

NPFL103: Information Retrieval – Assignment 2

Information Retrieval Frameworks

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

Goals and objectives

Goals and objectives

To get experience with open source toolkits for Information Retrieval and learn how to use them to deliver state-of-the-art results on the provided test collection (the test collection is the same as in the Assignment 1).

1. Learn about publicly available information retrieval toolkits and choose one of them.
2. Use the selected toolkit to experiment with various retrieval techniques, pre- and post-processing methods, and other enhancements.
3. Optimize the system on a test collection and a set of training topics.
4. Write a detailed report on your experiments.

Specification

Specification

- ▶ Learn about publicly available information retrieval toolkits, e.g.:
 - ▶ Lemur (<http://www.lemurproject.org/>)
 - ▶ Lucene (<http://lucene.apache.org/>)
 - ▶ Terrier (<http://terrier.org/>)
 - ▶ ...
- ▶ Choose one and install it.

Specification cont'd

1. Design and evaluate a baseline system as similar as possible to the baseline system from Assignment 1 (using the same test collections).
2. Tune the system (Run-1) by selecting the most effective techniques/options on the set of training topics (use Mean Average Precision as the main evaluation measure) and justify your decisions by conducting comparative experiments. The queries in this run must be constructed by automatic means based on topic titles only.
3. You can (optionally) submit one another run with absolutely no restrictions. You can perform manual query construction, use external data resources (thesauri) or third-party tools (e.g. word sense disambiguation).

Data

Test collection

Collection includes:

- ▶ 81,735 documents
- ▶ 50 topics (1–25 for training, 26–50 for testing)
- ▶ 10,145 relevance judgements for the training topics
- ▶ 10,462 relevance judgements for the test topics (not for students)

Topic example:

num: 10.2452/448-A

title: Nobelovy ceny za chemii

description: Najděte dokumenty o laureátech Nobelovy ceny za chemii a jejich konkrétní vědecké práci.

narrative: Relevantní dokumenty by měly obsahovat jména laureátů Nobelovy ceny za chemii a také poskytovat informace o jejich vědeckých výsledcích.

Document example:

docid: LN-20020306012

docnum: LN-20020306012

date: 03/06/02

geography: LONDÝN

text: O vyslání české polní nemocnice do mírových sil ISAF v Afghánistánu bylo v principu rozhodnuto. V Londýně to včera řekl britský ministr obrany Geoff Hoon. Jeho resortní kolega Jaroslav Tvrdík připomněl, že z české strany toto rozhodnutí ještě podléhá schválení vládou a parlamentem. Nemocnice by se podle Tvrdíka starala hlavně o vojáky mírových sil. "Protože se jedná o misi, jejímž hlavním úkolem je podpora nové civilní vlády v Afghánistánu, zapojila by se intenzivně i do plnění úkolů humanitárního či zdravotnického charakteru pro civilní obyvatelstvo." Hoon dodal, že experti obou zemí nyní v Kábulu řeší praktické záležitosti kolem plánovaného umístění nemocnice.

Document format example

```
<DOC>
<DOCID>LN-20020216003</DOCID>
<DOCNO>LN-20020216003</DOCNO>
<DATE>02/16/02</DATE>
<TITLE>
1 Kateřinu   Kateřina_ ;Y   NNFS4-----A---- 2 Atr
2 Neumannovou Neumannová_ ;S   NNFS4-----A---- 3 Obj
3 dělily     dělit_ :T   VpTP---XR-AA--- 0 Pred
4 od          od-1   RR--2----- 3 AuxP
5 druhého    druhý-1_ ^ (jiný) AAIS2----1A---- 6 Atr
6 bronzu     bronz   NNSI2----A---- 4 Adv
7 centimetry centimetr   NNIP1----A---- 6 Atr
</TITLE>
<TEXT>
1 Třicet     třicet`30   Cn-S1----- 3 Sb
2 centimetrů centimetr   NNIP2----A---- 1 Atr
3 chybělo    chybět_ :T_   VpNS---XR-AA--- 0 Pred
4 včera       včera   Db----- 3 Adv
5 nejlepší   dobrý   AAFS1----3A---- 7 Atr
6 české       český   AAFS6----1A---- 7 Atr
7 lyžařce    lyžařka_ ^ (*2) NNFS6----A---- 3 Obj
8 k           k-1   RR--3----- 7 AuxP
9 získání    získání_ ^ (*3at) NNNS3----A---- 8 Atr
10 medaile   medaile   NNFS2----A---- 9 Atr
</TEXT>
```

Topic format example

```
<top lang="cs">
<num>10.2452/448-AH</num>
<title>
 1 Novelovy Novelův UFP1M----- 2 Atr
 2 ceny cena-1_^(v_pen... NNFP1----A---- 0 ExD
 3 za za-1 RR--4----- 2 AuxP
 4 chemii chemie NNFS4----A---- 3 Atr
</title>
<desc>
 1 Najděte najít Vi-P---2--A---- 0 Pred
 2 dokumenty dokument NNIP4----A---- 1 Obj
 3 o o-1 RR--6----- 2 AuxP
 4 laureátech laureát NNMP6----A---- 3 Atr
 5 Nobelovy Nobelův_^(*) AUFS2M----- 6 Atr
 6 ceny cena-1_^(v_pen... NNFS2----A---- 4 Atr
 7 za za-1 RR--4----- 6 AuxP
 8 chemii chemie NNFS4----A---- 7 Atr
 9 a a-1 J^----- 7 Coord
10 jejich jeho_^(přivlast.) PSXXXXP3----- 13 Atr
11 konkrétní konkrétní AAFS4----1A--- 13 Atr
12 vědecké vědecký AAFS6----1A--- 13 Atr
13 práci práce_^(jako_č... NNFS6----A---- 9 Obj
14 .
.
Z:----- 0 AuxK
</desc>
...
```

Format of retrieval results and relevance assessments

sample-res.dat

```
10.2452/401-AH 0 LN-20020201065 0 0.53 run-0
10.2452/401-AH 0 LN-20020102011 1 0.51 run-0
10.2452/401-AH 0 LN-20020601039 2 0.47 run-0
10.2452/401-AH 0 LN-20020604081 3 0.35 run-0
10.2452/401-AH 0 LN-20020731020 4 0.29 run-0
10.2452/401-AH 0 MF-20020128004 5 0.28 run-0
10.2452/401-AH 0 LN-20020102051 6 0.28 run-0
10.2452/402-AH 0 LN-20020601039 0 0.67 run-0
10.2452/402-AH 0 LN-20020601076 1 0.52 run-0
10.2452/402-AH 0 LN-20020604072 2 0.34 run-0
```

Fields:

1. qid – query id, string
2. iter – iteration, integer (unused)
3. docno – document number, string
4. rank – rank, integer starting from 0
5. sim – similarity score
6. run_id – system/run identification

train-qrels.txt

```
10.2452/401-AH 0 LN-20020518024 0
10.2452/401-AH 0 LN-20020518030 0
10.2452/401-AH 0 LN-20020518054 0
10.2452/401-AH 0 LN-20020601039 1
10.2452/401-AH 0 LN-20020601076 0
10.2452/401-AH 0 LN-20020604072 0
10.2452/401-AH 0 LN-20020604081 1
10.2452/401-AH 0 LN-20020607062 0
10.2452/401-AH 0 LN-20020611002 0
10.2452/401-AH 0 LN-20020611069 0
10.2452/401-AH 0 LN-20020611130 0
10.2452/401-AH 0 LN-20020614032 0
10.2452/401-AH 0 LN-20020614068 0
```

Fields:

1. qid
2. iter
3. docno
4. rel – relevance {0,1}

Evaluation

Evaluation

- ▶ The evaluation tool is provided in the "eval" directory.
- ▶ Consult "eval/README" for building instructions.
- ▶ Evaluation is performed by running

```
./eval/trec_eval -M1000 qrels_train_cs.txt run-0_train_cs.res
```

which outputs summary of evaluation statistics:

- ▶ run_id – system/run identification
 - ▶ num_q – number of queries
 - ▶ num_ret – number of returned documents
 - ▶ num_rel – number of relevant documents
 - ▶ num_rel_ret – number of returned relevant documents
 - ▶ map – mean average precision (this is primary evaluation measure)
 - ...
 - ▶ iprec_at_recall_0.00 – Interpolated Recall at 0.00 recall
 - ...
 - ▶ P_10 – Precision of the 10 first documents (this is secondary evaluation measure)
 - ...
- ▶ For details see:
<http://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf>

Example results

runid	all	STANDARD
num_q	all	3
num_ret	all	1500
num_rel	all	561
num_rel_ret	all	131
map	all	0.1785
gm_map	all	0.1051
Rprec	all	0.2174
bpref	all	0.1981
recip_rank	all	0.4064
iprec_at_recall_0.00	all	0.4665
iprec_at_recall_0.10	all	0.3884
iprec_at_recall_0.20	all	0.3186
...		
iprec_at_recall_0.90	all	0.0312
iprec_at_recall_1.00	all	0.0312
P_5	all	0.2667
P_10	all	0.3000
P_15	all	0.3111
...		
P_500	all	0.0873
P_1000	all	0.0437

← Primary evaluation measure

← Secondary evaluation measure

Requirements

Specification details

You will have to deal with the following issues:

- a) extraction of terms from the input data (*data reading, tokenization, punctuation removal, ...*)
- b) equivalence classing of terms (*case folding, stemming, lemmatization, number normalization, ...*)
- c) removing stopwords (*none, frequency/POS/lexicon-based*)
- d) query construction (*automatic, manual*)
- e) topic specification fields used for query construction (*title, desc, narr*)
- f) term weighting (*boolean, natural, logarithm, log average, augmented*)
- g) document frequency weighting (*none, idf, probabilistic idf*)
- h) vector normalization (*none, cosine, pivoted*)
- i) similarity measurement (*cosine, BM25*)
- j) relevance feedback (*none, pseudo-relevance*)
- k) query expansion (*none, thesaurus-based*)

Submission

Run-0: baseline system

Implement and evaluate a baseline system (using the English and Czech data) with the following settings:

- ▶ tokenization: *whitespace+punctuation*
- ▶ class equivalence: *no*
- ▶ removing stopwords: *no*
- ▶ query construction: *all words from "title"*
- ▶ term weighting: *natural*
- ▶ document frequency weighting: *none*
- ▶ vector normalization: *cosine*
- ▶ similarity measurement: *cosine*
- ▶ relevance feedback: *none*
- ▶ query expansion: *none*

Run-1: your best system (constrained)

Select the best-performing method for solving each issue by optimizing the system on the set of training topics (both CS and EN) and justify your decisions by conducting comparative experiments for each solution.

- ▶ tokenization: ???
- ▶ class equivalence: ???
- ▶ removing stopwords: ???
- ▶ query construction: *based on "title" only*
- ▶ term weighting: ???
- ▶ document frequency weighting: ???
- ▶ vector normalization: ???
- ▶ similarity measurement: ???
- ▶ relevance feedback: ???
- ▶ query expansion: ???

Run-2: your best system (unconstrained)

Optionally, you can submit another system (for each language) with no restrictions:

- ▶ tokenization: ???
- ▶ class equivalence: ???
- ▶ removing stopwords: ???
- ▶ query construction: ???
- ▶ term weighting: ???
- ▶ document frequency weighting: ???
- ▶ vector normalization: ???
- ▶ similarity measurement: ???
- ▶ relevance feedback: ???
- ▶ query expansion: ???
- ▶ ...

Submission requirements

Create a single archive (e.g. "jan.novak-1.tgz") with:

- ▶ pdf file with your detailed report, pdf file with presentation slides
- ▶ source code of your system
- ▶ README with details how to build your system and run experiments
- ▶ results of at least two systems (run-0/1) on training and test topics for both Czech and English data

Submission will be done via email.

Grading: 0-100 points (+10 extra points for the performance of the unconstrained system "run-2").