Building an Information Retrieval System using University Data - M2

PRI 2023/24

GONÇALO ALMEIDA

ŽAN ŽLENDER

ILINA KIROVSKA

Milestone 1 recap

- ❑ Main idea is to build a search engine where students can get as much information about universities and various aspects related to them.

- ❑ Main data sources: QS WorldUniversity Rankings Kaggle dataset, Wikipedia page for each university and its city

- ❑ **The result of the Milestone 1 is a single JSON document containing a collection of 529 university documents**

Document schema

- 2024_rank
- 2023_rank
- institution_name
- country_code,
- country, size
- focus
- age
- status
- academic_reputation_score
- academic_reputation_rank
- employer_reputation_score
- employer_reputation_rank
- faculty_student_score
- faculty_student_rank
- age
- status
- academic_reputation_score
- academic_reputation_rank
- employer_reputation_score
- employer_reputation_rank
- faculty_student_score
- faculty_student_rank
- citations_per_faculty_score
- citations_per_faculty_rank
- international_students_score
- international_students_rank
- institution_name__wrong
- wikidata
- foundation_year
- overall_score
- international_research_network_score
- international_research_network_rank
- employment_outcomes_score
- employment_outcomes_rank
- wikipedia_text
- city_name
- city_wikipedia_text
- coordinates

Milestone 2

1. Create Solr schemas
2. Index documents
3. Query Information Needs
4. Retrieve the top 30 results
5. Evaluate results

Schema design

- Version 1: basic schema with no boosts and minimal use of filter
  - ASCII FoldingFilter
  - Lower Case Filter

Field	Indexed	Field type	
		Version 1	Version2
2024_rank	true	text	int
2023_rank	true	text	float
institution_name_-_wrong	false	text	text
institution_name	true	text	text
country_code	true	text	text
country	true	text	text
size	true	text	text
focus	true	text	text
age	true	text	text
status	true	text	text
academic_reputation_score	true	text	float
academic_reputation_rank	true	text	text
employer_reputation_score	true	text	float
employer_reputation_rank	true	text	text
faculty_student_score	true	text	float
faculty_student_rank	true	text	text

Field	Indexed	Field type	
		Version 1	Version2
citations_per_faculty_score	true	text	float
citations_per_faculty_rank	true	text	text
international_students_score	true	text	float
international_students_rank	true	text	text
international_research _network_score	true	text	float
international_research _network_rank	true	text	text
employment_outcomes_score	true	text	float
employment_outcomes_rank	true	text	text
overall_score	true	text	float
wikidata	false	text	text
wikipedia_text	true	text	wikipediaText
city_wikipedia_text	true	text	wikipediaText
foundation_date	true	text	date
coordinates	true	coordinates	coordinates

Schema design

- ❑ Version 2:
  - ASCII Folding Filter
  - Lower Case Filter
  - Classic Filter - strips periods from acronyms and "’s" from possessives
  - English Minimal Stem Filter - stems plural Englishwords to their singular form
  - Porter Stem Filter - applies the Porter Stemming Algorithm

Information need 1

- **Description:** Looking for universities that are top-ranked in computer science and are located in cities that have rich cultural heritage
- **Query 1:** wikipedia_text:"computer science" city_wikipedia_text:heritage 2024_rank:[1 TO 100]
- **Query 2:**wikipedia_text:"computer science"^3 AND city_wikipedia_text:heritage^2 AND 2024_rank:[* TO 200]^4
- **User priorities:** high ranking universities in the field of computer science where the city the university is located in also has a rich heritage

Query	P@10	AvP	R@10	F@10
Query 1	0.4	0.42	0.29	0.42
Query 2	1.0	1.0	0.34	0.67

Information need 2

- **Description:** Looking for universities located in the United Kingdom that have courses in biology and are ranked in the top 150
- **Query 1:** country: “United Kingdom“ country_code: UK wikipedia_text: biology 2024_rank:[1 TO 150]
- **Query 2:**country: "United Kingdom"^2 country_code: UK wikipedia_text:biology^2 2024_rank:[1 TO 150]^3
- **User priorities:** universities that have a rank of 150 or higher, giving those in the UK with courses in biology a lower priority

Query	P@10	AvP	R@10	F@10
Query 1	0.4	0.65	0.57	0.57
Query 2	0.9	0.9	0.64	0.78

Information need 3

- **Description:** Looking for universities in Germany that have a dental medicine faculty/dentistry and a large number of students
- **Query 1:** country: Germany country_code: DE size: large wikipedia_text: "dental medicine“
- **Query 2:** country: Germany^2 country_code: DE size: large wikipedia_text: dent*^2
- **User priorities:** universities located in Germany offering courses in dental medicine, giving those having a large number of students a lower priority

Query	P@10	AvP	R@10	F@10
Query 1	0.7	0.8	0.58	0.64
Query 2	1.0	1.0	0.56	0.71

Information need 4

- ❑ **Description:** Looking for universities that have a faculty of engineering or a faculty of science and are located in a city with a Mediterranean climate
- ❑ **Query 1:** wikipedia_text: “faculty of science” wikipedia_text: “faculty of engineering” city_wikipedia_text: “Mediterranean climate”
- ❑ **Query 2:** wikipedia_text: “faculty of science”^2 wikipedia_text: “faculty of engineering”^2 city_wikipedia_text: “Mediterranean climate”
- ❑ **User priorities:** universities offering courses in science and engineering, giving those in a city with a Mediterranean climate a lower priority than those with the wanted courses.

Query	P@10	AvP	R@10	F@10
Query 1	0.6	0.79	0.27	0.38
Query 2	1.0	0.89	0.43	0.61

Information need 5

- **Description:** Looking for top-ranked universities in the north of Europe with a focus on the Computer Science field
- **Query 1:** wikipedia_test: “Computer Science” city_wikipedia_text: “north Europe” 2024_rank:[1 TO 150]
- **Query 2:** wikipedia_test: “Comput* Science”^4 2024_rank:[1 TO 150] fq: !geofilt sfield=coordinates pt=61.069625,4.867638 d=1430868.94
- **User priorities:** top universities located in the north of Europe that have interest on the Computer Science field

Query	P@10	AvP	R@10	F@10
Query 1	0.3	0.37	0.5	0.38
Query 2	1.0	1.0	0.33	0.5

Information need 5

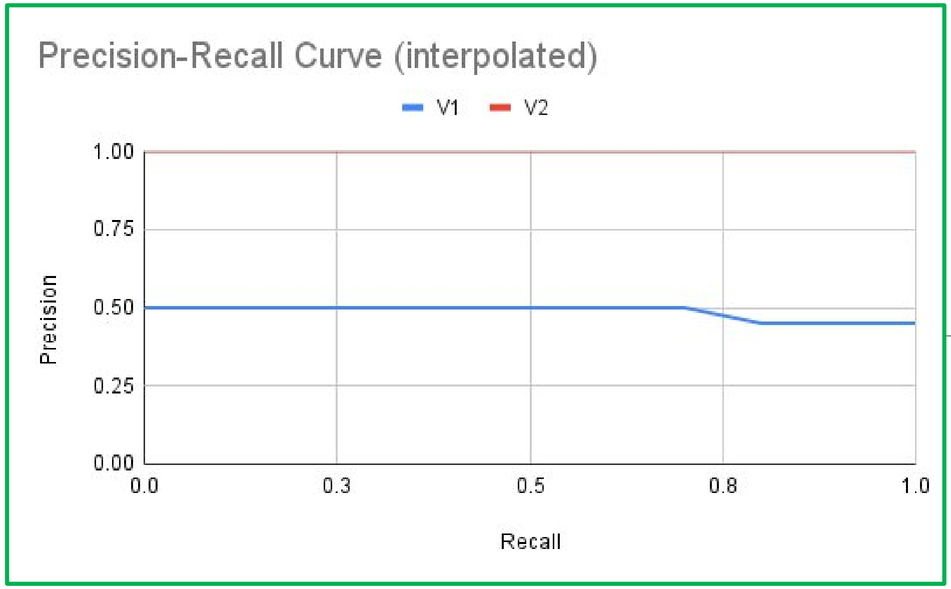
- **Description:** Looking for top-ranked universities in the north of Europe with a focus on the Computer Science field
- **Query 1:** wikipedia_test: “Computer Science” city_wikipedia_text: “north Europe” 2024_rank:[1 TO 150]
- **Query 2:** wikipedia_test: “Comput* Science”^4 2024_rank:[1 TO 150]
fq: !geofilt sfield=coordinates pt=61.069625,4.867638 d=1430868.94
- **User priorities:** top universities located in the north of Europe that have interest on the Computer Science field

Query	P@10	AvP	R@10	F@10
Query 1	0.3	0.37	0.5	0.38
Query 2	1.0	1.0	0.33	0.5

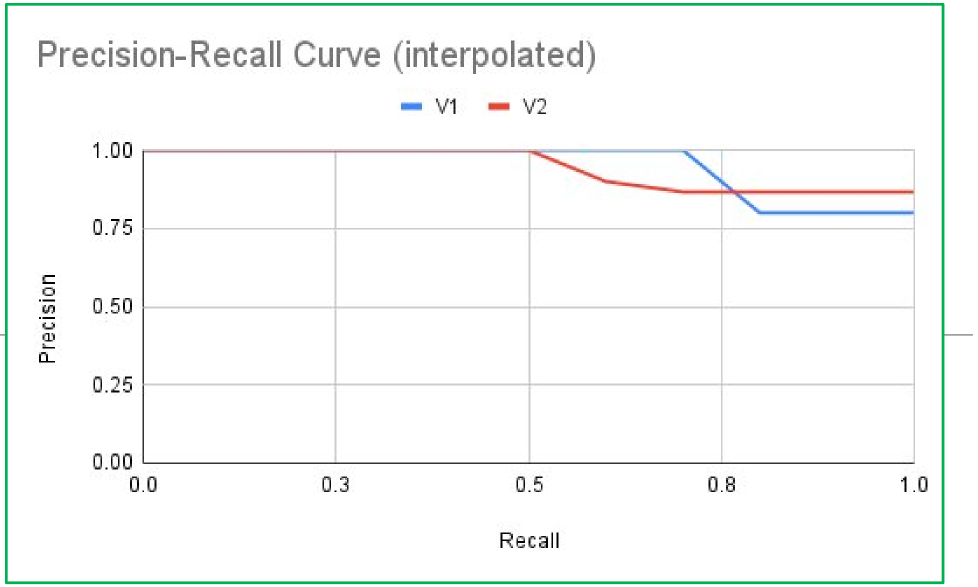
Evaluation discussion

- ❑ For any information retrieval system the relevance of the top results is of primary importance.
- ❑ This can be checked by calculating P@10.
- ❑ The previous results where P@10 = 1.0 for almost every boosted query lead to the conclusion that we have built a quite successful search engine.
- ❑ Additionally, the difference between Mean Average Precision for Version 1 and Version 2 of the search engine is another indicator that by using Solr filters and boosters the performance of the search engine has improved.

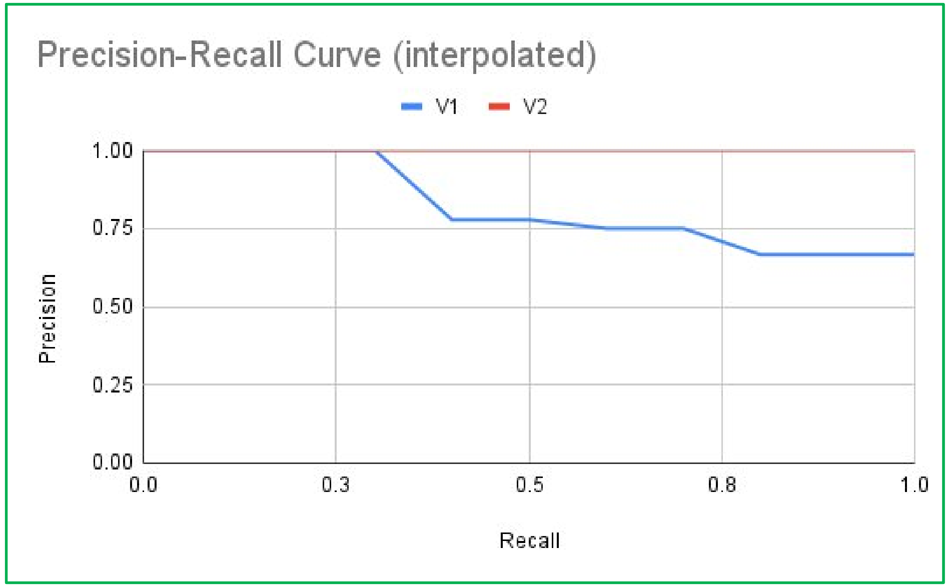
Metric	Version 1	Version 2
MAP	0.62	0.96



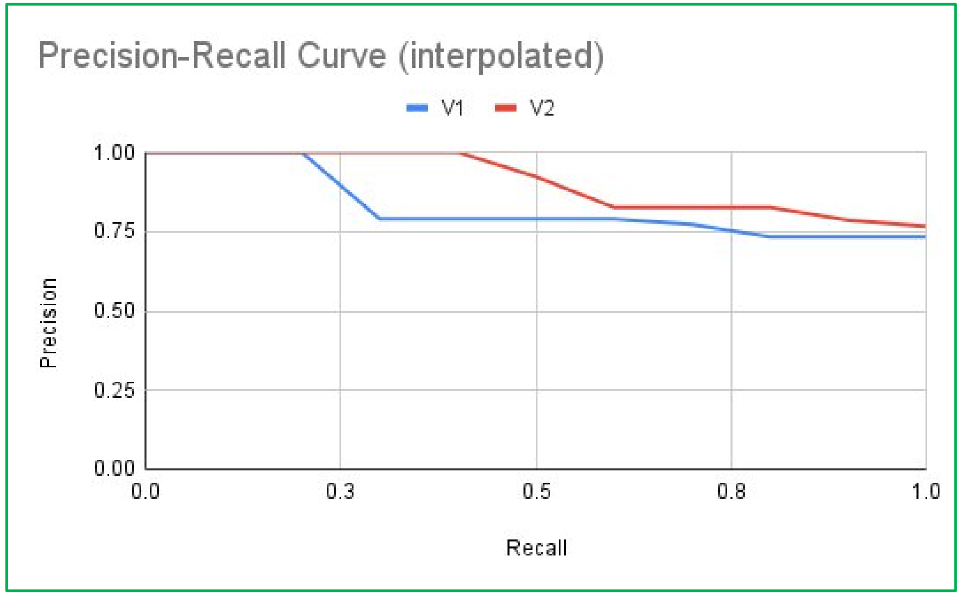
Information need 1



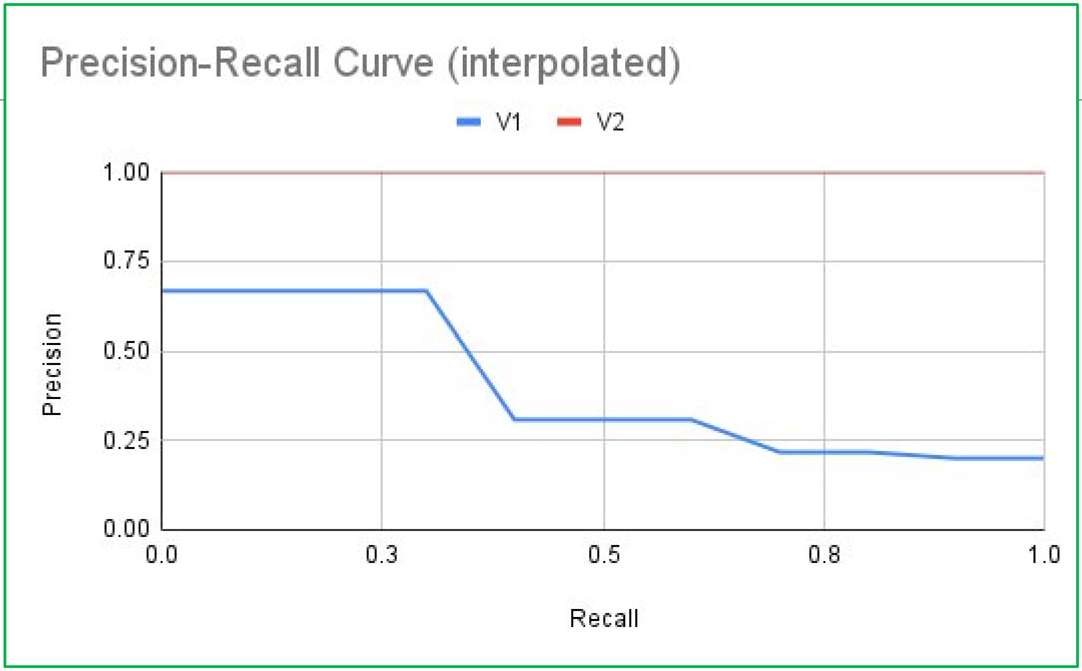
Information need 2



Information need 3



Information need 4



Information need 5

Conclusion

- Good understanding of Solr and it's indexing and querying capabilities

- Boosting queries gives much better results

- Proven that the current iimplementation of the university search engine works

Future work

In the future we will try to further improve the performance of the search engine using different methods, one of them being Natural Language Processing. Additionally we will focus on creating a user interface for the project.