Assignment-2 Report

Ameya
Aslah
Harsh

Dataset we have used is the bbc-news sports dataset. We have taken 20 documents from each sports category. They are stored in inputFiles.

001-020 : Cricket related news
021-040 : Athletics related news
041-060 : Football related news
061-080 : Rugby related news
081-100 : Tennis related news

Created two separate indexes, one that uses the WhiteSpaceAnalyzer to tokenize which tokenizes on whitespaces only. The other uses the EnglishAnalyzer which consists of StandardTokenizer, StandardFilter, EnglishPossessiveFilter, LowerCaseFilter, StopFilter, and PorterStemFilter. The indexes are stored in indexedFiles and indexedFilesSmart resp.

Queries used: "cricket", "athletics", "football", "rugby", "tennis".

Relevant documents for
cricket :- 001-020
athletics :- 021-040
football :- 041-060
rugby :- 061-080
tennis :- 081-100

Q1: "cricket"

basic index retrieves
017
016
011
005
014
015
004
012

p = 1
r = 8/20 = 0.4

smart index retrieves
009
017
014
016
004
011
005
012
015
002

p = 1
r = 10/20 = 0.5

Q2: "athletics"

basic index retrieves

031
030
021
034

p = 1
r = 4/20 = 0.2

smart index retrieves

027
030
033
023
021
037
026
031
035
040
034
039
028

p = 1
r = 13/20 = 0.65

Q_3: "football"

basic index retrieves

055
074
044
043

p = 3/4 = 0.75
r = 3/20 = 0.15

smart index retrieves

055
079
043
063
077
044
074
053
058

p = 5/9 = 0.56
r = 5/20 = 0.25

Q4: "rugby"

basic index retrieves

077
075
079
074
078
080
019
001

p = 6/8 = 0.75
r = 6/20 = 0.3

smart index retrieves

077
079
075
063
074
080
078
064
076
019
001

p = 9/11 = 0.82
r = 9/20 = 0.45

Q5: "tennis"

basic index retrieves

083
093
099
090

p = 1
r = 4/20 = 0.2

smart index retrieves

083
092
099
093
096
082
090

p = 1
r = 7/20 = 0.35

Here, we have assumed that as bbc has classified the dataset into 5 sports, the relevant documents for a query of the sport should be the same as those classified in that sport by bbc itself. we have calculated the precision and recall based on this assumption for each query.

As we can see the recall is better with the smart index and the precision is also better in most cases.