NPFL103: Information Retrieval – Assignment 1

Vector space models

Pavel Pecina

pecina@ufal.mff.cuni.cz

Lecturer

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

Goals and objectives

Goals and objectives

To get experience with implementing vector space models, text preprocessing, system tuning, and experimentation.

1. Develop an experimental IR system based on vector space model.
2. Experiment with methods for text processing, query construction, term weighting, similarity measurement, etc.
3. Optimize the system on the provided test collections
4. Write a detailed report on your experiments.
5. Prepare a presentation and present your results during the course practicals.

Specification

Specification

Use a programming language of your choice and implement an IR system based on vector space model which index the provided collection of documents and generate a ranked list of documents for each topic/query.

Example usage:

```
./run -q topics.xml -d documents.lst -r run -o sample.res
```

Where:

- `-q topics.xml` – a file including topics
- `-d documents.lst` – a file including document filenames
- `-r run` – a string identifying the experiment
- `-o sample.res` – an output file
- ... (additional parameters depending on your implementation)

Data

Test collection

	English	Czech
Documents	88,110	81,735
Training topics	25	
Test topics	25	
Training relevance judgements	10,145	12,739
Test relevance judgements	10,145	10,462

Topic example:

num: 10.2452/448-AH

title: Nobel Prizes for Chemistry

description: Find documents on winners of Nobel Prizes for Chemistry and their specific scientific studies.

narrative: Relevant documents should provide the name(s) of the winner(s) and give some information on their scientific results.

Topic format examples

```
<top lang="en">
<num>10.2452/448-AH</num>
<title>Nobel Prizes for Chemistry</title>
<desc>Find documents on winners of Nobel Prizes for Chemistry and their
specific scientific studies.</desc>
<narr>Relevant documents should provide the name(s) of the winner(s) and
give some information on their  scientific results.</narr>
</top>
```

```
<top lang="cs">
<num>10.2452/448-AH</num>
<title>Novelovy ceny za chemii</title>
<desc>Najděte dokumenty o laureátech Nobelovy ceny za chemii a jejich
konkrétní vědecké práci.</desc>
<narr>Relevantní dokumenty by měly obsahovat jména laureátů Nobelovy ceny
za chemii a také poskytovat informace o jejich vědeckých výsledcích.</narr>
</top>
```

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.

Czech document format (example)

```
<DOC>
<DOCID>LN-20020306012</DOCID>
<DOCNO>LN-20020306012</DOCNO>
<DATE>03/06/02</DATE>
<GEOGRAPHY>LONDÝN</GEOGRAPHY>
<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.
</TEXT>
</DOC>
```

English document format (example)

```
<DOC>
<DOCNO> LA031394-0395 </DOCNO>
<DOCID> 022565 </DOCID>
<SOURCE>
<P>
 Los Angeles Times
</P>
</SOURCE>
<DATE>
<P>
March 13, 1994, Sunday, Orange County Edition
</P>
</DATE>
<SECTION>
<P>
Sports; Part C; Page 13; Column 3; Sports Desk
</P>
</SECTION>
<LENGTH>
<P>
365 words
</P>
</LENGTH>
<HEADLINE>
<P>
BECKLEY ENJOYS HER ROOM WITH CHAMPIONSHIP VIEW
</P>
```

Format of retrieval results and relevance assessments

sample-results.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*)

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

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, e.g. for “run-0”:

```
./run -q topics-train_cs.xml -d documents_cs.lst -r run-0_cs -o run-0_train_cs.res
./run -q topics-train_en.xml -d documents_en.lst -r run-0_en -o run-0_train_en.res
./run -q topics-test_cs.xml -d documents_cs.lst -r run-0_cs -o run-0_test_cs.res
./run -q topics-test_en.xml -d documents_en.lst -r run-0_en -o run-0_test_en.res
```

Submission will be done via email.

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