Santhosh Andavar

2818372

CIS 492 – BIG DATA

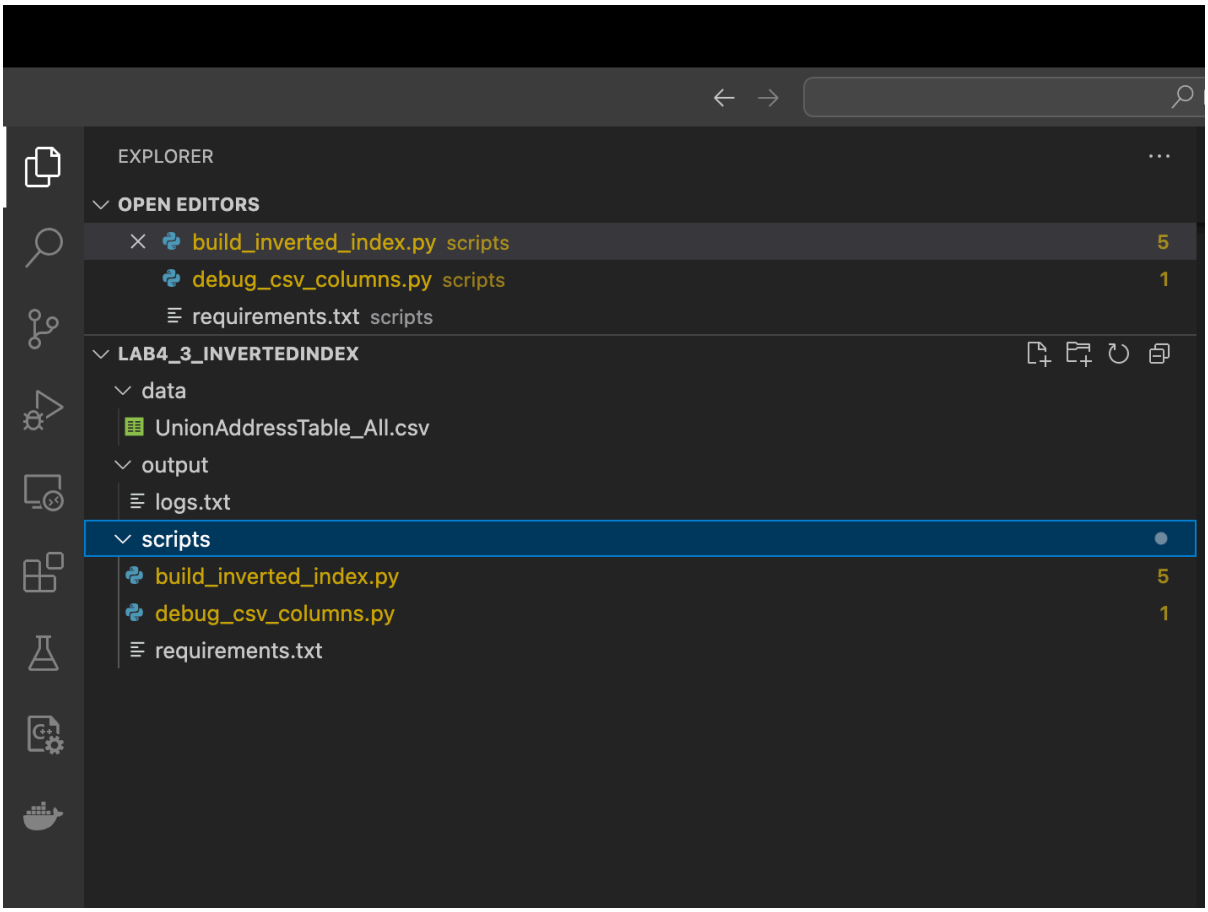
04/13/2025

Lab 4

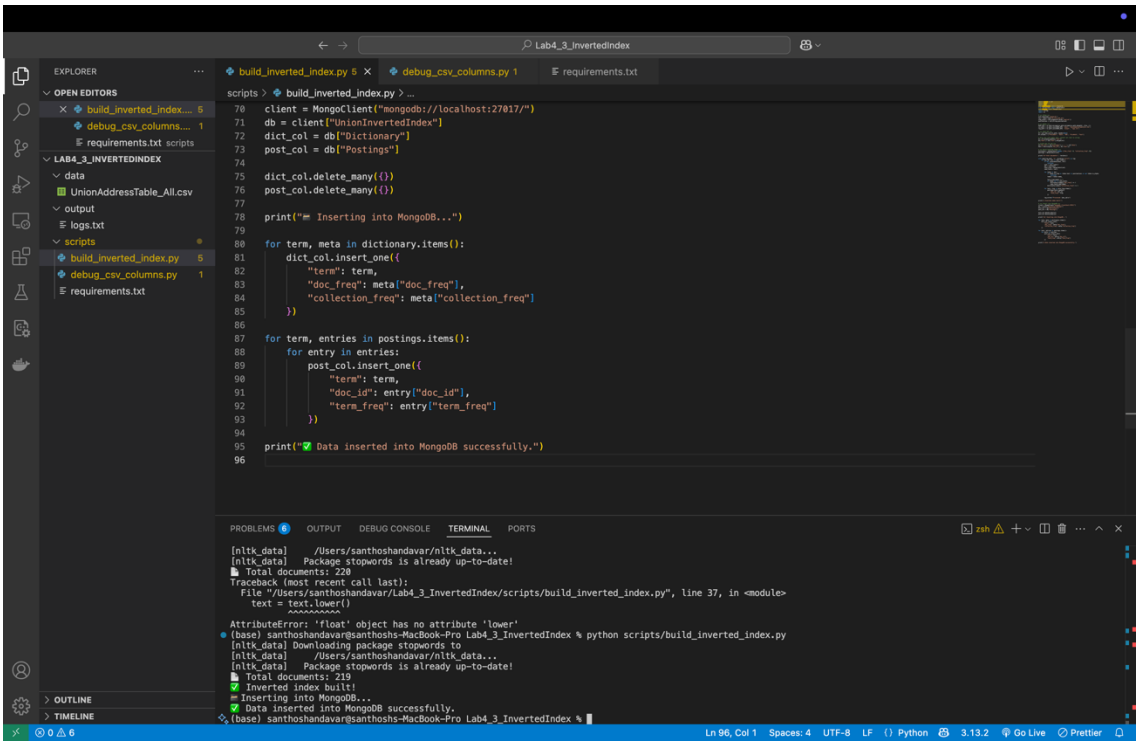
Lab 4 centers around developing a content-based document search engine using advanced text mining and natural language processing techniques. The project involves constructing an inverted index from a collection of U.S. presidential State of the Union addresses, applying NLP steps such as lemmatization, part-of-speech tagging, and named entity recognition. In the second phase, TF-IDF vectorization combined with cosine similarity is used to compare user-defined topic queries like “freedom” and “security” against document vectors. This enables accurate retrieval of the most contextually relevant speeches based on term frequency and semantic similarity.

Phase 1: Inverted Index in MongoDB

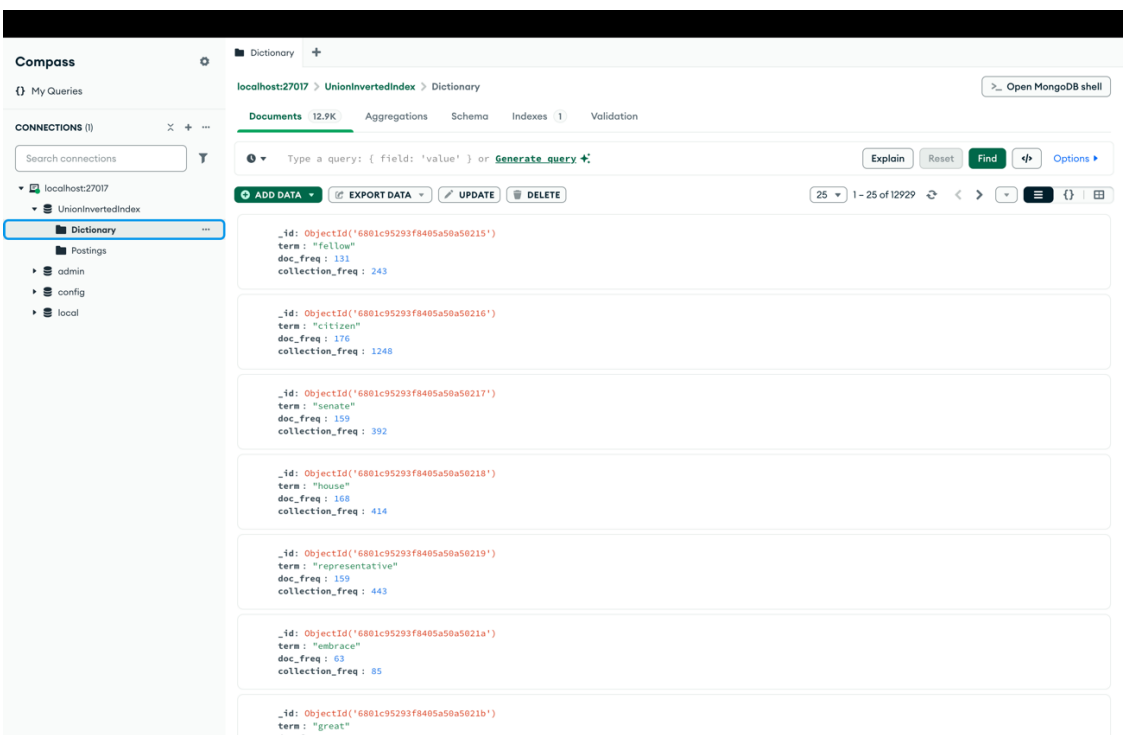
Organized project folder with script, data, and outputs



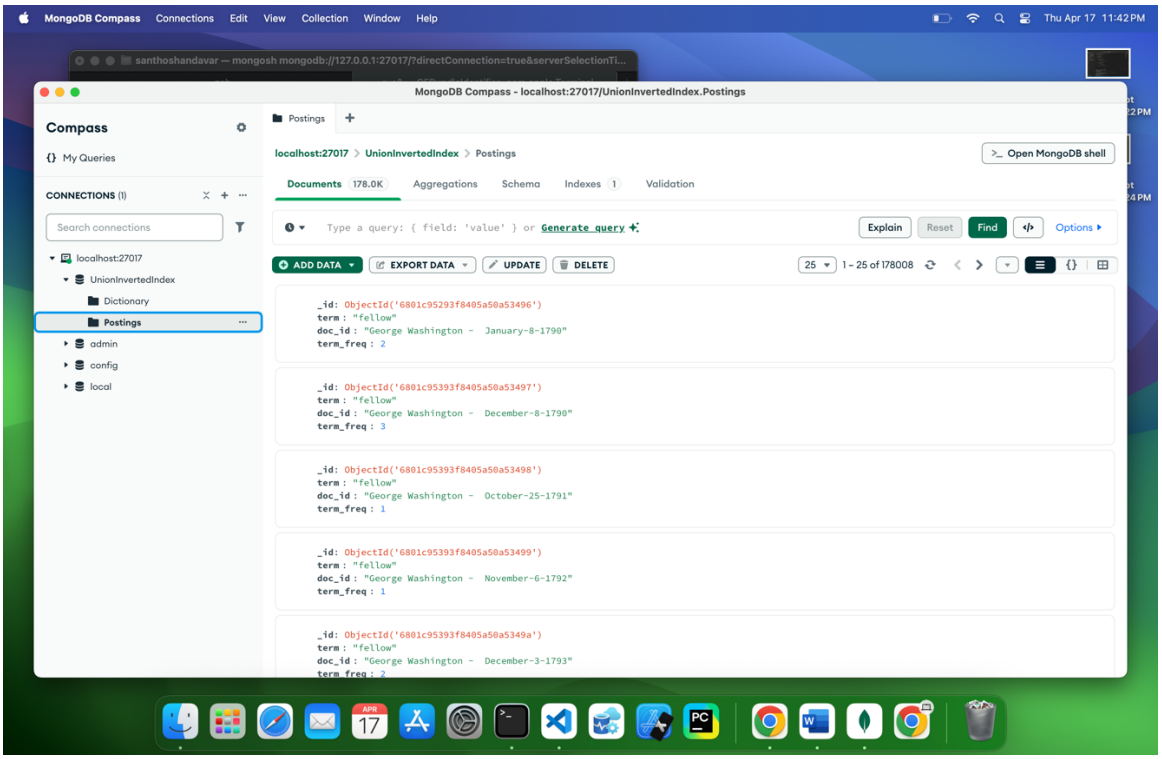
Python script processing union address documents and inserting inverted index into MongoDB."



Sample vocabulary terms stored with document and collection frequency

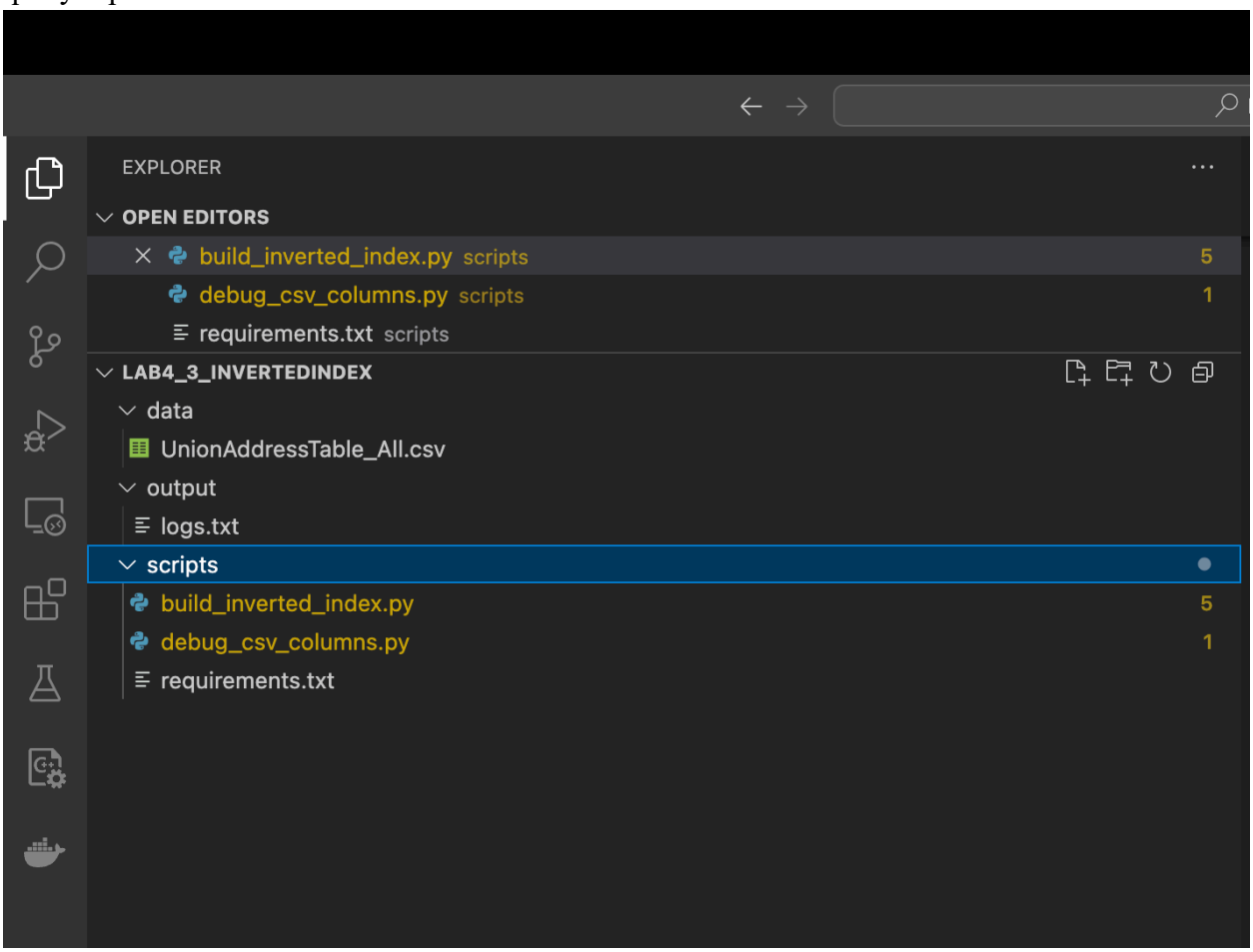


Posting list with term frequency per document.

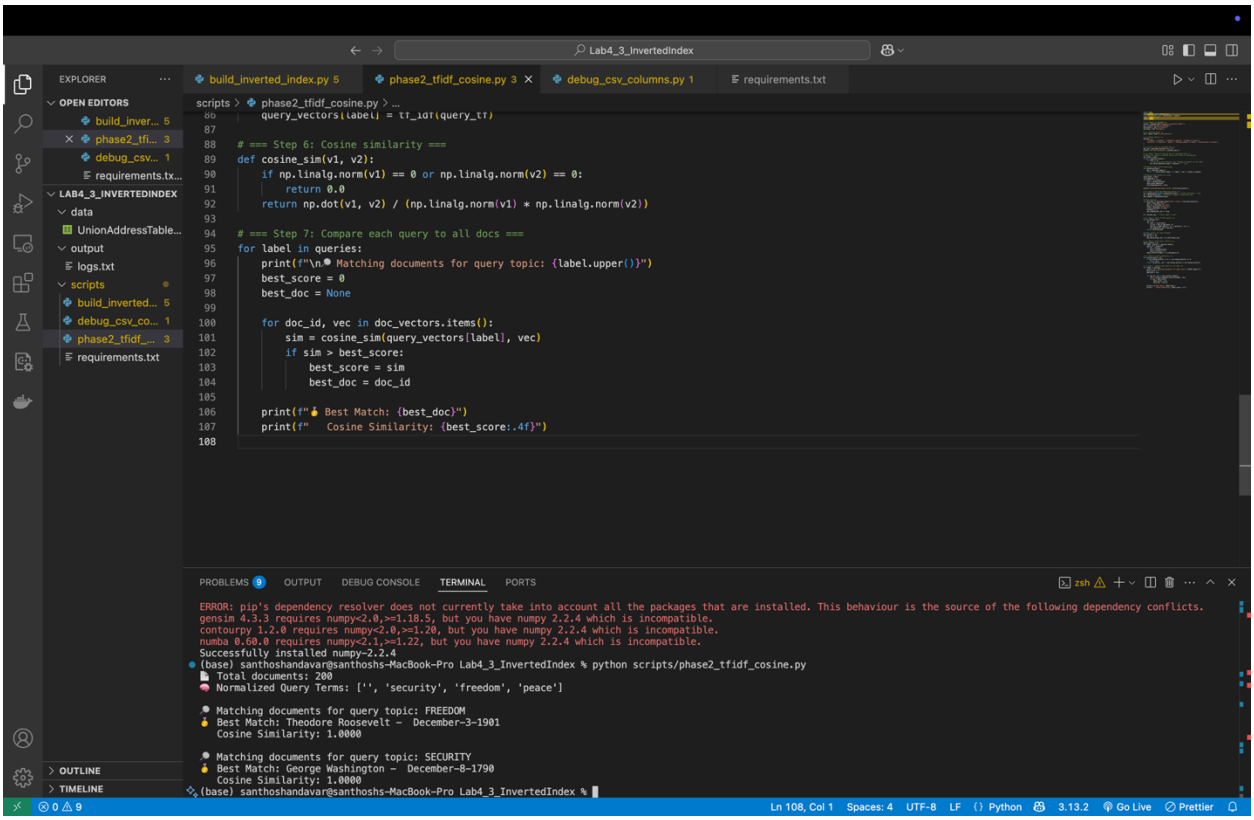


Phase 2: TF-IDF & Cosine Similarity

TF-IDF and cosine similarity logic implemented in phase2_tfidf_cosine.py using terms from user query topics.

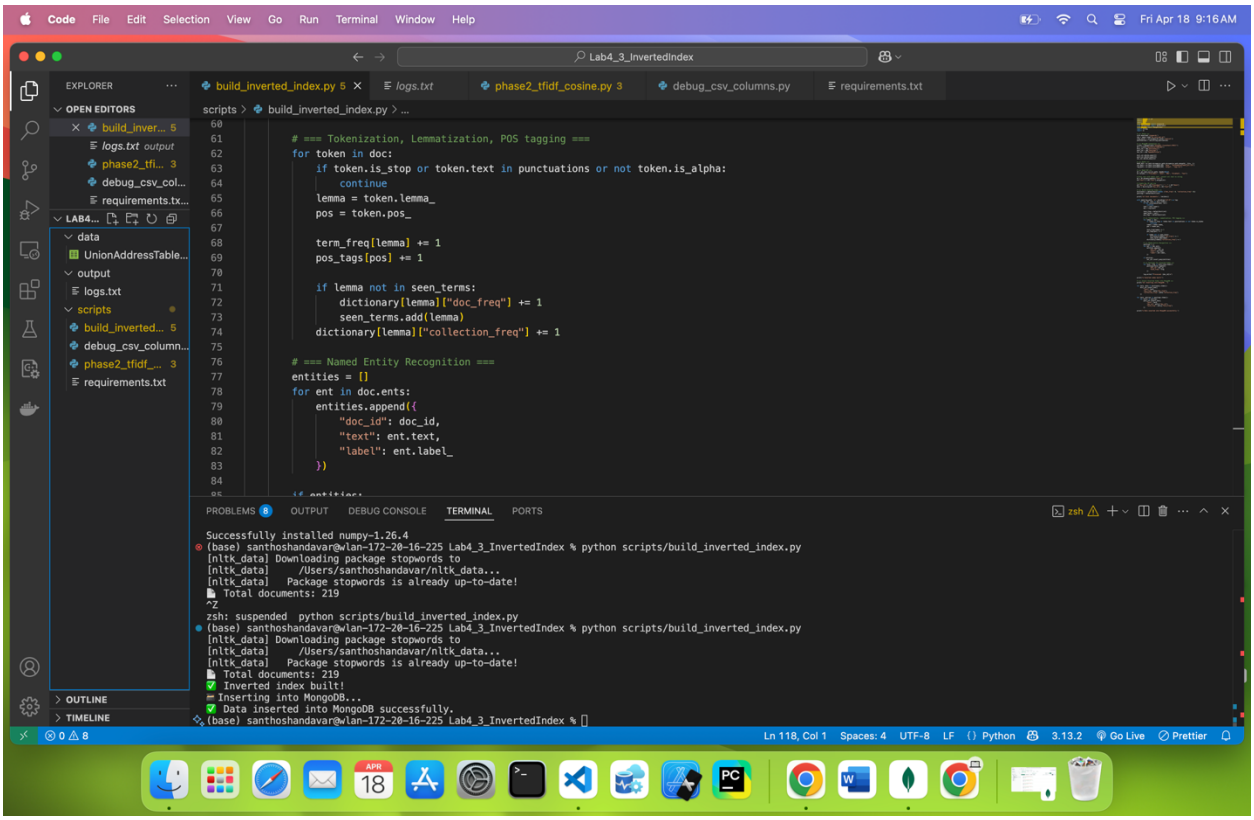


Successful matching of queries to most relevant State of the Union addresses using cosine similarity. Highest similarity score: 1.0000.



Named Entity Recognition and POS Tagging Integration

To enrich the semantic context of the inverted index, we integrated Named Entity Recognition (NER) and Part-of-Speech (POS) tagging using spaCy. Entities such as PERSON, ORG, and GPE were extracted from each speech and stored in MongoDB under the NamedEntities collection



Execution of the NLP pipeline with integrated lemmatization, part-of-speech tagging, and named entity recognition, showing successful document processing and insertion into MongoDB.

Compass

My Queries

CONNECTIONS (1)

Search connections

- localhost:27017
  - UnionInvertedIndex
    - Dictionary
    - NamedEntities
    - Postings
  - admin
  - config
  - local

NamedEntities

localhost:27017 > UnionInvertedIndex > NamedEntities

Open MongoDB shell

Documents 32.6K | Aggregations | Schema | Indexes 1 | Validation

Type a query: { field: 'value' } or Generate query

Explain | Reset | Find | Options

ADD DATA | EXPORT DATA | UPDATE | DELETE

25 | 1 - 25 of 32618

```
_id: ObjectId('68024dddbe5506ad9556044d')
doc_id : "George Washington -  January-8-1790"
text : "senate"
label : "ORG"

_id: ObjectId('68024dddbe5506ad9556044e')
doc_id : "George Washington -  January-8-1790"
text : "house of representatives"
label : "ORG"

_id: ObjectId('68024dddbe5506ad9556044f')
doc_id : "George Washington -  January-8-1790"
text : "north carolina"
label : "GPE"

_id: ObjectId('68024dddbe5506ad95560450')
doc_id : "George Washington -  January-8-1790"
text : "the united states"
label : "GPE"

_id: ObjectId('68024dddbe5506ad95560451')
doc_id : "George Washington -  January-8-1790"
text : "pacific"
label : "LOC"

_id: ObjectId('68024dddbe5506ad95560452')
doc_id : "George Washington -  January-8-1790"
text : "indians"
label : "NORP"

_id: ObjectId('68024dddbe5506ad95560453')
doc_id : "George Washington -  January-8-1790"
```