Zoekmachines Homework 4 Report

Mats Gonggrijp, Midas Amersfoort, Rutger Van Woerkom

October 2019

1 Introduction

During the building proces we had to make some considerations of how we ask for queries and present the data. In the following section we explain our considerations and why we made our choices. The main focus was simplicity for the user.

2 Considerations

The search engine has to be easy to use and helps users to make efficient queries. We have chosen for a simple prompt based system that first asks users for wether or not they want to use faceted search, their input will be 'y'=yes, or 'n'=no. If they choose 'n', then they are simply asked to input a query, that will be forwarded to the ElasticSearch API. If they choose 'y', they will be presented with another prompt and a set of possible features to choose from. They are asked for an input, which will be a string of numbers seperated by commas, where each number corresponds to a feature. This facet search will narrow down and make the process nmore efficient and accurate when the user already knows in what sort of category they are looking for.

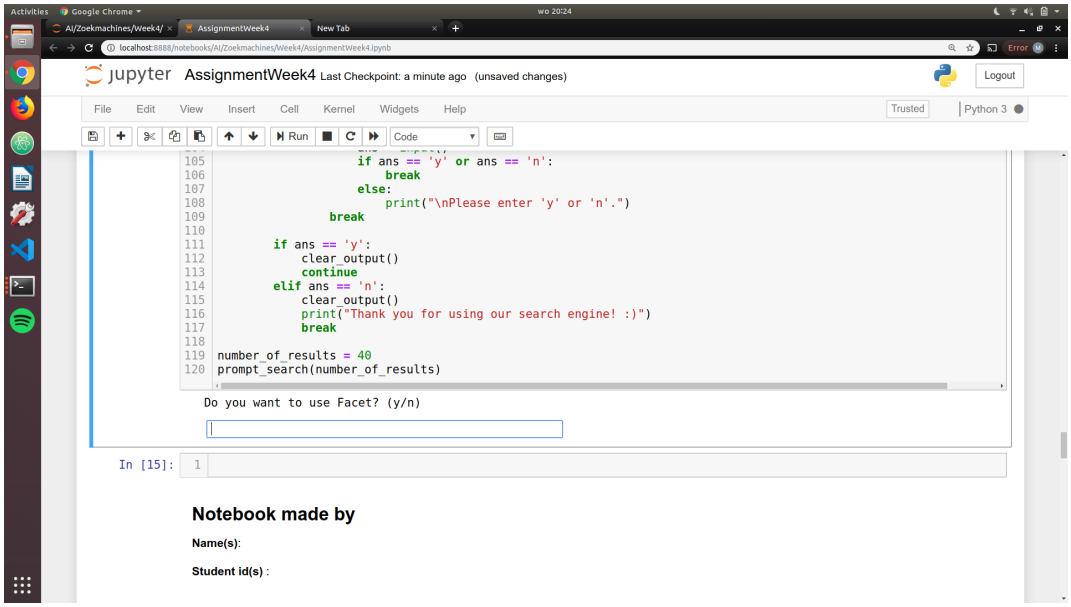
3 Output

The engine returns a ranked list of documents with some extra information to make it clear what is going on. For the text filled datasets that we use, we return the first 200 characters from the text content and the name of the author. This will help the user in deciding document relevancy. We return a list of the 10 most relevant documents.

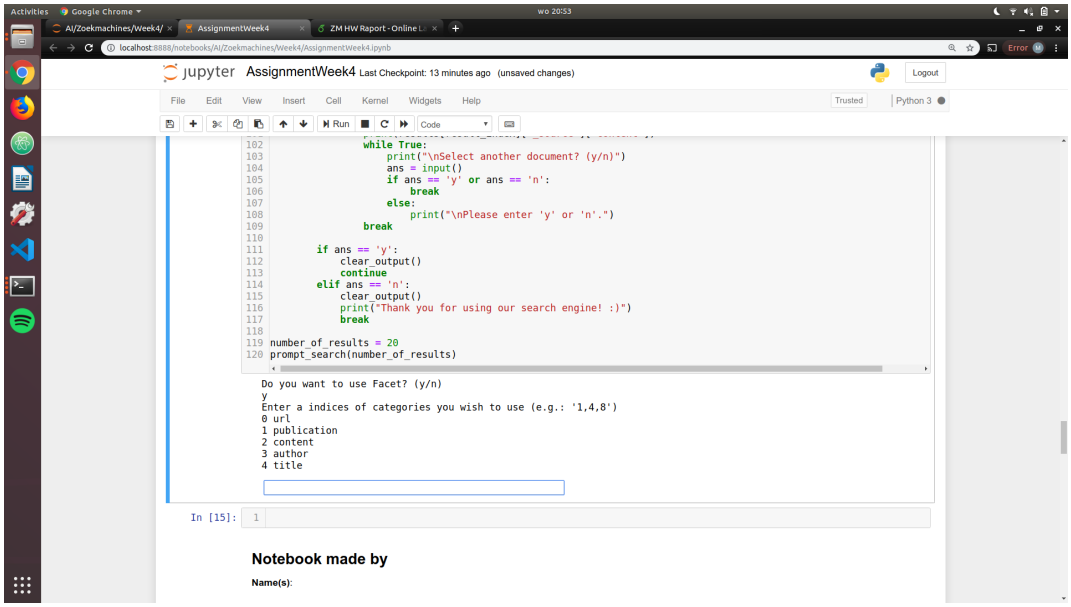
4 Demo

Below we show some screenshots of how the system works

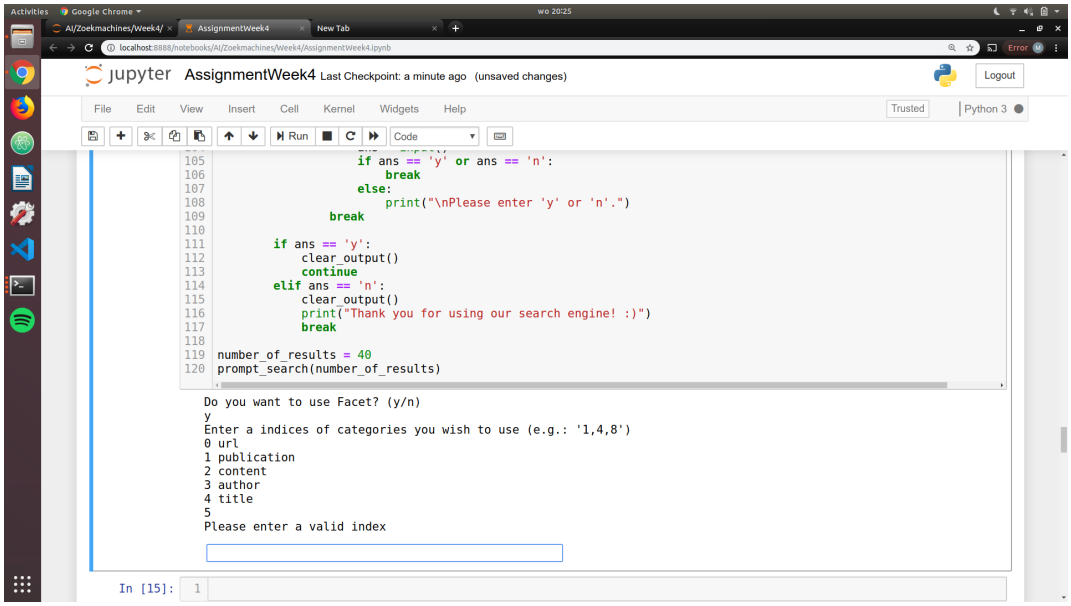
4.1 Ask user if they want to use facet search



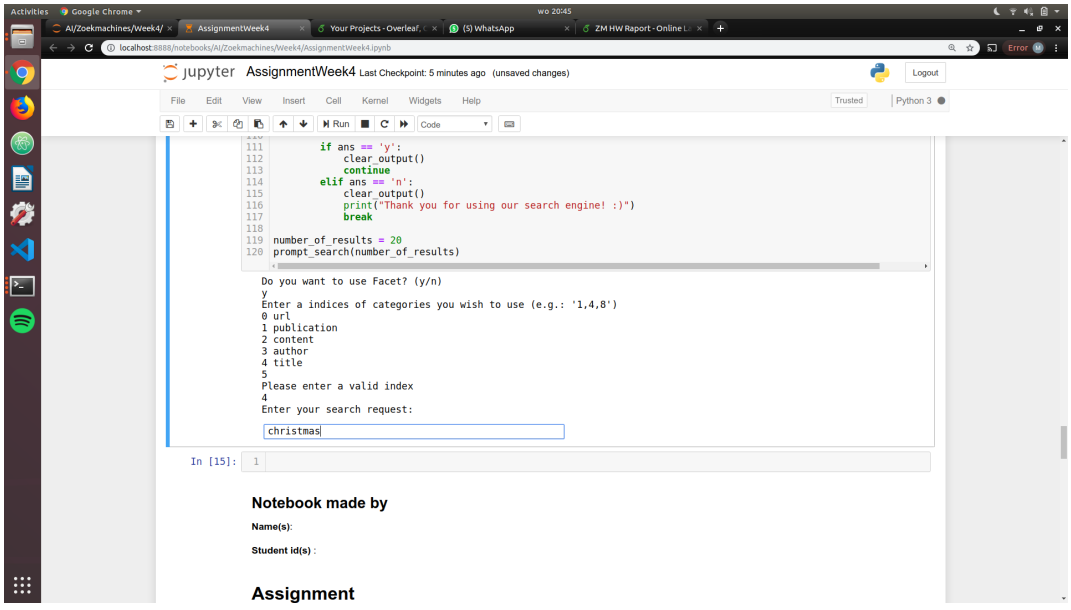
4.2 IF 'y', ask for facet set



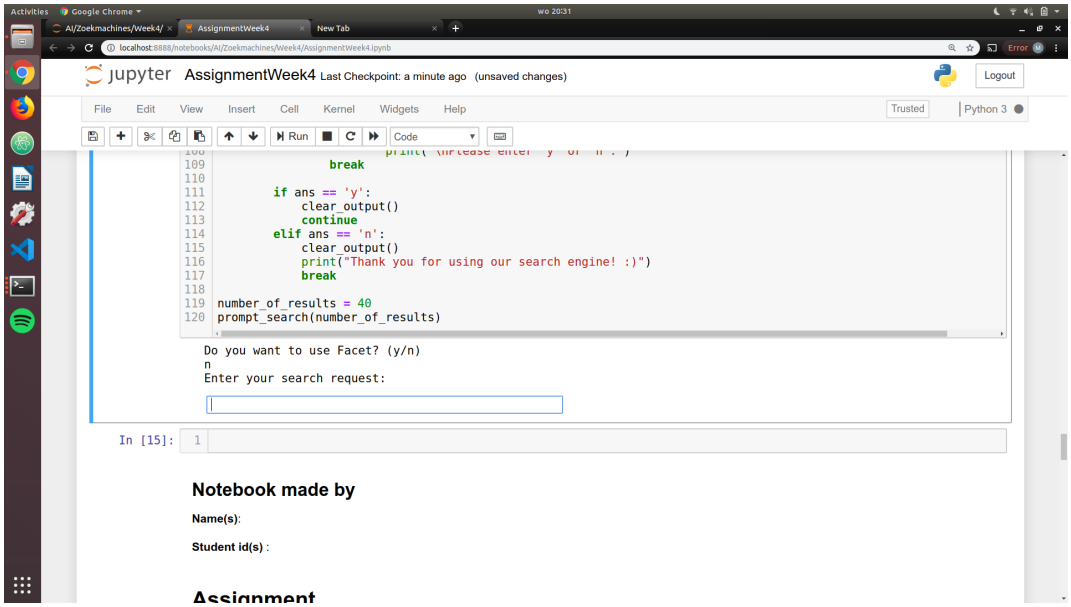
4.2.1 IF Facet set is incorrect



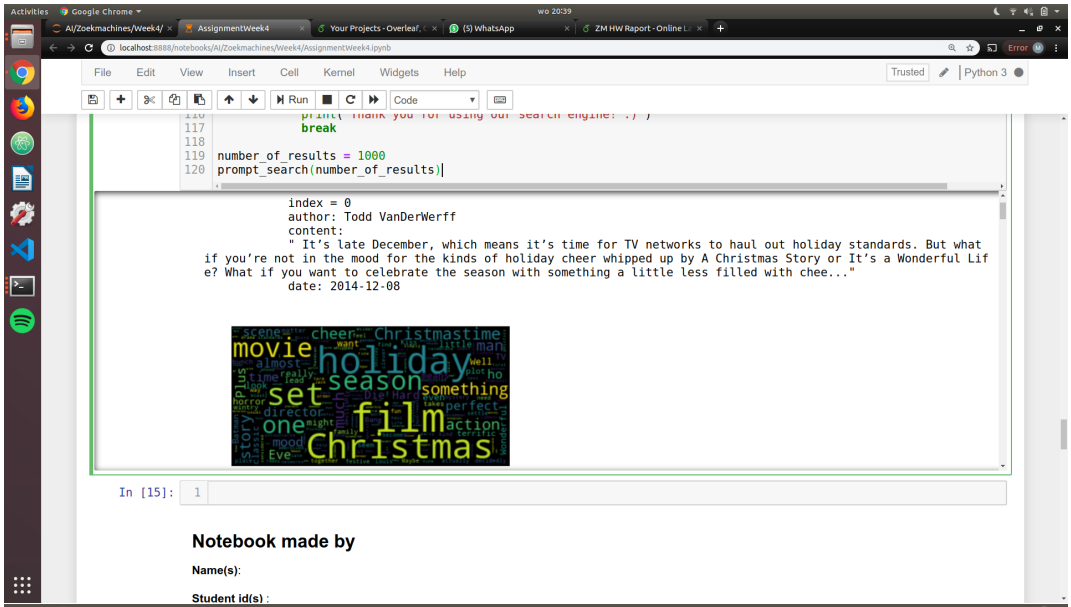
4.2.2 IF Facet set is correct

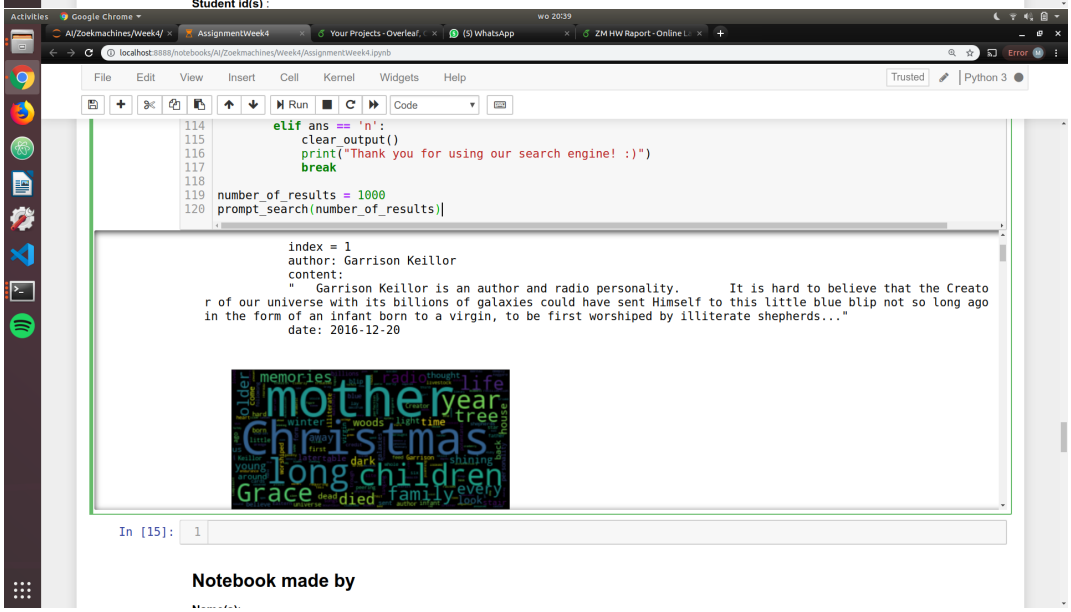


4.3 IF 'n', ask for simple query

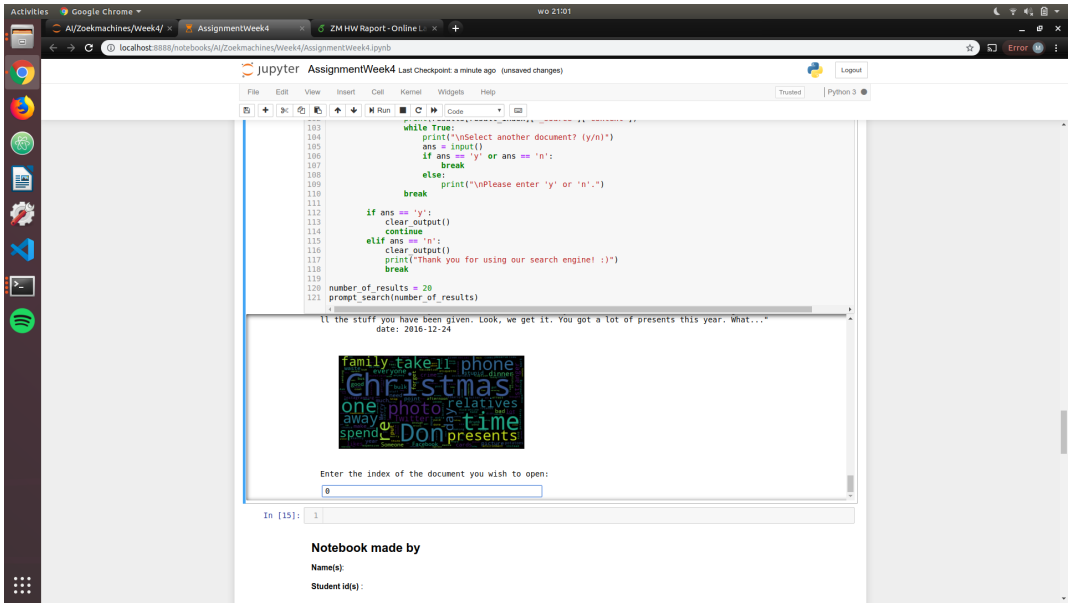


4.3.1 Return simple query results

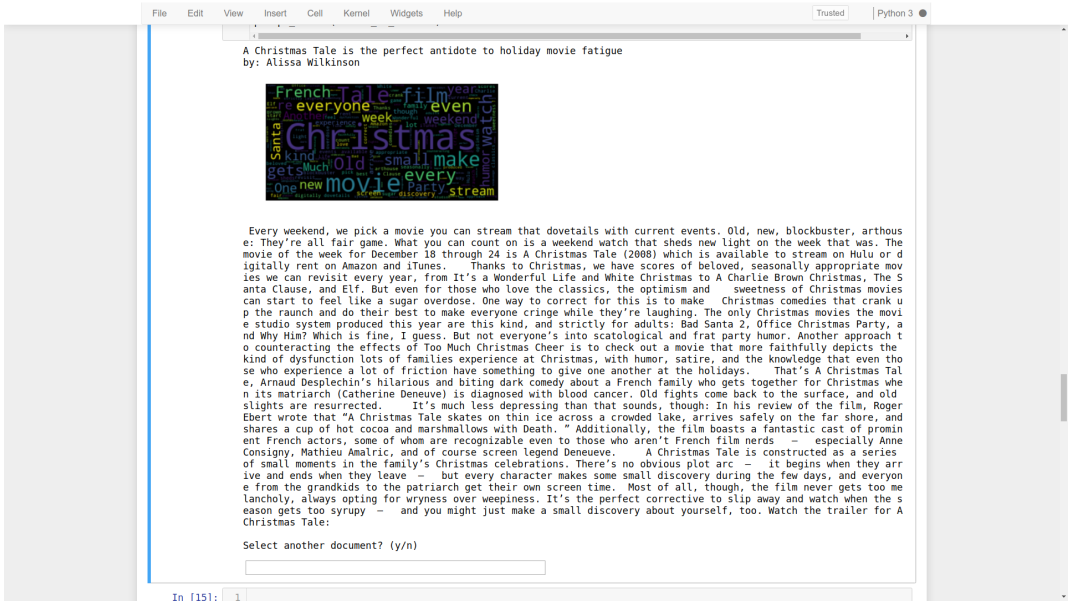




4.4 Return results, ask document selection



4.5 Return selected document content



5 Evaluation

The search engine works well for simple queries, but quickly becomes less accurate with more complicated constructions. The faceted search clearly helps with speed and accuracy, given that the user understands the categories.
Below are three examples of queries query: "brexit demonstrations 2017" Percentage of agreement: 80%

Cohen's kappa: 0.6000000000000001

Substantial agreement

doc_id	judge1	judge2
1	1	1
2	1	1
3	1	0
4	0	0
5	1	1
6	1	0
7	0	0
8	1	1
9	1	1
10	0	0

query: "wall in mexico demonstration" Percentage of agreement: 90%

Cohen's kappa: 0.8

Substantial agreement

doc_id	judge1	judge2
1	1	1
2	1	1
3	0	0
4	0	0
5	1	1
6	1	1
7	0	0
8	1	1
9	0	1
10	0	0

query: "christmas movies about cats and dogs" Percentage of agreement: 50%

Cohen's kappa: 0

No agreement

doc_id	judge1	judge2
1	0	1
2	1	1
3	1	0
4	1	0
5	0	1
6	1	1
7	0	0
8	0	0
9	0	1
10	0	0