Sina News Search Engine

Domain Specific Search Engine

Xinshuo Hu[1]

[1] Harbin Institute of Technology, Shenzhen
Department of Computer Science and Technology

10 July 2020

Outline

System Framework

Main Frameworks

- Crawler : Scrapy
- Database : MySQL
- Search Engine : Whoosh (with jieba Chinese Analyzer)
- Web Framework : Django

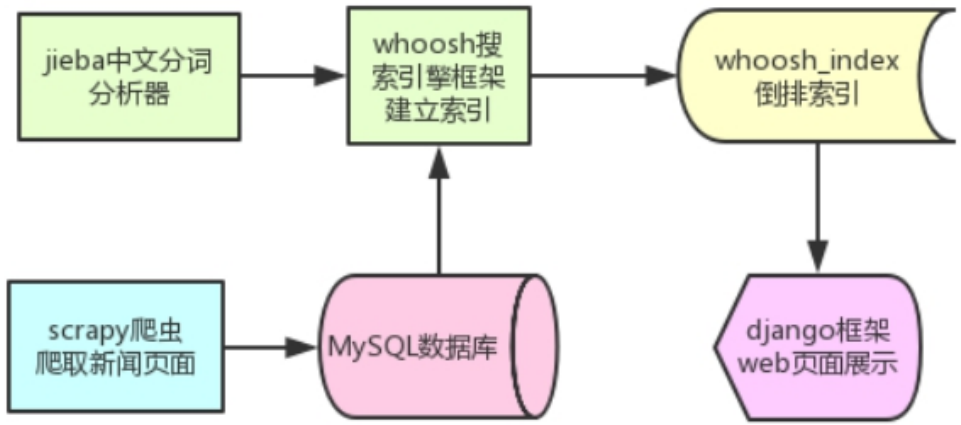

FIGURE – System Framework

Scrapy : Crawl for Sina News Webpages

- Crawl for Sina News Webpages under https ://news.sina.com.cn/
- Breadth First Search in network

Fields of Items

- Uniform Resource Locator of news webpage
- news title
- news press
- news publication time
- news content

Storing Method

- stored in MysQL database
- writed to a JSON file

Whoosh : Build Inversed Index

jieba

"Jieba" (Chinese for "to stutter") Chinese text segmentation : built to be the best Python Chinese word segmentation module
use ChineseAnalyzer fom jieba to analyze documents and queries

Fields of Schema in Whoosh, to build inversed index

- URL of news webpage : type of ID, to identify a webpage
- news title : type of TEXT, for searching
- news press : type of TEXT, for searching
- news publication time : type of TEXT, to score ranking list
- news content : type of TEXT, for searching

searcher for query

- search in fields of newsUrl, newsTitle, newsAuthor, newsContent
- limit of ranking list : 20
- scoring method : BM25F

Django : Render Webpage for Search Engine

- display searched documents nnumber and response time
- highlight the key words in query

FIGURE – search engine webpage

Evaluation

collect 8832 news webpages
evaluate top 5 search result for 10 random query

Precision

- For ranking list :
  (Mean Average Precison) MAP = 0.914861
- For non-rankinig list :
  Average Precision = 0.820000

Response Time

Mean Response Time = 0.087994