MiniSearch - Low Level Design

Search Engine LLD Overview

This document explains the Low Level Design (LLD) of a simple Search Engine project called MiniSearch. It indexes and searches over text, HTML, and Word documents.

Flow of the Search Engine

HOW IT WORKS (Flow):
1. User adds a document (file path).
2. The SearchEngine uses a DocumentReader (based on file type) to extract plain text.
3. A Document object is created and stored in a map.
4. The InvertedIndex indexes the words from the document.
5. On a search, the engine looks up the word in the index, builds result scores, and returns matching docs.

Class Descriptions

KEY CLASSES:
- SearchEngine: Coordinates document loading, storing, indexing, and searching.
- InvertedIndex: Maps word -> list of (docId, frequency). Core of the search engine.
- DocumentReader (interface): Strategy for reading different file types.
  - TextReader, HtmlReader, WordReader: Implementations for various formats.
- DocumentReaderFactory: Chooses the appropriate reader based on file extension.
- Document: Holds id, name, format, and extracted content.
- Posting: Represents one word's presence in a document.
- Result: Holds a docId and score for displaying search results.

Class Diagram

Refer to the next page for the class diagram.