Semestrální práce z předmětu KIV/IR

Indexer

Zdeněk Valeš

26.5. 2020

1 Zadání

V jazyce Java vytvořte indexer, který bude schopný indexovat zadané dokumenty a následně nad nimi provádět vyhledávání. K realizaci práce použijte připravená rozhraní.

2 Analýza

Aplikace se skládá ze dvou částí. První část tvoří jádro, které provádí indexaci a vyhledávání, druhou část tvoří jednoduché grafické rozhraní pro práci s indexerem.

2.1 Jádro

Jádro aplikace se skládá z indexu, textového preprocesoru a vyhledávače. Preprocesor je použit k převedení textu (obsah dokumentu nebo vyhledávací dotaz) na tokeny a ty pak do jejich základní formy.

2.1.1 Index

Indexované dokumenty jsou uložené v invertovaném indexu. Invertovaný index je mapa, která každému tokenu, který se v kolekci indexovaných dokumentů vyskytne, přiřadí seznam dokumentů (tzv. posting list), ve kterých se token alespoň jednou nachází. Výhodou této datové struktury oproti incidenční matici je zejména menší paměťová náročnost.

Posting list obsahuje kromě dokumentů také počet, kolikrát se token v určitém dokumentu vyskytuje. Tento údaj je později použit při ohodnocování nalezených výsledků.

2.1.2 Vyhledávač

Optimalizace: - předpočet TF-IDF pro termy a dokumenty - relativní cos similarity

2.2 Uživatelské rozhraní

3 Implementace

4 Závěr