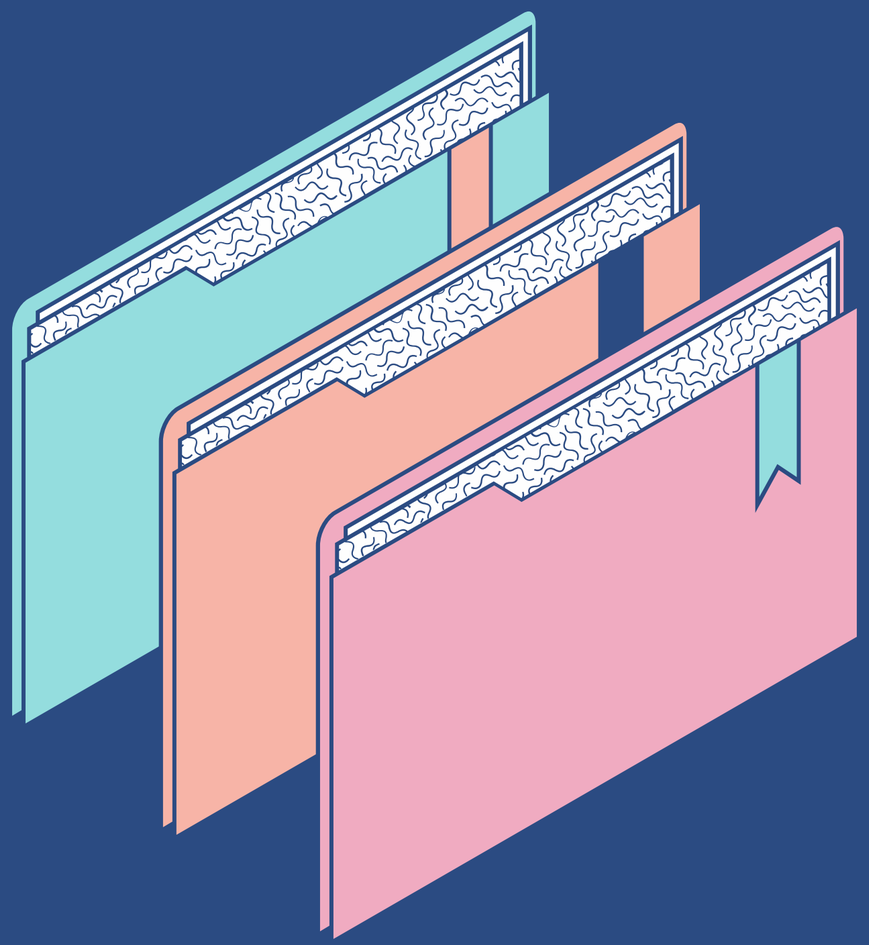
M.EIC PRI 2022/2023 G55

arXiv - Scientific Papers

Scientific Articles Search System

Beatriz Santos, up201905680
Sérgio Estêvão, up201905680
Sérgio da Gama, up201905680



Dataset

kaggle

Solr

JSON

Collection of scientific papers and their corresponding information from the website ARXIV

Collection

A paper defines a document which has the paper's information.

Most relevant fields:

- Title
- Summary
- Authors
- Areas
- Fields
- Subjects

Collection


```
{
  "link":["http://arxiv.org/abs/1606.02518v3"],
  "summary":"The multivariate normal density is a monotonic function of the distance to\nthe mean, and its ellipsoidal shape is due to the underlying Euclidean metric.\nWe suggest to repla
  "title":"A Locally Adaptive Normal Distribution",
  "authors":["Georgios Arvanitidis",
    "Lars Kai Hansen",
    "Søren Hauberg"],
  "date":"2016-06-08T00:00:00Z",
  "areas":["Statistics"],
  "fields":["Statistics"],
  "subjects":["Machine Learning"],
  "id":"22413",
  "_version_":1749478622659346433},
{
```

Indexing

Tokenizer:

ClassicTokenizerFactory

Filters:

ClassicFilterFactory
LowerCaseFilterFactory
ASCIIFoldingFilterFactory
PorterStemFilterFactory
StopFilterFactory
PhoneticFilterFactory
RemoveDuplicatesTokenFilterFactory
SynonymGraphFilterFactory
CommonGramsFilterFactory
BeiderMorseFilterFactory

Type	Fields	Indexed
titleType	Title	Yes
summary (personalized)	Sumamry	
names (personalized)	Authors, Areas, Fields, Subjects, Date	
date	Date	

Query 1

I am a developer creating a ML model to predict cars longevity based on their velocity.

Query 1

Query:

q: velocity
qf: link summary title authors date areas fields subjects
defType: edismax

Boosted:

q: velocity
qf: link summary^10 title^2 authors date areas fields subjects
defType: edismax

Query 1 - results boosted


```
{
  "link":["http://arxiv.org/abs/1802.07094v1"],
  "summary":"This paper documents the winning entry at the CVPR2017 vehicle velocity\nestimation challenge. Velocity estimation is an emerging
  "title":"Camera-based vehicle velocity estimation from monocular video",
  "authors":["Moritz Kampelmühler",
    "Michael G. Müller",
    "Christoph Feichtenhofer"],
  "date":"2018-02-20T00:00:00Z",
  "areas":["Computer Science"],
  "fields":["Computer Science"],
  "subjects":["Computer Vision and Pattern Recognition"],
  "id":"31202",
  "_version_":1749494417606049792},
{
  "link":["http://arxiv.org/abs/1705.09805v3"],
  "summary":"We propose position-velocity encoders (PVEs) which learn---without\nsupervision---to encode images to positions and velocities of
  "title":"PVEs: Position-Velocity Encoders for Unsupervised Learning of Structured\n  State Representations",
  "authors":["Rico Jonschkowski",
    "Roland Hafner",
    "Jonathan Scholz",
    "Martin Riedmiller"],
  "date":"2017-05-27T00:00:00Z",
  "areas":["Computer Science"],
  "fields":["Computer Science"],
  "subjects":["Robotics",
    "Computer Vision and Pattern Recognition",
    "Machine Learning"],
  "id":"12804",
  "_version_":1749493970372657152},
```

Query 1 - analysis

	Non-boosted	Boosted
Precision	1.0	1.0
Recalls	1.0	1.0

Query 2

I am a data analyst that is processing a dataset and I need to understand normal distributions.

Query 2

Query:

q: normal distribution
qf: link summary title authors date areas fields subjects
defType: edismax

Boosted:

q: normal distribution
qf: link summary^2 title^10 authors date areas fields subjects
bq: areas:Statistics^10
defType: edismax

Query 2 - results boosted


```
{
  "link":["http://arxiv.org/abs/1606.02518v3"],
  "summary":"The multivariate normal density is a monotonic function of the distance to\nthe mean, and its ellipsoidal shape is due
  "title":"A Locally Adaptive Normal Distribution",
  "authors":["Georgios Arvanitidis",
    "Lars Kai Hansen",
    "Søren Hauberg"],
  "date":"2016-06-08T00:00:00Z",
  "areas":["Statistics"],
  "fields":["Statistics"],
  "subjects":["Machine Learning"],
  "id":"22413",
  "_version_":1749494197137702912},
{
  "link":["http://arxiv.org/abs/1103.4789v3"],
  "summary":"We present the discrete infinite logistic normal distribution (DILN), a\nBayesian nonparametric prior for mixed members
  "title":"The Discrete Infinite Logistic Normal Distribution",
  "authors":["John Paisley",
    "Chong Wang",
    "David Blei"],
  "date":"2011-03-24T00:00:00Z",
  "areas":["Statistics"],
  "fields":["Statistics"],
  "subjects":["Machine Learning"],
  "id":"21576",
  "_version_":1749494175382896640},
{
```

Query 2 - analysis

	Non-boosted	Boosted
Precision	0.5	1.0
Recalls	0.625	0.9375

Query 3

I am a writer that wants to write a biography about Francis Bach, but I want to focus on his algorithmic work from 2008 until 2018 mostly, preferably from the year of 2015.

Query 3

Query:

q: Francis Bach algorithm
fq: date:[2014-01-01T00:00:00Z TO 2018-01-01T00:00:00Z}
qf: link summary title authors date areas fields subjects
defType: edismax

Boosted:

q: Francis Bach algorithm
qf: link summary^5 title authors^5 date areas fields subjects
fq: date:[2014-01-01T00:00:00Z TO 2018-01-01T00:00:00Z}
bf: if(and(gte(ms(date),ms(2015-01-01T00:00:00Z)),lt(ms(date),ms(2016-01-01T00:00:00Z))),10,0.1)
defType: edismax

Query 3 - results boosted


```
{
  "link":["http://arxiv.org/abs/1506.04908v3"],
  "summary":"We study supervised learning problems using clustering constraints to impose\nstructure on either features or samples, seeking t
  "title":"Learning with Clustering Structure",
  "authors":["Vincent Roulet",
    "Fajwel Fogel",
    "Alexandre d'Aspremont",
    "Francis Bach"],
  "date":"2015-06-16T00:00:00Z",
  "areas":["Computer Science"],
  "fields":["Computer Science"],
  "subjects":["Machine Learning"],
  "id":"32772",
  "_version_":1749494456557502464},
{
  "link":["http://arxiv.org/abs/1503.01563v1"],
  "summary":"Energy minimization has been an intensely studied core problem in computer\nvision. With growing image sizes (2D and 3D), it is
  "title":"Convex Optimization for Parallel Energy Minimization",
  "authors":["K. S. Sesh Kumar",
    "Alvaro Barbero",
    "Stefanie Jegelka",
    "Suvrit Sra",
    "Francis Bach"],
  "date":"2015-03-05T00:00:00Z",
  "areas":["Computer Science",
    "Mathematics"],
  "fields":["Computer Science",
    "Mathematics"],
  "subjects":["Computer Vision and Pattern Recognition",
    "Optimization and Control"],
  "id":"38998",
  "_version_":1749494613143453696},
{
```

Query 3 - analysis

	Non-boosted	Boosted
Precision	0.9	1.0
Recalls	0.9333	0.9333

Query 4

I am a researcher that wants to try some some new approaches related to my case study in linguistics.

Query 4

Query:

q: areas:(statistics) new approaches linguistics
qf: link summary title authors date areas fields subjects
defType: edismax

Boosted:

q: areas:(statistics) new approaches linguistics
qf: link summary title^2 authors date areas fields subjects
pf: title^10
defType: edismax

Query 4 - results boosted


```
{
  "link":["http://arxiv.org/abs/1411.3315v1"],
  "summary":"We propose a new computational approach for tracking and detecting\nstatistically significant linguistic shifts in the meaning and usage of wor
  "title":"Statistically Significant Detection of Linguistic Change",
  "authors":["Vivek Kulkarni",
    "Rami Al-Rfou",
    "Bryan Perozzi",
    "Steven Skiena"],
  "date":"2014-11-12T00:00:00Z",
  "areas":["Computer Science"],
  "fields":["Computer Science"],
  "subjects":["Computation and Language",
    "Information Retrieval",
    "Machine Learning"],
  "id":"9147",
  "_version_":1749493871850553344},
{
  "link":["http://arxiv.org/abs/1302.2569v1"],
  "summary":"We propose a new statistical model for computational linguistics. Rather than\ntrying to estimate directly the probability distribution of a ra
  "title":"Toric grammars: a new statistical approach to natural language modeling",
  "authors":["Olivier Catoni",
    "Thomas Mainguy"],
  "date":"2013-02-11T00:00:00Z",
  "areas":["Statistics",
    "Computer Science",
    "Mathematics"],
  "fields":["Statistics",
    "Computer Science",
    "Mathematics"],
  "subjects":["Machine Learning",
    "Computation and Language",
    "Probability"],
  "id":"8989",
  "_version_":1749493867142447104},
{
```

Query 4 - analysis

	Non-boosted	Boosted
Precision	0.5	0.6
Recalls	0.454545	1.0

Query 5

I am a student that wants to get all the papers related to economics and computer science, in the year of 2017.

Query 5

Query:

q: Computer Science economics
fq: date:[2017-01-01T00:00:00Z TO 2018-01-01T00:00:00Z}
qf: link summary title authors date areas fields subjects
defType: edismax

Boosted:

q: Computer Science economics
fq: date:[2017-01-01T00:00:00Z TO 2018-01-01T00:00:00Z}
qf: link summary title authors date areas^5 fields^5 subjects^5
defType: edismax

Query 5 - results boosted


```
{
  "link":["http://arxiv.org/abs/1701.08567v2"],
  "summary":"As we know, there is a controversy about the decision making under risk\nbetween economists and psychologists. We discuss to build a unified theory of\nrisky
  "title":"Decision structure of risky choice",
  "authors":["Lamb Wubin",
    "Naixin Ren"],
  "date":"2017-01-30T00:00:00Z",
  "areas":["Quantitative Finance",
    "Computer Science"],
  "fields":["Quantitative Finance",
    "Computer Science"],
  "subjects":["Economics",
    "Artificial Intelligence"],
  "id":"37267",
  "_version_":1749494566415761408},
{
  "link":["http://arxiv.org/abs/1702.02896v2"],
  "summary":"We consider the problem of using observational data to learn treatment\nassignment policies that satisfy certain constraints specified by a\npractitioner, suc
  "title":"Efficient Policy Learning",
  "authors":["Susan Athey",
    "Stefan Wager"],
  "date":"2017-02-09T00:00:00Z",
  "areas":["Mathematics",
    "Computer Science",
    "Economics",
    "Statistics"],
  "fields":["Mathematics",
    "Computer Science",
    "Economics",
    "Statistics"],
  "subjects":["Statistics Theory",
    "Machine Learning",
    "Econometrics",
    "Machine Learning",
    "Statistics Theory"],
  "id":"13127",
  "_version_":1749493979506802688},
{
```

Query 5 - analysis

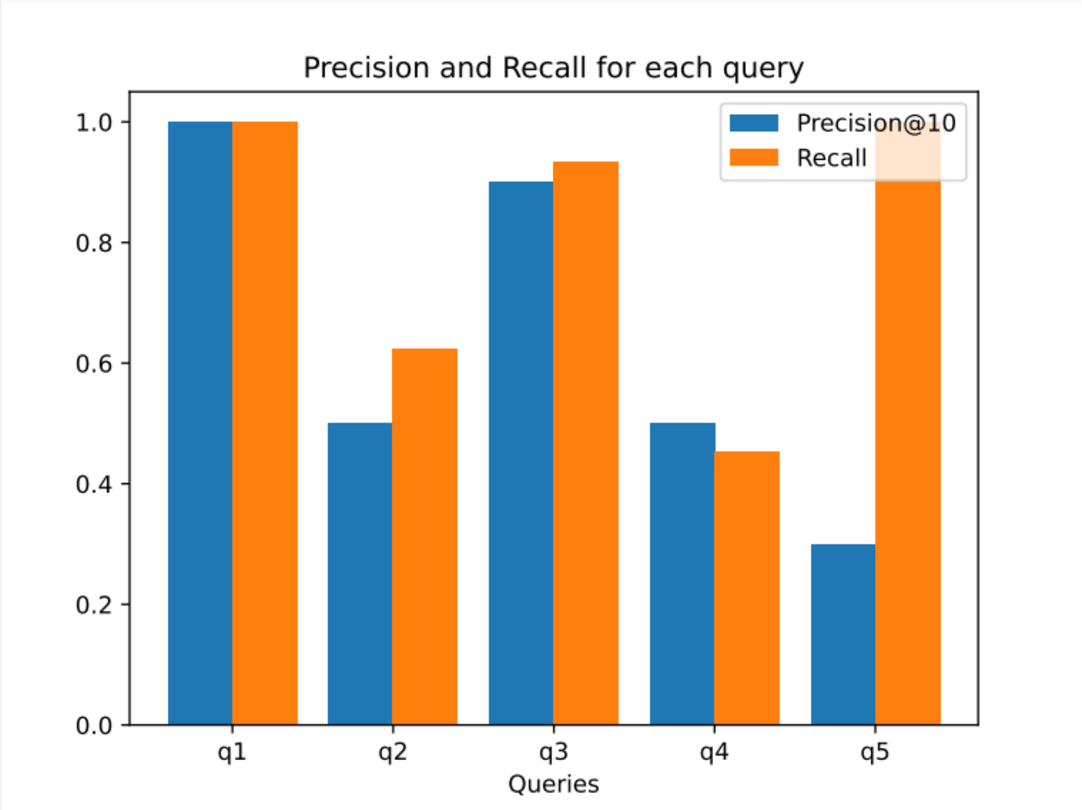
	Non-boosted	Boosted
Precision	0.3	0.9
Recalls	1.0	1.0

Overall results

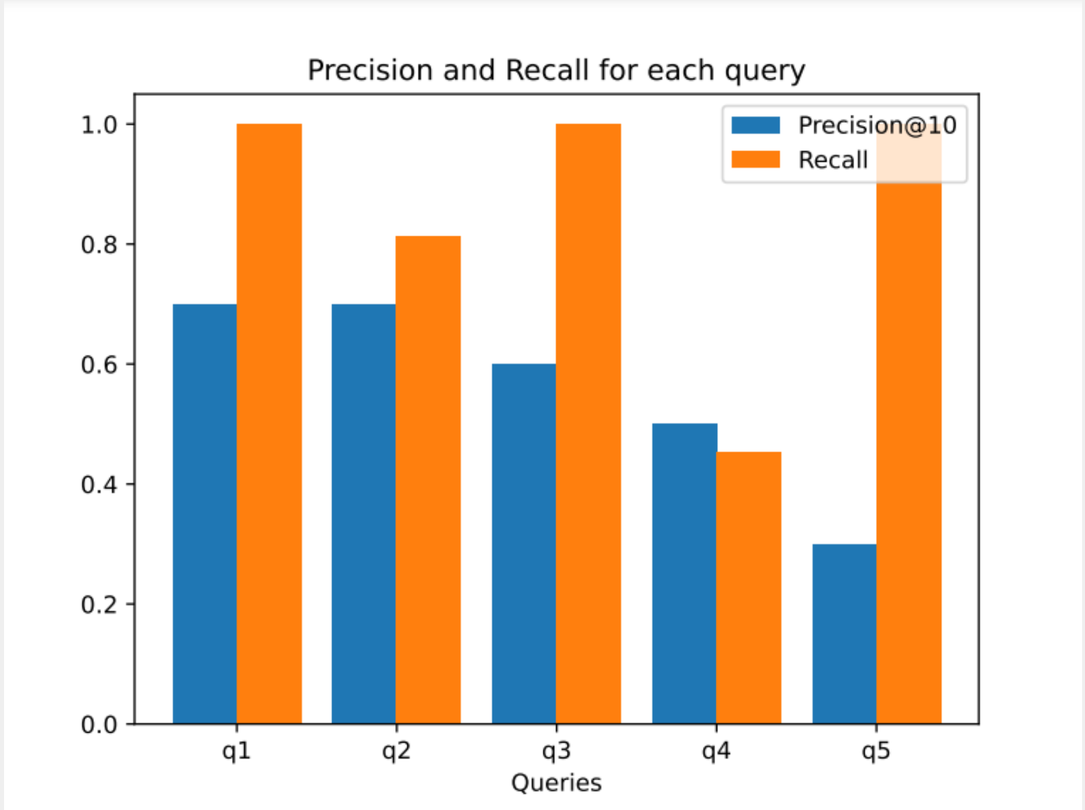
	Non-boosted	Boosted
Average Precision	0.6399	0.9
Average Recalls	0.8026	0.97417

Overall results - non-boosted

Best schema

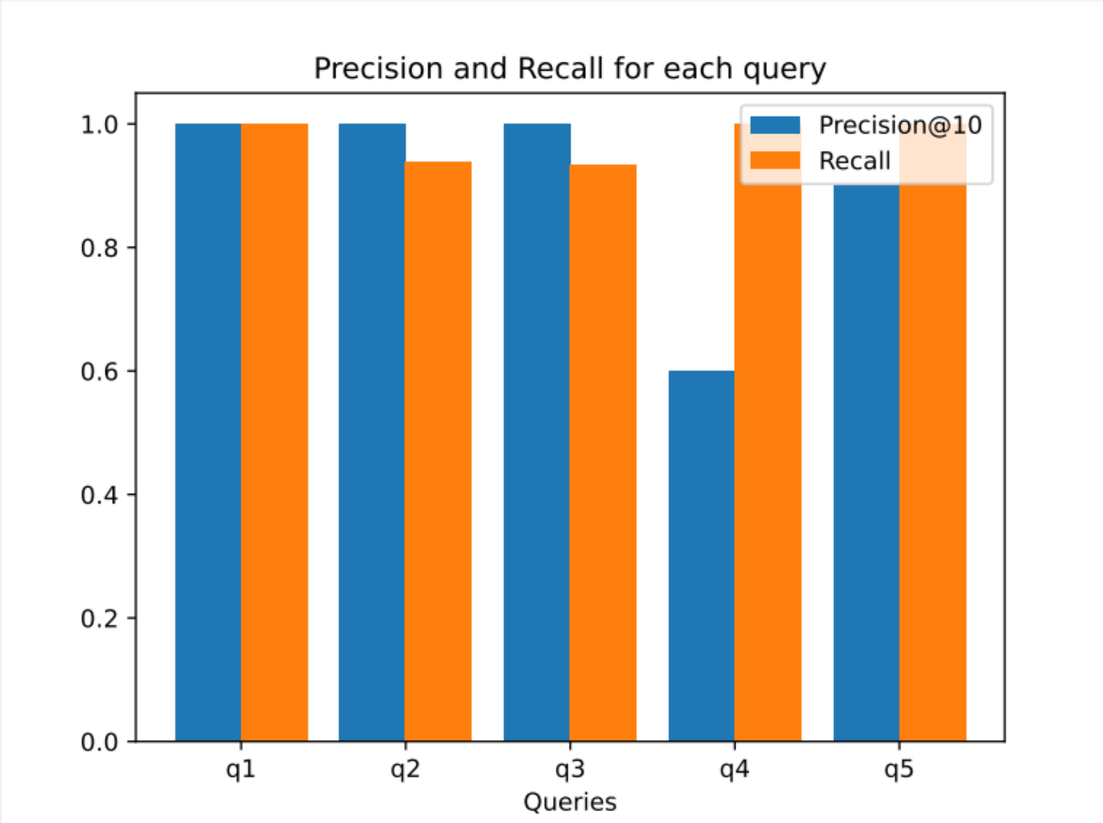


Worst schema

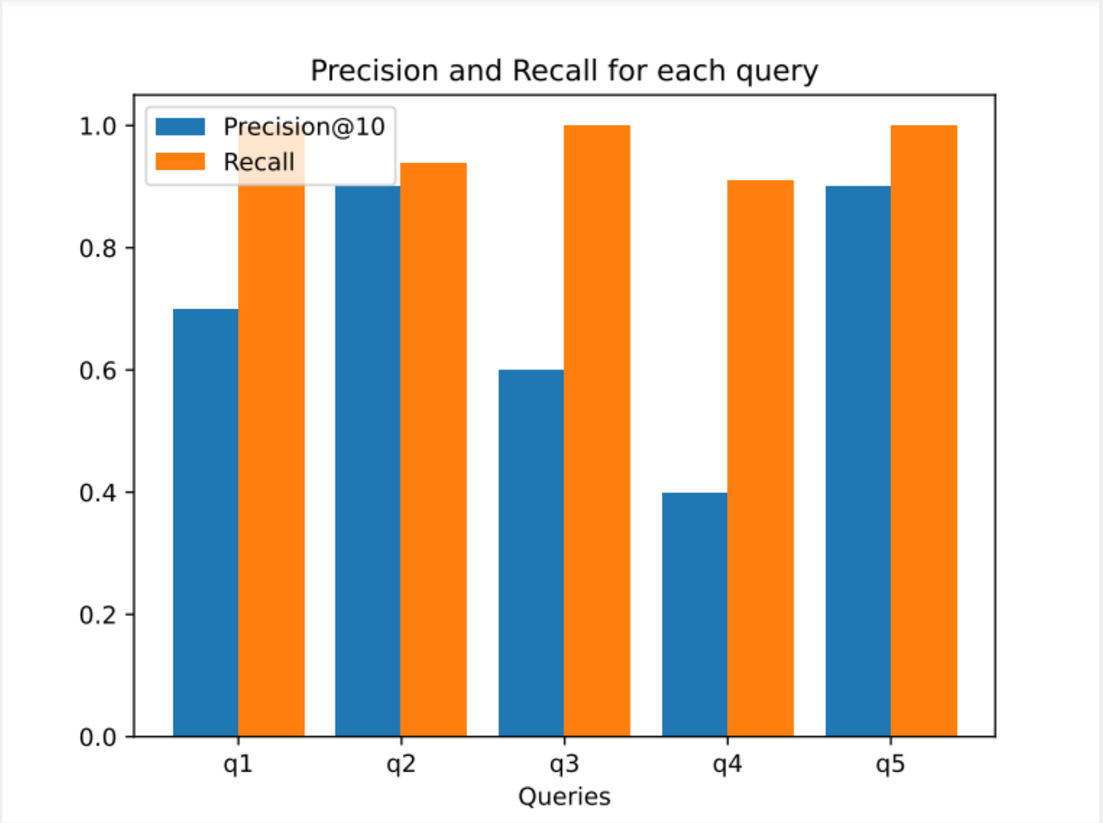


Overall results - boosted

Best schema



Worst schema



Questions?