# Author: Yunbin Tu

# System Design

The system is using flask and jinja, implementing a standard MVC framework.

#### Api

* “/” the home page for the service
* “/results” method=post using post request to retrieve matched queries from database, storing the result in the session.
* "/results/<int:page\_id>" method=post using post and a path parameter to get the coordinate page of the result, the function next\_page(page\_id) is called, retrive the result from the session.
* "/doc\_data/<doc\_id>" using get method and a path parameter doc\_id to get the coordinate detail of doc

#### Frontend

* Doc.html

A template showing the author, date, content etc. of a specific doc.

* Home.html

Home page for the app, contains a search box for the user input.

* Result.html

Result page, showing the result, has a next button if there’s more content in the next page.

#### Data Storage

Mongodb database was used to store index and wapo\_doc.

collection called "wapo\_docs"

- add a unique ascending index on the key "id"

- insert documents into the "wapo\_docs" collection

One collection called "vs\_index":

- add a unique ascending index on the key "term"

- insert posting lists (index\_list) into the "inverted\_index" collection\

Other collection called "doc\_len\_index"

- add a unique ascending index on the key "doc\_id"

- insert list of document vector length (index\_list) into the "doc\_len\_index" collection

Session was used to store the temporary query results.

# Description and test queries

## Small test file

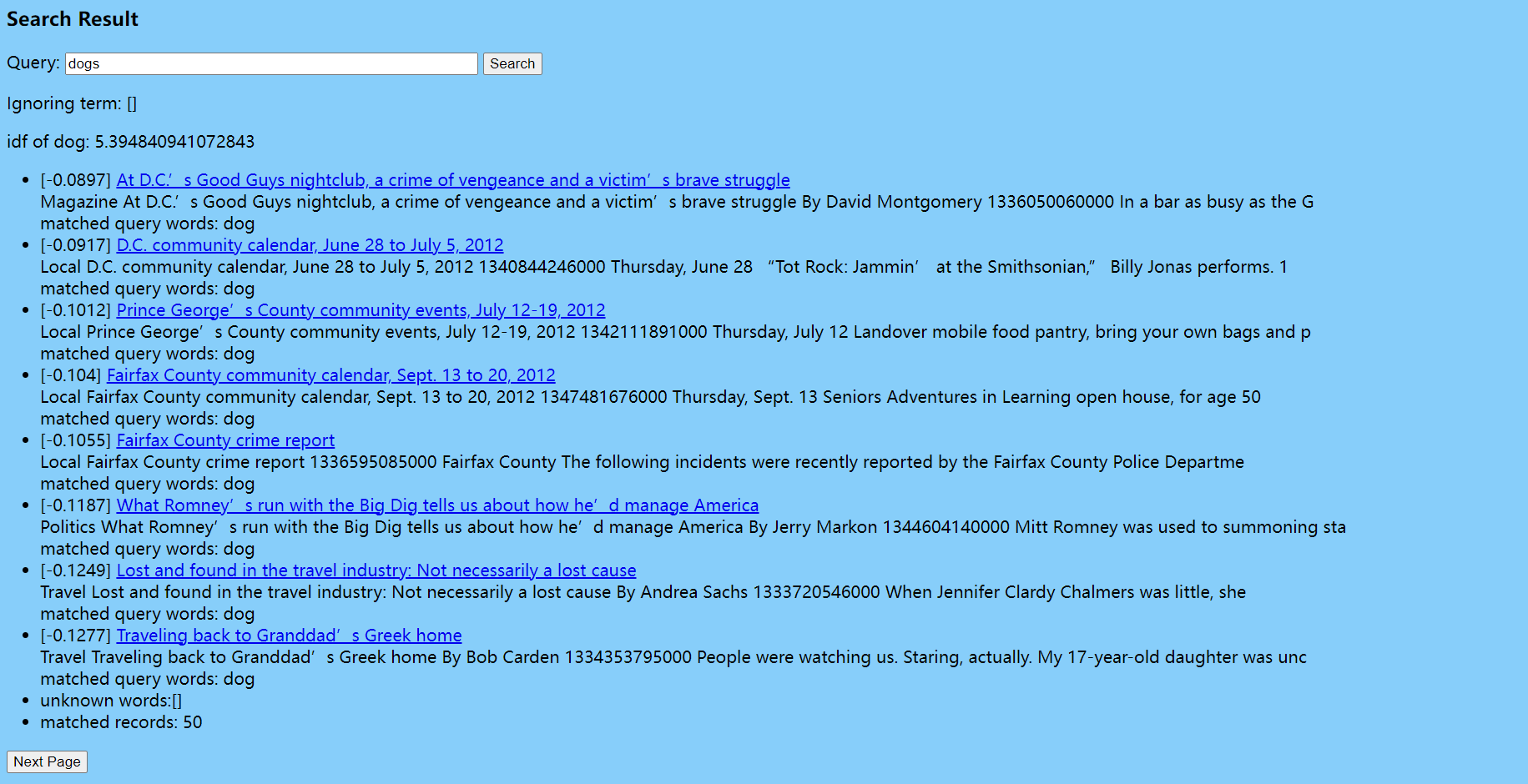
The first 10 records was selected from the original file, building a new file call smalltest.jl.

Then using the first 10000 records to test the actual behavior when scaling up.

## Test results

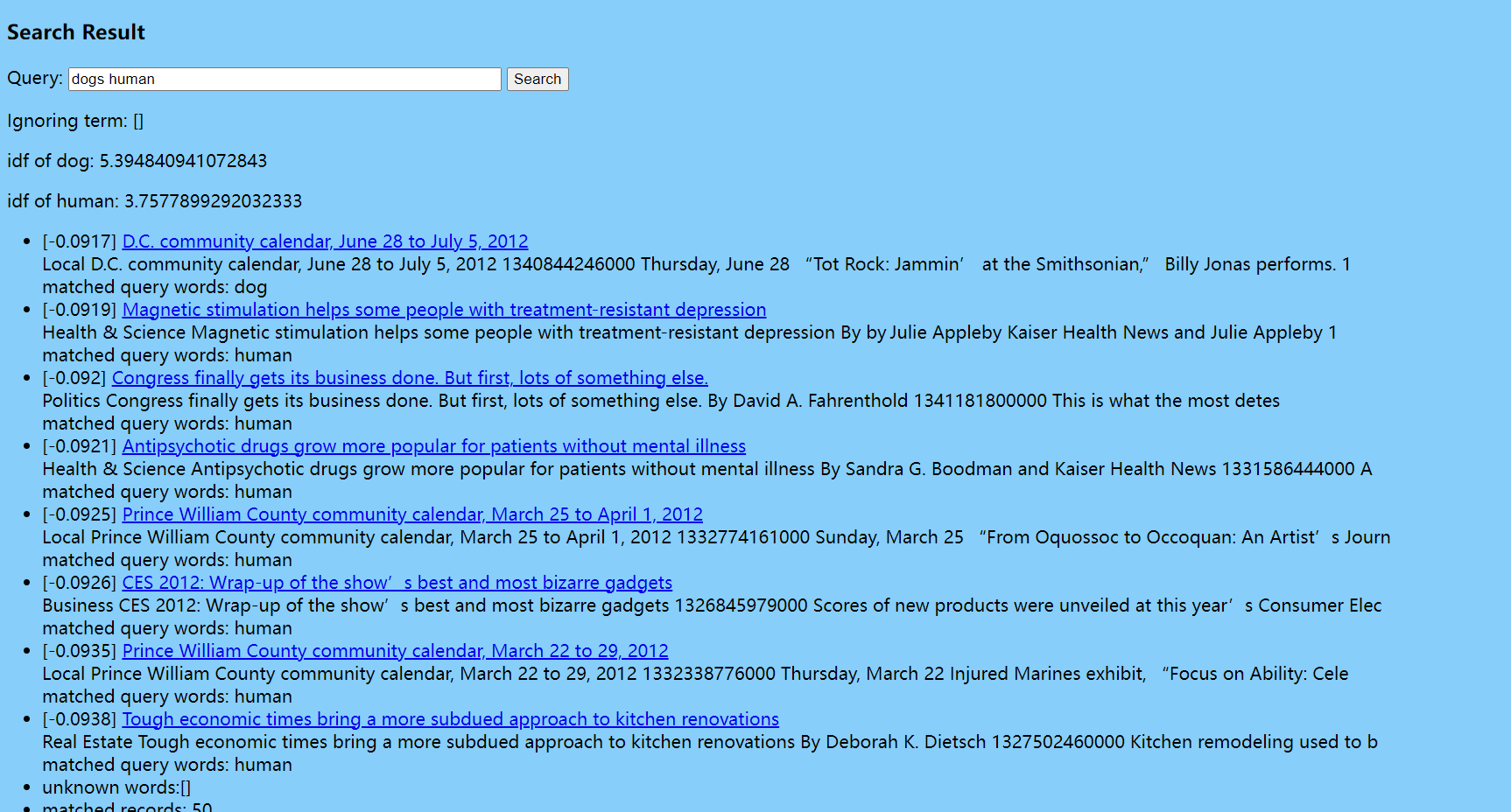
**Note that the parameter k in topk is selected to 50.**

### Query: dogs



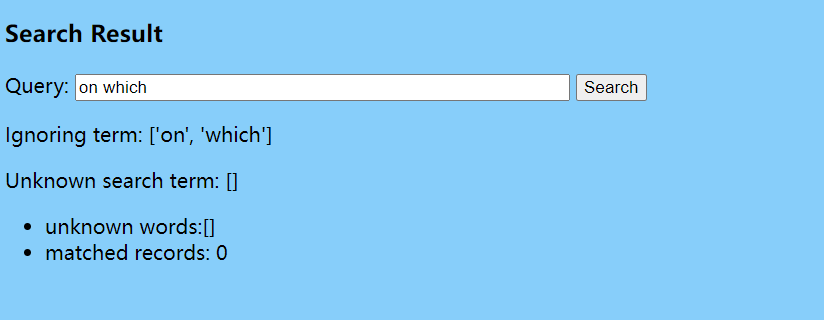
One word query working. The matching is listed from the highest similarity to the lowest. Only the top 50 mathcing records was selected.

### Query: dogs human



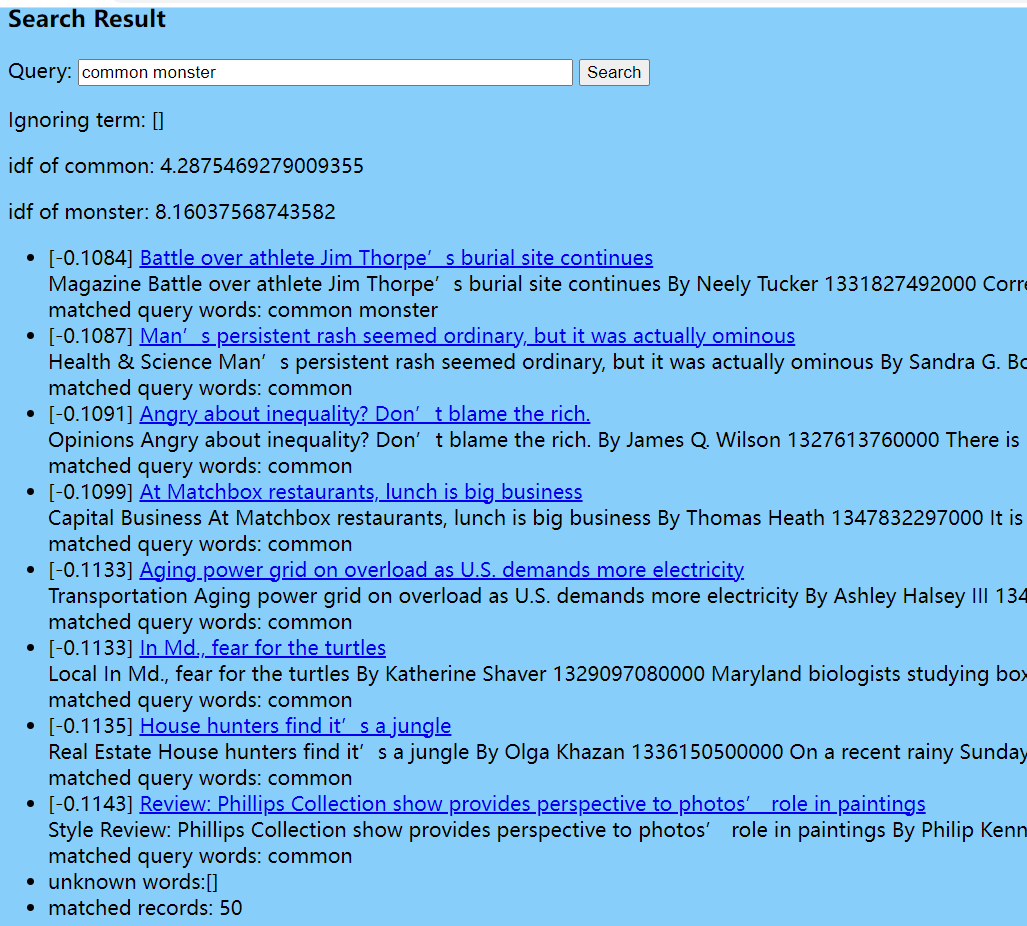
Notice that dog has higher idf than human, which means ‘dog’ is more informational than ‘human’, which is more or less expected.

### Query: on which



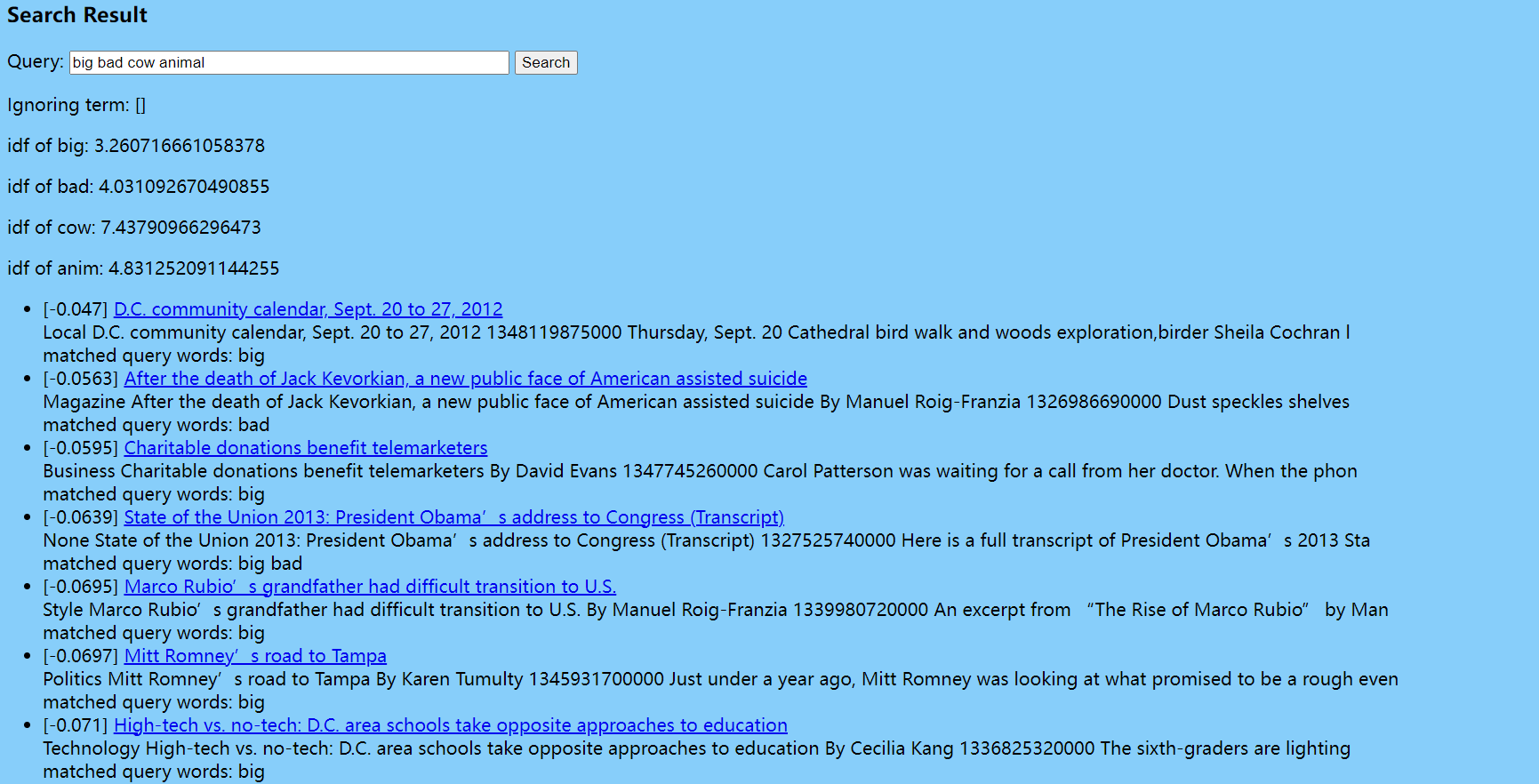
They are both stop words, so they are just negelected. And showing no results could be acceptable though we could present an error message to tell the user why.

### Query: common monster



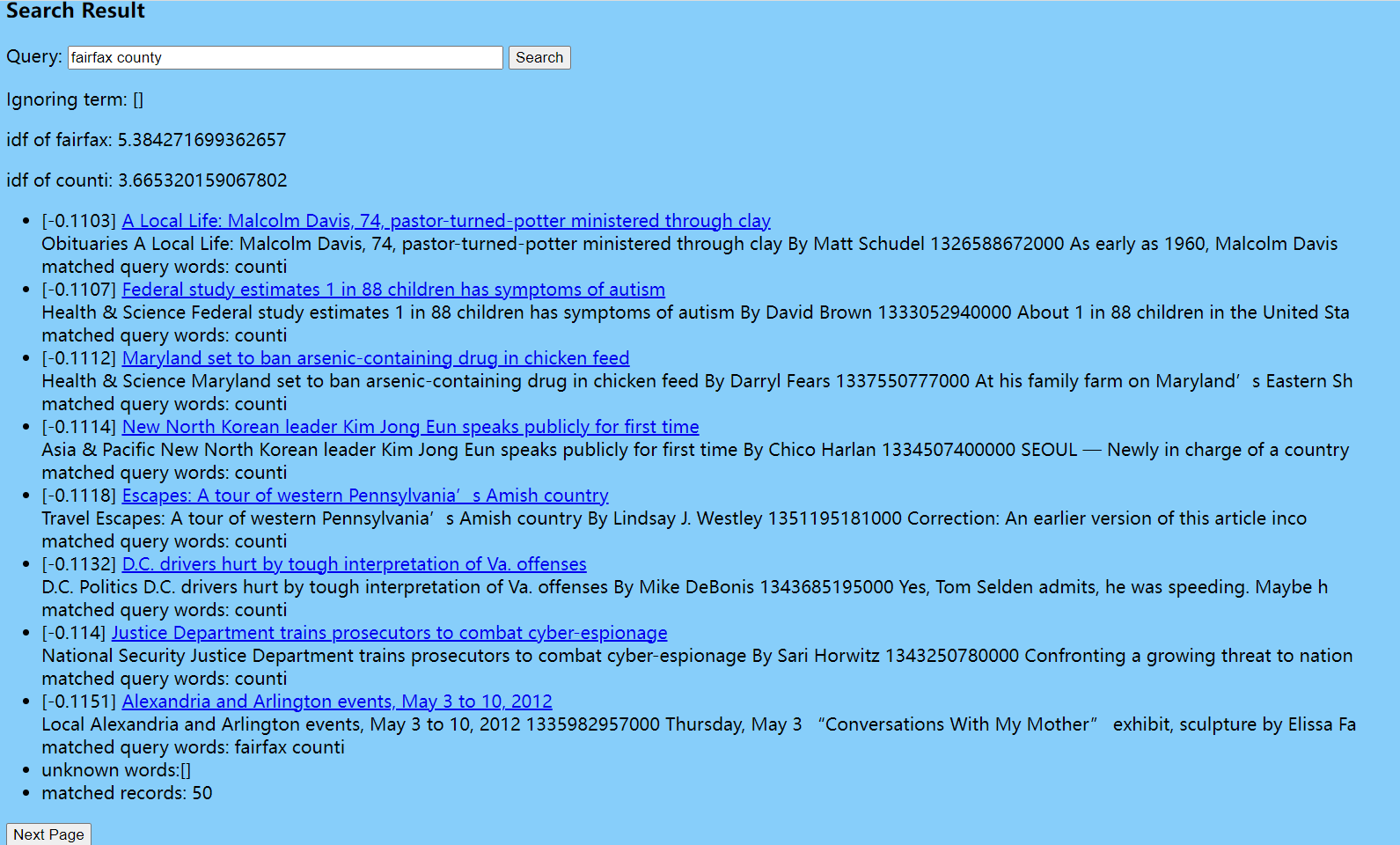
Monster is much more informational than common, so the idf is much higher.

### Query: big bad cow animal



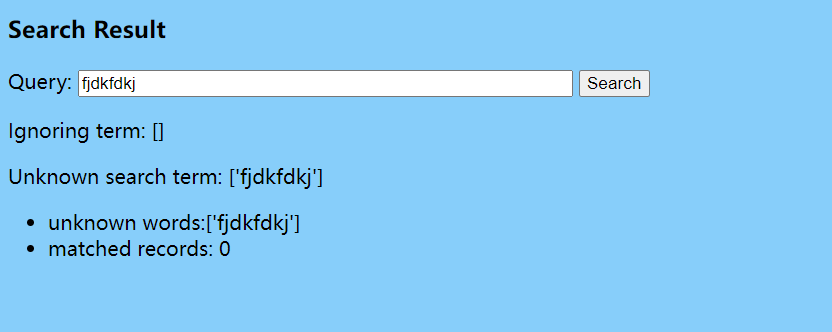
Working for 4 words queries, taking about 3-4 seconds to finish.

### Query: fairfax county



Notice that 2 matched words is working as expected as shown in the last record of the page.

### Query: fjdkfdkj



Shows unknown as expected