Search Engine

ULPGC

Authors: Kimberly Casimiro Torres, Víctor Gil Bernal, Jacob Jażdżyk and María Alonso León

Big Data

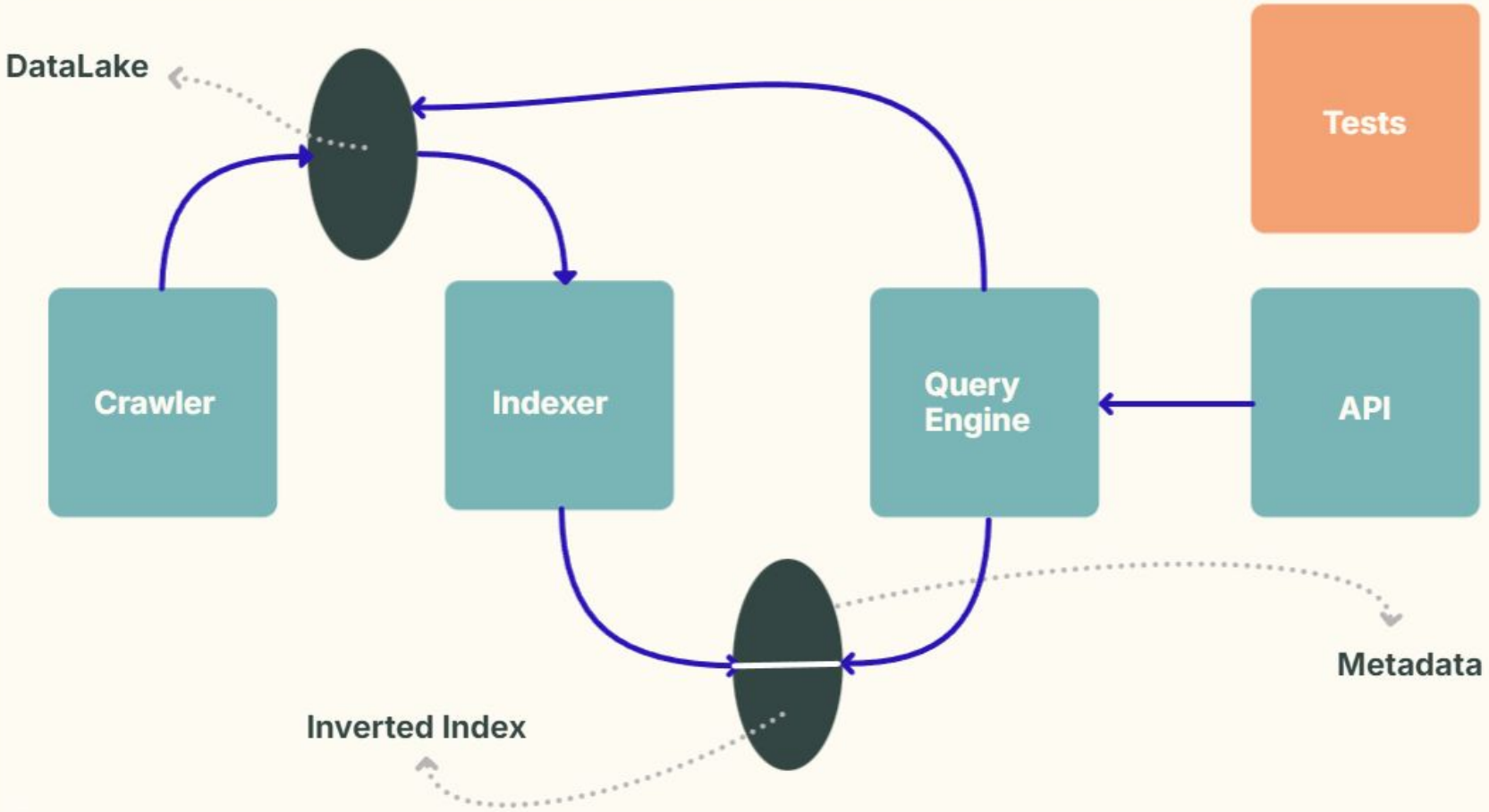
October 2024

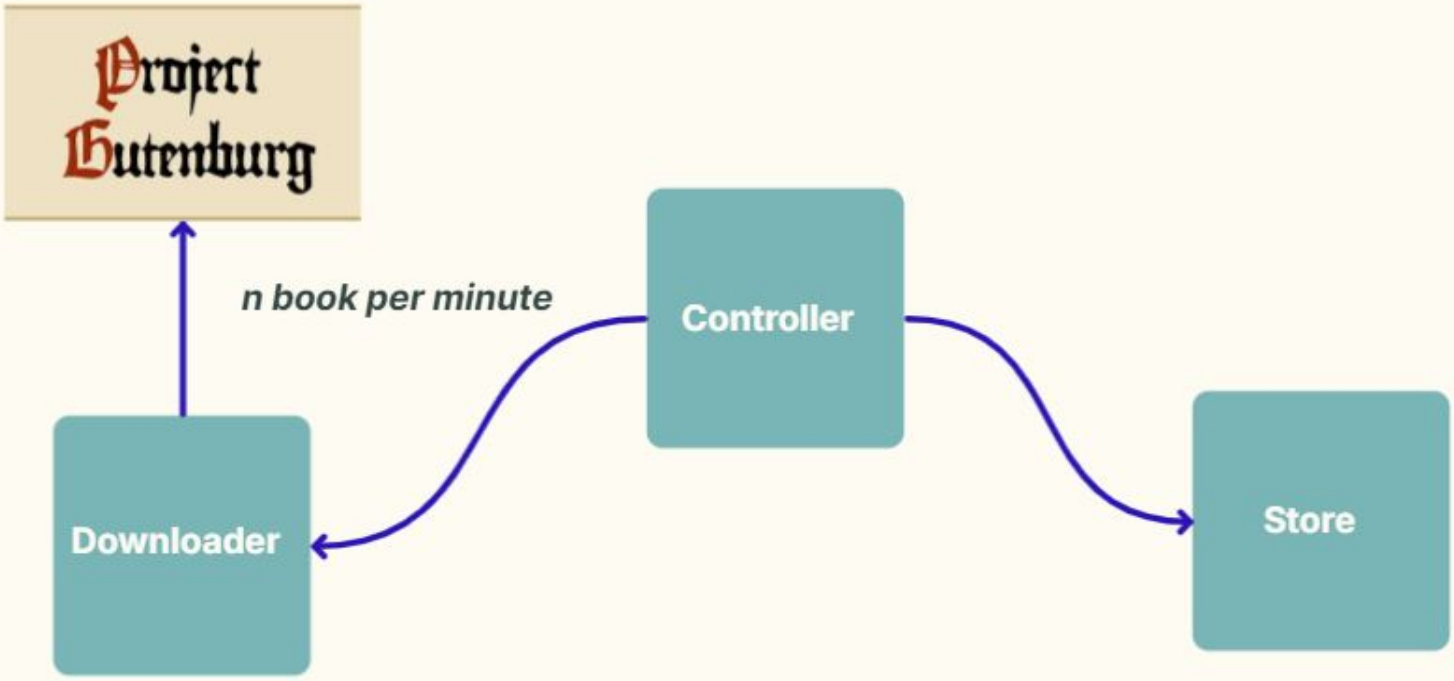
Data Science and Engineering

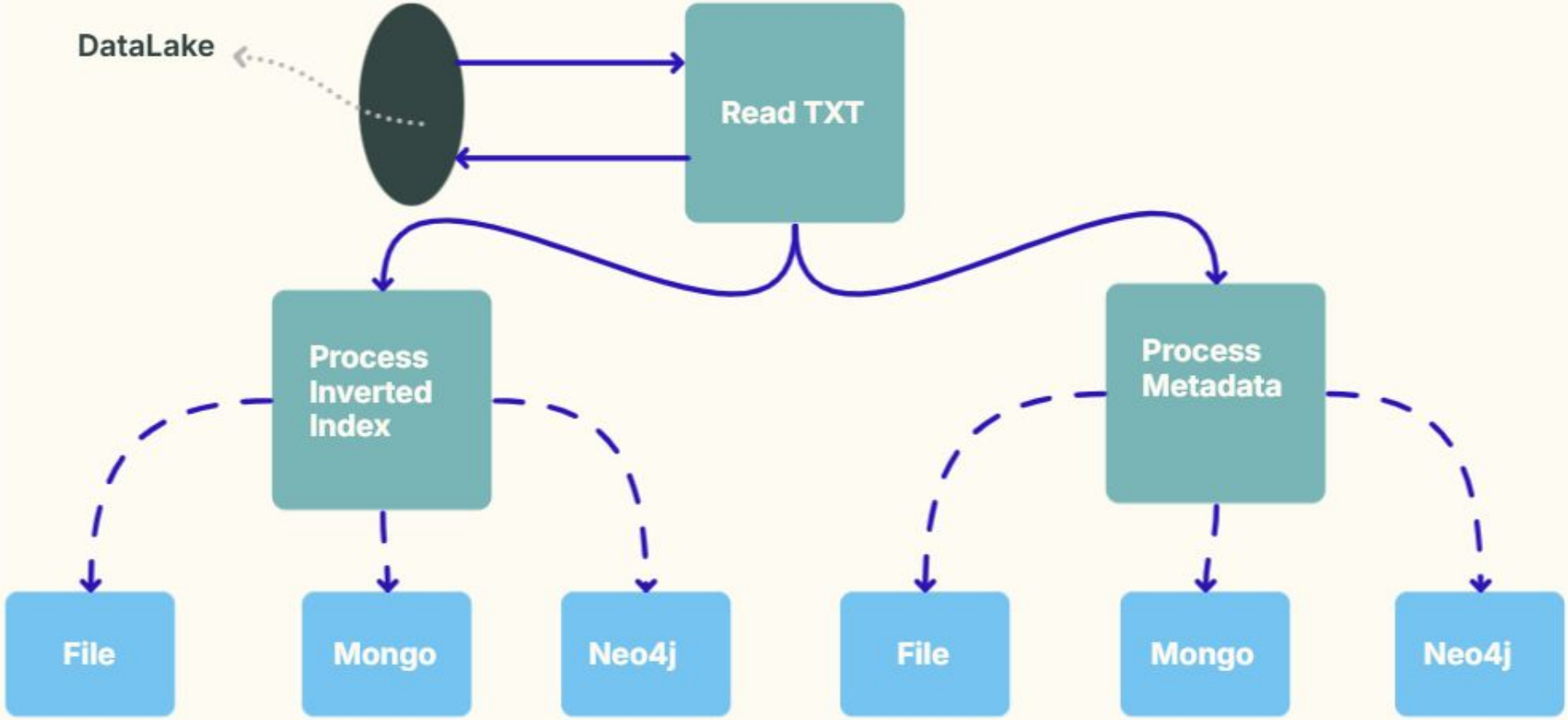
Index

A powerful book search engine designed to optimize performance and handle large datasets efficiently using cutting-edge data structures.

- **Performance Optimization**
  - We implemented improvements in query handling, enabling faster search in large data volumes
- **Appropriate Data Structures**
  - By comparing File System, MongoDB, and Neo4j, we selected the most efficient structure for each type of query (inverted index vs. metadata)
- **Modular Design with SOLID Principles**
  - By comparing File System, MongoDB, and Neo4j, we selected the most efficient structure for each type of query (inverted index vs. metadata)
- **Extensive Testing**
  - We conducted performance tests to identify bottlenecks and optimize processing and query times
- **Enhanced User Experience**
  - We developed a user-friendly and intuitive interface, improving the interaction with queries and results


DataLake
Tests
Crawler
Indexer
Query Engine
API
Metadata
Inverted Index


Project Gutenburg
n book per minute
Controller
Downloader
Store

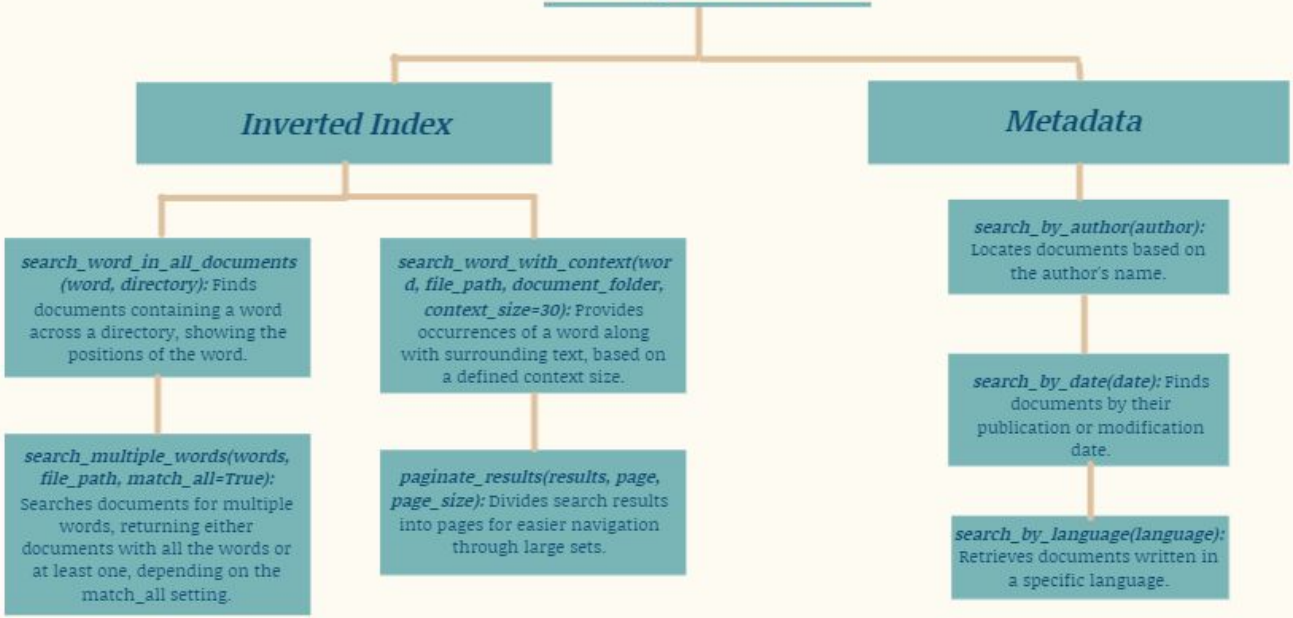

DataLake
Read TXT
Process Inverted Index
Process Metadata
File
Mongo
Neo4j
File
Mongo
Neo4j

File System

Mongo DB

Neo4j

Query Module



127.0.0.1:8000/files/search/?word=bill

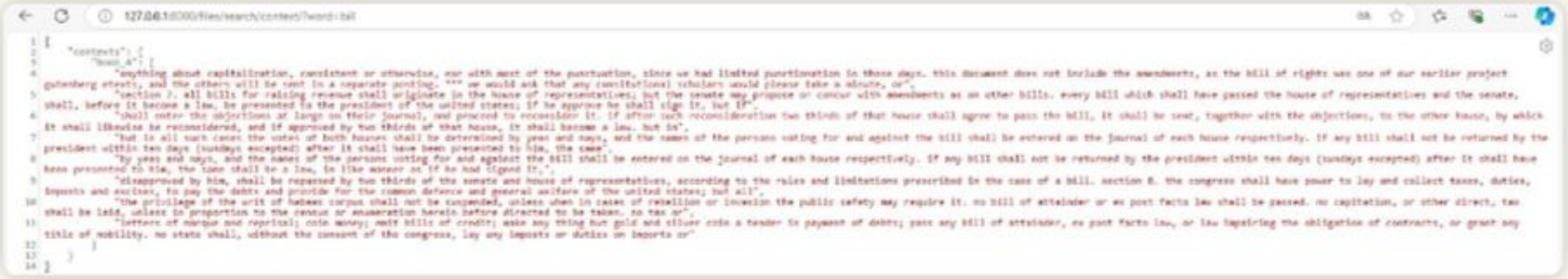

```
{
    "results": {
        "bill": {
            "book_4": {
                "positions": [
                    287,
                    1495,
                    1577,
                    1639,
                    1652,
                    1783,
                    2298,
                    2513
                ],
                "count": 8
            }
        }
    }
}
```

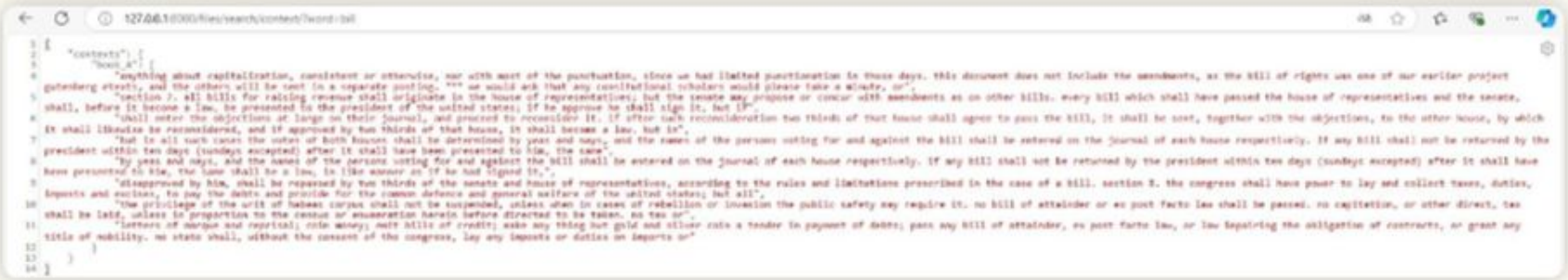

127.0.0.1:8000/neo4j/search/?word=bill

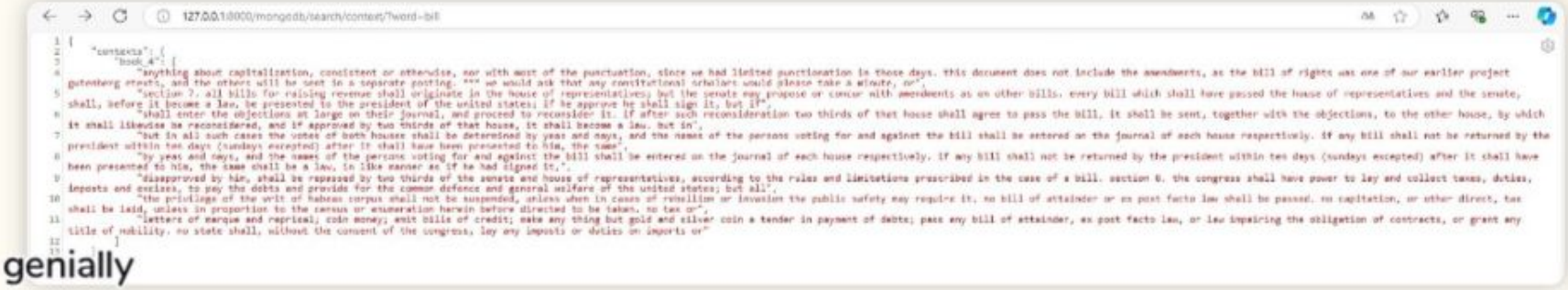

```
{
    "results for 'bill'": [
        [
            "book_4",
            [
                287,
                1495,
                1577,
                1639,
                1652,
                1783,
                2298,
                2513
            ],
            8
        ]
    ]
}
```


127.0.0.1:8000/mongodb/search/?word=bill

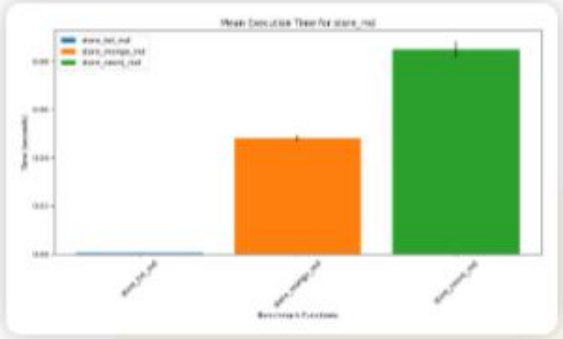

```
{
    "results for 'bill'": [
        {
            "book_id": "book_4",
            "positions": [
                [
                    [
                        287,
                        1495,
                        1577,
                        1639,
                        1652,
                        1783,
                        2298,
                        2513
                    ],
                    8
                ]
            ]
        }
    ]
}
```


127.0.0.1:8000/files/search/context/?word=bill
{
"contexts": {
"book_4": [
"anything about capitalization, consistent or otherwise, nor with most of the punctuation, since we had limited punctuation in those days. this document does not include the amendments, as the bill of rights was one of our earlier project gutenberg etexts, and the others will be sent in a separate posting. *** we would ask that any constitutional scholars would please take a minute, or",
"section 7. all bills for raising revenue shall originate in the house of representatives; but the senate may propose or concur with amendments as on other bills. every bill which shall have passed the house of representatives and the senate, shall, before it become a law, be presented to the president of the united states; if he approve he shall sign it, but if",
"shall enter the objections at large on their journal, and proceed to reconsider it. if after such reconsideration two thirds of that house shall agree to pass the bill, it shall be sent, together with the objections, to the other house, by which it shall likewise be reconsidered, and if approved by two thirds of that house, it shall become a law. but in",
"but in all such cases the votes of both houses shall be determined by yeas and nays, and the names of the persons voting for and against the bill shall be entered on the journal of each house respectively. if any bill shall not be returned by the president within ten days (sundays excepted) after it shall have been presented to him, the same",
"by yeas and nays, and the names of the persons voting for and against the bill shall be entered on the journal of each house respectively. if any bill shall not be returned by the president within ten days (sundays excepted) after it shall have been presented to him, the same shall be a law, in like manner as if he had signed it.",
"disapproved by him, shall be repassed by two thirds of the senate and house of representatives, according to the rules and limitations prescribed in the case of a bill. section 8. the congress shall have power to lay and collect taxes, duties, imposts and excises, to pay the debts and provide for the common defence and general welfare of the united states; but all",
"the privilege of the writ of habeas corpus shall not be suspended, unless when in cases of rebellion or invasion the public safety may require it. no bill of attainder or ex post facto law shall be passed. no capitation, or other direct, tax shall be laid, unless in proportion to the census or enumeration herein before directed to be taken. no tax or",
"letters of marque and reprisal; coin money; emit bills of credit; make any thing but gold and silver coin a tender in payment of debts; pass any bill of attainder, ex post facto law, or law impairing the obligation of contracts, or grant any title of nobility. no state shall, without the consent of the congress, lay any imposts or duties on imports or"
]
}
}

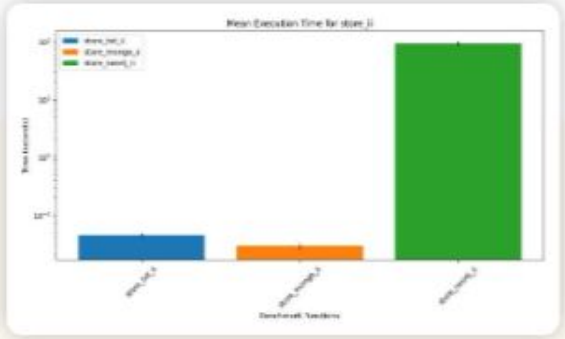

127.0.0.1:8000/files/search/context/?word=bill
{
"contexts": {
"book_4": [
"anything about capitalization, consistent or otherwise, nor with most of the punctuation, since we had limited punctuation in those days. this document does not include the amendments, as the bill of rights was one of our earlier project gutenberg etexts, and the others will be sent in a separate posting. *** we would ask that any constitutional scholars would please take a minute, or",
"section 7. all bills for raising revenue shall originate in the house of representatives; but the senate may propose or concur with amendments as on other bills. every bill which shall have passed the house of representatives and the senate, shall, before it become a law, be presented to the president of the united states; if he approve he shall sign it, but if",
"shall enter the objections at large on their journal, and proceed to reconsider it. if after such reconsideration two thirds of that house shall agree to pass the bill, it shall be sent, together with the objections, to the other house, by which it shall likewise be reconsidered, and if approved by two thirds of that house, it shall become a law. but in",
"but in all such cases the votes of both houses shall be determined by yeas and nays, and the names of the persons voting for and against the bill shall be entered on the journal of each house respectively. if any bill shall not be returned by the president within ten days (sundays excepted) after it shall have been presented to him, the same",
"by yeas and nays, and the names of the persons voting for and against the bill shall be entered on the journal of each house respectively. if any bill shall not be returned by the president within ten days (sundays excepted) after it shall have been presented to him, the same shall be a law, in like manner as if he had signed it.",
"disapproved by him, shall be repassed by two thirds of the senate and house of representatives, according to the rules and limitations prescribed in the case of a bill. section 8. the congress shall have power to lay and collect taxes, duties, imposts and excises, to pay the debts and provide for the common defence and general welfare of the united states; but all",
"the privilege of the writ of habeas corpus shall not be suspended, unless when in cases of rebellion or invasion the public safety may require it. no bill of attainder or ex post facto law shall be passed. no capitation, or other direct, tax shall be laid, unless in proportion to the census or enumeration herein before directed to be taken. no tax or",
"letters of marque and reprisal; coin money; emit bills of credit; make any thing but gold and silver coin a tender in payment of debts; pass any bill of attainder, ex post facto law, or law impairing the obligation of contracts, or grant any title of nobility. no state shall, without the consent of the congress, lay any imposts or duties on imports or"
]
}
}


127.0.0.1:8000/mongodb/search/context/?word=bill
{
"contexts": {
"book_4": [
"anything about capitalization, consistent or otherwise, nor with most of the punctuation, since we had limited punctuation in those days. this document does not include the amendments, as the bill of rights was one of our earlier project gutenberg etexts, and the others will be sent in a separate posting. *** we would ask that any constitutional scholars would please take a minute, or",
"section 7. all bills for raising revenue shall originate in the house of representatives; but the senate may propose or concur with amendments as on other bills. every bill which shall have passed the house of representatives and the senate, shall, before it become a law, be presented to the president of the united states; if he approve he shall sign it, but if",
"shall enter the objections at large on their journal, and proceed to reconsider it. if after such reconsideration two thirds of that house shall agree to pass the bill, it shall be sent, together with the objections, to the other house, by which it shall likewise be reconsidered, and if approved by two thirds of that house, it shall become a law. but in",
"but in all such cases the votes of both houses shall be determined by yeas and nays, and the names of the persons voting for and against the bill shall be entered on the journal of each house respectively. if any bill shall not be returned by the president within ten days (sundays excepted) after it shall have been presented to him, the same",
"by yeas and nays, and the names of the persons voting for and against the bill shall be entered on the journal of each house respectively. if any bill shall not be returned by the president within ten days (sundays excepted) after it shall have been presented to him, the same shall be a law, in like manner as if he had signed it.",
"disapproved by him, shall be repassed by two thirds of the senate and house of representatives, according to the rules and limitations prescribed in the case of a bill. section 8. the congress shall have power to lay and collect taxes, duties, imposts and excises, to pay the debts and provide for the common defence and general welfare of the united states; but all",
"the privilege of the writ of habeas corpus shall not be suspended, unless when in cases of rebellion or invasion the public safety may require it. no bill of attainder or ex post facto law shall be passed. no capitation, or other direct, tax shall be laid, unless in proportion to the census or enumeration herein before directed to be taken. no tax or",
"letters of marque and reprisal; coin money; emit bills of credit; make any thing but gold and silver coin a tender in payment of debts; pass any bill of attainder, ex post facto law, or law impairing the obligation of contracts, or grant any title of nobility. no state shall, without the consent of the congress, lay any imposts or duties on imports or"
]
}
}

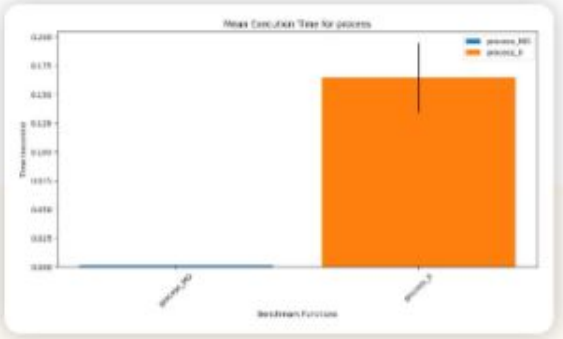
Time Results



Graph 1

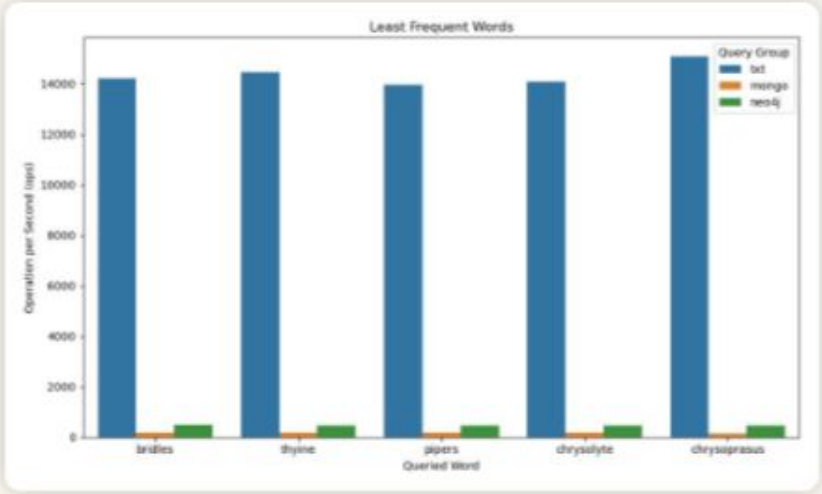


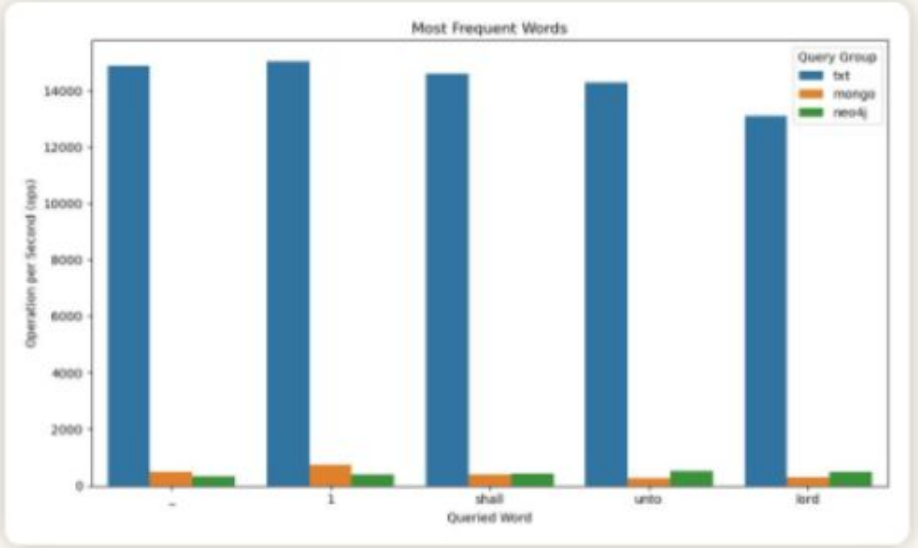
Graph 2



Graph 3

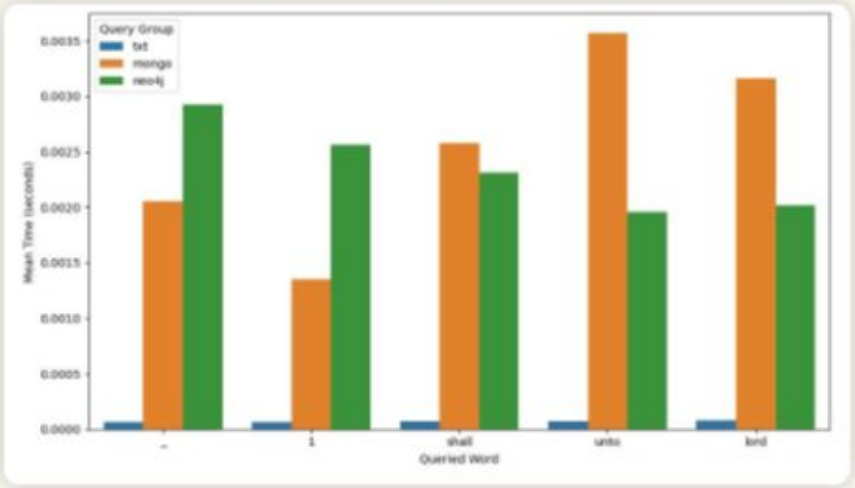
Least Frequent Words

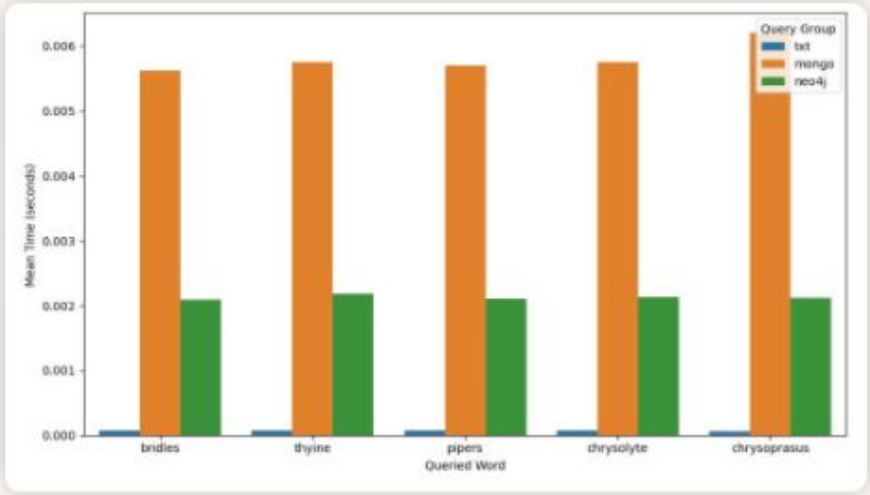




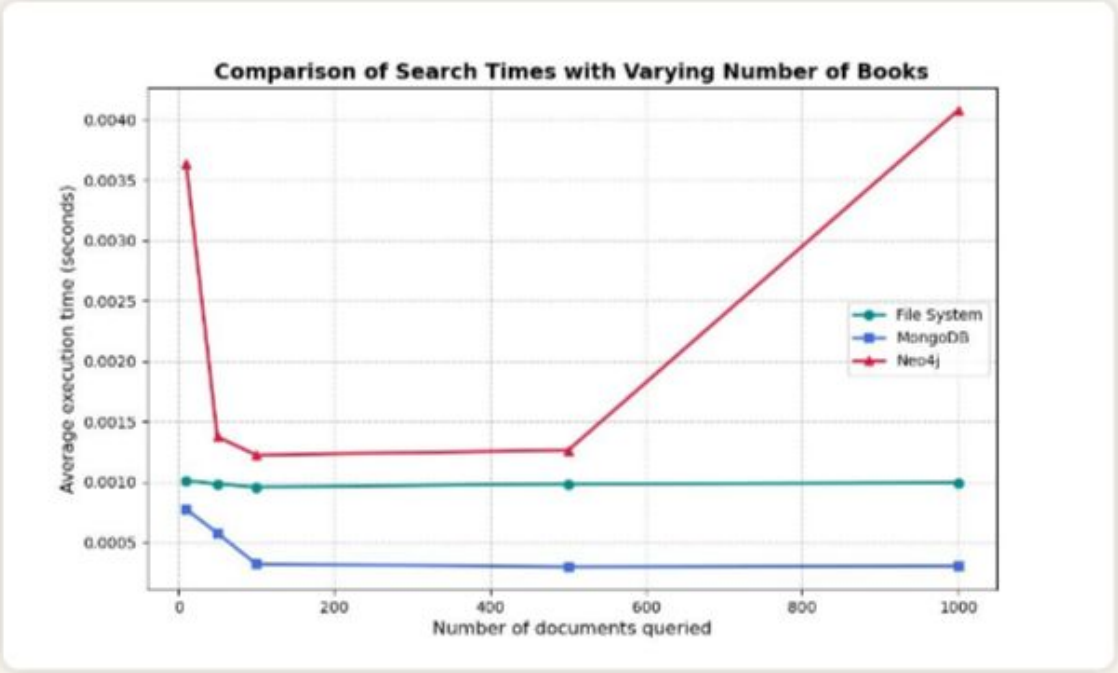
Most Frequent Words

Most Frequent Words





Least Frequent Words



Search Scalability

Performance Differences

Metadata Processing Eficiency

Scalability with Increased Data

- **Java**
  Implementing the system in Java to improve performance and better follow SOLID principles.
- **Docker**
  Using Docker to streamline deployment, ensure consistency, and enable easier scaling.
- **Creating the UI**
  Creating a UI will make the application more user-friendly.
- **Context Query**
  Enhancing the context query feature to make it more user-friendly and intuitive, similar to Google's search results.

Thank You