Update the dictionary

Giovanni Marchetto

Overview

- Updating dictionary
    - English based
    - Add and remove documents
    - Boolean and phrasal queries
    - Tokenization, normalization and stop word remotion

Structure

- Data structure
  - Dictionary
  - Posting list

- Queries
  - Boolean
  - Phrasal

- Operations
  - Tokenization
  - Normalization
  - StopWord

Dictionary

- Parameters
  - Dictionary hash map
    - Each term corresponds to a Posting List
  - Document list hash map
    - Each docID corresponds to a term’s list
  - Stop word list size

Dictionary

- Methods
  - Add document
    - Single or list of document
  - Remove document
    - Very efficient thank to document term list
    - If after the remotion of a document the Posting list of a term is empty, the term is removed
  - Change stop word list size
    - Load the correct stop word list

Posting List

- Posting list hash map
    - Each docID corresponds to a list of integer position (of the term)
    - Positional index
        - Permit phrasal query
- Add and remove posting
    - Add a single or a list of position of term

Boolean queries

- Boolean queries allowed
  - AND, OR and NOT
  - Terms are filtered in the same way as those in the document for consistent results
  - In CLI is not allowed the combination of queries
- Term query
  - To reach a single word posting list is a lot more efficient (exploit hash map)

Phrasal queries

- Phrasal query algorithm
  - Tokenization and normalization of phrase
  - Remotion of stop word
  - Matching of posting list between each two adjacent words
  - Reach the position of the start of the phrase
- Alternative Phrasal query
  - For the remotion of stop word it’s always use the extended list
  - Faster of longest phrases
  - More errors

Operations

- Tokenization
- Normalization
    - Lower case only
    - Remotion of symbols and numbers
        - Only "-" (half word) and "'" (plurals and verbs) remains
- Stop Word
    - Different size for the list of stop words
    - Also permit to not remove

Other choices

- Trim of Array List only with save of dictionary
  - during the addition of a document would result in a slowdown
- No stemming
  - A problem for phrasal queries
- CLI limitation
  - Save dictionary only when close CLI
  - Not alternative phrasal query (not a lot efficient)

Known problems

- More stop word removed, more possible error in phrasal queries results
- Plurals and singulars of a term are considered different
  - A partial solution not implemented: porter stemmer