social computing

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

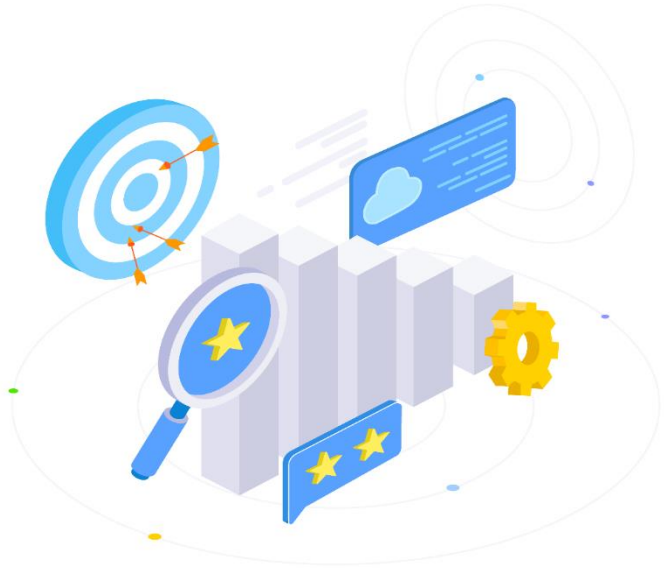
BTS-Books to Search

BTS

Neurons

- Farnaz Arghavan
- Divyangana Kothari
- Manoj Kumar Dara
- Julian Stülp

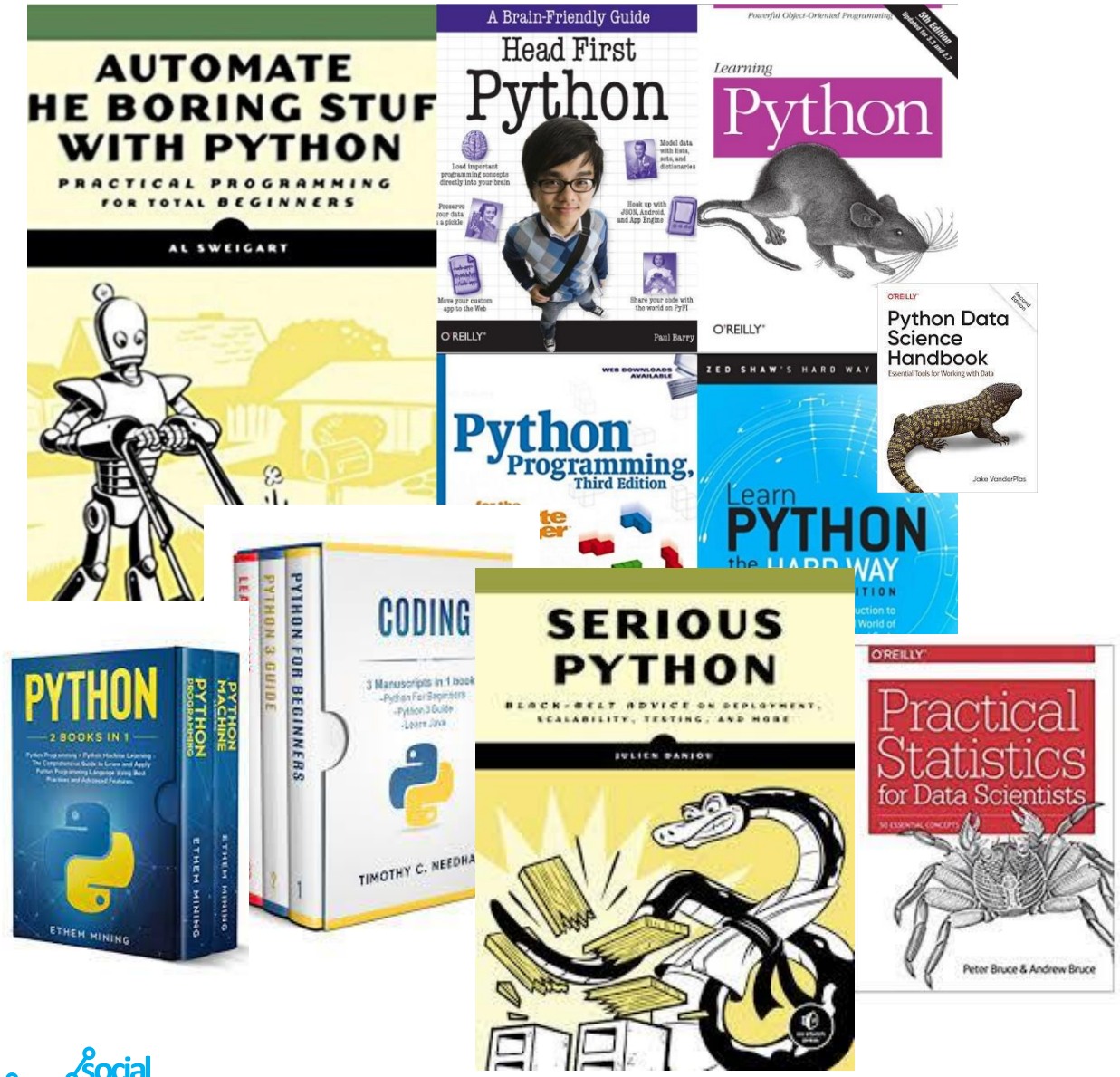
Overview



- What is **BTS**?
-

- Implementation
- Conceptual Solution
- Result
- Summary

What is **BTS**?



Despite the availability of instructional videos, books continue to be a crucial tool for learning new technology and obtaining knowledge. They've been with us for a long time and continue to play a vital role in our educational path.

Abstract

- A book recommendation system based on embedding, with interactive representations of recommended books.

What –Why – How

- The data used for this project includes book titles, which will be used to generate embeddings for each book.
- Data Information

```
RangeIndex: 1430 entries, 0 to 1429
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   title           1430 non-null   object
 1   author          1036 non-null   object
 2   Rating          1299 non-null   object
 3   complete_link   946 non-null    object
```

- The data is cleaned and preprocessed before being used to train the model to make sure it is accurate and useful.
- **MySQL DB** is used to store and retrieve datasets.

What –**Why** – How

To provide a personalized book recommendation system that takes into account the user's search query and **suggests books** that are semantically related to the query.

Action – Target pairs

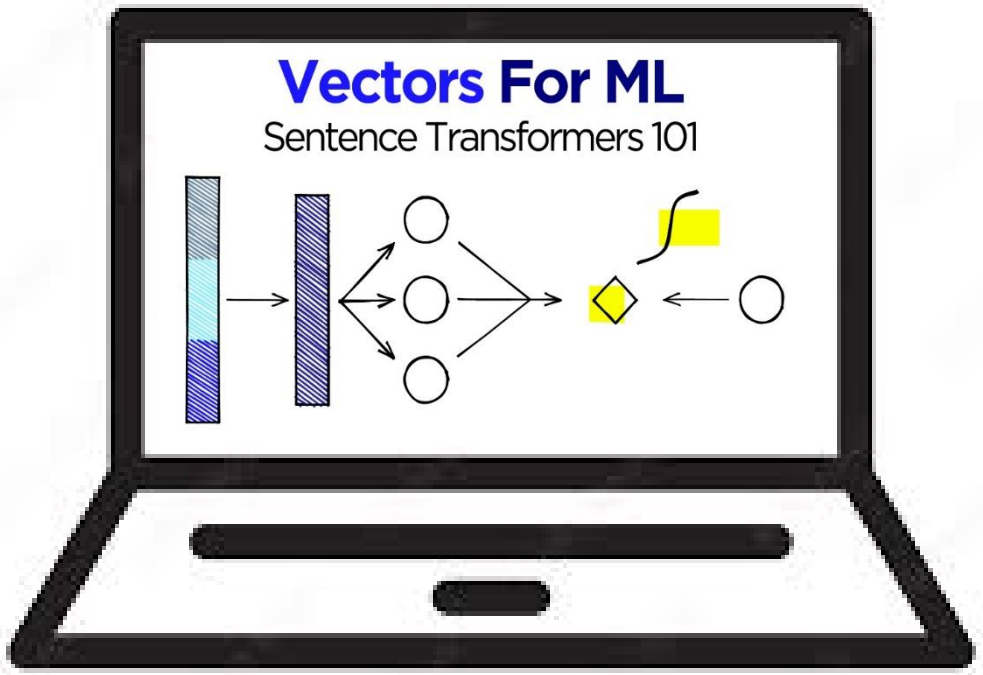
- **Discover features** of books by using embeddings
- **Locate** relevant books based on user's query
- **Compare similarities** between books,
- **Summarize trends** in the recommended books data
- **Derive** insights from the data for the recommendation system.

What –Why – **How**

- The book titles are used to generate embeddings for each book using **Sentence Transformer**.

- The embeddings are used to calculate the similarity between books using **Semantic Search**.

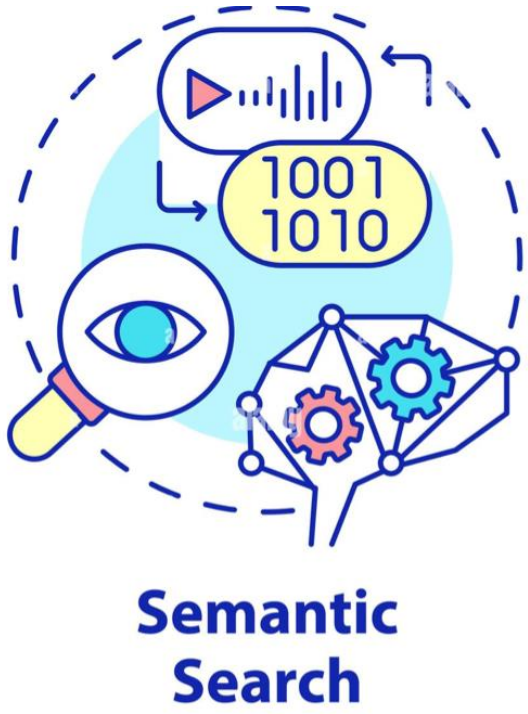
What –Why – **How**



Source :Sentence-Embeddings

- **Sentence Transformer** is a library that allows for the generation of sentence embeddings for natural language processing tasks.
- **BERT** (Bidirectional Encoder Representations from Transformers) is based neural network model that is used to generate sentence embeddings.
- It can be fine-tuned for various natural language processing tasks such as text **classification, question answering**, and more.

What –Why – **How**



- **Semantic search** uses natural language processing techniques such as word embeddings to understand the intent behind a user's query and return semantically related results.
- It takes into account context and synonyms to provide more accurate results.
- The **Sentence Transformer** library uses **cosine similarity** as the default similarity measure when generating embeddings.

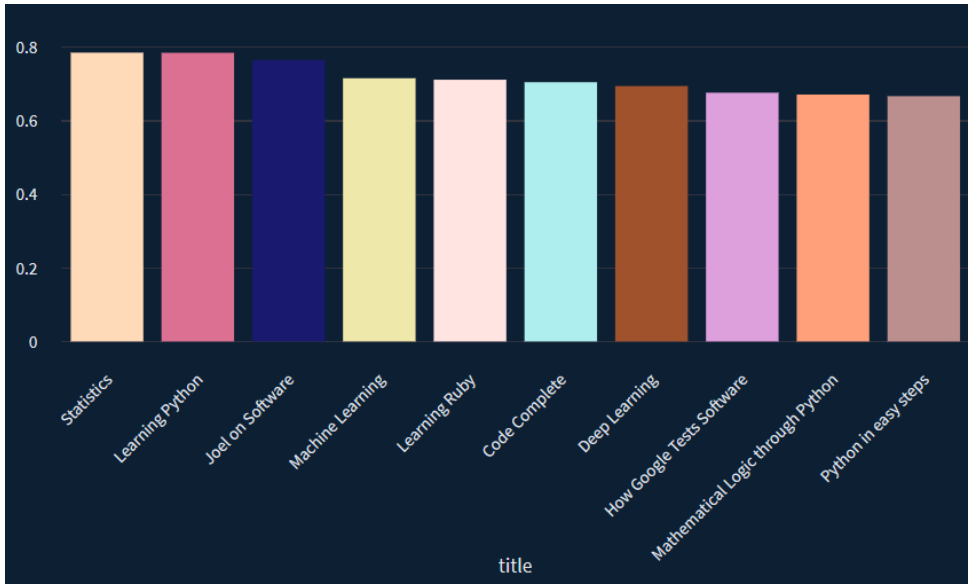
Design Choice

What (**Data**):	Book titles(Categorical – Nominal), ratings(Quantitative – Ordinal)
Why (**Tasks**):	Discover features, Locate, Compare similarities, Summarize trends , Derive
How (**Encode**):	Embeddings using Sentence Transformer and BERT model.
How (**Facet**):	Grouping the recommended books by ratings
How (**Reduce**)	Using Semantic search
How (**Manipulation**)	Similarity scores to compare

Data Visualization _Bar chart

What

- X-axis ‘Titles’ (One categorical)
- Y-axis ‘Similarity Score’ (One quantitative)



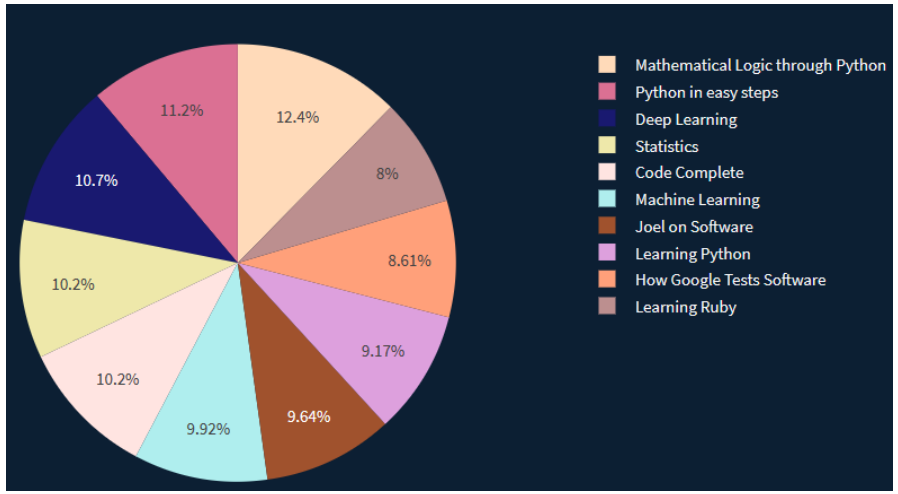
How

- **Channels**：
  - Position channel: The x-axis and y-axis will be used to position the bars.
  - Length channel: The length of the bars will represent the similarity scores.
  - Color channel: Different colors can be used to distinguish between different books
- **Marks**：
  - Bar marks: The bars themselves will represent the data.
  - Text marks: Labels will be used to identify the book titles on the x-axis.

Data Visualization _Pie chart

What (One categorical)

- Represent proportions of different ‘Titles’



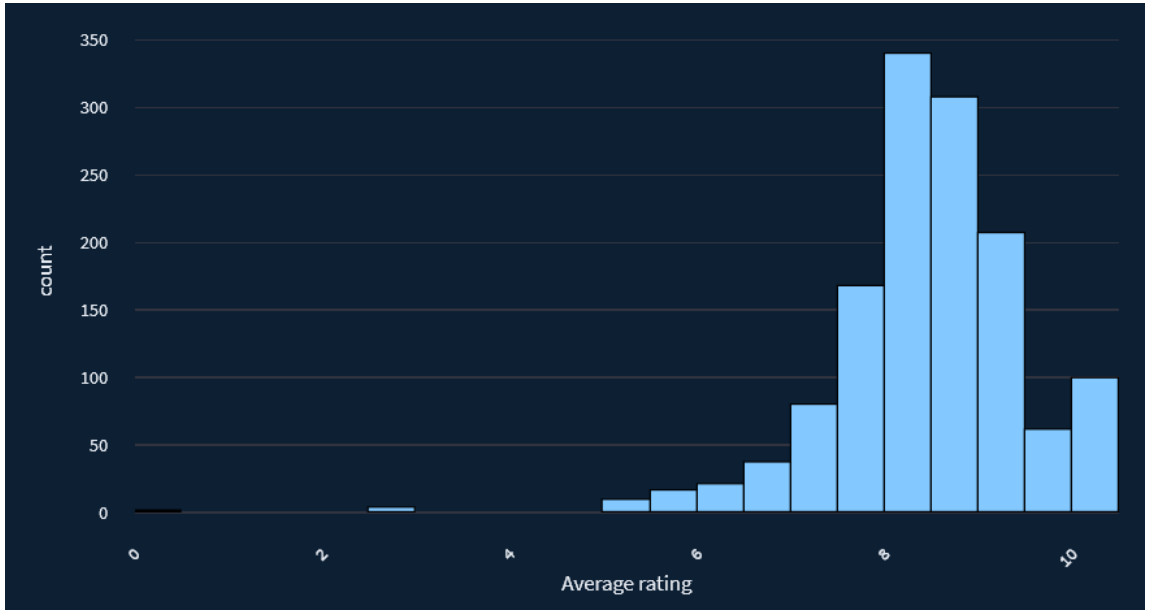
How

- **Channels** :
  - Size : The size of each wedge representing the rating of that title.
  - Color channel: Different colors can be used to distinguish between different books
- **Marks** :
  - The different sections of the pie chart representing each title

Data Visualization _ Histogram

What (2 Quantitative)

- Distribution of a dataset.



How

- **Channels :**

  X-axis represents the range of ratings in the dataset, Y-axis represents the frequency of those values.

- **Marks :**

  The bars of the histogram representing the frequency of a certain range of values in the dataset

Implementation

Data Collection from Kaggle

Pre-processing of data, including cleaning and tokenization

Encoding of data using Sentence Transformer and BERT model

Ranking of book titles based on similarity scores

Visualization of top 10 book titles

Output

Output:

Recommended book titles for the search query, with visual representation of similarity scores and book ratings.

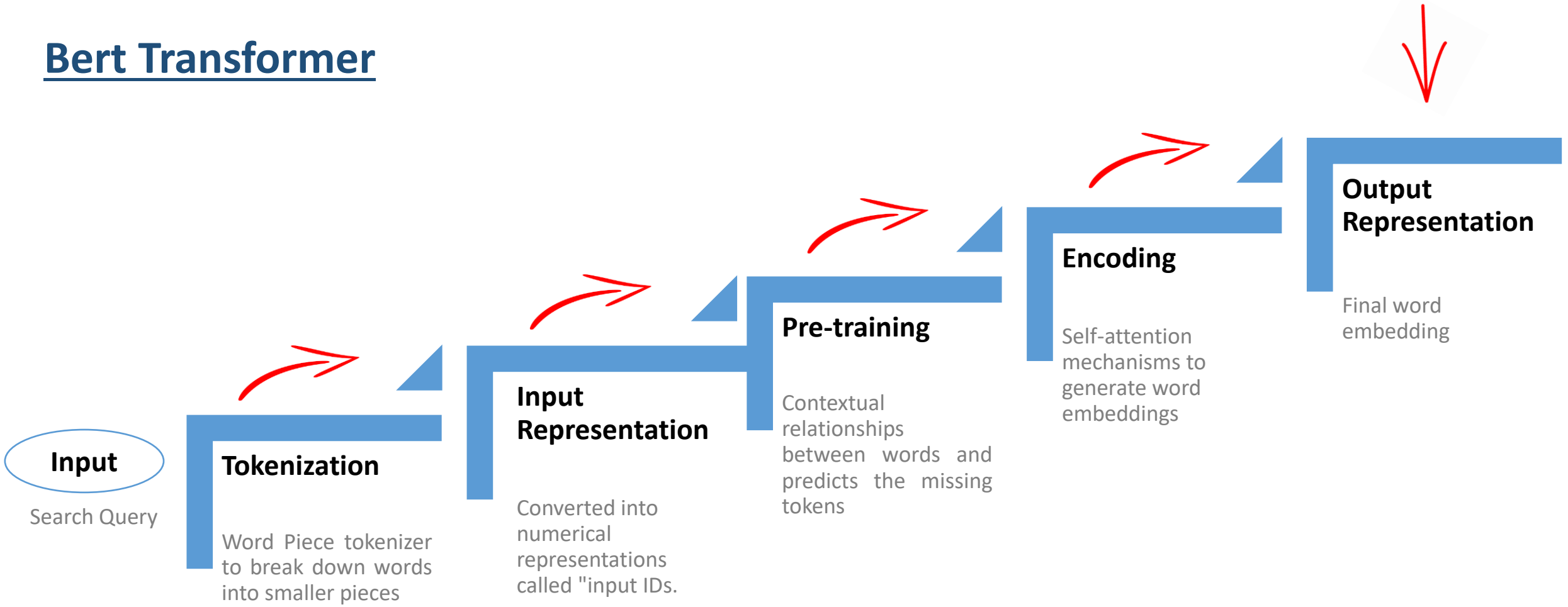
Conceptual Solution

- ✓Dataset 1
- ✓Dataset 2
- ✓Dataset 3

title	author	Rating	complete_link
Fundamentals of Data Engineering: Plan and Build Robust Data Systems	Joe Reis,Matt Housley	1	https://www.amazon.com/Fundamentals-Data-Engineering-Robust-Systems/
Effective Data Science Infrastructure: How to make data scientists productive	Ville Tuulos	1	https://www.amazon.com/Effective-Data-Science-Infrastructure-scientists/
Be Data Driven: How Organizations Can Harness the Power of Data	Jordan Morrow	1	https://www.amazon.com/Be-Data-Driven-Organizations-Harness/dp/13986
Python for Data Analysis: Data Wrangling with pandas NumPy and Jupyter	Wes McKinney	1	https://www.amazon.com/Python-Data-Analysis-Wrangling-Jupyter/dp/109
Data Science Ethics: Concepts Techniques and Cautionary Tales	David Martens	1	https://www.amazon.com/Data-Science-Ethics-Techniques-Cautionary/dp/0
Simplify Big Data Analytics with Amazon EMR: A beginners guide to learning and implementin	Sakti Mishra	1	https://www.amazon.com/Simplify-Big-Data-Analytics-Amazon/dp/1801071
SQL for Data Analytics: Harness the power of SQL to extract insights from data 3rd Edition	Jun Shan,Matt Goldwasser et al.	1	https://www.amazon.com/SQL-Data-Analytics-Harness-insights/dp/1801812
Communicating with Data: Making Your Case With Data	Carl Allchin	1	https://www.amazon.com/Communicating-Data-Making-Your-Case/dp/109

Conceptual Solution

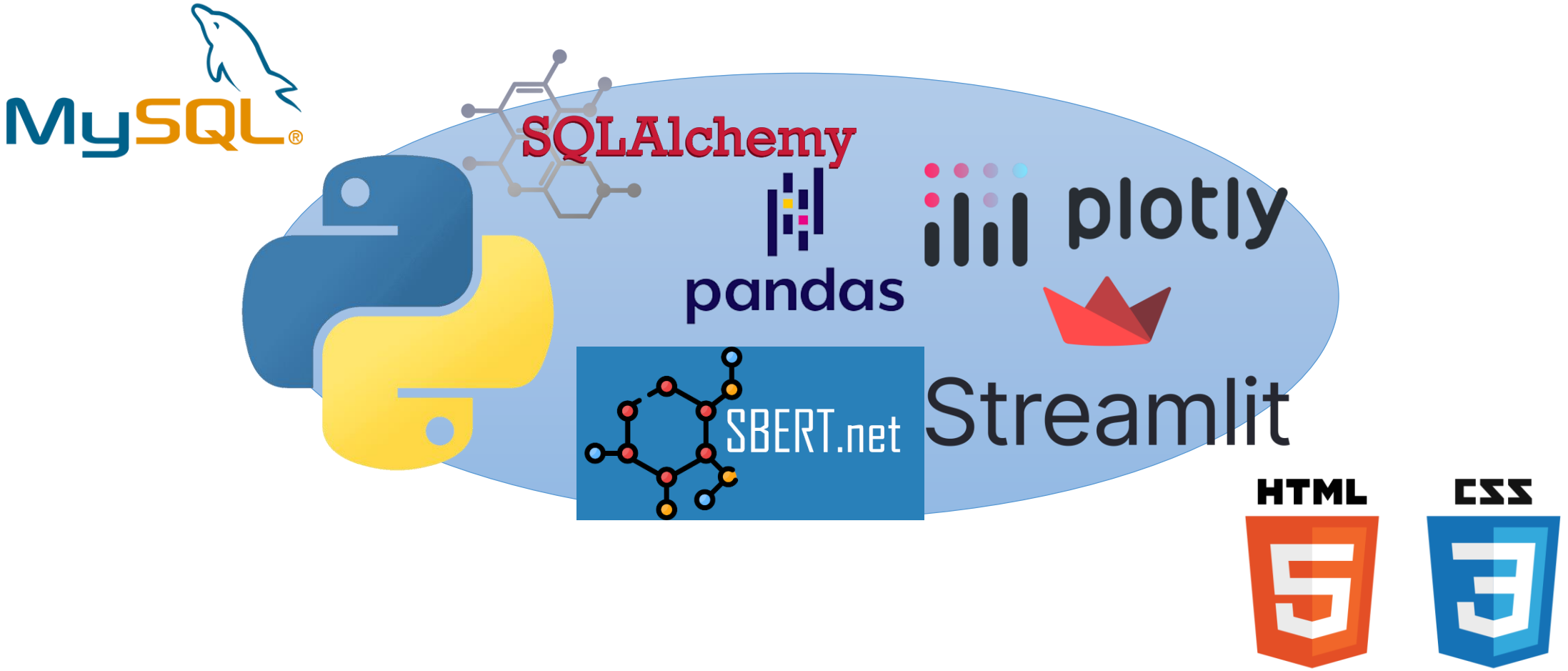
Bert Transformer



Source - [The Illustrated Transformer]

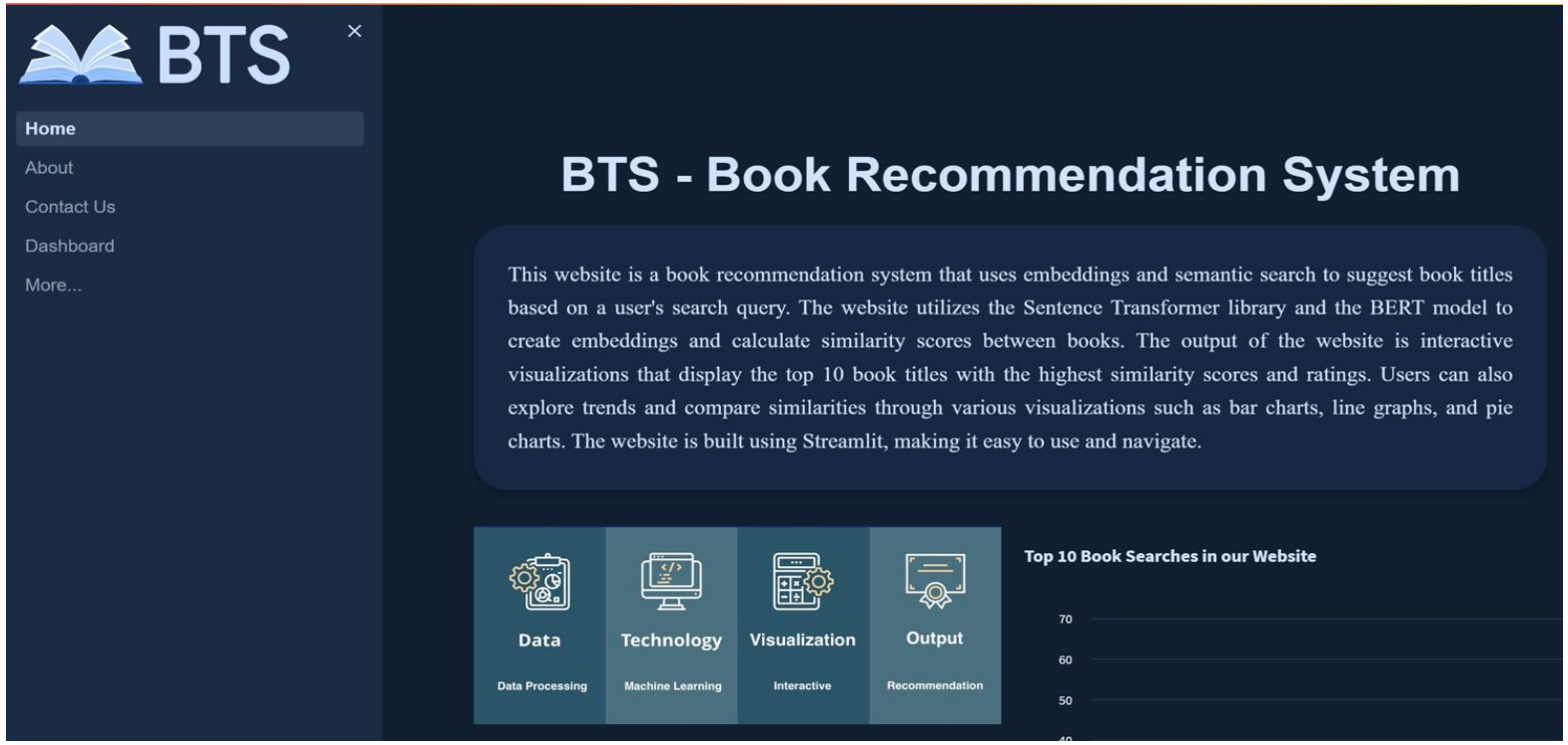
Conceptual Solution

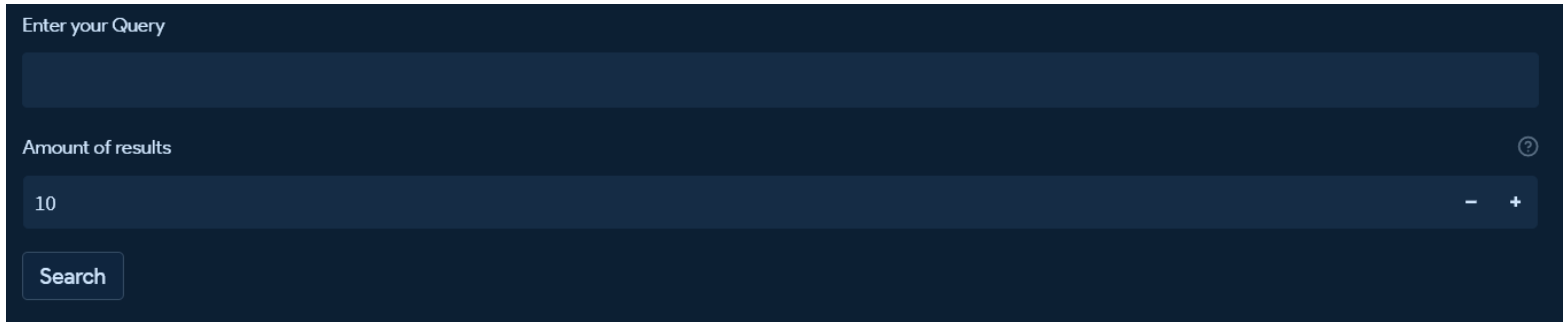
Technologies used:

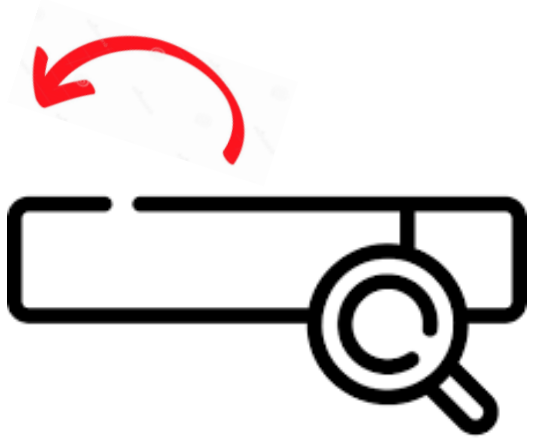


Results

Homepage







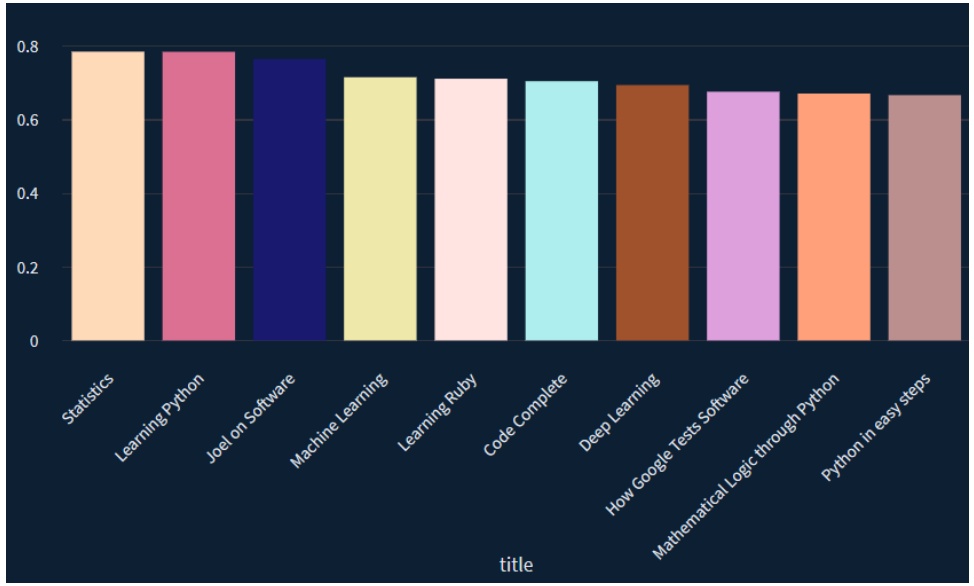
Results

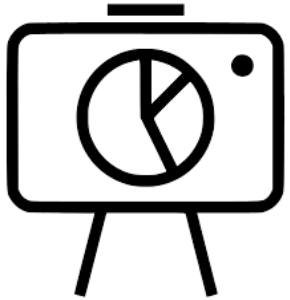
Line graph for top queries

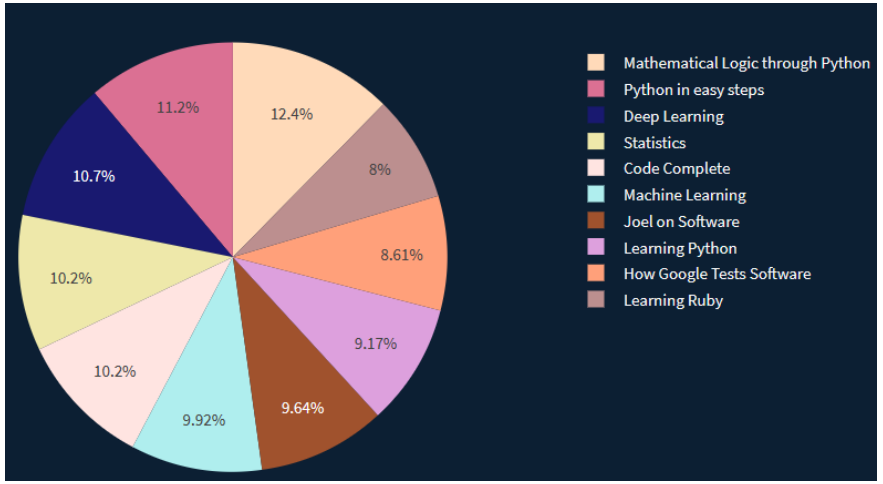
Table for top 10 books with scores

	score	title	rating
634	0.7858	Statistics	8.2500
1281	0.7854	Learning Python	7.4000
1140	0.7658	Joel on Software	7.7750
676	0.7162	Machine Learning	8.0000
1193	0.7121	Learning Ruby	6.4500
1200	0.7054	Code Complete	8.2250
1305	0.6949	Deep Learning	8.6250
1147	0.6768	How Google Tests Software	6.9500
64	0.6718	Mathematical Logic through Python	10.0000
317	0.6674	Python in easy steps	9.0000

Results

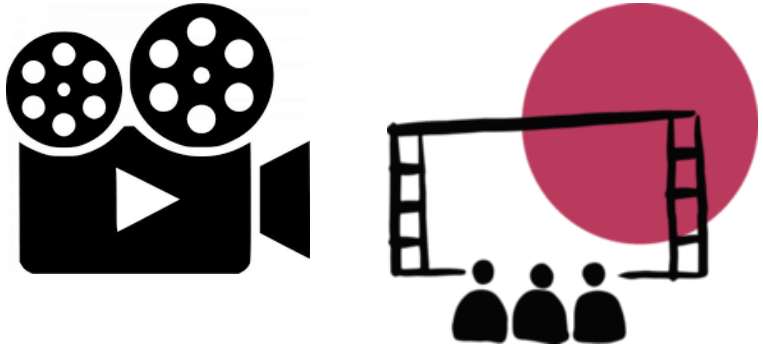

0.8
0.6
0.4
0.2
0
Statistics
Learning Python
Joel on Software
Machine Learning
Learning Ruby
Code Complete
Deep Learning
How Google Tests Software
Mathematical Logic through Python
Python in easy steps
title




11.2%
12.4%
10.7%
8%
8.61%
10.2%
9.17%
10.2%
9.92%
9.64%
Mathematical Logic through Python
Python in easy steps
Deep Learning
Statistics
Code Complete
Machine Learning
Joel on Software
Learning Python
How Google Tests Software
Learning Ruby

Results

Demo – 10 minute



Summary

Data – List of books with titles, ratings…

Providing a user-friendly book recommendation system,
and engaging way for users

Embeddings, Interactive Visualization, Web application

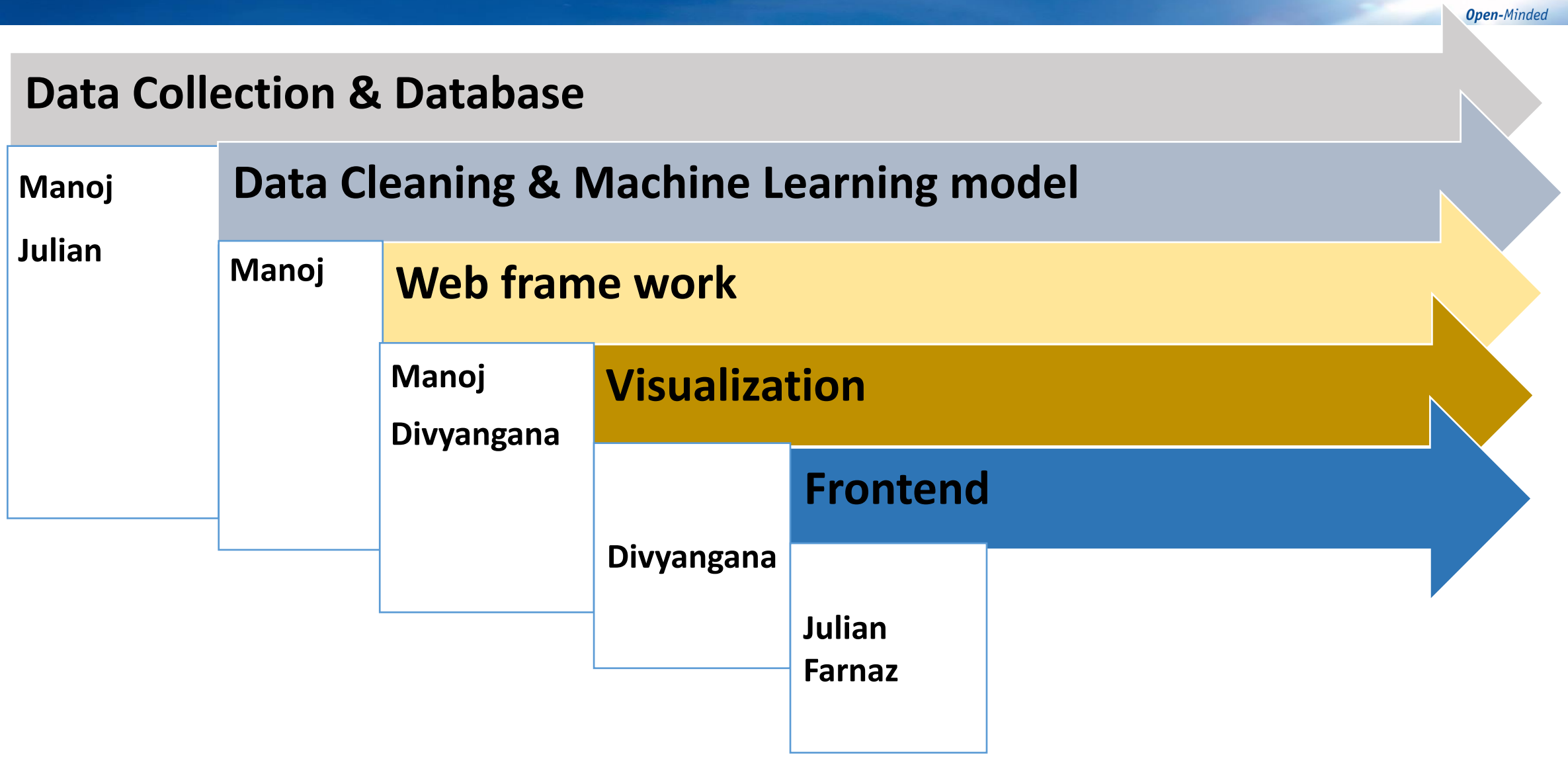
Students , Who are looking for recommendations on books

Summary

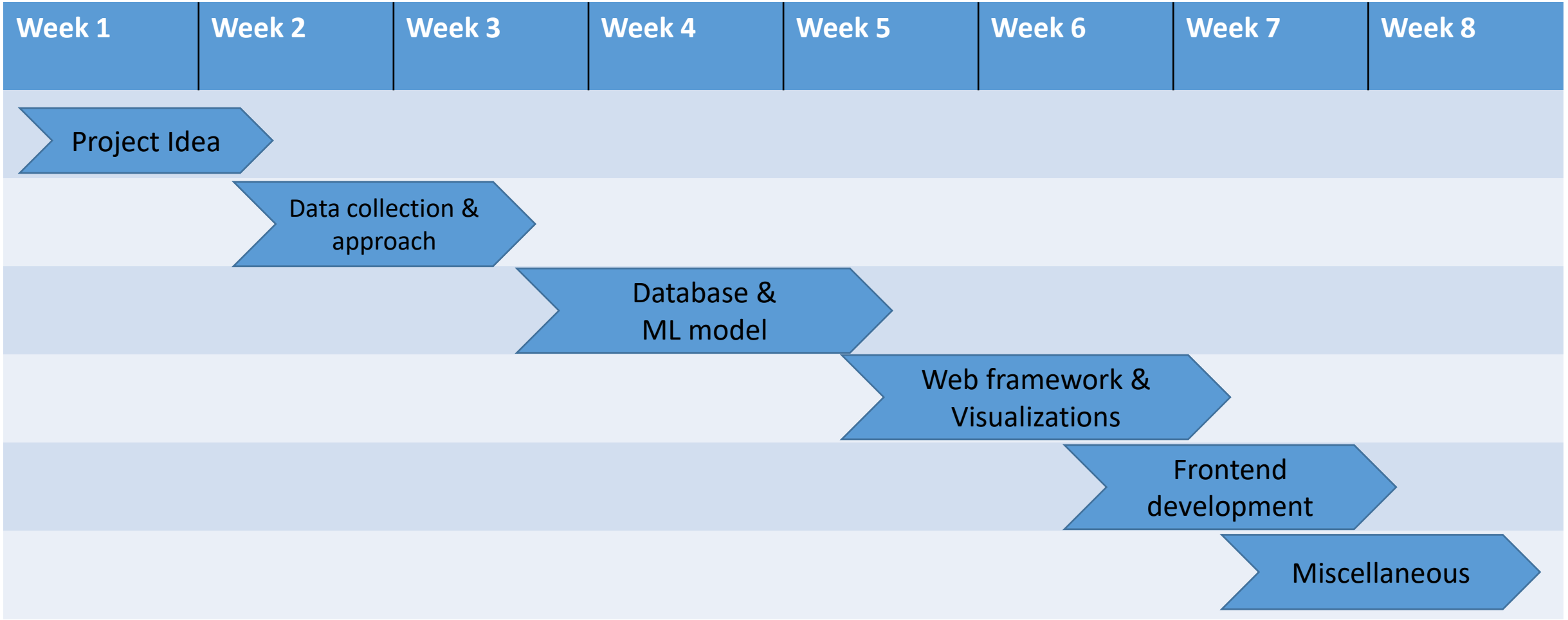
Future Improvements

- Improving the recommendation **additional features** such as book genres, description, and user preferences.
- **Expanding the database** of books to include a larger variety of titles.
- Improving the interface to make it **more user-friendly**.

Task Allocation



Schedule


Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Project Idea
Data collection & approach
Database & ML model
Web framework & Visualizations
Frontend development
Miscellaneous

References

GitHub https://github.com/Manojkdara/Books-to-Search-BTS-

ChatGPT https://chat.openai.com/chat

Pretrained models https://huggingface.co/models

Streamlit https://docs.streamlit.io/

MySQL https://www.w3schools.com/mysql/default.asp

Plotly https://www.geeksforgeeks.org/python-plotly-tutorial/

Dashboard https://startbootstrap.com/theme/sb-admin-2

Streamlit Tutorials https://www.youtube.com/playlist?list=PLuU3eVwK0I9PT48ZBYAHdKPFazhXg76h5

Lecture Slides Lecture notes of Prof. Dr. Mohamed Amine Chatti on **Learning Analytics**

thank you.