Custom Web Search Engine

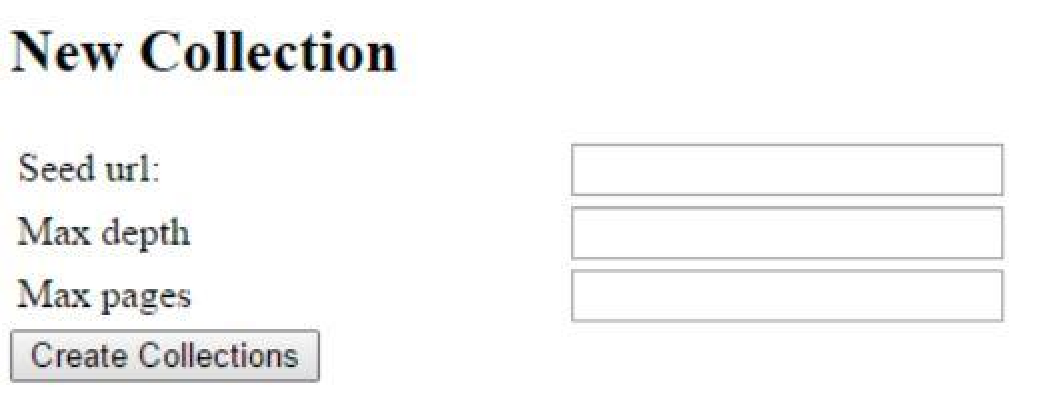
Emre Can Kucukoglu
Fatih Hafizoglu
Yigit Ilguner

Purpose of system

- Increase precision
- Giving user a configurable web search engine
- Optimize query processing
- Diversify search results

Corpus details

- On user demand
- User can generate corpus:



New Collection

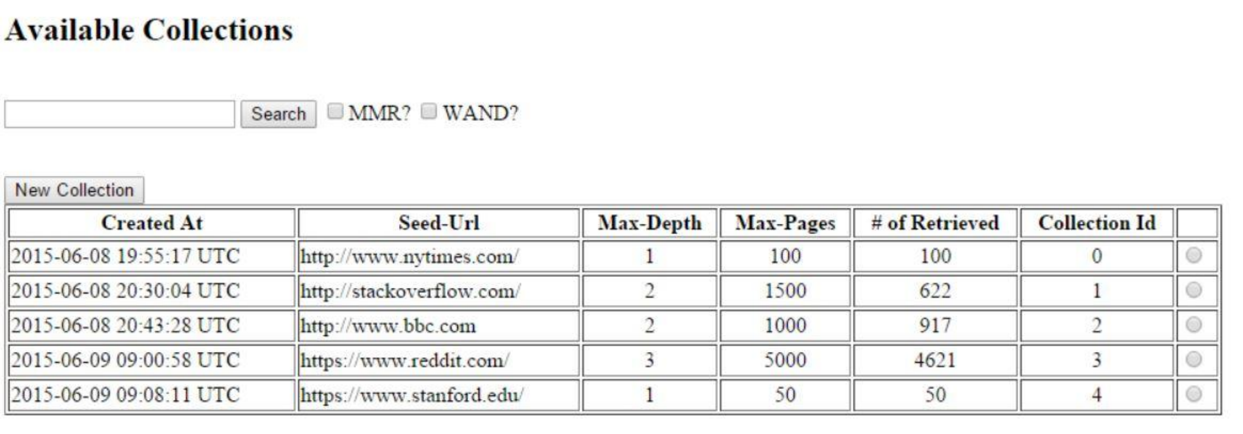
Seed url:

Max depth

Max pages

Corpus details

- On user demand
- User can select from existent corpuses



Available Collections

Created At	Seed-Url	Max-Depth	Max-Pages	# of Retrieved	Collection Id	
2015-06-08 19:55:17 UTC	http://www.nytimes.com/	1	100	100	0	○
2015-06-08 20:30:04 UTC	http://stackoverflow.com/	2	1500	622	1	○
2015-06-08 20:43:28 UTC	http://www.bbc.com	2	1000	917	2	○
2015-06-09 09:00:58 UTC	https://www.reddit.com/	3	5000	4621	3	○
2015-06-09 09:08:11 UTC	https://www.stanford.edu/	1	50	50	4	○

Processing steps

- Crawling
  - Text extraction
- Indexing
  - Stop-words elimination
  - Case-folding
- Query processing
  - Document-at-a-time processing
  - Weak AND optimization
  - Maximal marginal relevance

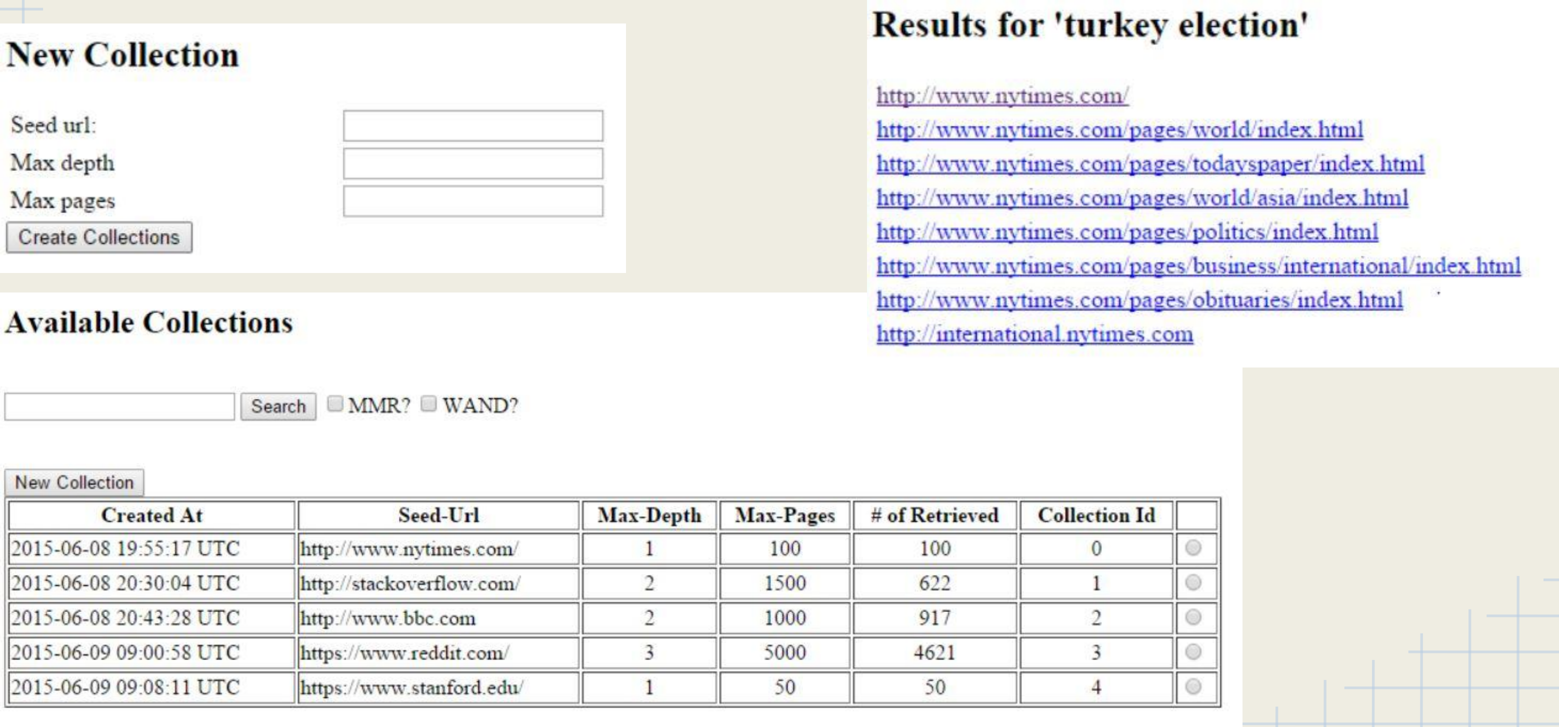
Used indices

- Inverted index

IR model

- Vector space model
- TF-IDF similarity calculation
- Cosine similarity

User interface



New Collection

Seed url: [ ]

Max depth [ ]

Max pages [ ]

Results for 'turkey election'

http://www.nytimes.com/

http://www.nytimes.com/pages/world/index.html

http://www.nytimes.com/pages/todayspaper/index.html

http://www.nytimes.com/pages/world/asia/index.html

http://www.nytimes.com/pages/politics/index.html

http://www.nytimes.com/pages/business/international/index.html

http://www.nytimes.com/pages/obituaries/index.html

http://international.nytimes.com

Available Collections

Created At	Seed-Url	Max-Depth	Max-Pages	# of Retrieved	Collection Id	
2015-06-08 19:55:17 UTC	http://www.nytimes.com/	1	100	100	0	○
2015-06-08 20:30:04 UTC	http://stackoverflow.com/	2	1500	622	1	○
2015-06-08 20:43:28 UTC	http://www.bbc.com	2	1000	917	2	○
2015-06-09 09:00:58 UTC	https://www.reddit.com/	3	5000	4621	3	○
2015-06-09 09:08:11 UTC	https://www.stanford.edu/	1	50	50	4	○

Used tools/technologies

- Crawler
- Indexer
  - Lucene indexer
- Query processor
  - DAAT
  - WAND
  - MMR
- User interface
  - Ruby on Rails web framework