Vibhor Mathur
Armando Pensado Valle
CS 221 Winter 2013

Project 3 Part 2

Original NDCG@5

mondego: 0.7322387343768747
machine learning: 0.034947681472593875
software engineering: 0.0
security: 0.0
student affairs: 0.0
graduate courses: 0.0
Crista Lopes: 0.4182791467639531
REST: 0.0
computer games: 0.0
information retrieval: 0.4057300173119513

Average Score: 0.1591195579925373

Final NDCG@5

mondego: 0.9238531025156478
machine learning: 0.1397907258903755
software engineering: 0.7303140311615123
security: 0.4057300173119513
student affairs: 0.8464077160964966
graduate courses: 0.7303140311615123
Crista Lopes: 0.5486978749788154
REST: 0.2559871398391726
computer games: 0.4057300173119513
information retrieval: 0.4057300173119513

Average Score: 0.5392554673579386

Techniques Used To Improve NDCG@5

General Techniques

- Added the following indexed fields (in addition to the fields used originally) to all indexed documents:
    - **urldomain**: The domain of the document's URL.
    - **stemtitle**: The document's stemmed title.
    - **stemcontent**: The document's stemmed content.
    - **contentheaders**: The content of the following HTML tags in a document
        - h1,h2,h3,h4,div[id*=title],div[class*=title],span[id*=title],span[class*=title]
    - **importantcontent**: The content of the following HTML tags in a document:
        - b,strong,em
    - **outgoingtext**: The text of outgoing links from the document.
    - **anchortext**: The text of incoming links to the document.
    - **stemanchortext**: The stemmed text of incoming links to the document.

We believe these fields are generic and can be used to improve the performance of a wide range of documents and queries.

Test-Query Set Specific Techniques

- Crawled missing pages
- Automatically optimized query-time field boosting weights against the test queries using hill-climb algorithm

Final Scoring Formula

The scoring formula consists of boosting queries on different document fields. The boosting scores are:

- **urldomain** (*default query*): 5.3

- **title** (*preference to occurrence at beginning of the field*): 240.5
- **stemtitle** (*preference to beginning of field*): 77.7
- **stemtile** (*proximity query*): 1.0

- **content** (*default query*): 116.2
- **stemcontent** (*default query*): 331.4
- **stemcontent** (*proximity query*): 91.0

- **contentheaders** (*default query*): 15.3
- **importantcontent** (*default query*): 129.9

- **anchortext** (*default query*): 17.6
- **stemanchortext** (*default query*): 74.6

Note: For the "stem" fields, the query text is also stemmed.
Note: Fields not listed had an effective weight of 0.