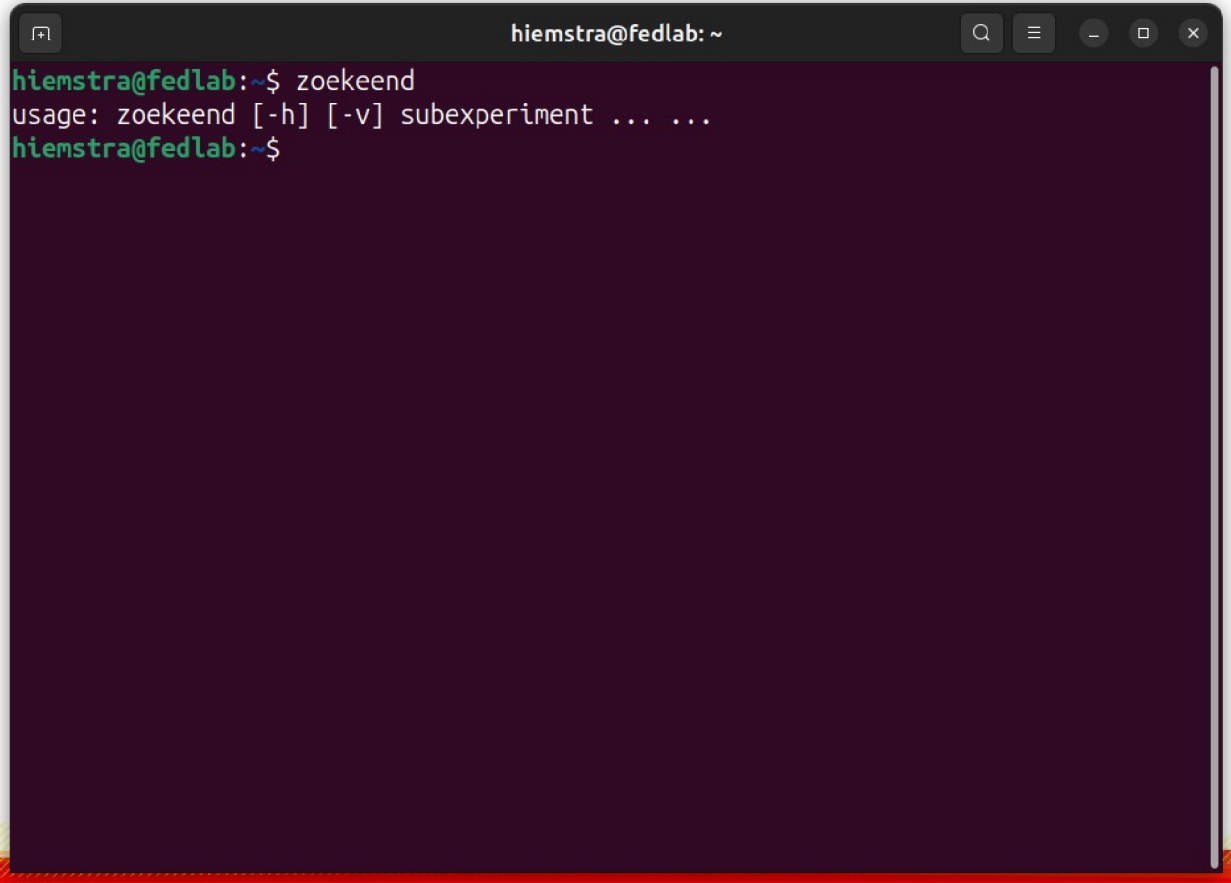
Zoekeend

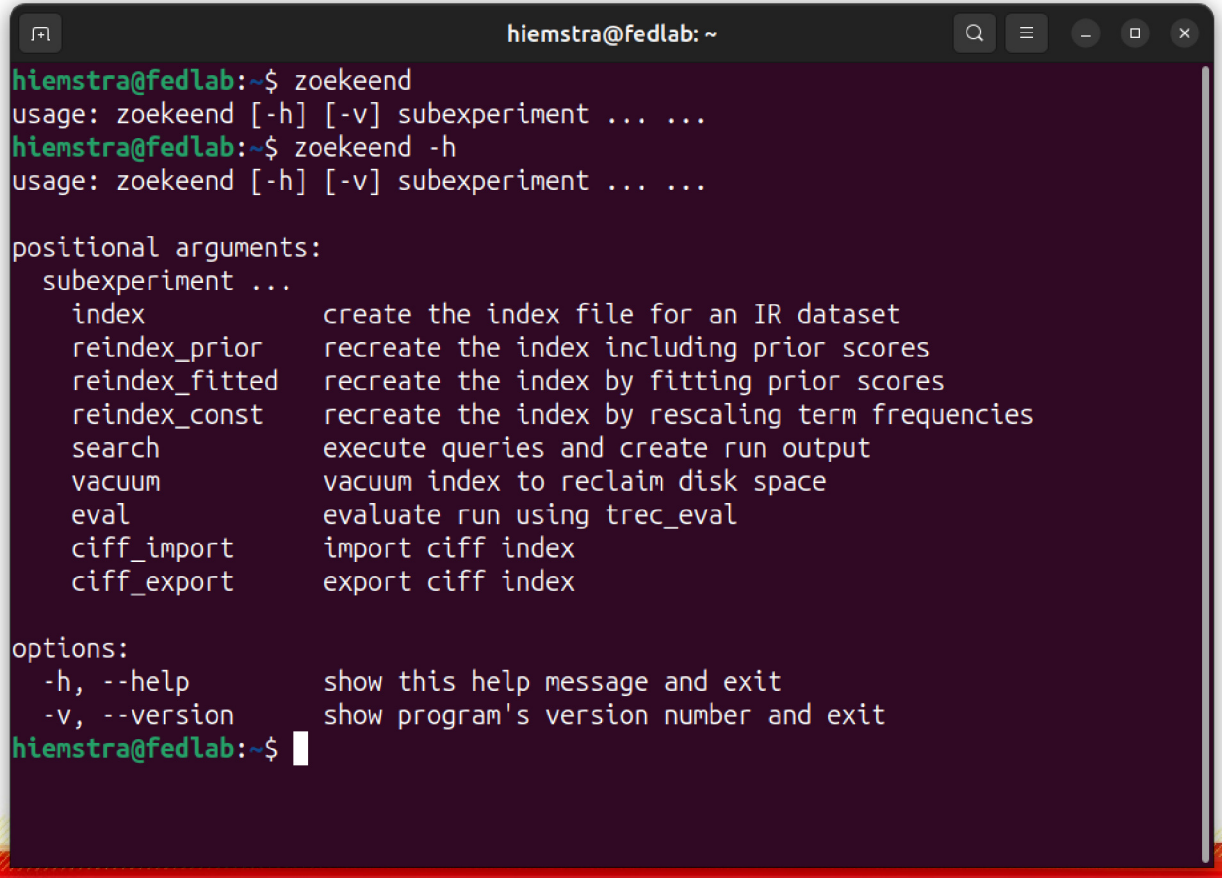
Seach engine experiments using DuckDB

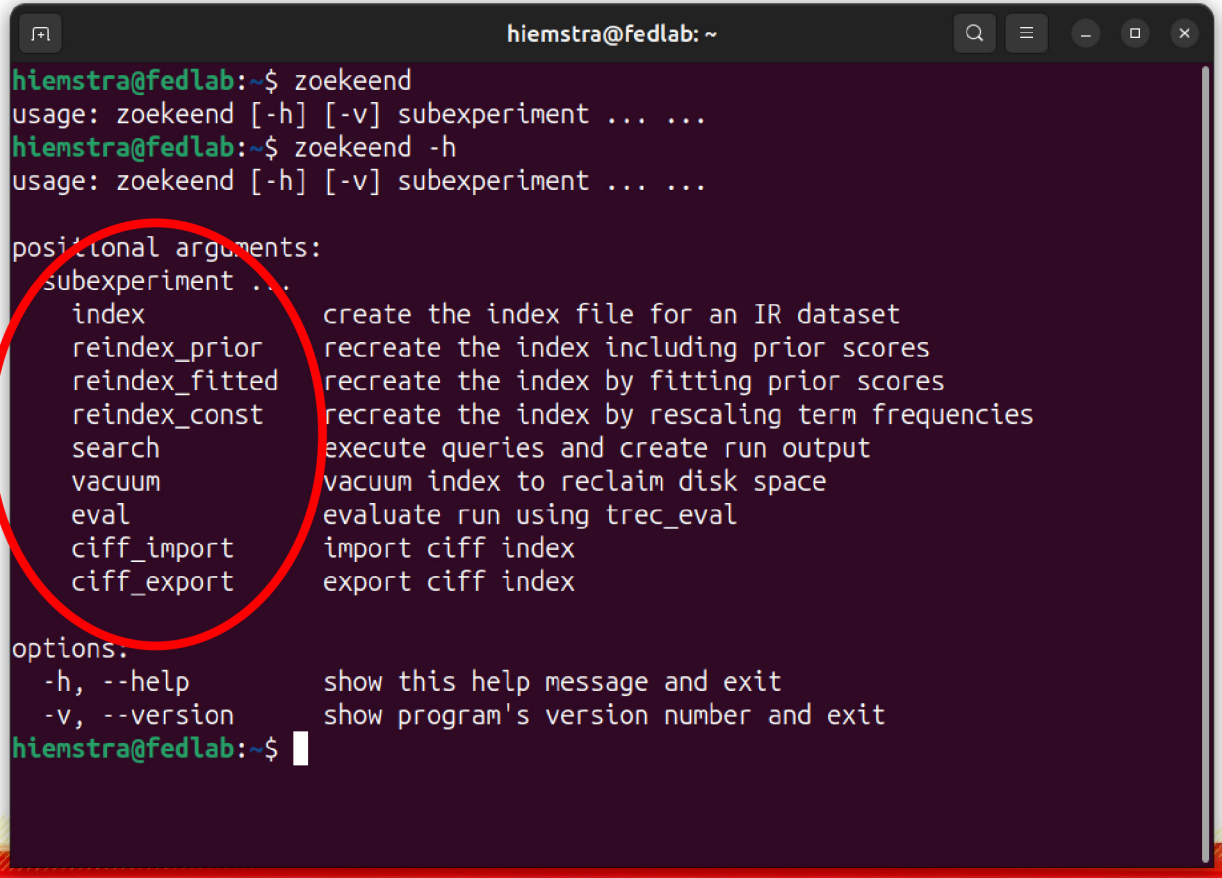
Djoerd Hiemstra

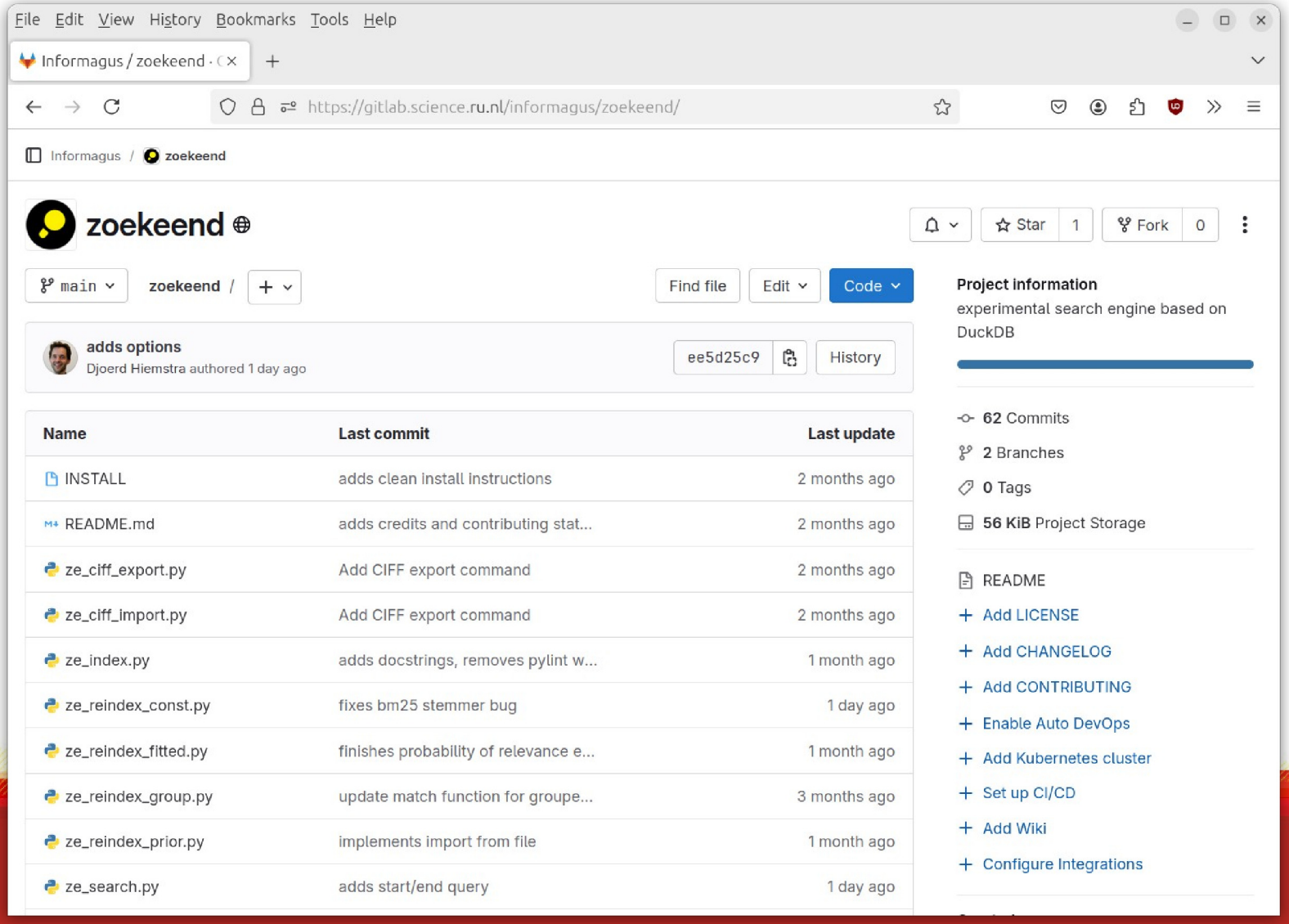
https://www.cs.ru.nl/~hiemstra

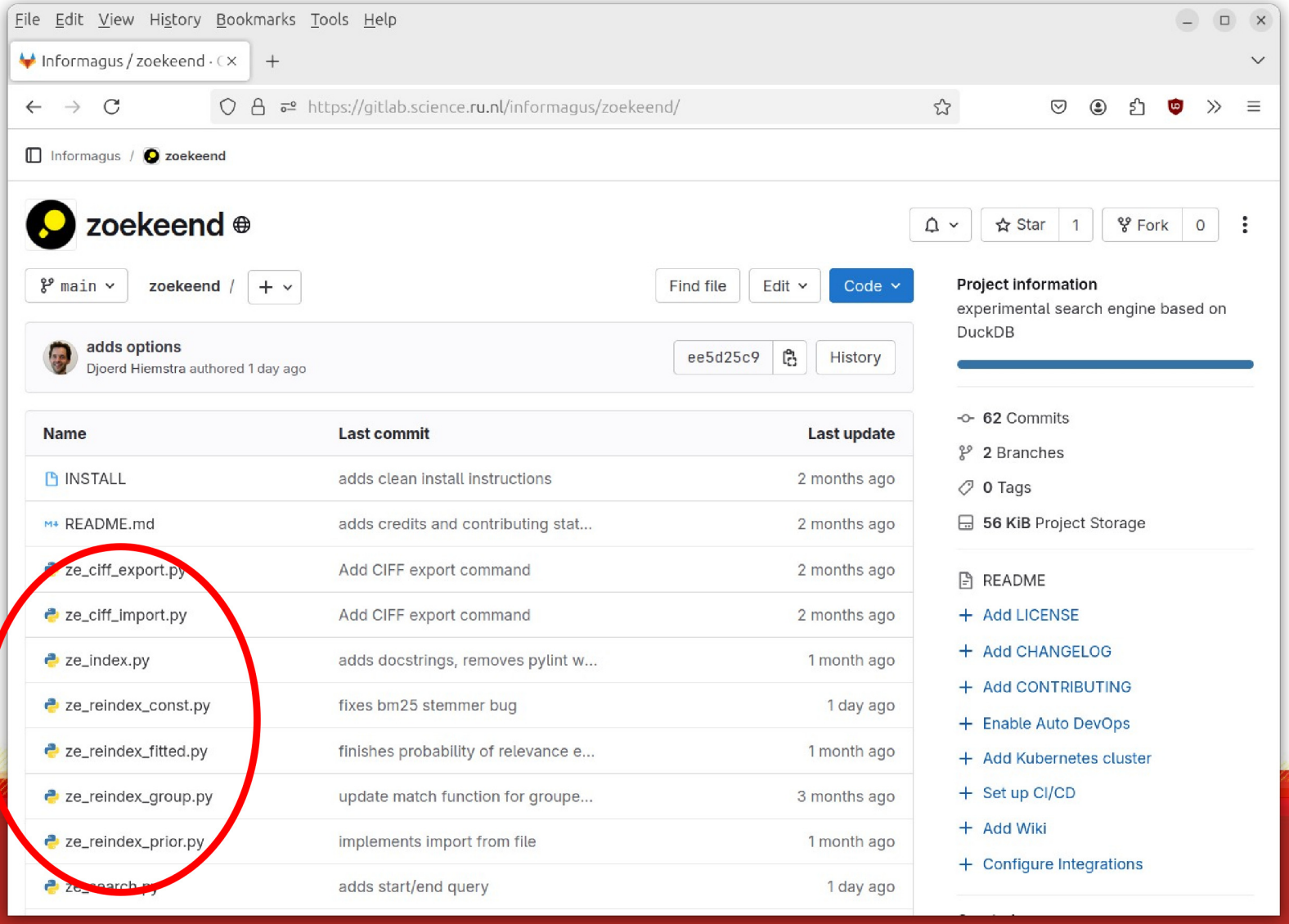
Radboud University


hiemstra@fedlab: ~
hiemstra@fedlab:~$ zoekeend
usage: zoekeend [-h] [-v] subexperiment ... ...
hiemstra@fedlab:~$


hiemstra@fedlab: ~
hiemstra@fedlab:~$ zoekeend
usage: zoekeend [-h] [-v] subexperiment ... ...
hiemstra@fedlab:~$ zoekeend -h
usage: zoekeend [-h] [-v] subexperiment ... ...
positional arguments:
subexperiment ...
index create the index file for an IR dataset
reindex_prior recreate the index including prior scores
reindex_fitted recreate the index by fitting prior scores
reindex_const recreate the index by rescaling term frequencies
search execute queries and create run output
vacuum vacuum index to reclaim disk space
eval evaluate run using trec_eval
ciff_import import ciff index
ciff_export export ciff index
options:
-h, --help show this help message and exit
-v, --version show program's version number and exit
hiemstra@fedlab:~$


hiemstra@fedlab: ~
hiemstra@fedlab:~$ zoekeend
usage: zoekeend [-h] [-v] subexperiment ... ...
hiemstra@fedlab:~$ zoekeend -h
usage: zoekeend [-h] [-v] subexperiment ... ...
positional arguments:
subexperiment ...
index create the index file for an IR dataset
reindex_prior recreate the index including prior scores
reindex_fitted recreate the index by fitting prior scores
reindex_const recreate the index by rescaling term frequencies
search execute queries and create run output
vacuum vacuum index to reclaim disk space
eval evaluate run using trec_eval
ciff_import import ciff index
ciff_export export ciff index
options:
-h, --help show this help message and exit
-v, --version show program's version number and exit
hiemstra@fedlab:~$

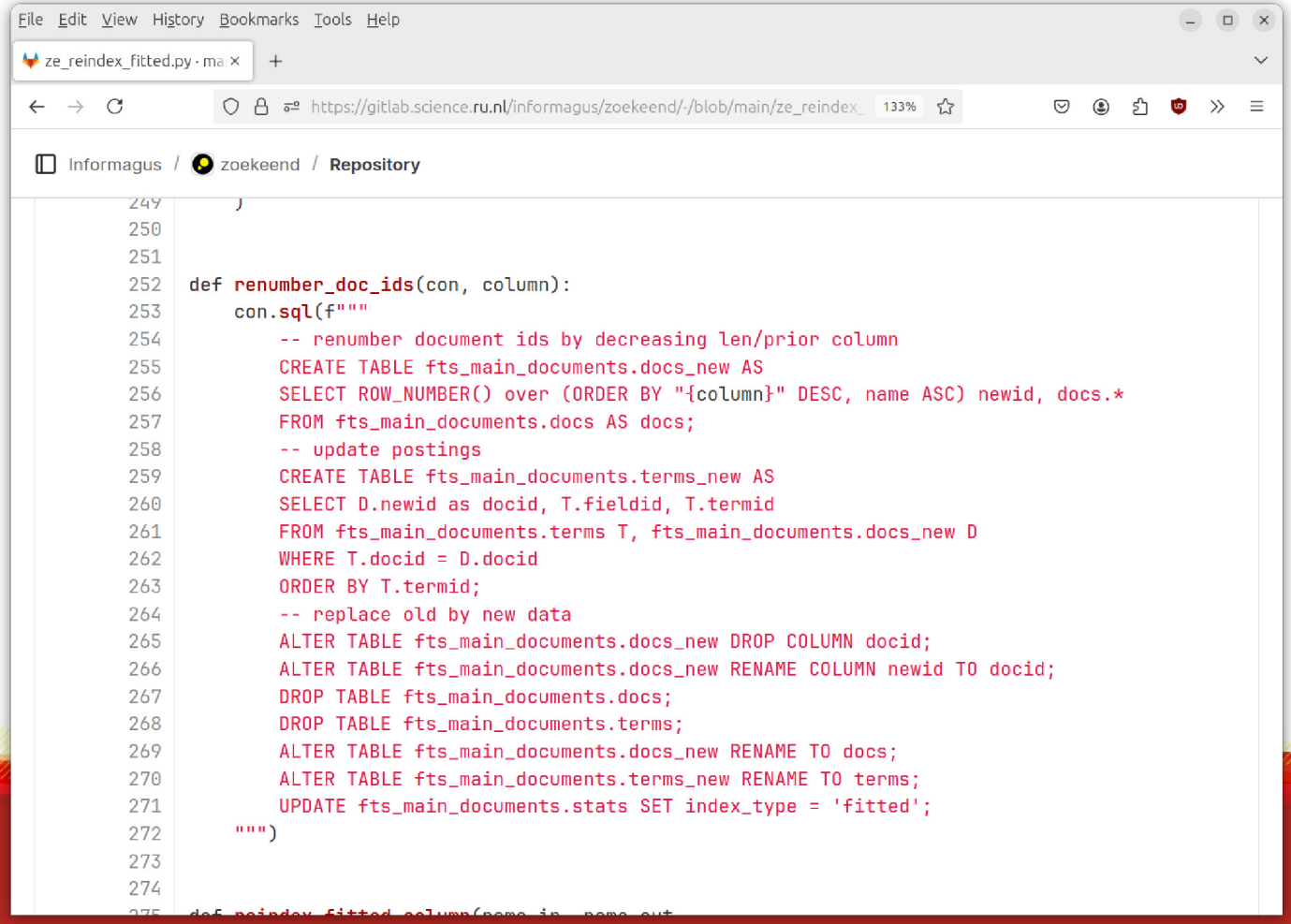

File Edit View History Bookmarks Tools Help
Informagus / zoekeend
https://gitlab.science.ru.nl/informagus/zoekeend/
Informagus / zoekeend
zoekeend
Star 1
Fork 0
main
zoekeend /
Find file
Edit
Code
adds options
Djoerd Hiemstra authored 1 day ago
ee5d25c9
History
Name
Last commit
Last update
INSTALL
adds clean install instructions
2 months ago
README.md
adds credits and contributing stat...
2 months ago
ze_ciff_export.py
Add CIFF export command
2 months ago
ze_ciff_import.py
Add CIFF export command
2 months ago
ze_index.py
adds docstrings, removes pylint w...
1 month ago
ze_reindex_const.py
fixes bm25 stemmer bug
1 day ago
ze_reindex_fitted.py
finishes probability of relevance e...
1 month ago
ze_reindex_group.py
update match function for groupe...
3 months ago
ze_reindex_prior.py
implements import from file
1 month ago
ze_search.py
adds start/end query
1 day ago
Project information
experimental search engine based on DuckDB
62 Commits
2 Branches
0 Tags
56 KiB Project Storage
README
Add LICENSE
Add CHANGELOG
Add CONTRIBUTING
Enable Auto DevOps
Add Kubernetes cluster
Set up CI/CD
Add Wiki
Configure Integrations


File Edit View History Bookmarks Tools Help
Informagus / zoekeend
https://gitlab.science.ru.nl/informagus/zoekeend/
Informagus / zoekeend
zoekeend
Star 1
Fork 0
main
zoekeend /
Find file
Edit
Code
adds options
Djoerd Hiemstra authored 1 day ago
ee5d25c9
History
Name
Last commit
Last update
INSTALL
adds clean install instructions
2 months ago
README.md
adds credits and contributing stat...
2 months ago
ze_ciff_export.py
Add CIFF export command
2 months ago
ze_ciff_import.py
Add CIFF export command
2 months ago
ze_index.py
adds docstrings, removes pylint w...
1 month ago
ze_reindex_const.py
fixes bm25 stemmer bug
1 day ago
ze_reindex_fitted.py
finishes probability of relevance e...
1 month ago
ze_reindex_group.py
update match function for groupe...
3 months ago
ze_reindex_prior.py
implements import from file
1 month ago
adds start/end query
1 day ago
Project information
experimental search engine based on DuckDB
62 Commits
2 Branches
0 Tags
56 KiB Project Storage
README
Add LICENSE
Add CHANGELOG
Add CONTRIBUTING
Enable Auto DevOps
Add Kubernetes cluster
Set up CI/CD
Add Wiki
Configure Integrations

My secret

```
import argparse
```

The *real* secret: SQL!




```
249     )
250
251
252 def renumber_doc_ids(con, column):
253     con.sql(f"""
254         -- renumber document ids by decreasing len/prior column
255         CREATE TABLE fts_main_documents.docs_new AS
256         SELECT ROW_NUMBER() over (ORDER BY "{column}" DESC, name ASC) newid, docs.*
257         FROM fts_main_documents.docs AS docs;
258         -- update postings
259         CREATE TABLE fts_main_documents.terms_new AS
260         SELECT D.newid as docid, T.fieldid, T.termid
261         FROM fts_main_documents.terms T, fts_main_documents.docs_new D
262         WHERE T.docid = D.docid
263         ORDER BY T.termid;
264         -- replace old by new data
265         ALTER TABLE fts_main_documents.docs_new DROP COLUMN docid;
266         ALTER TABLE fts_main_documents.docs_new RENAME COLUMN newid TO docid;
267         DROP TABLE fts_main_documents.docs;
268         DROP TABLE fts_main_documents.terms;
269         ALTER TABLE fts_main_documents.docs_new RENAME TO docs;
270         ALTER TABLE fts_main_documents.terms_new RENAME TO terms;
271         UPDATE fts_main_documents.stats SET index_type = 'fitted';
272     """)
273
274
275 def reindex_fitted_column(name_in, name_out
```

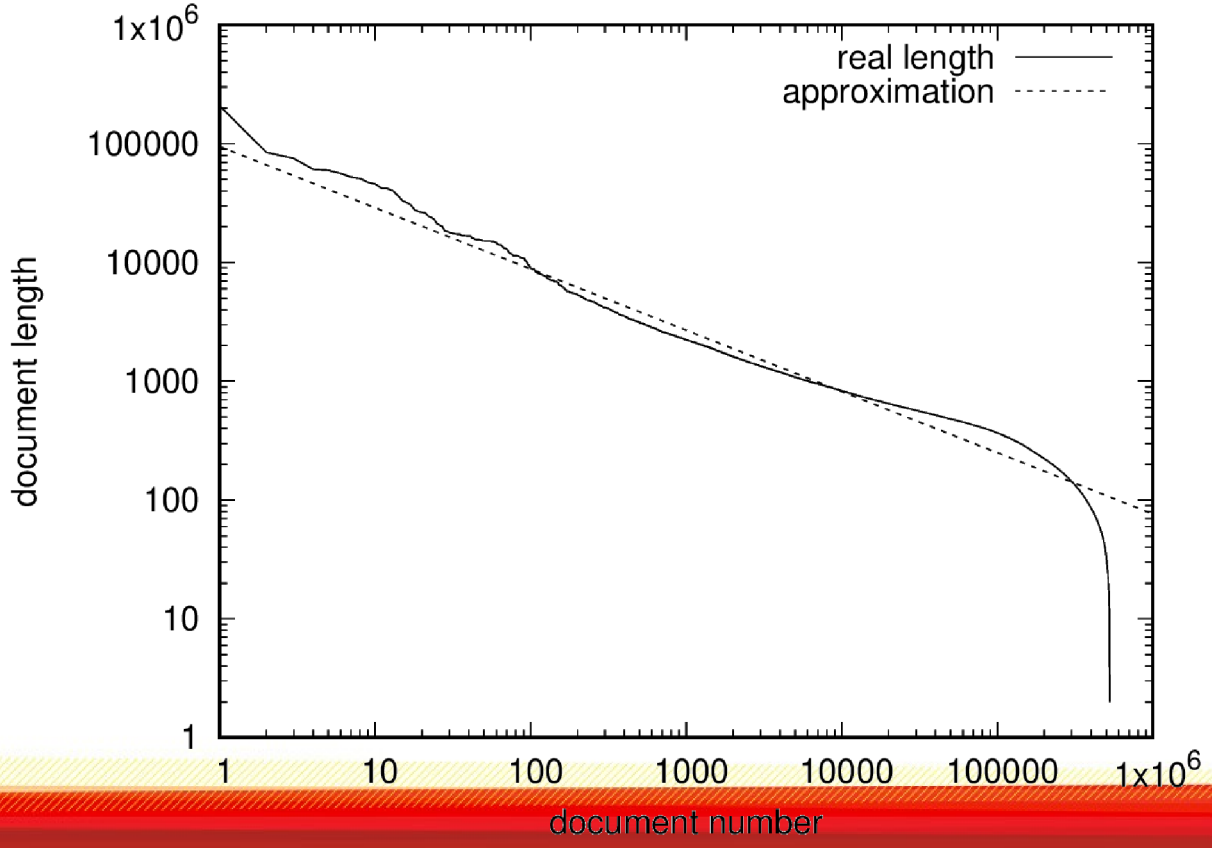
No more document lengths!

- **Challenge**: Search engines like Lucene, Terrier (and DuckDB FTS) store all document lengths.
  - Scanned for every query!
  - Grows linearly with collection!

- **Solution**: Remove document lengths
  - … but approximate

```
ALTER TABLE docs DROP COLUMN len;
```

Idea: The score-fitted index (data: Robust'04)



Results

collection	experiment	MAP	$nDCG_{10}$
Robust'04	BM25 original	0.221	0.448
Robust'04	BM25 fitted	0.219	0.446
Robust'04	LM original	0.220	0.424
Robust'04	LM fitted	0.219	0.425
MS MARCO	BM25 original	0.307	0.400
MS MARCO	BM25 fitted	0.301	0.405
MS MARCO	LM original	0.330	0.408
MS MARCO	LM fitted	0.327	0.417
WT10g	BM25 original	0.185	0.276
WT10g	BM25 fitted	0.189	0.287
WT10g	LM original	0.174	0.251
WT10g	LM fitted	0.173	0.251

Table 1: Results of original and fitted approaches

Other Zoekeend results

- SQL Mastery Assignments, Information Modelling & Databases, 2024/2025
- Timo van Straaten, Integrating Static Index Pruning methods into Zoekeend, *Bsc Thesis, Radboud University*, January 2025
- Yannick Voncken, Adverserial Machine Learning for indexing, *BSc thesis* (to be submitted)

@djoerd@idf.social