IR 634 Information Retrieval Final project

Sainath Dutkar

Aqsa Sheikh

Istabraq almusally

1) Spell Check Context based

Lucene Spell checker-

At a high level, it works by taking an existing field from the main Lucene index and builds a secondary index designed specifically for rapid look up of candidate suggestions. This index, for those who are curious, is built by creating character-based n-grams of the words from the original field. At query time, the word to be checked is appropriately analyzed and then searched against this secondary index. Assuming one or more hits are returned, the candidate word is then compared to the original word using a String distance measure (see org.apache.lucene.search.spell.String Distance).

The Implemented Spell Check Concept

We have implemented the Norvig's Spell Checker by following the candidate model.

Candidate Model: First a new concept: a **simple edit** to a word is a deletion (remove one letter), a transposition (swap two adjacent letters), a replacement (change one letter to another) or an insertion (add a letter). For a word of length *n*, there will be *n* deletions, *n*-1 transpositions, 26*n* alterations, and 26(*n*+1) insertions, for a total of 54*n*+25.

The Basic Idea of the Program

Step 1 : Creating a dictionary with Words as Key and Weight as value

Step 2 : Check if the is word present in Dictionary-- if yes return the word

Step 3 : Get all Possible Edits

Edits => Returns the list of words

Step 1 : deletes (remove 1 letter) - gives N

Step 2 : Transpose (Swap adjusted letters) - gives N-1

Step 3 : Replace (Change one letter to another) - gives 26N

Step 4 : Insert (Add a letter) - gives 26N+1

Step 4 : Create a Candidate list

Step 5 : Traverse all possible edits comparing with Dictionary list and storing the matches in a Candidate MAP

Step 6 : Return the most weighted word

Step 7 : If no match is found. Run the edit function on the list of edit words once.

Step 8 : Repeat step 5 and Step 6

Step 9 : If no possible hit found -> return the word

Feature no.2 Snippet

Generally, Snippet is 2-3 line description of the main information of a website just found below the URL on the main search page of a search engine.

We have implemented this feature to display a brief summary of the information in the document.

Approach

1) Take the document from the file
2) Enter the query search
3) Find the query and pick up the desired sentence length after the query search word
4) These words form the Snippet of the document.

Precision

Precision – How many relevant documents are in the search result.

Precision = ($Number\ of\ relevant\ documents\ retrieved$)/($Number\ of\ documents\ retrieved$)

Query set	TREC EVALAUATION	OUR CODE EVALUATION
Query 1	3/30 =10%	4/30 =13.33%
Query 2	10/30 = 33.33%	8/30 =26.66%
Query 3	10/30 =33.33%	6/30 =20%
Query 4	3/30 = 10%	7/30 =23.33%
Query 5	11/30 =36.66%	14/30 =46.66%
Avg	**24.664**	**25.99**

Impact of the changes on the working of Lucene System

2) The feature Spell check lets you corrects a misspelled word and provides a correct word from the inbuilt dictionary.
3) The feature Snippet gives you a rough idea of the main information in the document without reading the whole document.