Search System
Formula 1

Diogo Nunes | up201808546 Jéssica Nascimento | up201806723 João Vítor Fernandes | up201806724
Faculdade de Engenharia da Universidade do Porto | Processamento e Recuperação de Informação | Group 23

Milestone 2 Schema

File	Number of Documents
Races	1058
Drivers	853
Constructors	209
Circuits	77
Seasons	72
Pages	19

Table 1: Number of result documents in each file

- Result Documents about races
- Search fields: name, race_text and qualifying_text
- Schema Filters:
  - ASCIIFoldingFilterFactory
  - LowerCaseFilterFactory
  - EnglishMinimalStemFilterFactory
  - ManagedSynonymGraphFilterFactory
  - FlattenGraphFilterFactory
- Schema Tokenizers:
  - StandardTokenizerFactory

Milestone 3 Schema

- Result Documents about races, **drivers**, **circuits**, **constructors**, **season** and **pages** about the F1 cars, regulations, etc.

- Search fields: name, **race_text**, **qualifying_text**, **driver_text**,**circuit_text**, **constructor_text**, **season_text**, **page_text**, **firstName**, **lastName**, **nationality**, **year** and **location**

- Schema Filters:
  - ASCIIFoldingFilterFactory
  - LowerCaseFilterFactory
  - EnglishMinimalStemFilterFactory
  - ManagedSynonymGraphFilterFactory
  - FlattenGraphFilterFactory
  - **StopFilterFactory**

- Schema Tokenizers:
  - StandardTokenizerFactory

Configuration of cores

Schemaless (Solr's default schema) and **Schema** (presented in the previous slide):

- fl: *, score
- defType: edismax
- qf: race_text qualifying_text driver_text constructor_text circuit_text season_text page_text name nationality firstName lastName year location
- pf: race_text qualifying_text driver_text constructor_text circuit_text season_text page_text name nationality firstName lastName year location
- ps: 10

Note: **q**: field to input the query of the user; **q.op**: AND or OR; **fq**: filter query; **fl**: field list; **qf**: query fields with optional boosts; **pf**: phrase boosted fields (gives boost based on proximity of searched words); **ps**: phrase slop (maximum amount of tokens that a search result might have between searched words).

Configuration of cores

Schema (presented in the previous slide) with boost:

- fl: *, score
- defType: edismax
- qf: race_text^10 qualifying_text driver_text constructor_text circuit_text season_text page_text name^100 nationality^50 firstName^25 lastName^50 year^25 location^25
- pf: race_text^10 qualifying_text driver_text constructor_text circuit_text season_text page_text name^100 nationality^50 firstName^25 lastName^50 year^25 location^25
- ps: 10

Note: **q**: field to input the query of the user; **q.op**: AND or OR; **fq**: filter query; **fl**: field list; **qf**: query fields with optional boosts; **pf**: phrase boosted fields (gives boost based on proximity of searched words); **ps**: phrase slop (maximum amount of tokens that a search result might have between searched words).

Search results: Query 1

Youngest drivers to ever win a race or a championship

Schemaless	1	R	R	R	R	N	N	R	N	R	N
	2	R	R	R	R	R	R	R	N	R	N
Schema	1	R	R	R	R	N	R	N	N	R	N
	2	R	R	R	R	R	N	R	R	R	N
Schema + Boost	1	R	R	R	R	N	R	N	N	R	N
	2	N	R	R	R	R	N	R	N	R	R

Table 2: Result for query 1

1: Searching only on races
- q: youngest driver win
- q.op: AND
- fq: category:race

2: Searching on all files
- q: youngest driver win

1: Searching only on races
2: Searching on all files
R: Relevant
N: Not relevant

Evaluation of Query 1

Metric	Schemaless	Schema	Schema + Boost
Average Precision	0.90	0.92	0.92
Precision at 10	0.60	0.60	0.60
Recall at 10	1.0	1.0	1.0

Table 3: Evaluation when searching only on races for Q1 (1)

Metric	Schemaless	Schema	Schema + Boost
Average Precision	0.99	0.95	0.69
Precision at 10	0.80	0.80	0.70
Recall at 10	1.0	1.0	1.0

Table 4: Evaluation when searching on all files for Q1 (2)

Evaluation of Query 1

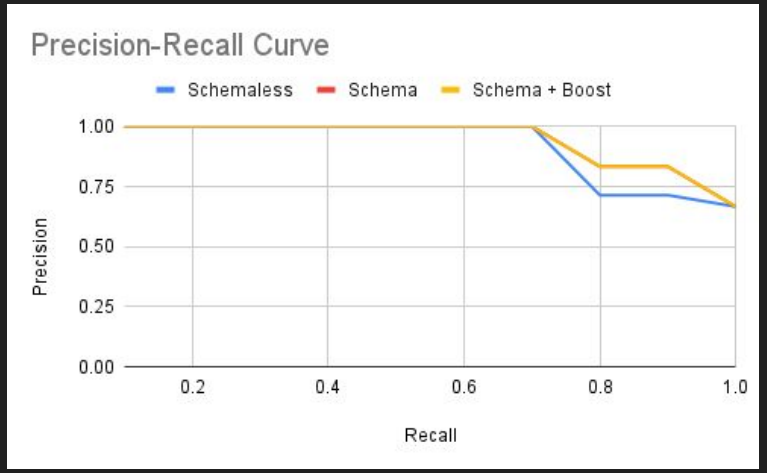

Figure 1: Evaluation when searching only on races for Q1 (1)

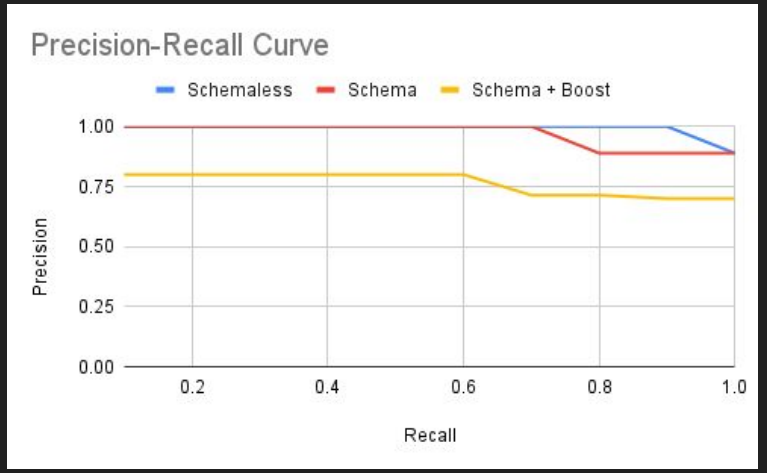

Figure 2: Evaluation when searching on all files for Q1 (2)

Search results: Query 2

Incidents involving a certain driver

Schemaless	1	R	R	R	R	N	R	R	R	N	N
	2	R	R	N	R	R	R	R	R	R	R
Schema	1	R	R	R	N	R	R	N	R	R	R
	2	R	R	R	R	R	N	R	R	R	R
Schema + Boost	1	R	R	R	R	R	R	N	R	R	N
	2	R	R	R	N	R	R	N	R	R	N

Table 5: Result for query 2

1: Searching only on races
- q: “incident Verstappen"~10 "crashVerstappen"~10 "accident Verstappen"~10 "collision Verstappen"~10 "contact withVerstappen"~10
- q.op: OR
- fq: category:race

2: Searching on all files
- q: “incident Verstappen"~10 "crashVerstappen"~10 "accident Verstappen"~10 "collision Verstappen"~10 "contact withVerstappen"~10

1: Searching only on races
2: Searching on all files

R: Relevant
N: Not relevant

Evaluation of Query 2

Metric	Schemaless	Schema	Schema + Boost
Average Precision	0.94	0.87	0.97
Precision at 10	0.70	0.80	0.80
Recall at 10	1.0	1.0	1.0

Table 6: Evaluation when searching only on races for Q2 (1)

Metric	Schemaless	Schema	Schema + Boost
Average Precision	0.88	0.95	0.88
Precision at 10	0.90	0.90	0.70
Recall at 10	1.0	1.0	1.0

Table 7: Evaluation when searching on all files for Q2 (2)

Evaluation of Query 2

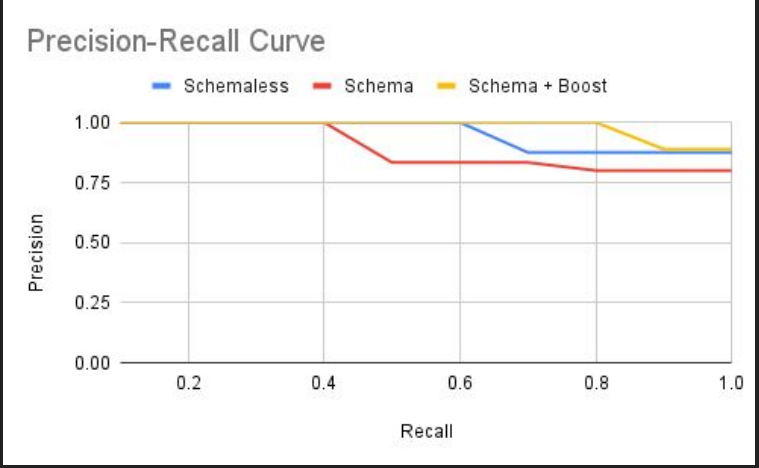

Figure 3: Evaluation when searching only on races for Q2 (1)

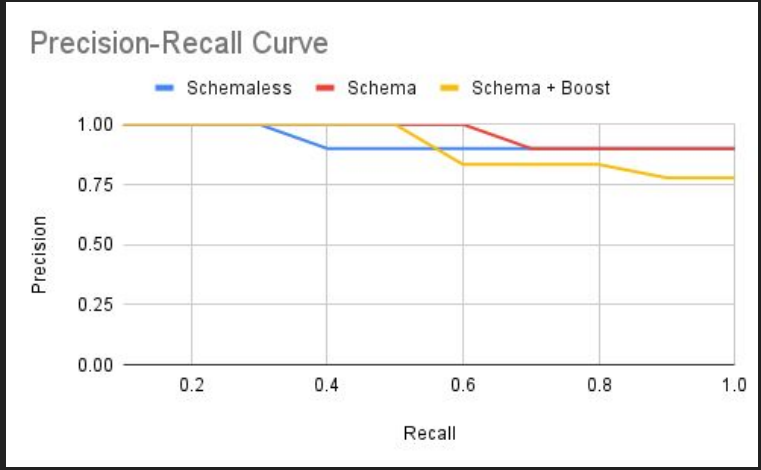

Figure 4: Evaluation when searching on all files for Q2 (2)

Search results: Query 3

Overtakes of a certain driver

Schemaless	1	R	N	N	N	R	N	N	R	N	N
	2	N	R	R	R	N	N	N	N	N	N
Schema	1	R	R	N	N	R	R	N	R	R	N
	2	R	R	R	N	R	N	R	N	N	R
Schema + Boost	1	R	R	N	R	R	R	R	R	R	N
	2	R	R	R	R	N	N	R	R	N	N

Table 8: Result for query 3

1: Searching only on races
- q: "Vettel overtake"~10 "Vettel pass"~10
- q.op: AND
- fq: category:race

2: Searching on all files
- q: "Vettel overtake"~10 "Vettel pass"~10

1: Searching only on races
2: Searching on all files

R: Relevant
N: Not relevant

Evaluation of Query 3

Metric	Schemaless	Schema	Schema + Boost
Average Precision	0.59	0.76	0.88
Precision at 10	0.30	0.60	0.80
Recall at 10	1.0	1.0	1.0

Table 9: Evaluation when searching only on race for Q3 (1)

Metric	Schemaless	Schema	Schema + Boost
Average Precision	0.64	0.85	0.91
Precision at 10	0.30	0.60	0.60
Recall at 10	1.0	1.0	1.0

Table 10: Evaluation when searching on all files for Q3 (2)

Evaluation of Query 3

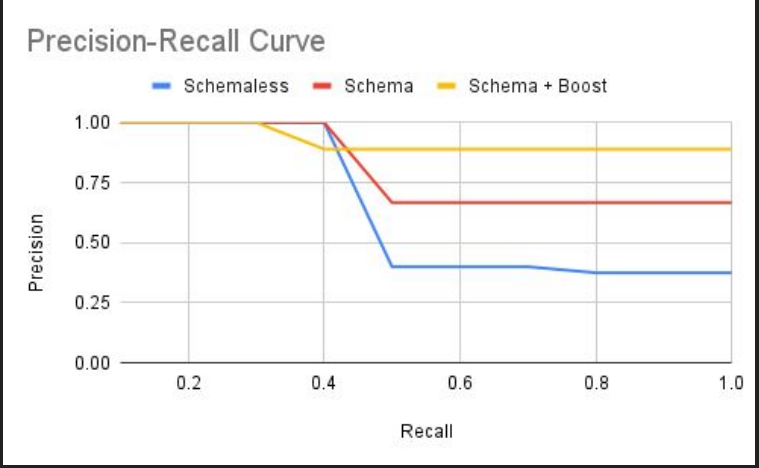

Figure 5: Evaluation when searching only on races for Q3 (1)

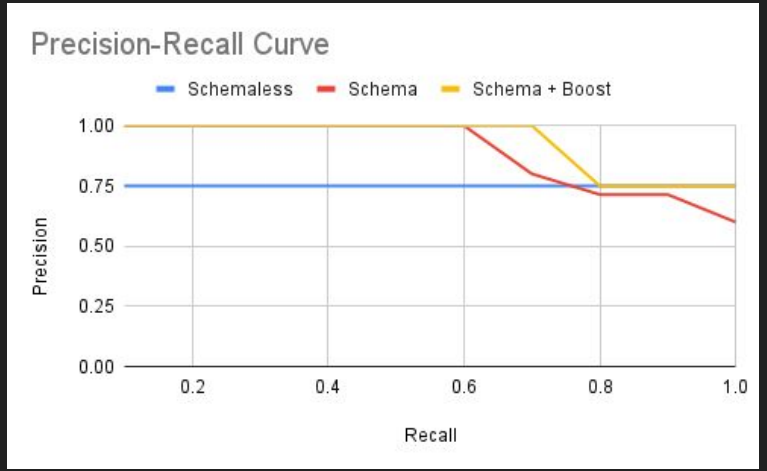

Figure 6: Evaluation when searching on all files for Q3 (2)

Search results: Query 4

Teammates of a certain driver

Schemaless	N	N	N	N	N	N	N	N	N	N
Schema	R	N	R	R	R	R	N	R	N	R
Schema + Boost	R	N	R	R	R	R	N	R	N	R

Table 11: Result for query 4

Searching on all files:

- q: "Raikkonen teammate"~10
- fq: category:driver

R: Relevant
N: Not relevant

Evaluation of Query 4

Metric	Schemaless	Schema	Schema + Boost
Average Precision	0	0.79	0.79
Precision at 10	0	0.7	0.7
Recall at 10	0	1.0	1.0

Table 12: Evaluation when searching only on race for Q3 (1)

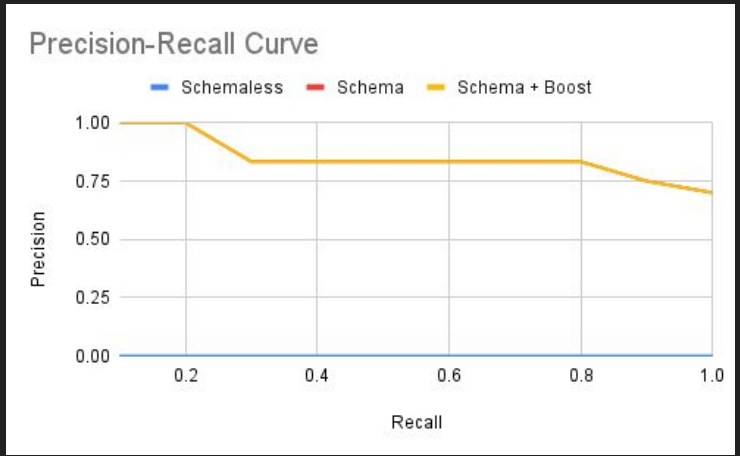

Figure 7: Evaluation when searching on all files for Q3 (2)

Conclusion and Future Work

We constructed a Search System capable of being asked/queried about almost every detail about Formula 1 with relative good precision.

Although our schema and boosts sometimes doesn't improve the search results, we believe it's a schema that, in general, improves our Search System performance.

Some features that our Search System may incorporate later:

- Web interface connected to Solr;
- Use strategies to find related documents inside other documents, so we can easily read about topics that might appear and the user want to clarify;
- Propose new ranking signals using the existing information (e.g. PageRank signal bases on citation data).

Thank you!

Group 23

Diogo Nunes | up201808546 Jéssica Nascimento | up201806723 João Vítor Fernandes | up201806724