DATA*i2i* Industrial Data Science workshop - 2023

Capstone projects presentation

TextBook Clustering

Team
T Lakshmi kousalya
Korada Harika
Naladeega Amrutha
Simma Hymavathi
K Jhansi Lakshmi

DATAi2i
Innovate | Insight | Impact

01
Why do we need this?

Problem Objective

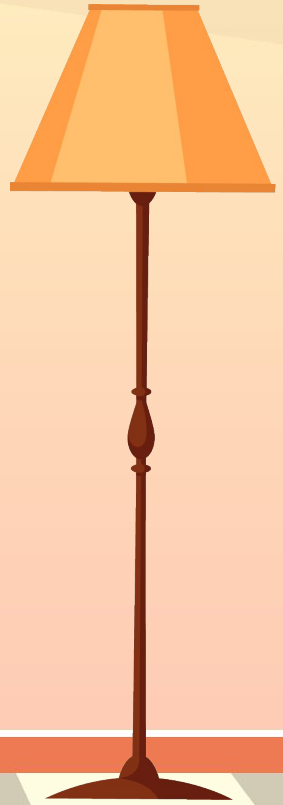
Objective :

The objective of the project is to develop a website that can automatically organize and categorize textbooks based on their content, topics, or subject areas.

Benefit :

Allows users to navigate through the collection of textbooks and find material related to specific subject more effectively, saving time and effort in finding specific information.

02
Execution Overview



Webpage sample

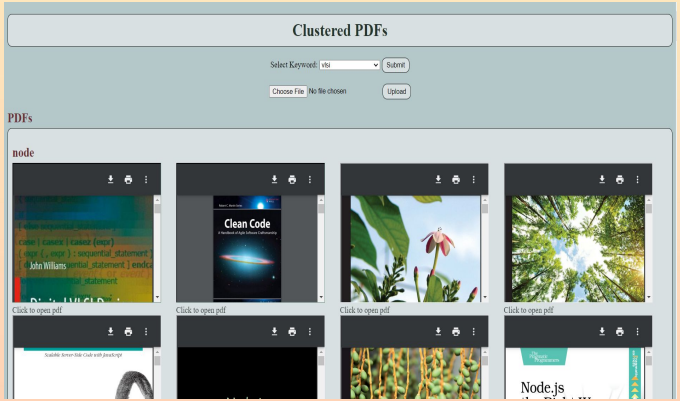
Tools:

libraries :

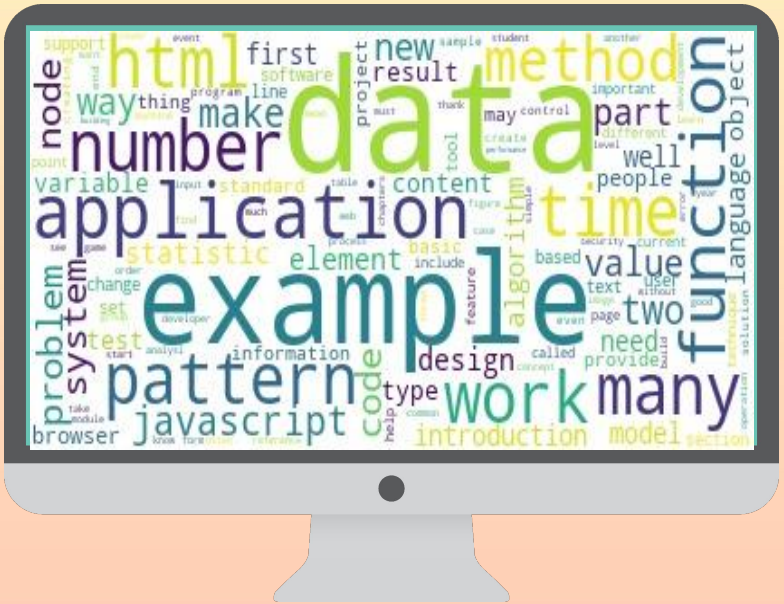
- pandas, numpy, matplotlib : basic data transformation and visualization
- PyPDF2 : text extraction
- dataprep : text cleaning
- nltk : text processing and analysis
- gensim : model development

framework : Flask - flexible way to develop webpage

frontend : html,css

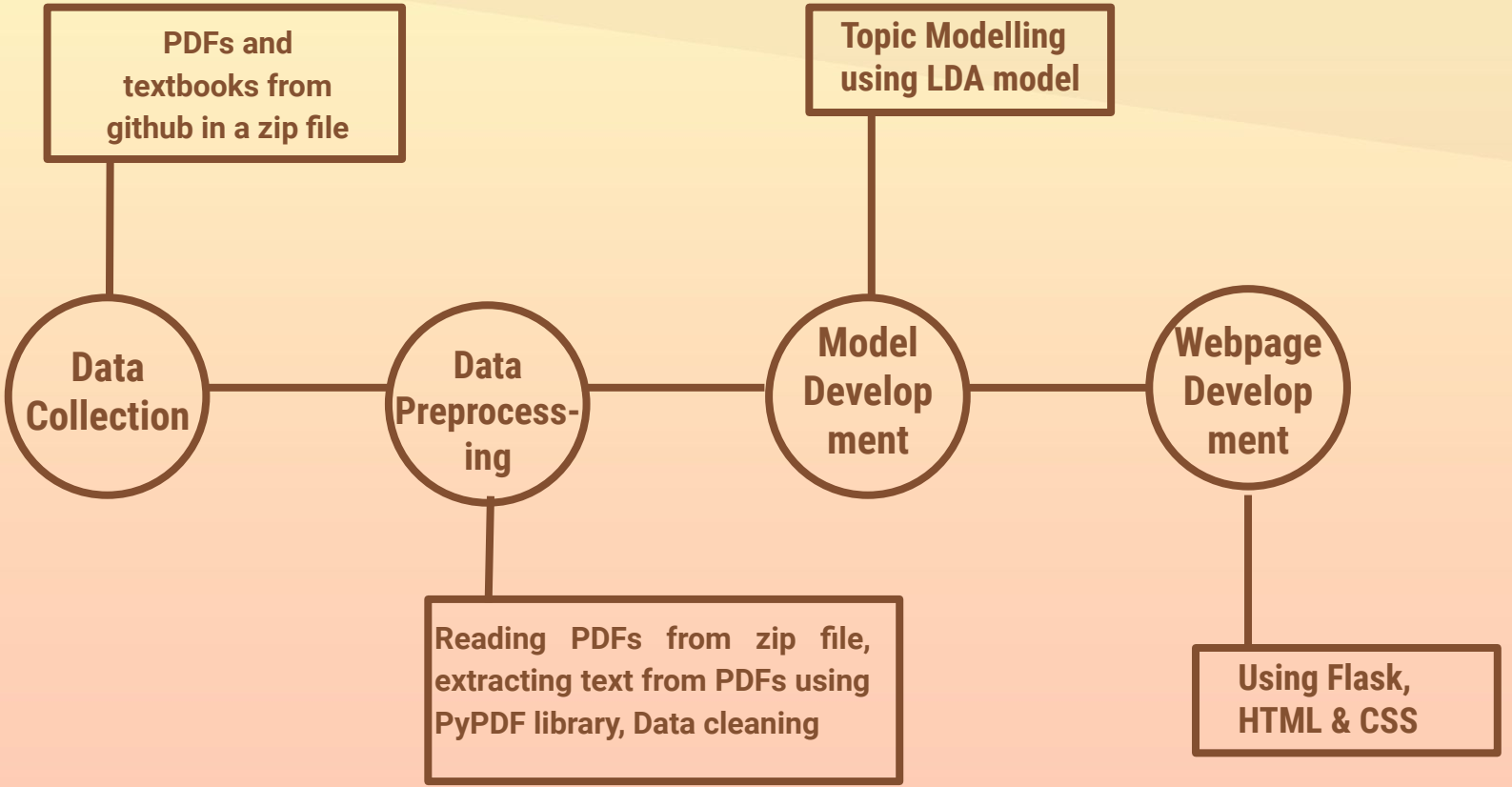


03
Analytical Approach



The data had textbook PDFs of different subject downloaded from GitHub, and our own textbooks

We can observe that the topics vary a lot ranging from data science to electrical systems


PDFs and textbooks from github in a zip file
Topic Modelling using LDA model
Data Collection
Data Preprocess-ing
Model Develop ment
Webpage Develop ment
Reading PDFs from zip file, extracting text from PDFs using PyPDF library, Data cleaning
Using Flask, HTML & CSS

Data Collection : Around 103 PDF files are collected from GitHub and other websites
Sources : GitHub, mrcet.com

Data Preprocessing :

- The first 25 pages text is extracted from PDFs using PyPDF2 library
- Basic data cleaning like removing digits, removing urls, lowercase the text is done using dataprep library
- Stop-words are removed and a corpus is build which is later converted into “bag of words (BOW)

	Folder	PDF	Text	cleaned_text
0	Cluster champs data/vlsi	Digital VLSI Design with Verilog (John William...	digital vlsi design verilogjohn williams digit...	[digital, vlsi, design, verilogjohn, williams,...
1	Cluster champs data/vlsi	VLSI Design _ GSK.pdf	sri chandrasekharendra saraswathi viswa mahavi...	[sri, chandrasekharendra, saraswathi, viswa, m...
2	Cluster champs data/mpmc	VIJAYARAGHAVAN_mp _mc notes.pdf	dr vijayarghava n microprocessor microcontroll...	[dr, vijayarghava, n, microprocessor, microcon...
3	Cluster champs data/vlsi	digital-integrated-circuits-a-design-perspecti...	table contents digital integrated circuits des...	[table, contents, digital, integrated, circuit...
4	Cluster champs data/mpmc	mpmc digtal notes.pdf	microprocessors microcontrollers lecture notes...	[microprocessors, microcontrollers, lecture, n...
...	...	...	...	...
98	Cluster champs data/web development	[JavaScript The Definitive Guide Activate Your...	javascript definitive guidesixth edition javas...	[javascript, definitive, guidesixth, edition, ...
99	Cluster champs data/statistics	sts(15).pdf	think stats probability statistics programmers...	[think, stats, probability, statistics, progra...
100	Cluster champs data/signal processing	DSP Sample Chapter_01_09_19 (1).pdf	digital signal processingusing arm cortex base...	[digital, signal, processingusing, arm, cortex...
101	Cluster champs data/statistics	sts(4).pdf	robertv hogg allent craig theuniversity ofiowa...	[robertv, hogg, allent, craig, theuniversity, ...
102	Cluster champs data/power systems	Power System Analysis (John Grainger, Jr.,Will...	powe r system analysis mcgraw hill series elec...	[powe, r, system, analysis, mcgraw, hill, seri...

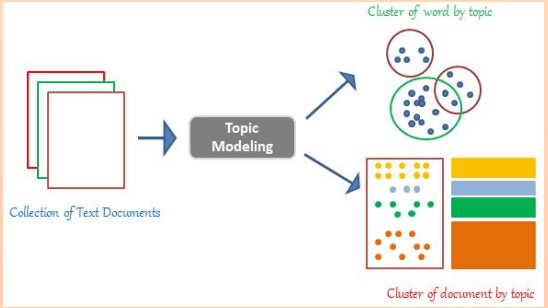
103 rows × 4 columns

Model Development :

- Topic modelling approach is used to group textbooks with similar topics together
- The Latent Dirichlet Allocation (LDA) is trained on the corpus to extract topics of the documents

```
Topic 0: security (Probability: 0.0208)
Topic 1: motors (Probability: 0.0123)
Topic 2: digital (Probability: 0.0201)
Topic 3: statistics (Probability: 0.0138)
Topic 4: winding (Probability: 0.0157)
Topic 5: javascript (Probability: 0.0121)
Topic 6: html (Probability: 0.0460)
Topic 7: memory (Probability: 0.0167)
Topic 8: statistics (Probability: 0.0136)
Topic 9: node (Probability: 0.0325)
Topic 10: performance (Probability: 0.0085)
Topic 11: energy (Probability: 0.0313)
Topic 12: angularjs (Probability: 0.0160)
```

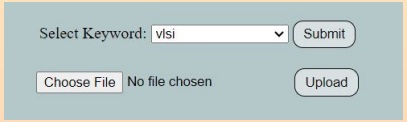
- LDA assumes that each document is generated by a statistical generative process. That is, each document is a mix of topics, and each topic is a mix of words.

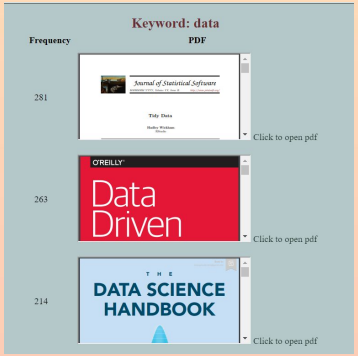


Webpage Development :

- Flask is a lightweight web framework for building web applications in Python.
- It provides a simple and flexible way to develop web-based projects, from small applications to complex websites.
- HTML is used to provide the structure of the page and CSS is used for the visual layout.

- The webpage allows users to upload a PDF file and the page returns which cluster/topic the book belongs to.
- The webpage also has keyword search which enables users to pick a keyword and get the frequency of the keyword in the documents.





04
Technical Challenges

1. **K-Means clustering:** We used K-means clustering, but the clusters had combinations of different topics

	Folder	PDF	cluster_labels
10	Cluster champs data/software development	[Cyberspace and Cybersecurity [Print Replica] ...	5
0	Cluster champs data/vlsi	Digital VLSI Design with Verilog (John William...	5
13	Cluster champs data/power systems	Principles of Power System (V K Mehta, Rohit M...	3
24	Cluster champs data/power systems	Power System Analysis Power System Analysis (N...	2

2. **Package Versions :** The required package versions are different than the system versions **Solution**: Use a separate virtual environment for each project
3. **No. of topics** : Less number of topics merged different topics together, too many topics gave unnecessary topics **Solution** : Experiment with different values

```
Topic 6: 0.043*"(" + 0.040*"$" + 0.033*"%" + 0.032*"'" + 0.032*"*" + 0.030*"&" + 0.029*"'" + 0.027*":" + 0.024*"!" + 0.020
*"#"
Topic 7: 0.054*"." + 0.048*"," + 0.014*"<" + 0.014*">" + 0.011*":" + 0.007*"'" + 0.007*")" + 0.007*"(" + 0.007*"html5" + 0.00
6*"chapter"
```