Data Processing and Storage Hands-on

1 Introduction

In this hands-on you will index the TREC 1-2 test collection. You will use different data processing options (stop-word removal and stemming) and assess their influence on search quality.

2 Setup

The same as in the previous hands-on.

3 Indexing

To build an index of the TREC 1-2 collection, follow these steps:

1. Get acquainted with the "Indexing" part of the page `http://terrier.org/docs/v4.2/quickstart_experiments.html`

2. Enter the Terrier folder
```
cd <Terrier folder>
```

3. Remove the current `etc/collection.spec` file
```
rm etc/collection.spec
```

4. Create a new `etc/collection.spec` file and fill it with paths to files that need to be indexed.
```
bin/trec_setup.sh <path to TREC dataset (containing both volumes)>
```

5. Remove current index from `var/index`
```
rm var/index/*
```

6. Create a new index (it will be stored in `var/index`)
```
bin/trec_terrier.sh -i
```

7. Run some search method, following the previous tutorial. This time your newly built index will be used.
In the `etc/terrier.properties` file, the `TrecQueryTags.skip` property should be set to `DESC,NARR,head,dom,smry,con,fac,def`:
```
TrecQueryTags.skip=DESC,NARR,head,dom,smry,con,fac,def
```

4 Data Processing

To accomplish this part, you will need the following information:

- All Terrier properties are stored in the `etc/terrier.properties` file.

- To change the index path from `var/index` to something else, set property `terrier.index.path=<new index path>`
- To change the data processing options, use property `termpipelines=<comma-separated list of data processing options>`
  The options are:
  - Empty list (`termpipelines=`) – no data processing is performed.
  - `Stopwords` – if present, removes stop-words for indexing and retrieval. The stop-word list can be found in `share/stopword-list.txt`.
  - `PorterStemmer`, `SnowballStemmer` – if present, uses the corresponding stemming for indexing and retrieval.

  Note that if you use certain options for indexing, you must use the same options for retrieval! This is because the data processing options are applied not only to documents, but also to queries. We will discuss this in more detail next week.

Now you need to build indices using different data processing options and evaluate their effect on search quality:

1. Build indices with the following data processing options:
   - No data processing
   - Remove stop-words only
   - Use the Porter stemming only
   - Use the snowball stemming only
2. Measure the size of each index (in KB or MB).
3. Run the BM25 search method using each of the created indices. **Use the same data processing options for retrieval as you used for indexing!**
4. Use trec_eval to evaluate search quality for different indices.
5. Fill in Table 1 with your results. You will then report this table in the next assignment.

	Index size	P@10	MAP	RR
No processing				
No stop-words				
Porter stemming				
Snowball stemming				

Table 1: Index size and search quality (BM25) for different data processing options.