CS410 Final Project Report – Fall 2023

Topic mining with LLM

Background of selection of topic: Scientific and technical publishing is in crisis and the heart of the problem is the amount of scientific and technical papers have surpassed human capability for reading, interpretation, and synthesis. One of the responses to this problem is to use text mining of articles instead of reading them. Therefore, I decided to use LLM to see how I can use it for topic mining and find top and important keywords from technical papers and college course lectures that can help students and researchers digest the most important content from text. I used several of my publications for this analysis and I evaluated the result per my expectation.

Dataset: I have used CS410 text retrieval (week 1 to week 6) and CS410 text mining (week 7 to week 12) course transcript as coursera lectures examples for topic mining. Furthermore, I have used four of my past technical publications (conference and journal) as a technical paper analysis. You can find these datasets in the "dataset folder" in project repo.

Codes: I have used Jupyter Notebook (Python language) for this project. I have run the code in my local drive and uploaded these codes in github. You can download dataset and codes from project repo and run those in your local drive for testing.

There is total separate code script for this project:

Doc_to_text.ipynb: This script converts doc documents to txt format with some data cleaning process in order to convert technical publication to text for topic mining analysis by LLM. I have used the "textract" library on this script.

Topic_LLM.ipynb: You need to install the "langchain" and "openai" library in order to use this script. Another requirements are set up an OpenAi account (https://platform.openai.com/) and create a secret key (https://platform.openai.com/api-keys) to communicate with OpenAI API. You will get $5 credit when registering an account. When you obtain your secret key, you should add it to your system environment variables. Here are the steps that you should follow (https://www.immersivelimit.com/tutorials/adding-your-openai-api-key-to-system-environment-variables) to make your device ready to communicate with OpenAI. I have imported several functions from the "langchain" library such as text_splitter, create_extraction_chain, ChatPromptTemplate, and HumanMessagePromptTemplate. First, I used "*gpt-3.5-turbo-0613*" but the suggested topics were not long enough therefore I used "*gpt-4-0613"* for main topic analysis alongside gpt 3.5 to reduce the cost. This script load documents and text from the directory folder and splitted to chunks with chunk overlaps. Use templates as training for OpenAI in order to generate topics from documents that are sent to OpenAI. Main functions on template are: system_message_prompt, human_message_template and chat_prompt_combin (. After communication with the API, the system generates topic names with a description that describes the topic. I saved these topics in csv format.
(https://python.langchain.com/docs/modules/model_io/prompts/prompt_templates/)

Find_keyword.ipynb: I used "nltk" and "stopwords" on this script. Generated topics from previous script are in csv format. This script converts the found topics and converts them to txt. Unfortunately, I could not remove all unnecessary words by stopwords therefore I wrote a code line to replace some common words (such as This and The) and punctuation. Later I used "stopswords" to remove them from text to make it ready for maximum likelihood estimation analysis. I used a dictionary to count the word frequency in filtered topic text that followed by another code that calculate maximum likelihood estimation. Finally, I plotted the top 30 or top 15 results as keywords.

Topics Output: I uploaded and saved the topics that were found from text documents by "Topic_LLM.ipynb" in this folder.

Filtered Topics: This folder contains topics in text format and filtered topics (stop words have been removed).

Keywords: Finally, word frequencies and maximum likelihood estimation for each technical papers and CS410 course transcripts have been saved in this folder.

Results: Topic founds for "In-Plane Vibration Mode Shapes for Rotating Disks – Exact Solution" technical paper:

topic_name	description
Rotating Disks	This topic covers the development of an analytical method to determine the modal vibration characteristics of high-speed rotatin
Applications of Rotating Disks	This section explores the various applications of rotating disks in engineering systems such as flywheels, torsional disk dampers, tu
Analysis of Rotating Disks	This topic reviews past research efforts on the in-plane vibration analysis of rotating annular disks. It discusses the works of variou
Governing Equations for Rotating Disks	This section presents the governing equations for a two-dimensional homogeneous, elastic, and isotropic disk rotating about its a
Modal Displacements and Stresses	This topic discusses the radial and tangential displacement and modal radial and shear stresses. It introduces non-dimensionalized
Modal Analysis	This section discusses the determination of modal information, the importance of satisfying boundary conditions, and the derivati
Comparison and Validation of Analytical Procedure	This topic presents a comparison of the results from the analysis with established results reported by other researchers. It validate
Conclusion	The conclusion summarizes the research, emphasizing the development of an analytical method for predicting in-plane natural fre

Word frequency for "In-Plane Vibration Mode Shapes for Rotating Disks – Exact Solution" technical paper:

count_topic_mode_shape_rotating_disk_paper - Notepad

File Edit Format View Help

```
rotating: 12
disks: 11
discusses: 7
modal: 6
analysis: 6
vibration: 5
equations: 5
topic: 4
analytical: 4
annular: 4
disk: 4
research: 4
presents: 4
stresses: 4
applications: 3
section: 3
also: 3
in-plane: 3
importance: 3
displacements: 3
```

maximum likelihood estimation for “In-Plane Vibration Mode Shapes for Rotating Disks – Exact Solution” technical paper:

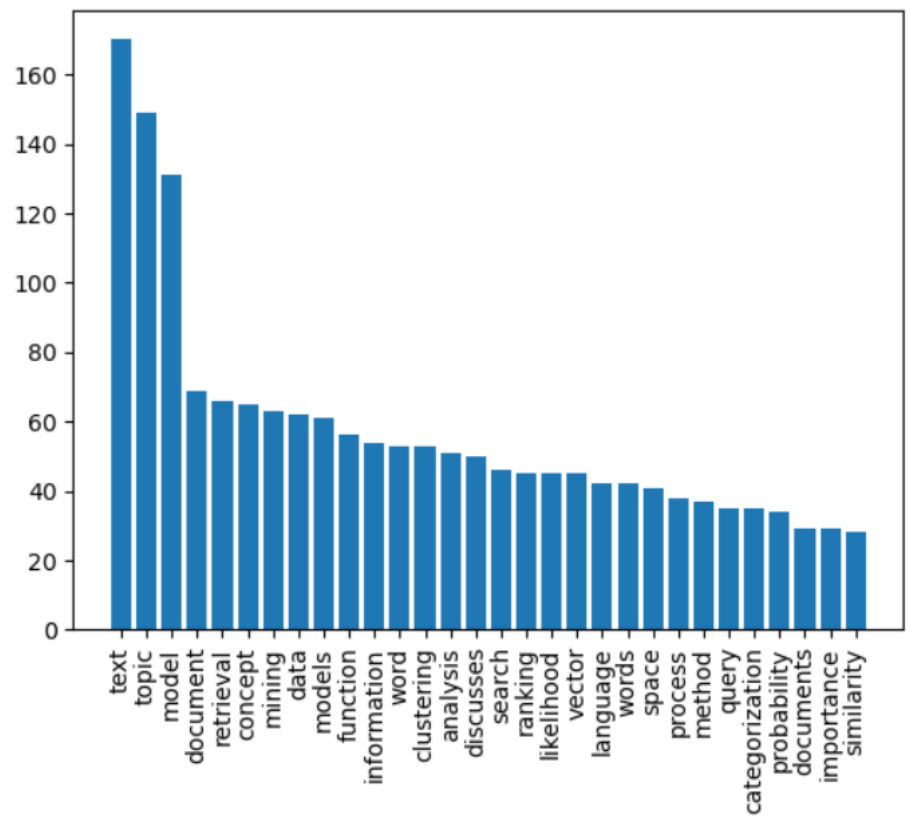
p_topic_mode_shape_rotating_disk_paper - Notepad

File Edit Format View Help

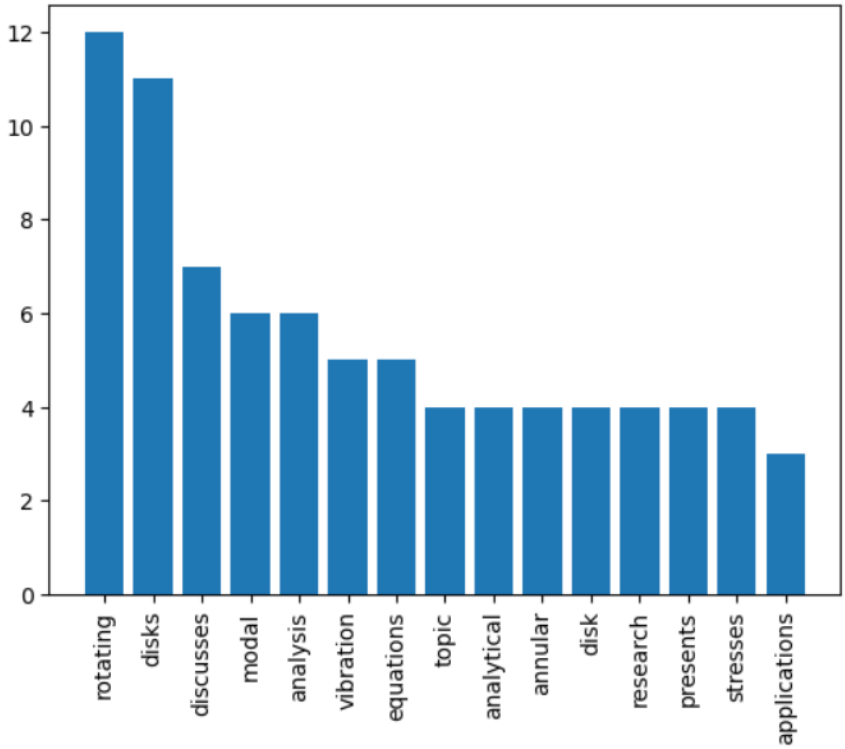
```
rotating: 0.054545
disks: 0.05
discusses: 0.031818
modal: 0.027273
analysis: 0.027273
vibration: 0.022727
equations: 0.022727
topic: 0.018182
analytical: 0.018182
annular: 0.018182
disk: 0.018182
research: 0.018182
presents: 0.018182
stresses: 0.018182
applications: 0.013636
section: 0.013636
also: 0.013636
in-plane: 0.013636
importance: 0.013636
displacements: 0.013636
```

Keywords visualization:

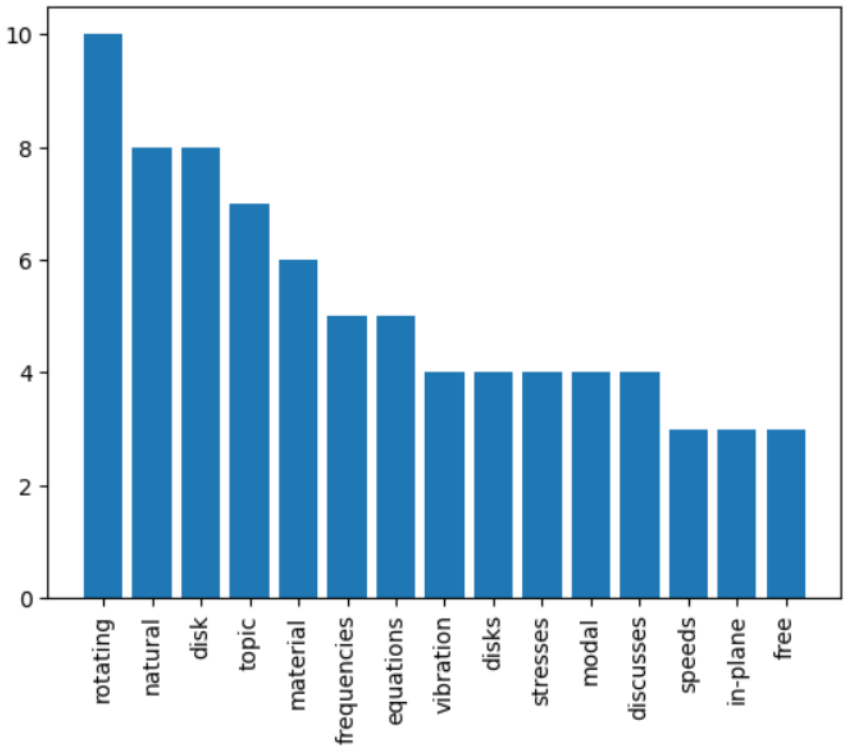
CS410 course:



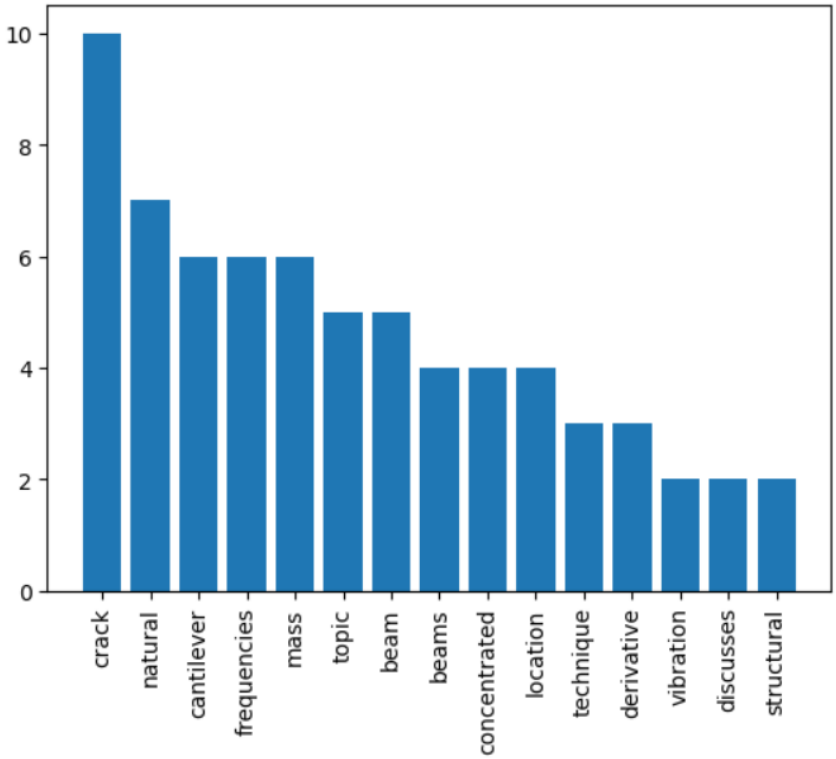
“In-Plane Vibration Mode Shapes for Rotating Disks – Exact Solution” technical paper:



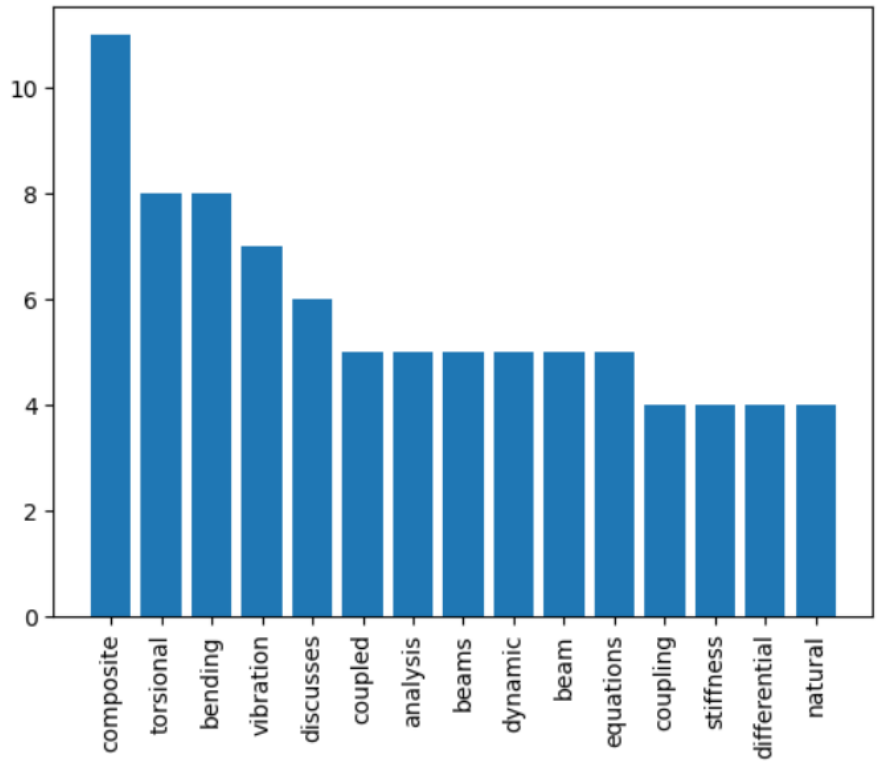
“Influence of Embedded Material on Natural Frequencies of Double Segment Rotating Disk” technical paper:



“Analytical Verification for Vibration Analysis Technique used in Determination of Cracking in Cantilever Beams” technical paper:



“Coupled Flexural and Torsional Vibration Analysis of Composite Beams” technical paper:



Discussion and system evaluation:

In general, I was impressed by using LLM and OpenAI for topic analysis. That can be replaced statistical topic mining such as LDA and PLSA in future. In order to evaluate the accuracy and performance of the system that I established for text mining to help researchers, engineers, and graduate students to have a reliable way to get fast and quick topics and content from technical publication without reading them, we will check if the system could predict the keywords that I provided to the publisher. Key words on the following technical paper are: [in-plane, free vibration, rotating disks, compound disks, annular thin disks, mode shapes, critical speeds, natural frequencies, medium with discontinuity (I will remove this since I did not discuss it in paper)]. Therefore, there are 8 total keywords for this paper. Most of them are bigram words but I used unigram word distribution in my analysis.

Journal of Applied Nonlinear Dynamics

https://lhscientificpublishing.com/Journals/JAND-Default.aspx

Influence of Embedded Material on Natural Frequencies of Double Segment Rotating Disk

Ehsan Sarfaraz† and Hamid R. Hamidzadeh

Department of Mechanical and Manufacturing Engineering, Tennessee State University, Nashville, TN 37209, USA

Submission Info

Communicated by Valentin Afraimovich

Keywords

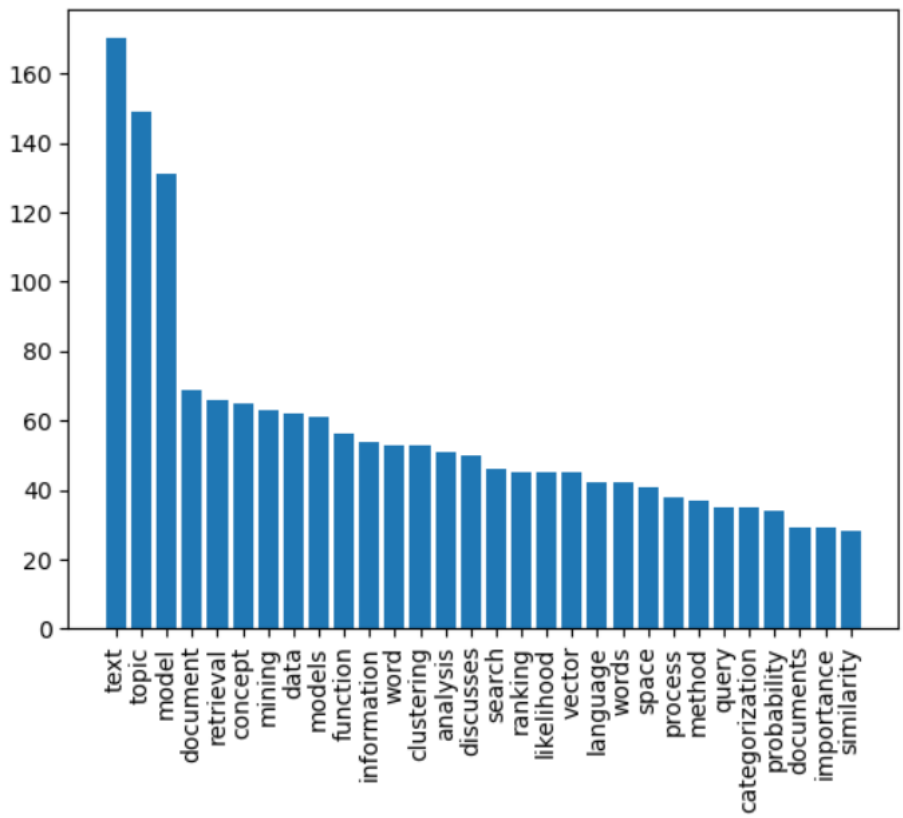
In-Plane
Free vibration
Rotating disks
Compound disks
Annular thin disks
Mode shapes
Critical speeds
Natural frequencies
Medium with discontinuity

Abstract

An analytical method is presented to determine the effect of adding different materials at one of the edges of an annular rotating disk on its in-plane natural frequencies and critical speeds. The proposed analysis is based on the linear in-plane free vibration of a compound disk with material discontinuity, by adopting the two-dimensional plane stress theory. The frequency equation was achieved by satisfying the compatibilities of the displacements and stresses at the interfaces of the different segments. The materials used in each segments of the disk are assumed to be homogenous, elastic, and isotropic. Furthermore, the annular disk is considered to be clamped at the inner side and free at the outer edge with a radius ratio of 0.3, and rotates with a constant angular speed. The variation of non-dimensional natural frequencies in fixed coordinates for different modes and different segment radiuses at the inner or outer side with respect to speed of rotation are computed. Presented results indicated that by adding additional segment, undesirable natural frequencies of the rotating disk can be modified to be within the acceptable range.

As you can see in top 15 key words extracted from system, in-plane, free vibration, rotating disk, and natural frequencies are there. Critical speeds and mode shapes are very close to modal and speeds. I have not seen compound or thin disk on keyword, but I have seen material and disks as unigram word. I think is almost 80 percent close to my expectation which is very good to find such quick keyword to use it as query to find relevant technical papers.

CS410 course analysis:



Here it is, the most word that was discussed in this class is "text" and I was expecting. Furthermore, the most other topics and keywords for this class are model, document, retrieval, mining, function, ranking, word, clustering, likelihood, vector, query, probability, similarity, and categorization. As you can see these are the topics and keywords that help students to more focus to get study and learn the subject.

Project commitment: I spent 24 hours to complete this project (research, gather the data, clean data, codes and script, and documentation).

Reference:

https://textract.readthedocs.io/en/stable/installation.html

https://pypi.org/project/langchain/

https://platform.openai.com/docs/overview

https://github.com/gkamradt/langchain-tutorials/tree/main

https://github.com/coursera-dl/coursera-dl

https://pypi.org/project/langchain/