Query expansion based on similarity metrics

Group - 07

Muhammed Durakovic - Implementation of the algorithm
Aleksandar Iliskovic - Research and preprocessing

Link: https://github.com/ailiskovic/AIR

Introduction

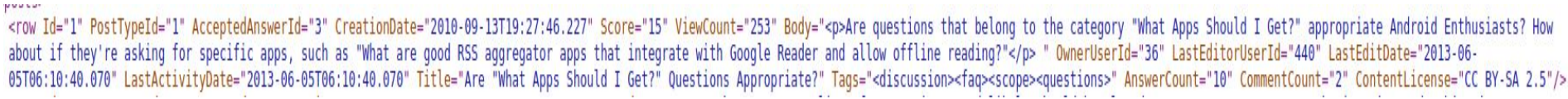
- query expansion for the stack exchange sites (e.g. StackOverflow)
- improve IR results
- extend query by given number of elements

Example:

- original query: “code developers”
- number of elements to extend: 2
- new query “pure code device developers”

Data and Methods

- posts of all stack exchange web sites, raw data from Posts.xml file
- for each post data are extracted from body attribute
- necessary preprocessing steps
- creating a dictionary



Results

- split words to vectors
- run similarity measures for query and dictionary words
- take a words with the highest similarity

```
Query:code developers
[('pure', 0.62208736), ('code', 1.0), ('device', 0.557385), ('developers', 1.0)]
```

Conclusion

- new query should improve the IR results
- limited to Posts.xml file
- not possible to have query which words are not in dictionary