BigData Assignment 2 Report

Andrei Pavlov

April 15, 2025

GitHub Repository

https://github.com/IAndermanI/big-data-assignment2-2025

1 Methodology

1.1 Data Preparation

- Collect plain text documents in `data/` directory
- Filename format: `<doc_id>.<doc_title>.txt`
- Upload to HDFS for distributed processing

1.2 MapReduce Pipelines

1.2.1 First Pipeline (Inverted Index)

Mapper:

```
# Input: (doc_id, text)
for term in tokenize(text):
    emit((term, doc_id), 1)
```

Reducer:

- Aggregate term frequencies (TF)
- Store in Cassandra tables:
  - `inverted_index(term, doc_id, freq)`
  - `doc_length(doc_id, length)`

1.2.2 Second Pipeline (Term Statistics)

Mapper:

```
# Use sets for unique term-doc pairs
emit((term, doc_id), None)
```

Reducer:

- Calculate document frequency (DF)
- Compute collection statistics:
  - Total documents (n)
  - Average document length (avg_doc_length)
- Store in `statistics(term, df, n, avg_doc_length)`

1.3 Retrieval (PySpark & BM25)

Implementation workflow:

1. Load data from Cassandra tables
2. Filter query terms using Spark RDDs
3. Compute BM25 scores:

$$\mathrm{BM25}(d,q)=\sum_{t\in q}\ln\left(\frac{n-\mathrm{df}(t)+0.5}{\mathrm{df}(t)+0.5}\right)\times\frac{\mathrm{freq}(t,d)(k+1)}{\mathrm{freq}(t,d)+k\left(1-b+b\frac{|d|}{\text{avg\_doc\_length}}\right)}$$

4. Rank and return top 10 results

1.4 Deployment

- **Indexing:** Hadoop Streaming MapReduce

  ```
  bash index.sh
  ```

- **Search:** Spark submit job

  ```
  spark-submit --class QueryProcessor search.sh
  ```