ESSIR
The 14th European Summer School on Information Retrieval

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval
MSCA ITN/ETN No. 860721

Group work on Medical IR

Georgios Peikos, Wojciech Kusa, Annisa Maulida Ningtyas

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE
TU WIEN

Ground Rules

Ask anything you need, whenever you want

If you have a question, please ask it out loud

During the hands-on, report on the combinations you try and the performance achieved

Agenda

Introduction to search engines

Indexing

Query Language

Document Ranking and Evaluation

Resources

Hands – on

Experimenting with a Search Engine

Collection of documents

Queries

Lexicon-based ranking models

Neural models

Experimental Evaluation Measures

Human annotated documents – ground truth – labels – qrels

Example of Search Engines

PyTerrier/Terrier

Lucene/Anserini/Pyserini

Elastic Search

INQUERY/Lemur/Indri

SMART

MG4J

TIREx (SIGIR 2023)

How many of you have experience with one of these search engines (or another)?

PyTerrier

PyTerrier is a Python framework.

It relies on Terrier information retrieval toolkit, for indexing and retrieval.

Agenda

Introduction to search engines

Indexing

Query Language

Document Ranking and Evaluation

Resources

Hands – on

Indexing – What can we index?

TREC Collections

- TRECCollectionIndexer

Dataframes

- DFIndexer

Raw files

- FilesIndexer

Dictionaries

- IterDictIndexer

```
<DOC>
<DOCNO>FT923-12914</DOCNO>
<PROFILE>_AN-CGPA3ADFFT</PROFILE>
<DATE>920716
</DATE>
<HEADLINE>
FT  16 JUL 92 / Carrington sees no progress on Bosnia: London Peace Talks
</HEADLINE>
<BYLINE>
   By JUDY DEMPSEY and ROBERT MAUTHNER
</BYLINE>
<DATELINE>
   LONDON
</DATELINE>
<TEXT>
PEACE TALKS on Bosnia-Hercegovina made no progress in London yesterday but
negotiators will attempt to resume their efforts today.
...
Mr Hurd is anxious to use the British presidency of the EC to boost efforts
to find a peaceful solution to the Yugoslav crisis.
</TEXT>
<PUB>The Financial Times
</PUB>
<PAGE>
London Page 2
</PAGE>
</DOC>
```

Indexing - Data Structures – What do we get?

Lexicon

Records the list of all unique terms and their statistics.

Document Index

Records the statistics of all documents.

Inverted Index

Records the mapping between terms and documents. Contains many posting lists.

MetaIndex

Records document metadata, e.g., documents text. (useful for re-ranking)

Direct Index

Records terms for each document.

Indexing – Configuration – What can we control?

PyTerrier indexing configuration:

- Languages and tokenization, also supports pre-tokenization
- Stemming or stopwords removal

```
indexer.setProperty("termpipelines", "Porterstemmer")
```

https://pyterrier.readthedocs.io/en/latest/terrier-indexing.html (Bottom of the page)
http://terrier.org/docs/current/javadoc/org/terrier/indexing/tokenisation/package-summary.html

Agenda

Introduction to search engines

Indexing

Query Language

Document Ranking and Evaluation

Resources

Hands – on

Query Language

Flexible query language

term 1 term 2

+term1 -term2

“term 1 term 2”

```
<top>

<num> Number: 347
<title> Wildlife Extinction

<desc> Description:
The spotted owl episode in America highlighted U.S. efforts
to prevent the extinction of wildlife species.  What is not
well known is the effort of other countries to prevent the
demise of species native to their countries.  What other
countries have begun efforts to prevent such declines?

<narr> Narrative:
A relevant item will specify the country, the involved
species, and steps taken to save the species.

</top>
```

https://github.com/terrier-org/terrier-core/blob/5.x/doc/querylanguage.md

Agenda

Introduction to search engines

Indexing

Query Language

Document Ranking and Evaluation

Resources

Hands – on

Retrieval Models

Lexicon-based Models

BM25, Divergence-from-randomness models, etc.

http://terrier.org/docs/current/javadoc/org/terrier/matching/models/package-summary.html

Dense Retrievers & Neural re-rankers

ColBERT, monoT5

https://pyterrier.readthedocs.io/en/latest/neural.html

Evaluation

Ir_measures

- Python package

Retrieved but unjudged documents

- Consider them irrelevant (most common approach)
- Consider them relevant (uncommon approach)
- Exclude them from evaluation (condensed list evaluation)

```
301 0 CR93E-10279 0
301 0 CR93E-10505 0
301 0 CR93E-1282 1
301 0 CR93E-1850 0
301 0 CR93E-1860 0
301 0 CR93E-1952 0
301 0 CR93E-2191 0
301 0 CR93E-2473 0
301 0 CR93E-3103 1
301 0 CR93E-3284 0
301 0 CR93E-38 0
301 0 CR93E-392 0
301 0 CR93E-4648 0
```

https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)

https://pyterrier.readthedocs.io/en/latest/experiments.html#evaluation-measures-objects

https://amitness.com/2020/08/information-retrieval-evaluation/?fbclid=IwAR1kM-U5BJml01FY5CYtAf_CC5trlYt9plFOWLeiZmrGWLqZg6NS5ZGIrAw

Experimentation

Select the Index and the model

```
tfidf = pt.BatchRetrieve(index, wmodel="BM25")
```

Retrieve using a query

```
results = tfidf.transform("term term")
```

Evaluate using the qrels and your result list

```
eval_results = pt.Utils.evaluate(results, dataset_qrels, metrics=[P@5,P@10], perquery=False)
```

Experimentation

All in one

```
pt.Experiment(
[tfidf,BM25],
 path_queries,
 path_qrels,
eval_metrics=[“AP(rel=2)@5", "nDCG@10"])
```

Complex Pipelines: Operators

Employing these operators allow us to create complex retrieval pipelines.

Operator	Meaning
>>	Then - chaining pipes
+	Linear combination of scores
*	Scalar factoring of scores
&	Document Set Intersection
|	Document Set Union
%	Apply rank cutoff
^	Concatenate run with another
**	Feature Union
~	Cache transformer result

https://pyterrier.readthedocs.io/en/latest/operators.html

Agenda

Introduction to search engines

Indexing

Query Language

Document Ranking and Evaluation

Resources

Hands – on

Further Resources

PyTerrier

https://github.com/terrier-org/ecir2021tutorial

Tutorial in Medical Search

https://github.com/ielab/health-search-tutorial

Interesting readings

https://link-springer-com.unimib.idm.oclc.org/article/10.1007/s10791-015-9277-8

https://dl-acm-org.unimib.idm.oclc.org/doi/pdf/10.1145/3462476

Agenda

Introduction to search engines

Indexing

Query Language

Document Ranking and Evaluation

Resources

Hands – on

Hands-on

Running a retrieval pipeline

We will see step-by-step how we can perform a retrieval pipeline

So, open the colab.