CS5615 Information Retrieval

Programming Project

by

V B Wickramasinghe - 148245f

Part 1: Spelling Correction

1. **Implementation -** Bigram based. In the indexing phase all words are broken down into their bigrams and stored in a Java HashMap where bigrams are the keys and the **set of words** in the corpus involving the given bigram is the value. Corrections are provided by breaking down the given misspelled word into its bigrams and returning the words in the corpus that involves the calculated bigrams.

   Words-Corrections -

Misspelled words	Corrections
concider	consider, coincide, considerate, unconsidered, coincided, coincides, inconsiderable, cider, considered ….
hierchy	hierarchy, handkerchief, handkerchiefs, kerchief, kerchiefs, hypertrophied, hypertrophies, merchantability, archery ...
sorces	successors, sources, predecessors, sordes, processors, successor, forces, resources, sores, professors, circumstances, sovereignties ...

   cs276.pe1.spell.SpellingScorer output -

   cs276.pe1.spell.KGramSpellingCorrector: 158/270 in 4566ms

   Sources of error - With following additional output captured from the results,

   Total misspelled words input : 270

   First suggestion correct : 158

   Not first suggestion : 90

   Not found : 22

   There are 90 words that actually had the correct suggestion in the corrections list but wasn't the first suggestion so it was the biggest source of error. A better scoring function for the results

would yield better results. A larger corpus would help in reducing the number of 'not founds'.

2. Implementation with edit distance

Words-Corrections -

Misspelled words	Corrections
concider	consider, coincide, coincided, coincides, considers, concede, confide, confided, concise, conceded, cider, considered, coincident, reconsider...
hierchy	hierarchy, thierry, anarchy, thiersch, fiercely, pierced, frenchy, pierce, heresy, hereby, itchy, percha, chiefly, monarchy, highway, cheerily, hitched, history...
sorces	sources, sordes, forces, sores, shores, porches, spores, soirees, forceps, scores, soles, source, stores, torches, stories, soames, sorts, sommes ...

cs276.pe1.spell.SpellingScorer output -

cs276.pe1.spell.KGramWithEditDistanceSpellingCorrector: 204/270 in 11063ms

Sources of error - With following additional output captured from the results,

Total misspelled words input : 270

First suggestion correct : 204

Not first suggestion : 44

Not found : 22

Before the not first suggestion error occured 90 times now its 44 therefore edit distance have improved the results.

3.

- Both correctors were modified but not at the same time.
    - To test if an improvement could be achieved with any

of them.

- Added a sort(ascending and descending) on word frequency to both correctors final output.
- Impact was negative. I think the word frequencies in the corpus are not a good parameter to take into account in spell checks.

Part 2: Lucene

1. Returns a list of documents, but contains a lot of duplicates.

   Eg: Title: Fabbrica dei tedeschi, La (2008)

   Author: Marco

   Title: DOA: A Coroner's Fairy Tale (2001)

   Author: Michael Marco

   Title: Hirtt&#65533;m&#65533;tt&#65533;m&#65533;t (1971)

   Author: Marco Bognomini

   Title: "'Allo 'Allo!" (1982) {(#8.3)}

   Author: Marco van Hoof <k_luifje7@hotmail.com>

   Title: "'Allo 'Allo!" (1982) {(#8.4)}

   Author: Marco van Hoof <k_luifje7@hotmail.com>

2. No results were returned.

3. Returns the same result multiple times. Eg:

   Title: "1-800-Missing" (2003) {Fugitive (#3.8)}

   Plot: Dylan Anders, a convict about to go on trial for murder, kills two police officers and escapes. He then steals a Dr. Laurel Bennett's car and kidnaps her. Jess, Nicole, Antonio and Janey are assigned to the case and Jess and Nicole eventually rescue Dr. Bennett from a

burning building. Meanwhile, Nicole urges Jess to get to know Janey better and the three women urge Antonio to keep his job instead of taking a new job in airline security. Back on the case, the mystery only begins and it is eventually learned that Dr. Tanner is helping Anders instead putting the FBI off his trail. It culminates at a drugstore where one of the FBI's own is fatally wounded in a shootout with Anders.

Author: Anonymous

Title: "1-800-Missing" (2003) {John Doe (#2.15)}

Plot: After arresting a dangerous man on the run, Jess and Nicole find a man shot in the head hanging on to life in the trunk of the car he was trying to steal. The man has amnesia and doesn't remember anything. The bullet leads the team to believe that he's connected in some way to the disappearance of young heiress Victoria Farlow and they must decide if "John Doe" is innocent or guilty while trying to figure out his identity and the circumstances of his near-death.

Author: van_whistler@hotmail.co.uk

Spelling correction for Lucene

Lucene doesn’t do any spell checking automatically. So the query for titles containing ‘trmmy’ returns no documents. But Lucene does provide spell check facilities in **org.apache.lucene.search.spell.SpellChecker** class.

Because lucene doesn’t do automatic spell checking my spell checker does better. For example in Luke ‘tst’ query doesn’t return anything, But with my spell checker it returns all movie titles with test in them.