SMS based FAQ Retrieval

Monolingual FAQ retrieval in English

FIRE 2012
17.12.2012

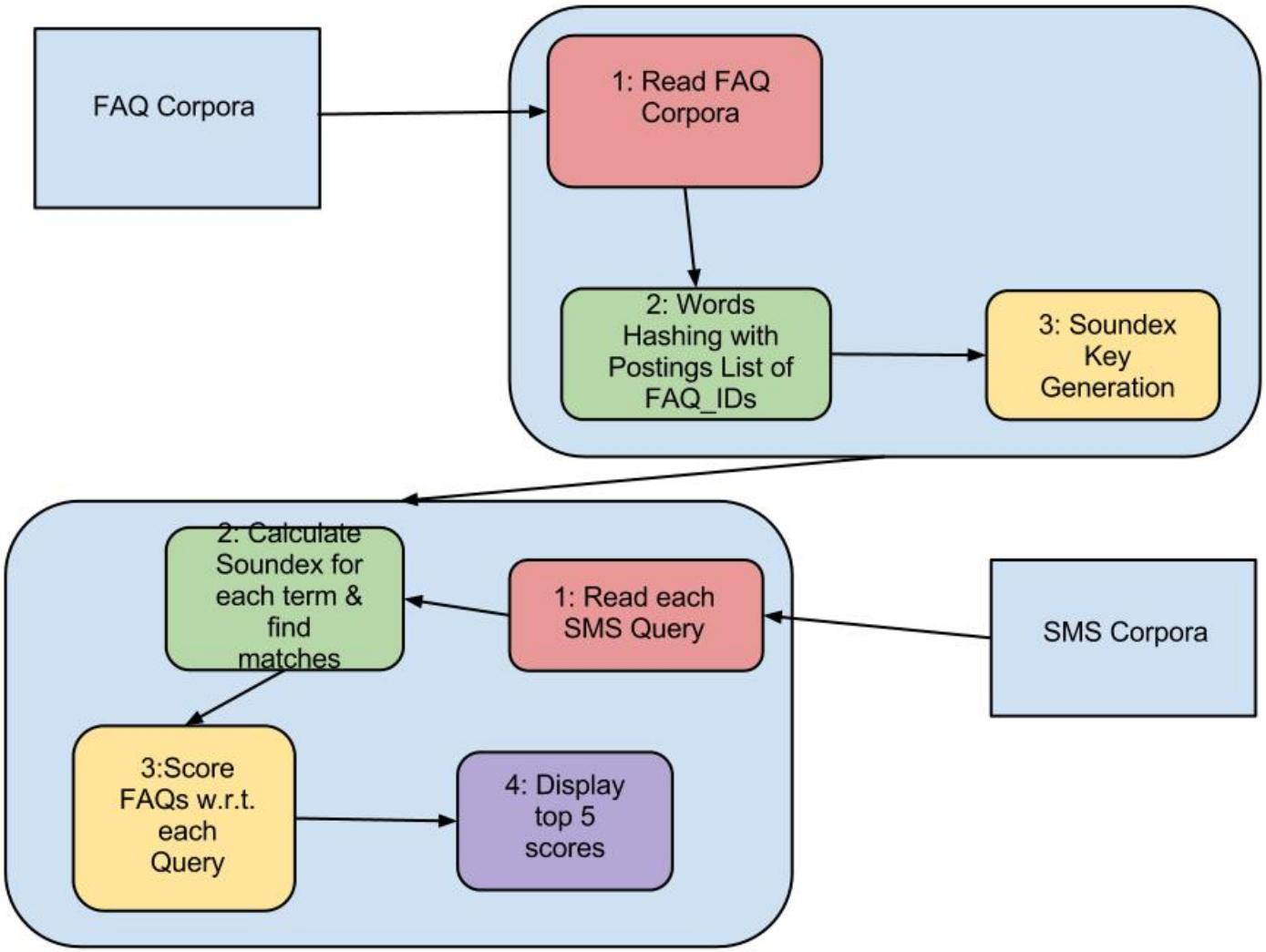
Monolingual FAQ Retrieval

- Finding the top 5 matches for an SMS query from a FAQ corpora where both the SMS query and the FAQs are in English

Overview

- Use Phonetic Search(Soundex Algorithm) to match terms in SMS query with terms in FAQs
- Assign Scores to FAQs based on matching terms
- List 5 FAQs with highest scores as predicted matches

Components


FAQ Corpora
1: Read FAQ Corpora
2: Words Hashing with Postings List of FAQ_IDs
3: Soundex Key Generation
2: Calculate Soundex for each term & find matches
1: Read each SMS Query
SMS Corpora
3:Score FAQs w.r.t. each Query
4: Display top 5 scores

Phonetic Search

- Indexing words by their sound when pronounced in English

Soundex Code

- Soundex was originally developed by Russell and Odell and underwent some modifications thereafter

- Soundex code for a string consists of a letter followed by three numbers

Soundex Code

- The first letter is same as that in the original string and is always retained. For the following characters:
- a e i o u y h w => remove
- b v f p => 1
- c g j k q s x z => 2
- d t => 3
- l => 4
- m n => 5
- r => 6

Soundex Code

- If two or more letters with the same number were adjacent in the original string, or adjacent except for any intervening h and w, then omit all but the first

- Return the first four characters padded with zero

SMS Language

- misspellings
- Intentionally or accidentally omitted characters in words
- Vowels are frequently dropped

Scores assigned to FAQs

$$f.score \leftarrow f.score + idf_m * (b + \log(tf_{m,f}))$$

f = FAQ

m = a term in FAQ corpora whose soundex code matches a term t in the SMS query

idf_m = log(#domains / #domains where m occurs)

b = val1 if m is an exact with t else b = val2
value of val1 was varied and quality of predictions noted

$tf_{m,f}$ = frequency of matching term m in FAQ f

Training Set Results

Table 1: Number of correct matches for various values of $val1$

$val1$	Correct Cases	Wrong Cases
5	822	607
10	824	605
20	824	605
50	819	610
70	817	612

val2 = 1 Number of predictions that matched the domain of query
~ 1105 / 1429

Test set results

- ***** FIRE 2011 SMS TASK EVALUATION REPORT *****
- No. of In-domain Queries :726
- No. of Out of Domain Queries:1007
- In Domain correct:363/726 (0.5)
- Out of Domain correct:0/1007 (0.0)
- Total Score: 0.20946336
- Mean Reciprocal Rank (MRR): 0.5395123

Analysis

- val1 signifies the extra weightage given to terms that exactly match the terms in SMS query.
- If too much of importance is assigned to exact matches then we do not allow enough discount for spelling errors in SMS text.
- If too little importance is given to exact matches then non-matching FAQs containing terms similar to those in the SMS query may be scored higher than actual matching ones

Analysis

- In some cases the first word of the term may also be different for example – fertilizer replaced by phaertilizer etc.
- Soundex code does not match in those cases. However not treating the first character of the term separately leads to poorer quality of matches.

Analysis

- The results obtained for out of domain queries do not differ much from those for in-domain queries in terms of scores (values and distribution) making it difficult to predict which queries are out of domain

- The predicted results sometimes are in a relevant domain

Example of Out of Domain Query

- <SMS_TEXT>special trains 4 dwali</SMS_TEXT>
- <MATCHES>
- <ENGLISH>NONE</ENGLISH>

Ranked second in predicted matches

- FAQ ENG INDIAN RAILWAYS 212
- <QUESTION>Where Can I get the details of Luxury Trains running in Indian Railways?</QUESTION>
- <ANSWER> To get the detail of Luxury Trains click on this link http://www.indianrail.gov.in/scenic_rly.html
- </ANSWER>

Future Work

- Distinguishing in-domain and out-of-domain queries
- Trying out other phonetic search algorithms like Metaphone, Double Metaphone etc. for better performance

Language and Resources

- Design was implemented in Java
- JDOM API s for parsing XML data
- Soundex Algorithm and tf-idf scoring from Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze

Acknowledgements

I am thankful to the FIRE 2012 organizers and Dr. Pabitra Mitra for advising and encouraging me in this work.

Thank you