

Adaptive Filtering Software: Additional Data Sets

Contents

1. [Corpus sizes](#)
2. [RCV1-v2](#)
3. [WIPO-alpha](#)

Besides the small data sets included into this distribution, we have intensively tested this software on two large multi-megabyte data sets. That was the best approximation to "real life" that we could our hands on. You do probably do not need these data, since you likely have a lot of your own. Nonetheless, you may want to obtain those data sets if you want to compare the performance of our software on your data with that on the corpora that we used, and to see, in particular, whether the parameter setting that performed best on our test data are also good in your environment.

Corpus sizes

The two large corpora we intensively used are the following: RCV1-v2 and WIPO-alpha. The main parameters of the corpora are as follows:

Corpus	Training doc #	Test doc #	Topic #	Impact # in the full inverted index of the training set	Avg doc length, the training set	Average # impacts per document in the training set	# terms	Basic kNN operation count
RCV1-v2	23149	781265	103	1757801	120.9	75.9	47152	9.45953e+10
WIPO-alpha	46318	28923	573	23913938	3300.5	516.3	644216	7.00541e+11

The "original operation count" in the table above counts "elementary operations" ($a+=b*c$ operation pairs) performed during all neighborhood finding calls in a complete filtering run with the kNN classifier using the non-truncated index. This includes classifier training, scoring the training set for a report, and scoring the test set. This number is presented to give some idea of the cost of classifying the corpus.

RCV1-v2

This corpus is based on almost a year's worth of Reuters news articles. It was built in cooperation between Reuters and David Lewis.

The corpus is to be interpreted as a fully judged corpus: that is, it is considered that we know that each document belongs only to those categories to which the QREL file explicitly says it belongs. Thus, no negative judgment ("does not belong" information) needs to be explicitly present in the QREL file.

You can contact Dr. Lewis via <http://www.daviddlewis.com/> for information on how to obtain a CD with the corpus in its original format from Reuters. You may have to enter into a legal agreement with Reuters.

Once you have Reuters CD, you need to convert the data to the Lemur input format using the script `lyr1/lyr1.sh` included in the Adaptive Filtering Software Distribution. This is a fairly time-consuming task for an average PC. You may need to modify the script as appropriate for your file system layout. In case of problems, contact the developers at vmenkov@apl.rutgers.edu.

For more information, please refer to the references in the main page of this documentation set, [Information on the RCV1-v2 corpus \(LYRL\)](#).

WIPO-alpha

Dataset

These files have been generated from the wipo-alpha dataset, obtained from <http://www.wipo.int/ibis/datasets/index.html>. It is an English-language patent text dataset from the World Intellectual Property Organization (WIPO).

Below I describe the pre-processing that was carried out before feeding the data to the classifier. On our research system (`mms-02.rutgers.edu`) the data were installed in the directory `/home/share/wipo/wipo-alpha-lemur1`.

The original dataset, in XML format can be found in `/home/share/wipo/wipo-alpha`. I downloaded the corpus from the WIPO site, and then excluded 10 documents with WIPO syntax errors or invalid category ids. The list of excluded files can be found in `/home/share/wipo/move-away.sh`

The document count in the original WIPO-alpha data set was as follows:

```
Training collection: 46,324 docs.
Test collection:    28,926 docs.
```

After excluding the invalid files, the following number remained:

```
Training collection: 46,318 docs.
Test collection:    28,923 docs.
```

Without stemming and stopword removal:

```
-rw-r--r-- 1 vmenkov root 3533585051 Jun 28 03:22 all.parsed
-rw-r--r-- 1 vmenkov root 1565261686 Jun 28 03:15 test.parsed
-rw-r--r-- 1 vmenkov root 1968323365 Jun 28 03:08 train.parsed
```

After term filtering (see below), before stemming and stopword removal:

```
-rw-r--r-- 1 vmenkov root 1496391832 Jun 29 20:15 test.sgml
-rw-r--r-- 1 vmenkov root 1894622370 Jun 29 19:58 train.sgml
```

After term filtering, stemming, and stopword removal:

```
-rw-r--r--    1 vmenkov  root      1883730630 Jun 30 00:48 all.parsed
-rw-r--r--    1 vmenkov  root      839384544 Jun 29 23:39 test.parsed
-rw-r--r--    1 vmenkov  root      1044346086 Jun 29 20:56 train.parsed
```

Document text

Stage 1: compiling the document text.

The text of each document in the corpus has been formed by concatenating the Title, Abstract, Claims, and Text fields of the patent (the "ti", "cl", "ab", "txt" elements in the original XML files).

Stage 2: term filtering.

It was found out that if the corpus formed at stage 1 were to be used to build a Lemur index, even with the SMART stopwords removal and Porter stemming, the number of unique terms in the resulting index would be unacceptably large: 1.25 million in the training set, and more than 2 million in the entire corpus (training+test).

On investigation, it was found that certain documents in the corpus contain a large amount (up to 2 megabytes) of, apparently, computer-generated text describing, for example, nucleotide sequences in DNA, or aminoacid sequences in proteins. There are also certain number of documents where some word-separating spaces in the input text have been lost during the OCR conversion.

To bring the vocabulary to a manageable size, I applied the following filtering rules to the corpus:

- removing all terms containing any digits;
- removing all terms longer than 25 characters;
- removing all terms consisting solely of letters ACTG that are longer than 4 characters.

This technique has brought the total number of the unique terms (post-stemming and post-stopword removal) to 0.64 million in the training set and 1.1 million in the entire corpus, at the cost of losing only 3-4% of the overall "text".

Stage 3: Regular Lemur parsing.

At this stage, we used stopwords removal (429-term SMART stopwords list) and Porter parsing, as implemented in Lemur's ParseToFile utility.

QRELS.

The corpus is to be interpreted as a fully judged corpus.

Each document contains one main IPC (International Patent Classification) code, and, possible, some additional IPC code.

The classifications are hierarchical and very deep; e.g., D21H01120:

Section	D: textiles, paper
Class	D21: paper & cellulose
Subclass	D21H: pulp compositions, preparation...
Main group	D21H 11: Pulp or paper, comprising cellulose or lignocellulose fibres of natural origin only
Subgroup	D21H 11/20: Chemically or biochemically modified fibres.

In the English collection we are given (wipo-alpha), the main IPC symbols represent 8 sections (A to H), 114 classes, and 451 subclasses. Secondary IPC symbols may occasionally belong to classes or subclasses not represented among the main IPC symbols.

In the training collection, with 46,324 docs total, each subclass (among those found in the main IPC symbols) contains between 20 and 2000 documents (on average, 103 docs per subclass); in the test collection, half as much.

I have chosen to create topics based on the Section, Class, and Subclass components present in IPC symbols. Thus, a patent with the main IPC symbol "D21H01120" will be assign to the topics "D", "D21", and "D21H", plus similarly defined topics based on the additional IPC symbols that the patent may contain.

The resulting QREL files contain 187689 judgments for the training set and 117765 judgments for the test set.

Topic list ("query")

I have made the decision to form the topic list for classification (the "query" file) only from the topics based on the main IPC symbols. This ensures that each topic listed in the query file has some non-empty representation in both training and test sets, and avoids the situation (occurring e.g. in RCV1-v2) of having an occasional topic with no representation at all in either the training set or the test set.

This of course means that the QREL files contain occasional lines entries referring to topics not included in the "query list" (901 out of 187689 entries in train.qrel, and 373 out of 117765 entries in test.qrel). These entries are ignored by our Adaptive Filtering application.

Conversion scripts

All the conversion scripts (some of which are actually Java programs) used for the pre-processing described above can be found in the Adaptive Filtering Software distribution, in subdirectory `wipo`. The Java code can be compiled using Apache Ant, with `wipo/build.xml` as the build file. The main Java class is `wipo.WIPOConvert`. Please contact the developer at `vmenkov@apl.rutgers.edu` for further details.

Sample config file

A sample configuration file for using this test set can be found in `/home/vmenkov/runs-wipo/knn1`