ASSIGNMENT REPORT

Information retrieval (IR) systems help users find relevant documents efficiently. This assignment implements three key IR techniques:

1. **Boolean Retrieval** – Using an **inverted index** to process Boolean queries.
2. **Spell Correction** – Using **Soundex, Edit Distance, and N-gram similarity** to correct spelling errors.
3. **BSBI Algorithm** – Implementing **disk-based inverted indexing** to handle large datasets.

The objective is to build efficient IR components and analyze their performance.

Experiment 1: Boolean Retrieval using an Inverted Index

**Problem Statement**

We are given a collection of documents (bool\_docs.json). Our goal is to:

1. **Preprocess documents** (remove stopwords, apply stemming & lemmatization).
2. **Build an inverted index** to store word-to-document mappings.
3. **Process Boolean queries** from bool\_queries.json using **AND, OR, NOT operators**.

**Implementation Details**

* Used **NLTK** for preprocessing (stemming, lemmatization, stop-word removal).
* Constructed an **inverted index** (dictionary {word: [doc\_ids]}).
* Implemented **query processing** using set operations.

Experiment 2: Spell Correction Techniques

**Problem Statement**

We implemented **three spell correction methods** to fix misspelled words in queries:

1. **Soundex Algorithm** – Matches words based on phonetic encoding.
2. **Edit Distance (Levenshtein Distance)** – Finds words with the fewest edits.
3. **N-gram Approach (Jaccard Similarity)** – Matches words based on character-level similarity.

**Implementation Details**

Each method compared misspelled words from spell\_queries.json with words in dictionary.txt.

**Experiment 3: Block Sort-Based Indexing (BSBI)**

**Problem Statement**

* **BSBI avoids loading large datasets into RAM** by processing **blocks** of bsbi\_docs.json.
* It **creates & merges** small inverted indexes to handle **millions of documents efficiently**.

**Implementation Details**

* Processed blocks of size 1 and 1000.
* Constructed **block-wise inverted indexes** (index\_block\_x.json).
* Merged them into a **final inverted index** (final\_inverted\_index.json).