Santhosh Andavar

2818372

CIS 492(Big Data)

04/20/2025

Lab4 part 2

In the second part of the lab, we computed the cosine similarity between 10 selected documents from the State of the Union dataset using the TF-IDF vector space model.

Steps:

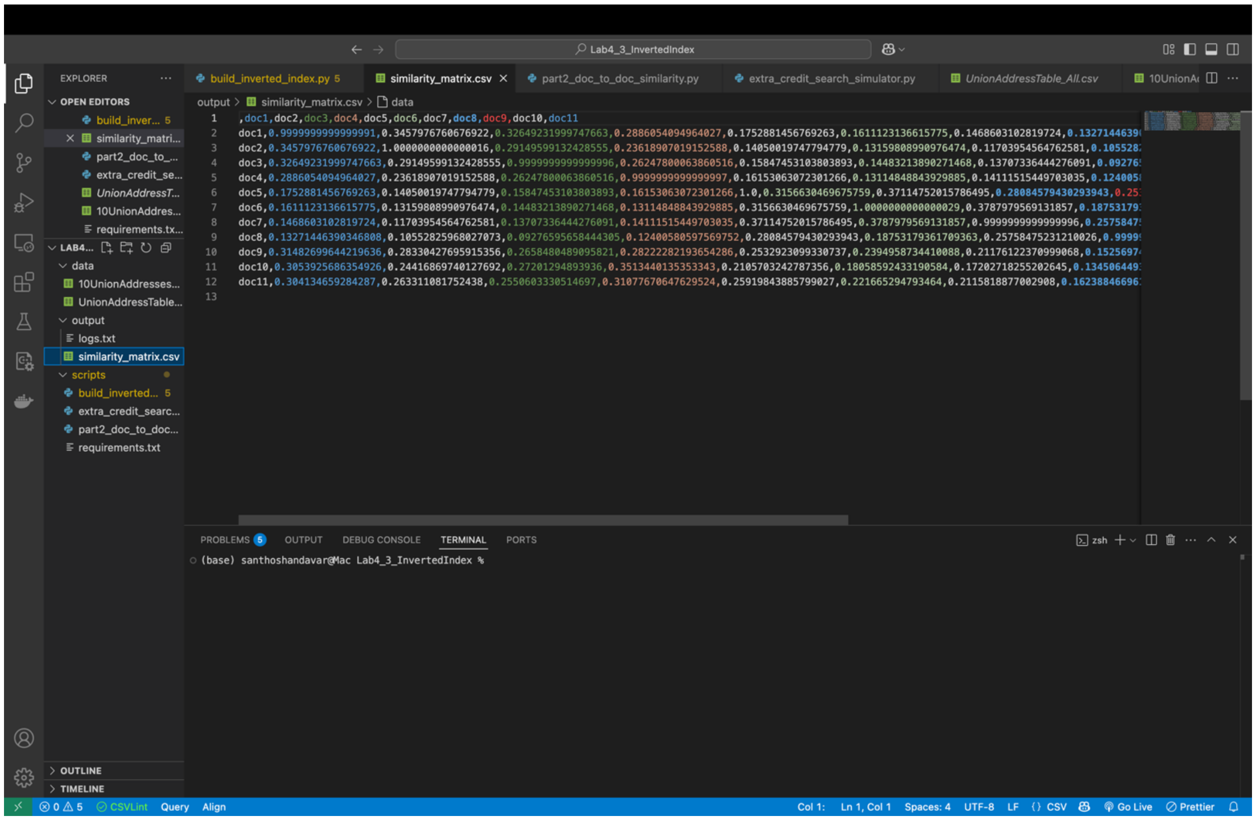
- Used TfidfVectorizer from scikit-learn to compute document vectors
- Computed pairwise cosine similarity across documents
- Saved a 10x10 similarity matrix to CSV format

Query: `freedom peace economy`

Top 5 Similar Documents:

- doc5
- doc2
- doc6
- doc3
- doc7

Cosine similarity matrix showing pairwise document similarity across the 10 State of the Union addresses using TF-IDF vectors.



```
,doc1,doc2,doc3,doc4,doc5,doc6,doc7,doc8,doc9,doc10,doc11
doc1,0.9999999999999991,0.3457976760676922,0.32649231999747663,0.2886054094964027,0.1752881456769263,0.1611123136615775,0.1468603102819724,0.132714463
doc2,0.3457976760676922,1.0000000000000016,0.29149599132428555,0.23618907019152588,0.14050019747794779,0.13159808990976474,0.11703954564762581,0.105528
doc3,0.32649231999747663,0.29149599132428555,0.9999999999999996,0.26247800063860516,0.15847453103803893,0.14483213890271468,0.13707336444276091,0.09276
doc4,0.2886054094964027,0.23618907019152588,0.26247800063860516,0.9999999999999997,0.16153063072301266,0.13114848843929885,0.14111515449703035,0.124005
doc5,0.1752881456769263,0.14050019747794779,0.15847453103803893,0.16153063072301266,1.0,0.3156630469675759,0.37114752015786495,0.28084579430293943,0.25
doc6,0.1611123136615775,0.13159808990976474,0.14483213890271468,0.13114848843929885,0.3156630469675759,1.0000000000000029,0.3787979569131857,0.18753179
doc7,0.1468603102819724,0.11703954564762581,0.13707336444276091,0.14111515449703035,0.37114752015786495,0.3787979569131857,0.9999999999999996,0.2575847
doc8,0.13271446390346808,0.10552825968027073,0.09276595658444305,0.12400580597569752,0.28084579430293943,0.18753179361709363,0.25758475231210026,0.9999
doc9,0.31482699644219636,0.28330427695915356,0.2658480489095821,0.28222282193654286,0.2532923099330737,0.23949587344410088,0.21176122370999068,0.152569
doc10,0.3053925686354926,0.24416869740127692,0.27201294893936,0.3513440135353343,0.2105703242787356,0.18058592433190584,0.17202718255202645,0.13450644
doc11,0.304134659284287,0.263311081752438,0.2550603330514697,0.31077670647629524,0.25919843885799027,0.221665294793464,0.2115818877002908,0.1623884669
```

```
(base) santhoshandavar@Mac Lab4_3_InvertedIndex %
```

Extra Credit: Web-Based Search Engine with User Input

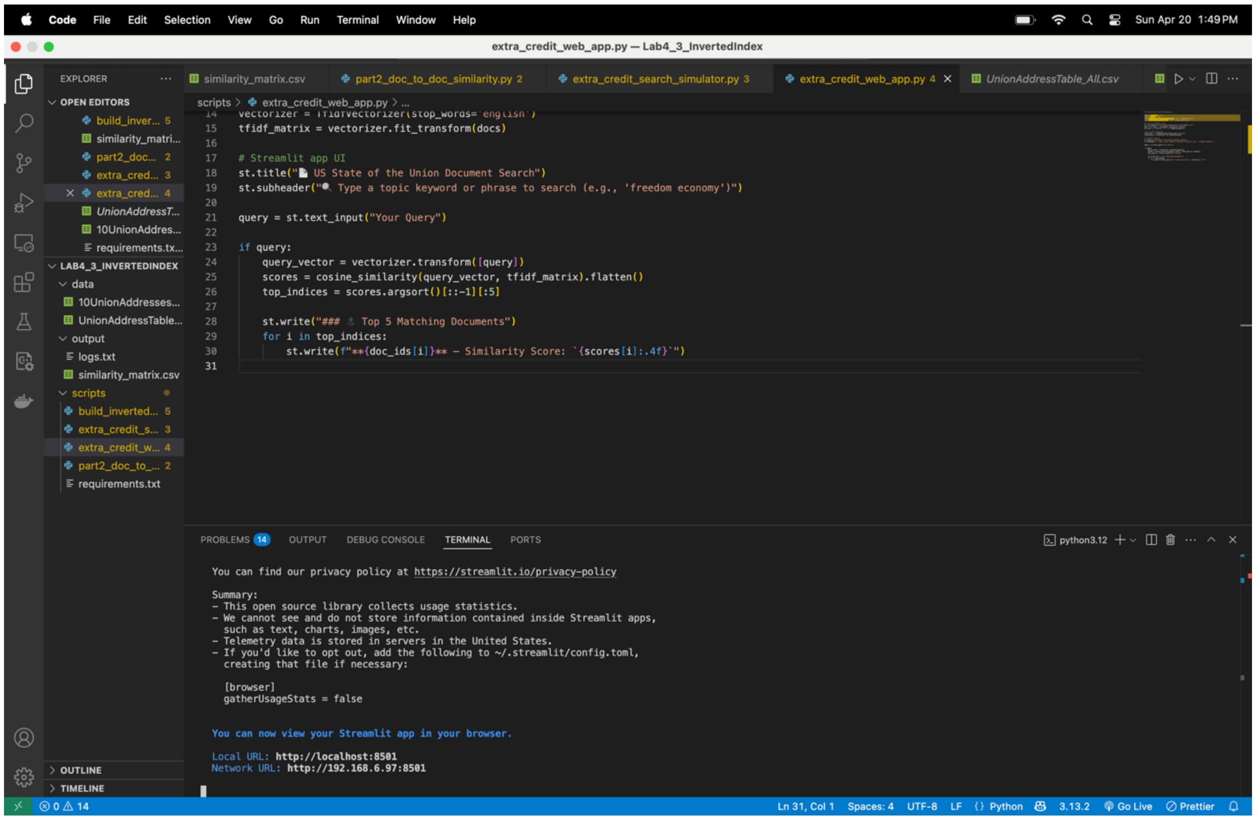
A Streamlit web application was developed to simulate a real-time search engine. Users can enter a query, and the system returns the top 5 most relevant documents using TF-IDF and cosine similarity.

- Input box captures keywords like "freedom peace economy"
- Matches query vector to document vectors
- Displays most similar speeches and scores

Queries Used for Screenshots:

1. `freedom rights democracy`
2. `economy inflation growth`
3. `war peace treaty`
4. `education science innovation`
5. `tax revenue budget`

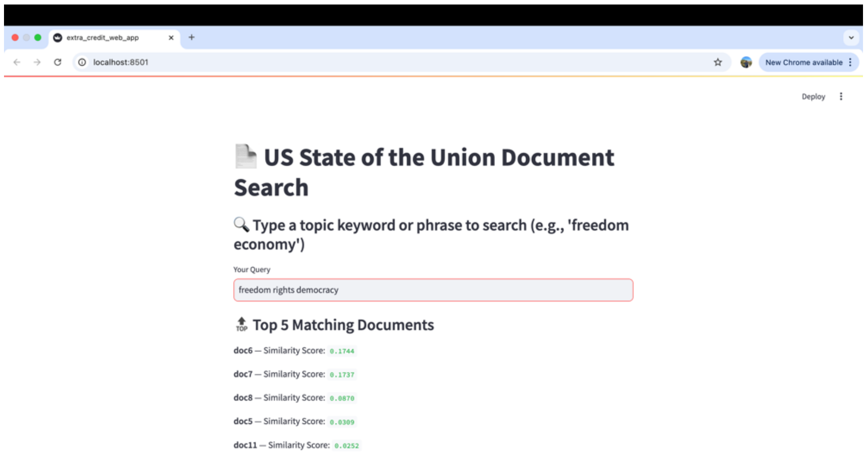
Real-time Streamlit app matching user queries against speech documents using cosine similarity.



```python
vectorizer = TfidfVectorizer(stop_words='english')
tfidf_matrix = vectorizer.fit_transform(docs)

# Streamlit app UI
st.title("📄 US State of the Union Document Search")
st.subheader("🔍 Type a topic keyword or phrase to search (e.g., 'freedom economy')")

query = st.text_input("Your Query")

if query:
    query_vector = vectorizer.transform([query])
    scores = cosine_similarity(query_vector, tfidf_matrix).flatten()
    top_indices = scores.argsort()[::-1][:5]

    st.write("### 🔝 Top 5 Matching Documents")
    for i in top_indices:
        st.write(f"**{doc_ids[i]}** — Similarity Score: `{scores[i]:.4f}`")
```

```
You can find our privacy policy at https://streamlit.io/privacy-policy

Summary:
- This open source library collects usage statistics.
- We cannot see and do not store information contained inside Streamlit apps,
  such as text, charts, images, etc.
- Telemetry data is stored in servers in the United States.
- If you'd like to opt out, add the following to ~/.streamlit/config.toml,
  creating that file if necessary:

  [browser]
  gatherUsageStats = false


You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.6.97:8501
```

"Freedom rights democracy" — shows documents discussing civil liberties and governance.



📄 US State of the Union Document Search

🔍 Type a topic keyword or phrase to search (e.g., 'freedom economy')

Your Query

freedom rights democracy

🔝 Top 5 Matching Documents

doc6 — Similarity Score: `0.1744`

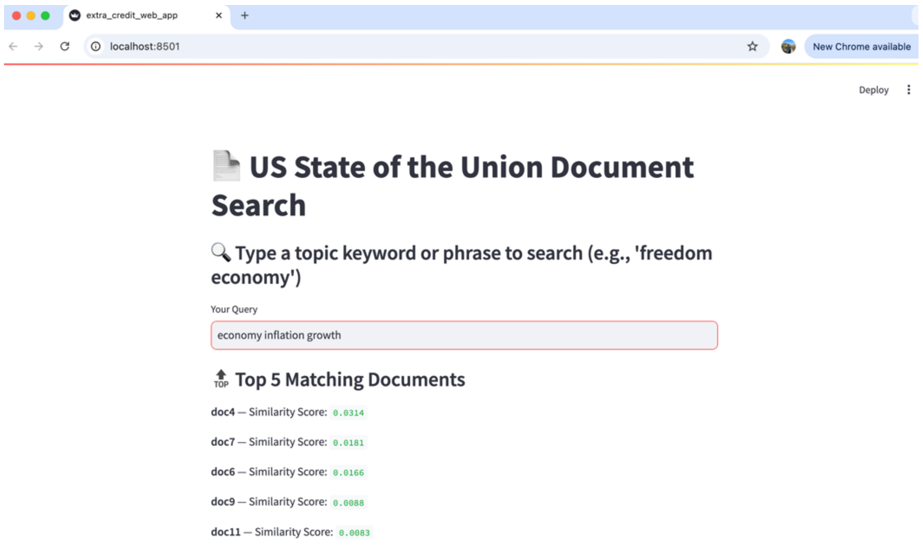
doc7 — Similarity Score: `0.1737`

doc8 — Similarity Score: `0.0870`

doc5 — Similarity Score: `0.0309`

doc11 — Similarity Score: `0.0252`

"Economy inflation growth" — returns economic-focused speeches.



US State of the Union Document Search

Type a topic keyword or phrase to search (e.g., 'freedom economy')

Your Query

economy inflation growth

Top 5 Matching Documents

doc4 — Similarity Score: `0.0314`

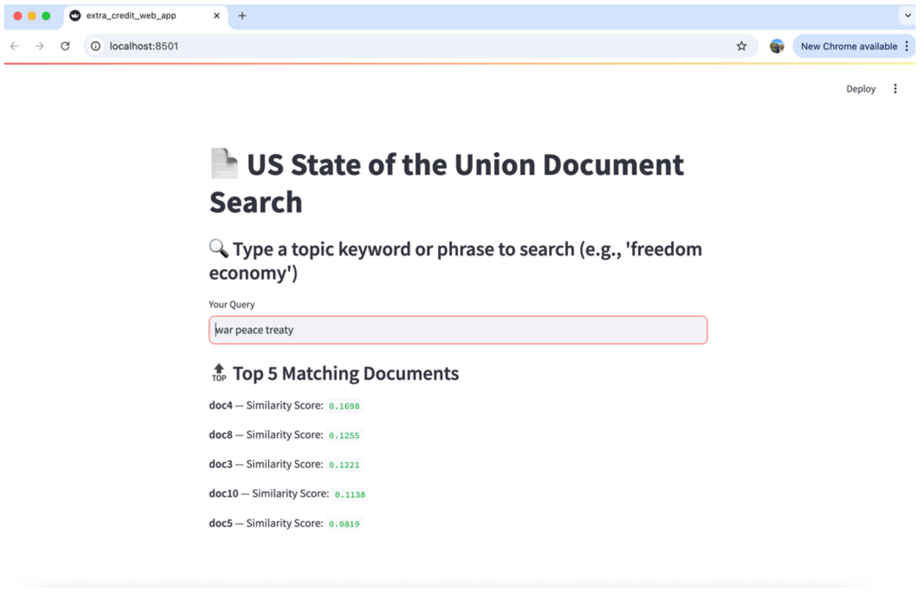
doc7 — Similarity Score: `0.0181`

doc6 — Similarity Score: `0.0166`

doc9 — Similarity Score: `0.0088`

doc11 — Similarity Score: `0.0083`

″War peace treaty" — matches documents addressing defense and foreign relations.



US State of the Union Document Search

Type a topic keyword or phrase to search (e.g., 'freedom economy')

Your Query

war peace treaty

Top 5 Matching Documents

doc4 — Similarity Score: `0.1698`

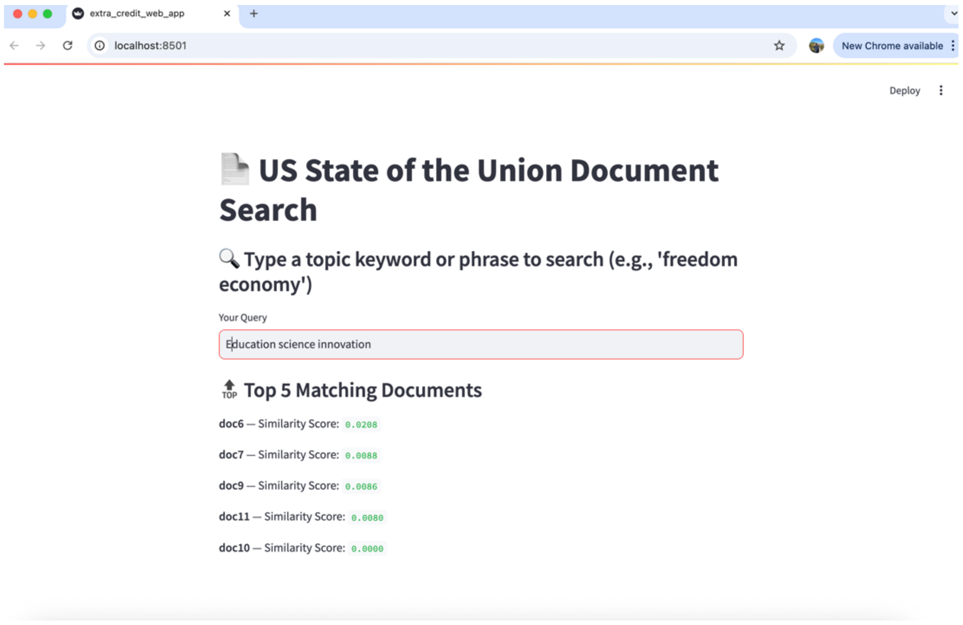
doc8 — Similarity Score: `0.1255`

doc3 — Similarity Score: `0.1221`

doc10 — Similarity Score: `0.1138`

doc5 — Similarity Score: `0.0819`

"Education science innovation" — highlights educational and innovation content



📄 US State of the Union Document Search

🔍 Type a topic keyword or phrase to search (e.g., 'freedom economy')

Your Query

Education science innovation

🔝 Top 5 Matching Documents

doc6 — Similarity Score: `0.0208`

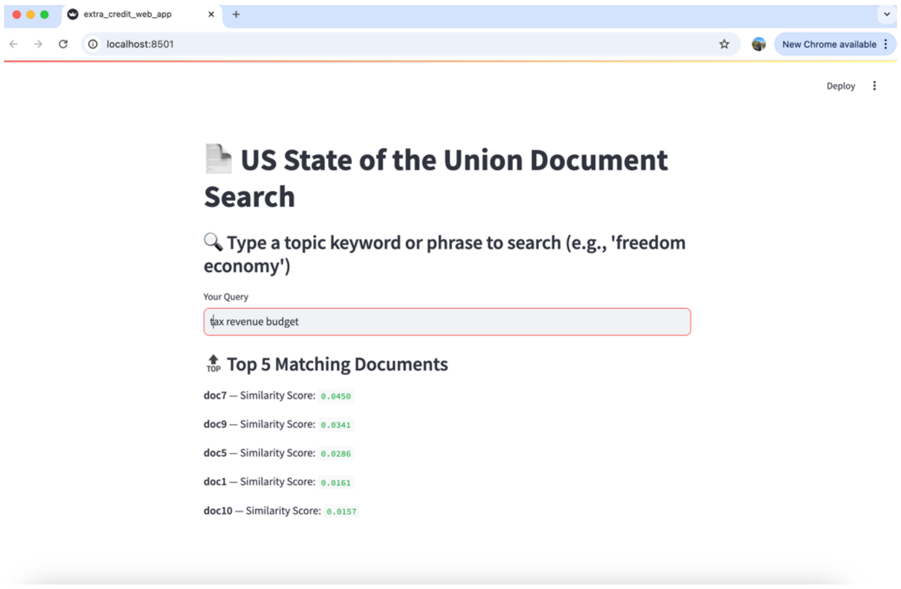
doc7 — Similarity Score: `0.0088`

doc9 — Similarity Score: `0.0086`

doc11 — Similarity Score: `0.0080`

doc10 — Similarity Score: `0.0000`

"Tax revenue budget" — captures fiscal discussions in presidential addresses.



📄 US State of the Union Document Search

🔍 Type a topic keyword or phrase to search (e.g., 'freedom economy')

Your Query

tax revenue budget

🔝 Top 5 Matching Documents

doc7 — Similarity Score: `0.0450`

doc9 — Similarity Score: `0.0341`

doc5 — Similarity Score: `0.0286`

doc1 — Similarity Score: `0.0161`

doc10 — Similarity Score: `0.0157`

Conclusion:

This lab successfully demonstrated building an inverted index and computing document similarity using TF-IDF and cosine similarity. The integration of NLP techniques and a web-based query system allowed us to explore real-time content-based search with extra credit features, making this a comprehensive document search pipeline.