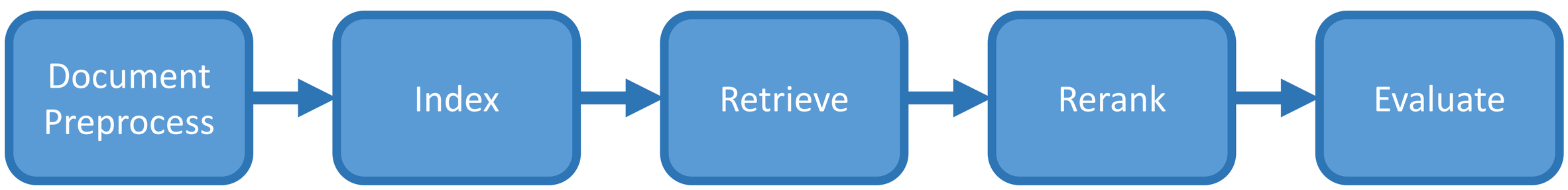
Patapasco: A Python Framework for Cross-Language Information Retrieval Experiments

Cash Costello Eugene Yang Dawn Lawrie James Mayfield
Human Language Technology Center of Excellence, Johns Hopkins University




```yaml
run:
  name: hc4_zho_test

documents:
  input:
    format: irds
    lang: zho
    path: hc4/zh/test
    fields: title+text
  process:
    normalize:
      lowercase: true
    stem: false
    stopwords: lucene
    strict_check: true
    tokenize: spacy

database:
  name: sqlite
  output: true

index:
  name: lucene
  output: true

topics:
  input:
    format: irds
    lang: eng
    source: original
    path: hc4/zh/test
  fields: title

queries:
  output: true
  process:
    normalize:
      lowercase: true
    stem: false
    stopwords: lucene
    tokenize: spacy
  psq:
    lang: eng
    normalize:
      lowercase: true
      report: false
    path: ./eng_zho_transtable.dict
    stem: false
    stopwords: lucene
    threshold: 0.97

retrieve:
  b: 0.4
  k1: 0.9
  name: bm25
  number: 1000
  output: retrieve
  psq: true

score:
  input:
    format: irds
    path: hc4/zh/test
```

Select dataset

Select query

Example Configuration file

While there are high-quality software frameworks for information retrieval experimentation, they do not explicitly support cross-language information retrieval (CLIR). **To fill this gap, we have created Patapsco, a Python CLIR framework.** This framework specifically addresses the complexity that comes with running experiments in multiple languages. Patapsco is designed to be **extensible to many language pairs**, to be **scalable to large document collections**, and to support **reproducible experiments** driven by a configuration file.

	CLEF Persian			CLEF Russian			NTCIR Chinese		
	MAP	nDCG	R@1k	MAP	nDCG	R@1k	MAP	nDCG	R@1k
PSQ	0.1370	0.3651	0.4832	0.2879	0.4393	0.7441	0.1964	0.3752	0.5867
QT	0.2511	0.5340	0.6945	0.3857	0.5527	0.9268	0.2186	0.3953	0.6201
DT	0.3420	0.6808	0.8501	0.3408	0.5151	0.8881	0.3413	0.5627	0.8110
HT	0.4201	0.7476	0.9175	0.3975	0.5623	0.9401	0.4810	0.6840	0.9125

Retrieval Effectiveness Results by Using Patapsco on English Queries and Persian/Russian/Chinese Documents with PSQ, Query Machine Translation (QT), Document Machine Translation (DT), Human Query Translation (HT)

TREC 2022 NeuCLIR Track

https://neuclir.github.io/

- Documents: Chinese / Persian / Russian (released)
- Queries: English (TBA in June)
- Submission due in July
- **Please consider participation! Patapsco makes it easy!**

github.com/hltcoe/patapsco

human language technology center of excellence

JOHNS HOPKINS UNIVERSITY