperative investigations , psychosocial data (type and location of surgery) , histopathologic diagnoses , and postoperative complications .

two - year follow - up data cover seizure situation , aeds , and psychosocial data .

the 5- , 10- , and 15-year follow - ups are structured telephone interviews regarding seizure situation , aeds , psychosocial aspects , and driving . in this study , we analyzed seizure outcome and aed medication 5 and 10 years after resective epilepsy surgery in patients who had 5- and 10-year follow - ups in 2005 to 2007 (and hence were operated on in 2000 to 2002 and 1995 to 1997) .

the cohort comprises the 327 patients who had resective surgery during these time periods . in 2005 to 2007 , 144/176 patients operated on in 1995 to 1997 (98/116 adults and 46/60 children 18 years) had a 10-year follow - up , and 134/151 patients operated on in 2000 to 2002 (92/103 adults and 42/48 children 18 years) had a 5-year follow - up .

seventeen patients were reoperated before long - term follow - up and there were 11 deaths .

twenty - one patients (6.4%) were lost to follow - up (for details , see figure e-1 on the neurology web site at www.neurology.org) , as a control group , consecutive patients who underwent presurgical investigations during the same time periods but were not operated were identified at 3 of the 6 operating centers (gteborg , uppsala , and lund) .

eighty adults and 13 children out of 94 adults and 13 children underwent a cross - sectional long - term follow - up in 2008 after a mean of 9.3 years (adults) and 8.8 years (children) using the same structured telephone interview as for the surgical group .

thirteen adult patients had died (4 epilepsy - related deaths) and 1 was lost to follow - up (1%) .

reasons for not having surgery were nonconclusive workup (n = 41) , multifocality (n = 27) , patient declined surgery (n = 12) , seizure onset within eloquent cortex (n = 11) , and neuropsychological reasons (n = 2) .

seizure freedom (without or with aura , international league against epilepsy [ilae] class i and ii) is reported for the year preceding the follow - up except for patients with sustained seizure freedom since 1990 , which is considered permanent for patients with continuing seizures or seizure relapse postoperatively , the mean monthly seizure frequency in the last year of follow - up is categorized as follows : 75% reduction in seizure frequency ; 50%-74% reduction in seizure frequency ; 0%-49% reduction in seizure frequency , for comparison between 2 groups , mann - whitney u test for continuous variables , and mantel - haenszel test for ordered categorical variables .

fisher exact test was used for dichotomous variables , mann - whitney u test for continuous variables , and mantel - haenszel test for ordered categorical variables .

logistic regression analysis was performed for each independent variable to predict seizure - free outcome .

a forward stepwise multiple logistic regression was used to select independent predictors to outcome . only univariate predictors attaining a p value of < 0.10

this study was approved by the regional board of medical ethics at the university of gothenburg .

Summarized Article

This study investigates the long-term outcomes of resective epilepsy surgery in Sweden, comparing patients who underwent surgery with those who did not.

****Key findings:****

- **Seizure Freedom:**** Resective surgery significantly increased the likelihood of long-term seizure freedom compared to non-operated patients (41% of adults and 44% of children vs. none in the control group).
- **AED Withdrawal:**** A significant portion of seizure-free patients (43% of adults and 86% of children) discontinued antiepileptic drugs (AEDs) after 10 years, compared to none in the control group.
- **Resection Type:**** Temporal lobe resections (TLR) showed the best long-term seizure outcomes, with 63% of adults and 60% of children becoming seizure-free.
- **Predictors:**** High baseline seizure frequency and longer epilepsy duration were negatively associated with long-term seizure freedom, while positive MRI findings were positively associated.
- **Control Group:**** The control group was composed of patients who underwent presurgical evaluation but were deemed ineligible for surgery. While not ideal, they share many baseline characteristics with the operated group.

****Strengths of the study:****

- **Prospective and Longitudinal:**** The study uses a large, prospective, and long-term follow-up design, providing robust data.
- **Population-Based:**** The study uses a Swedish national registry, providing a representative sample of patients.
- **Standardized Outcome Measures:**** The study utilizes a consistent protocol for assessing seizure outcomes.

****Limitations:****

- **Lack of Masking:**** The study lacks blinding, which could introduce bias in outcome assessment.
- **Control Group Limitations:**** While the control group is representative of patients evaluated for surgery but not operated, it is not ideal for comparison.

****Overall:**** This study provides strong evidence for the long-term efficacy of resective epilepsy surgery in reducing seizures and allowing for AED discontinuation, highlighting the importance of early identification of surgical candidates.

[Go Back](#)

Future Improvements

Enhanced Preprocessing: Implement more advanced text preprocessing techniques like stemming or lemmatization.

Model Fine-Tuning: Train and fine-tune the summarization model on domain-specific data for improved accuracy.

User Interface: Improve the UI for better user experience and support for more file types.

Caching and Optimization: Optimize API calls and application performance.

Conclusion

The PubMed Summarization Project encapsulates a comprehensive approach to simplifying the review process of scientific literature by harnessing the power of advanced text summarization technologies. Through meticulous data exploration and preparation, we ensured that the dataset was ready for effective summarization. The development of a user-friendly web application using Flask has provided an accessible platform for users to quickly and efficiently summarize lengthy PubMed articles. By integrating a cutting-edge API-based summarization model like Gemini, the project not only enhances the accessibility of scientific knowledge but also accelerates research and review workflows.

This project serves as a testament to the potential of combining data science with modern web technologies to create tools that can significantly impact research communities. While the current implementation provides a robust foundation, there is ample scope for future improvements, such as fine-tuning the summarization model for specific domains, enhancing the preprocessing pipeline, and expanding the application's capabilities to support a broader range of document formats and languages.

Overall, this project stands as a valuable resource for researchers, academics, and professionals who seek to keep pace with the ever-growing volume of scientific literature. By continuing to build on this work, we can further bridge the gap between extensive research data and its efficient, digestible presentation.