U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

# Retrieving and Processing Information on Company Reviews

**Professor:**

Sara Fernandes

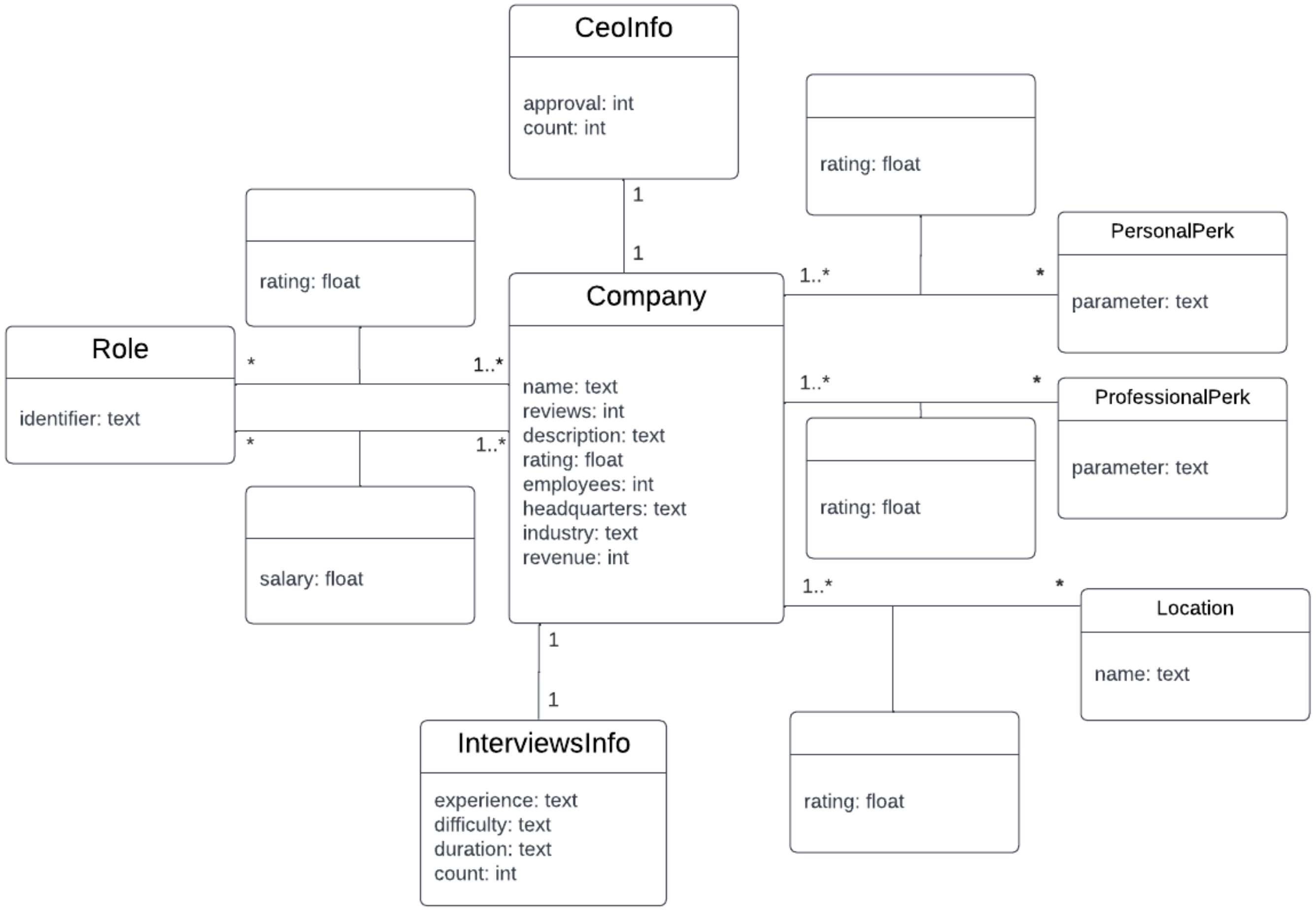
**PRI, group 69**

Bruno Rosendo, up201906334

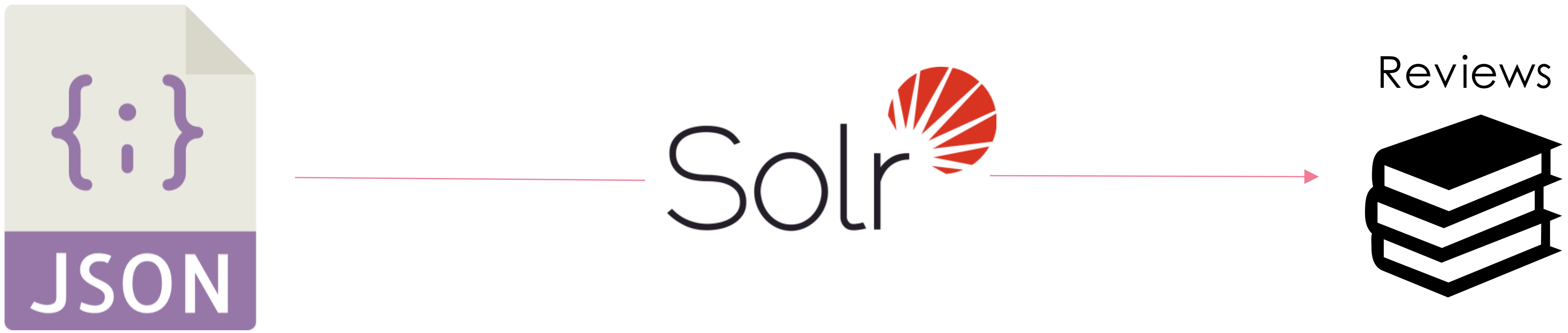
João Mesquita , up201906682

Rui Alves, up201905853

# Conceptual Model of the Dataset



# Document Based System



Company reviews dataset in JSON format

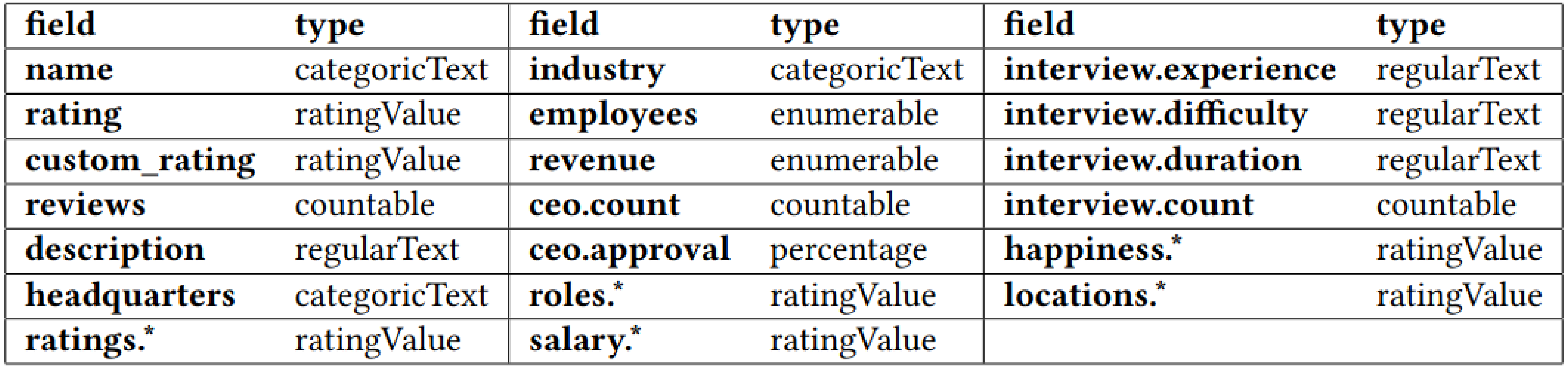
Tool selected for information retrieval

Single collection of company reviews (core)

# *Indexing: Defined Types*

- **regularText:** written text, usually carrying a considerable amount of words.
  - StandardTokenizerFactory
  - ASCIIFoldingFilterFactory, LowerCaseFilterFactory, PorterStemFilterFactory, StopFilterFactory, SynonymGraphFilterFactory, RemoveDuplicatesTokenFilterFactory
- **categoricText:** textual fields representing categories, such as names of entities or genres.
  - StandardTokenizerFactory
  - ASCIIFoldingFilterFactory, LowerCaseFilterFactory
- **ratingValue** and **percentage:** *double* values used for the different ratings and percentages the dataset.
- **countable** and **enumerable:** *integer* values used for whole numbers with real meaning (e.g. number of ratings) or fields holding categorical information in the form of numbers.

# *Indexing: Schema Fields*



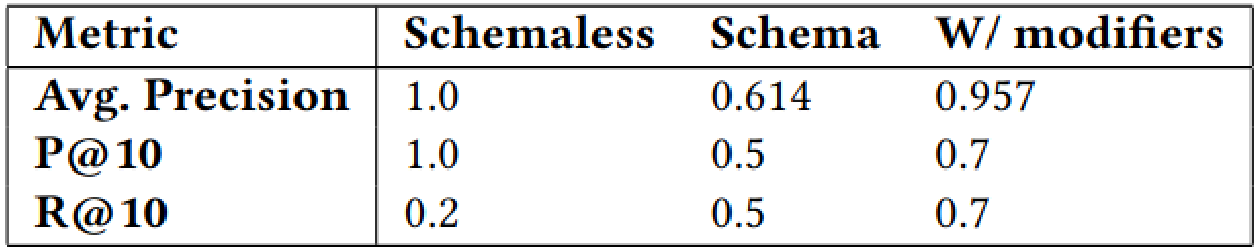
| field | type | field | type | field | type |
|---|---|---|---|---|---|
| **name** | categoricText | **industry** | categoricText | **interview.experience** | regularText |
| **rating** | ratingValue | **employees** | enumerable | **interview.difficulty** | regularText |
| **custom_rating** | ratingValue | **revenue** | enumerable | **interview.duration** | regularText |
| **reviews** | countable | **ceo.count** | countable | **interview.count** | countable |
| **description** | regularText | **ceo.approval** | percentage | **happiness.*** | ratingValue |
| **headquarters** | categoricText | **roles.*** | ratingValue | **locations.*** | ratingValue |
| **ratings.*** | ratingValue | **salary.*** | ratingValue | | |

- The *reviews.count*, *ceo.count*, and *interview.count* fields do not require **indexing**, since they do not represent information useful for the user to query by.
- Since every field from the dataset has relevant information for the user, the ***stored*** attribute is enabled in all of them.

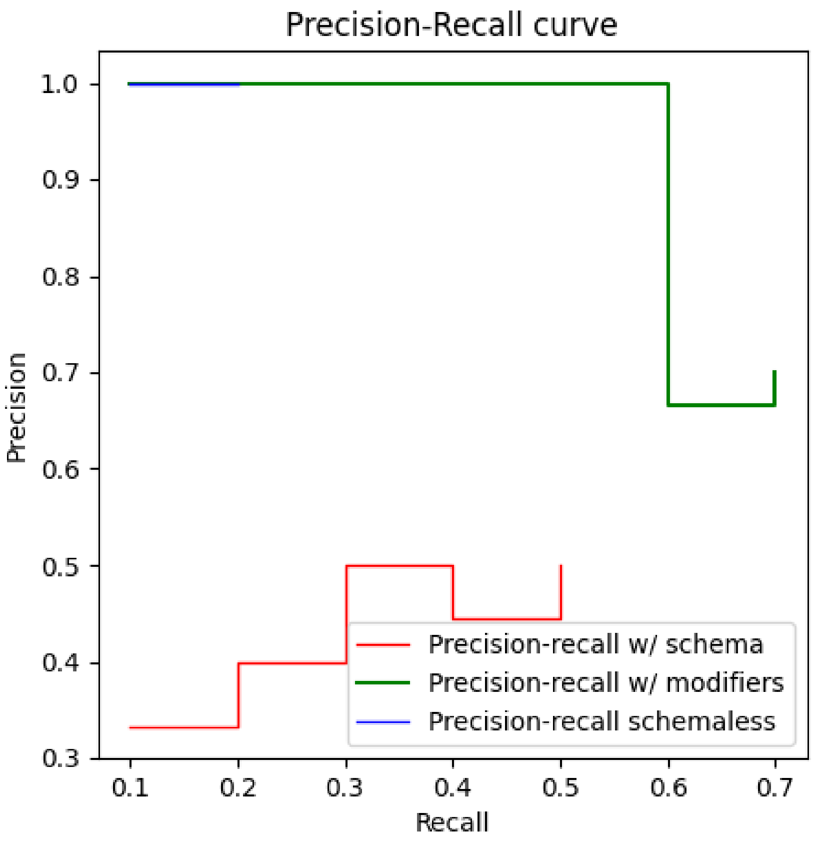
# Live Demo, let's jump into Solr!

# Evaluation of Search Results

- A user wants to find companies working in the industry of telecommunications.
- **Query:** Telecommunications

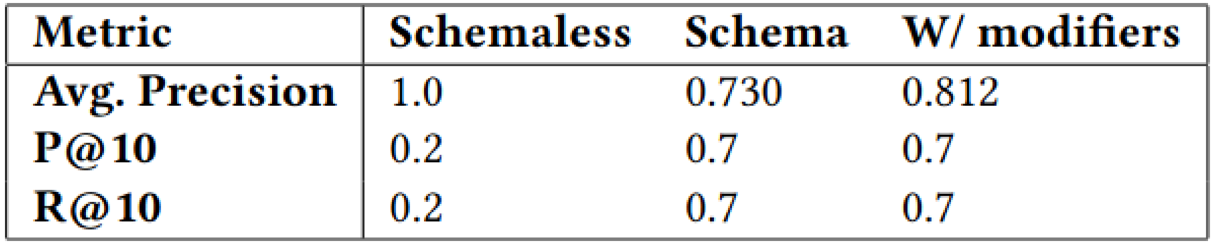


| Metric | Schemaless | Schema | W/ modifiers |
|---|---|---|---|
| Avg. Precision | 1.0 | 0.614 | 0.957 |
| P@10 | 1.0 | 0.5 | 0.7 |
| R@10 | 0.2 | 0.5 | 0.7 |

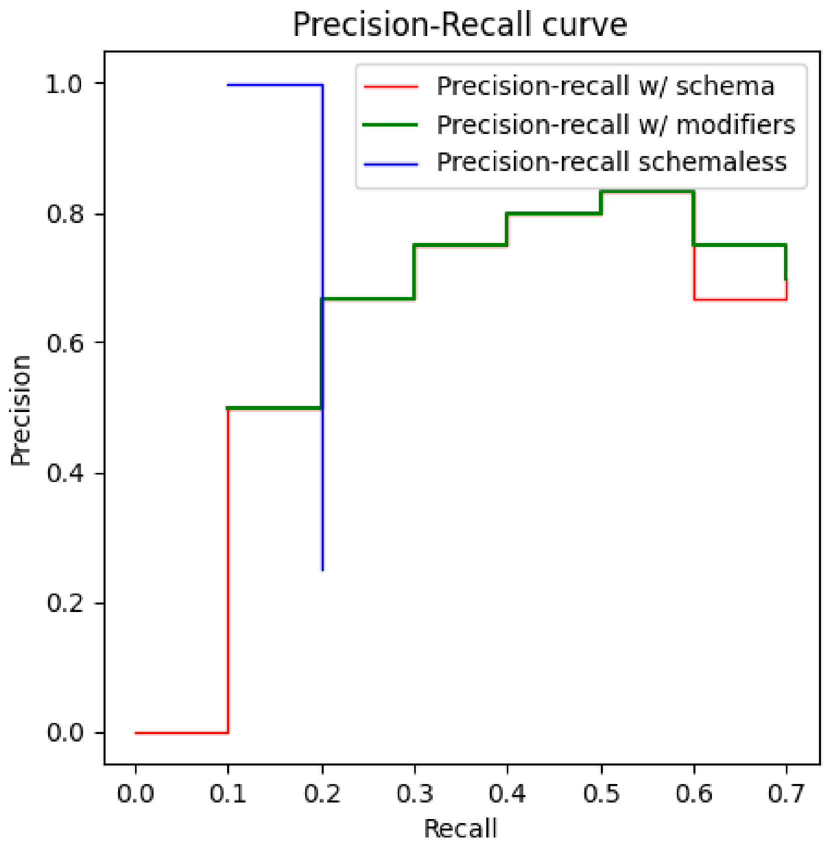


# Evaluation of Search Results

- A user wants to find companies working in the financial technology industry that have more than 5000 employees (categorical value of 8 or more).
- **Query:** Financial Technology (employees: [8 TO *])

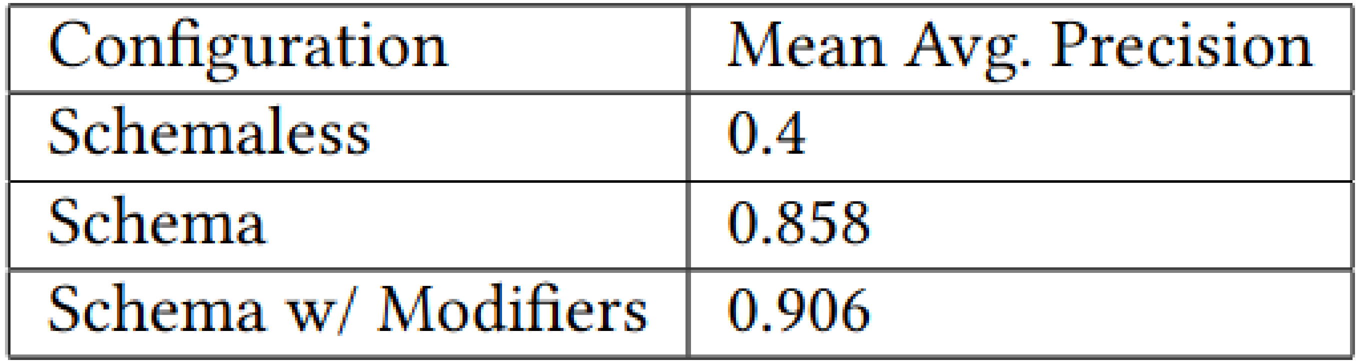


| Metric | Schemaless | Schema | W/ modifiers |
|---|---|---|---|
| Avg. Precision | 1.0 | 0.730 | 0.812 |
| P@10 | 0.2 | 0.7 | 0.7 |
| R@10 | 0.2 | 0.7 | 0.7 |



# *Evaluation of Search Results*

- Mean average precision for the three systems tested in the evaluation process:



| Configuration | Mean Avg. Precision |
|---|---|
| Schemaless | 0.4 |
| Schema | 0.858 |
| Schema w/ Modifiers | 0.906 |

# Future Work

## Review Iteration

Review if the work done on Iteration 2 is ready for the following stages

## Improve Search System

Experiment with more boosts and modifiers. Create additional information needs.

## Product enhancement for the end user

Develop a friendly user interface to interact with the system