Ian Taylor

CS7263: Information Retrieval

Summer 2025

web crawling

The `crawler.py` script starts with the seed URL for the textbook and crawls recursively. It keeps track of visited and discovered URLS, and saves them as .json files.

Here is a snippet of a scraped website:


```
{
  "url": "http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-xml-retrieval-1.html",
  "title": "Evaluation of XML retrieval",
  "body": "\n\n\n\n\nEvaluation of XML retrieval\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n Next: Text-centric vs. data-centric XML\n Up: XML retrieval\n Previous: A vector space model\n    Contents \n Index\n\n\n\n\n \n\nEvaluation of XML retrieval
  ...
```


The next step is to handle the body output and save the processed contents to a new file.

text preprocessing

For this step, we use `nltk` to tokenize and remove stopwords. Any HTML tags are removed via `beautifulsoup`, cases are folded to lowercase, extra spaces and newlines are removed, and punctuation is stripped. The processed file is also saved as .json in a different directory.

Here is a snippet of a processed site:


```
[
  "evaluation",
  "of",
  "xml",
  "retrieval",
  "next",
  "textcentric",
  "vs",
  "datacentric",
  "xml",
  "up",
```

```
  "xml",
  "retrieval",
  "previous",
  "a",
  "vector",
  "space",
  "model",
  "contents",
  ...
```

statistics

Now that the URLs are processed and words are tokenized, we can determine some statistics about the corpus we just created.

```
Total number of documents: 291
Total number of tokens: 185619
Number of unique words: 9469
Average page length (in words): 637.87
```

Top 30 most frequent words (with collection and document frequencies):

```
token | collection freq | document freq
the | 11642 | 290
of | 7108 | 289
a | 4907 | 280
in | 4717 | 290
and | 4218 | 283
to | 4010 | 290
is | 3720 | 287
for | 2564 | 277
we | 2016 | 242
that | 1858 | 258
as | 1627 | 260
are | 1391 | 241
this | 1389 | 251
an | 1109 | 287
be | 1095 | 235
document | 1057 | 190
documents | 1048 | 196
with | 1035 | 239
index | 1029 | 288
on | 991 | 241
query | 987 | 166
by | 856 | 228
can | 799 | 214
retrieval | 784 | 136
```

```
at | 751 | 288
page | 748 | 286
it | 746 | 225
not | 732 | 211
from | 709 | 209
or | 655 | 208
```

Top 30 most frequent words after removing stop words:

```
document: 1057
documents: 1048
index: 1029
query: 987
retrieval: 784
page: 748
information: 645
term: 643
terms: 642
classification: 639
text: 595
next: 592
web: 573
clustering: 558
figure: 541
contents: 539
previous: 538
search: 538
model: 522
may: 521
section: 462
two: 461
case: 455
set: 447
one: 444
example: 433
number: 427
relevance: 414
vector: 413
book: 377
```