



Journey to a
Relevant Search

Haystack Conference - Sept 29th, 2021



Leonardo Wajnsztok

GitHub: @leotok

Linkedin: @leonardowajnsztok

About me

- Senior Software Engineer (Search & Recommendation) at OLX Brasil
- Master's Degree in NLP at PUC-Rio (in progress)
- Fun Fact: former member of RioBotz combat robot team



OLX is the biggest classifieds platform in Brazil

+6M

daily active users
visiting our platform

+500k


new ads inserted
every day






16M indexed ads


+150M

search result pages
presented everyday

How does OLX work?



 Meus Anúncios  Chat  Notificações  Leonardo 



Brasil > RJ

DDD 21 - Rio de Janeiro e região, 18.171
DDD 22 - Norte do Estado e Região dos Lagos, 1.742
DDD 24 - Serra, Angra dos Reis e região, 1.622

Busca por categorias

< Todas as categorias

< Eletrônicos e celulares

Celulares e telefonia


Tipo

- ☐ Iphone
- ☐ Asus
- ☐ Samsung
- ☐ LG
- ☐ Nokia
- ☐ Motorola e Lenovo
- ☐ Sony
- ☐ Blackberry
- ☐ HTC
- ☐ Outras marcas
- ☐ Telefones e aparelhos de fax
- ☐ Acessórios

Novo/Usado

- ☐ Novo
- ☐ Usado

Preço



1 - 50 de 21.535 resultados

"iphone" - Celulares e telefonia no Rio de Janeiro

Pagamento e entrega



Todos os anúncios

Tipos de anúncio

Todos os anúncios

Ordenar por

Mais Relevantes



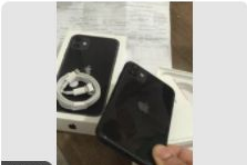
iPhone XR

● Online

Petrópolis, Independência - DDD 24

R\$ 3.100

Hoje 15:44



iPhone 11

R\$ 3.100

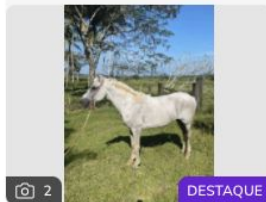
Hoje 15:44

All sort of ads



Playstation 4 Slim de 1tb 1000gb.

Valinhos - SP



Vendo cavalo mangalarga

Maceió - AL

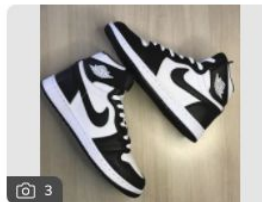
DESTAQUE



HONDA HR-V 1.8 16V FLEX EX 4P AUTOMÁTICO

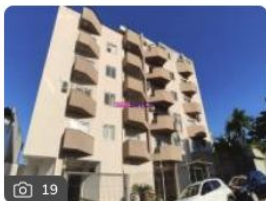
26.520 km | Câmbio: Automático | Flex

Fortaleza - CE



Nike Air Jordan 1 Preto com branco

Palmas - TO



APARTAMENTO MOBILIADO EM COQUEIROS DE 2 QUARTO S, 1 GARAGEM E ELEVADOR!

À venda | 2 quartos | 69m² | Condomínio: R\$ 288 | 1 vaga

Florianópolis - SC



Helicoptero Robinson 44 R44

Pinhais - PR

Problems and challenges

Ads/items are unique

- Harder to use historical information for feedback loop

Both online and offline transactions

- Harder to make sure an item was purchased

Many ads for a single "product"

- Quality vs Democracy

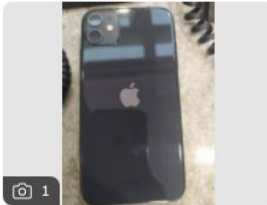


Problems and challenges



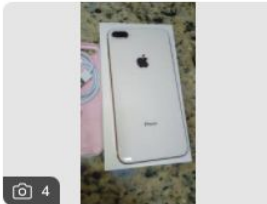
iPhone 8 Plus Cinza Espacial 64gb ótimo estado (Aceito cartã

Niterói, Itaipu - DDD 21



Iphone

Rio de Janeiro, Campo Grande - DDD 21



Vendo

Rio de Janeiro, Campo Grande - DDD 21

Translation:
"for sale"



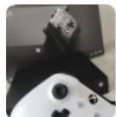
Past x Present

ps4



2017

1 - 20 de 58.608 resultados



Xbox One X 2100 à vista ou
12 X 200,00 AC Trocas !!!!!!!

R\$2200,00

Brasília



Hoverboard

R\$550,00

Rio de Janeiro



Hoverboard

R\$550,00

Rio de Janeiro



Street Fighter V PS4 Game
Original P HD

R\$49,00

Vila Velha



Xbox series X Novo lacrado
com Nota Fiscal 5500 AC
cartao 12 x 500

R\$,00

Brasília

2021

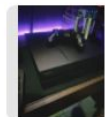
1 - 20 de 58.608 resultados



Ps4 fat 1 terá

R\$1599,00

São Paulo



PS4 fat + jogos

R\$1600,00

Viana



Ps4 Slim 1tb + 2 Controles e 1R\$1850,00
jogo

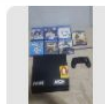
Santo André



Aluga-se PS4

R\$50,00

Boa Vista



PS4 Fat 1 TB.

R\$1700,00

Marinópolis

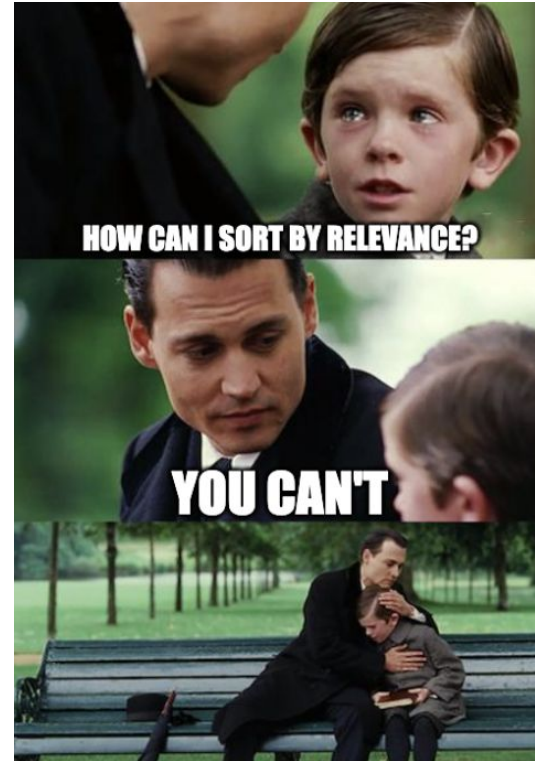
Why was it so bad in 2017?

Sort by

Date

Date

Price



Beginning of the journey

Convince stakeholders

Relevance testing

Baseline DCG



Beginning of the journey

Simon Jacquin
a search evaluation platform

Experiments

Splainer

Compare

New experiment

Show metric:

DCG_log2

At:

@10

	Id ↕ Enter	Description Enter Descripti	Rank ID Enter Rank ID	Sort Type Enter !	Suggest Enter Suggi	N° docs Enter	Index Enter ID	Queryset Enter Queryset	Completed ↕ Yes	Metric ↕ Enter
<input type="checkbox"/>	306	Smart Recency Podium Top queries	search-smart- recency_on	recency_relevance		10	evaluation_ 20200109	115 top queries removendo filtros -> Foram extraídas do ano de 2019 - Android	✓	DCG_log2 @10: 0.9485
<input type="checkbox"/>	305	Baseline Recência para Top queries	cdrelrank_rk0000	recency		10	evaluation_ 20200109	115 top queries removendo filtros -> Foram extraídas do ano de 2019 - Android	✓	DCG_log2 @10: 0.5359
<input type="checkbox"/>	303	Podium H (test)	cdrelrank- podium_podium- h	relevance		10	evaluation_ 20200109	1000 queries de teste -> Foram extraídas da randomicamente do queryset de teste. Ano de 2019 - Android	✓	DCG_log2 @10: 0.8641
<input type="checkbox"/>	302	Podium H	cdrelrank- podium_podium- h	relevance		10	evaluation_ 20200109	2002 queries -> Foram extraídas do ano de 2019 - Android	✓	DCG_log2 @10: 0.8369

Simon or Quepid?

Simon Jacquin

a search evaluation platform

Experiments

Show metric:

At:

DCG_log2

@10

	<div>Id</div>	<div>Description</div>	<div>Rank ID</div>	<div>Sort Type</div>	<div>Suggest</div>	<div>N° docs</div>	<div>Index</div>	<div>Queryset</div>	<div>Completed</div>	<div>Metric</div>
	<div>Enter</div>	<div>Enter Descripti</div>	<div>Enter Rank ID</div>	<div>Enter T</div>	<div>Enter Suggi</div>	<div>Enter</div>	<div>Enter ID</div>	<div>Enter Queryset</div>	<div>Yes</div>	<div>Enter</div>
<input type="checkbox"/>	306	Smart Recency Podium Top queries	search-smart-recency_on	recency_relevance		10	evaluation_20200109	115 top queries removendo filtros -> Forum extraídas do ano de 2019 - Android	<input checked="" type="checkbox"/>	DCG_log2 @10: 0.9485
<input type="checkbox"/>	305	Baseline Recência para Top queries	cdreirank_rk0000	recency		10	evaluation_20200109	115 top queries removendo filtros -> Forum extraídas do ano de 2019 - Android	<input checked="" type="checkbox"/>	DCG_log2 @10: 0.5399
<input type="checkbox"/>	303	Podium H (test)	cdreirank-podium_podium-h	relevance		10	evaluation_20200109	1000 queries de teste -> Forum extraídas da randomicamente do queryset de teste. Ano de 2019 - Android	<input checked="" type="checkbox"/>	DCG_log2 @10: 0.8641
<input type="checkbox"/>	302	Podium H	cdreirank-podium_podium-h	relevance		10	evaluation_20200109	2002 queries -> Forum extraídas do ano de 2019 - Android	<input checked="" type="checkbox"/>	DCG_log2 @10: 0.8369

← →

www.quepid.com

↺

Quepid

Relevancy cases

Organizations

Custom scorers

test@example.com

92

average

Current case

Movies! —Try 13

Select scorer

Create snapshot

Compare snapshots

Share case

Developer Settings

Add a query to this case

Add query

Collapse all

Sort

Manual

Name

Score

Errors

90

action

73 Results

86

rambo

4 Results

Score All

Explain Missing Documents


Toggle Notes

Set Threshold

Move Query

Delete Query

9



Rambo

overview:...When governments fail to act on behalf of captive missionaries, ex-Green Beret John James Rambo sets aside his peaceful existence along the Salween River in a war-torn region of Thailand to take action

release_date:2008-01-24

Rank: #1

Matches

Math.min of

queryBoost

titleBoost

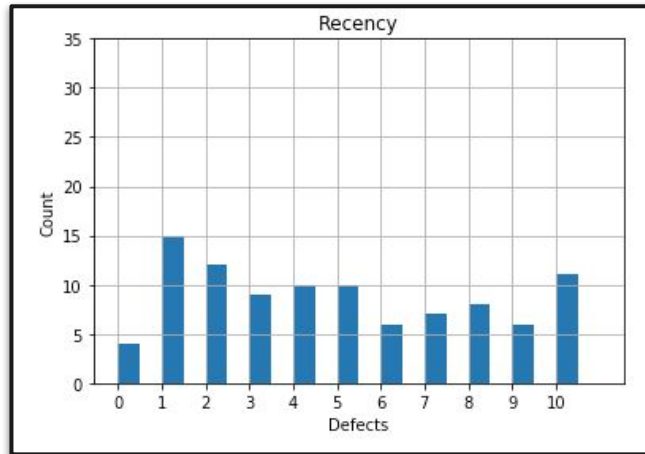
Show 1 More

First relevance iteration!

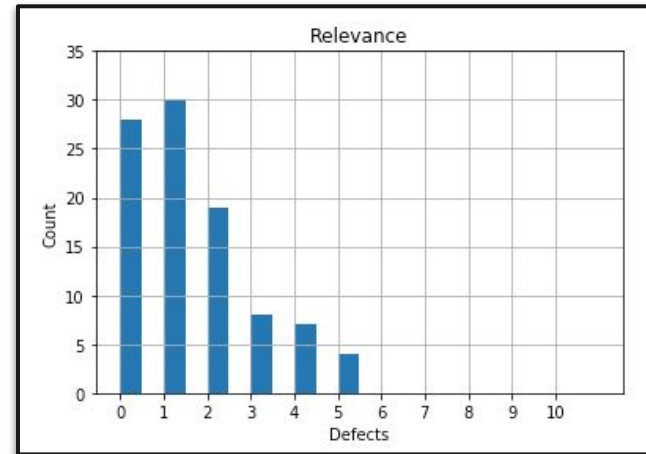
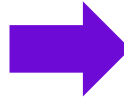
Elasticsearch: multi_match (BM25): $_score * \text{recency decay factor}$

-68.4% defect rate in a sample of search queries

Number of search queries with X irrelevant items



Number of irrelevant items in a search result



Number of irrelevant items in a search result

First "post-relevance" problem...

~30 fields in a document


Handcrafted fields boosts

Millions of possibilities to try

```
"multi_match": {  
  "query": "iphone 8",  
  "type": "cross_fields",  
  "operator": "and",  
  "fields": [  
    "title^3.0",  
    "description^1.5",  
    "model^70.0",  
    "category^35.0",  
    "brand^10.0",  
    "other^1.5",  
    "attribute^1.0",  
    "operation^0.0",  
    "adjective^0.0",  
    "state^10.0",  
    "city^35.0",  
    "neighbourhood^70.0",  
    "address^1.0",  
    "neighbor_cities^3.0",  
    "neighbor_neighbourhoods^2.0",  
    "poi^0.0",  
    ...  
  ]  
}
```

Small blanket problem

Translation:
sneakers




ASX 2.0 Automática 2014
Impecável

R\$ 52.000

Hoje às 09:26


DESTAQUE



Boneca em porcelana antiga
alemã.

R\$ 100

Ontem às 11:08



Honda Fit Lx Automático

R\$ 44.900

Ontem às 08:52

Ads from these neighborhoods:

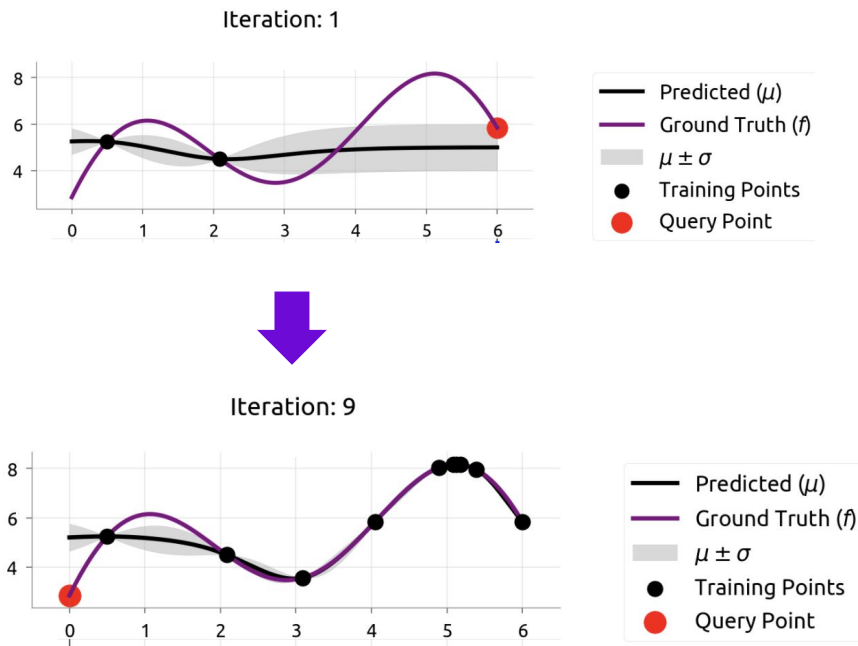
- Vila Tennis Clube
- Pinheiros Tennis Village
- Delta Ville Tennis Club

What about Hyperparameter Optimization?

Relevance judgements
to optimize fields boosts

Reorder ads optimizing NDCG

Optuna library



<https://broutonlab.com/blog/efficient-hyperparameter-optimization-with-optuna-framework>
<https://distill.pub/2020/bayesian-optimization/>

Boosts after optimizing

Old boosts:

```
"multi_match": {
  "query": "iphone 8",
  "type": "cross_fields",
  "operator": "and",
  "fields": [
    "title^3.0",
    "description^1.5",
    "model^70.0",
    "category^35.0",
    "brand^10.0",
    "attribute^1.0",
    "operation^0.0",
    "adjective^0.0",
    "state^10.0",
    "city^35.0",
    "neighbourhood^70.0",
    "address^1.0",
    "neighbor_cities^3.0",
    "neighbor_neighbourhoods^2.0",
    "poi^0.0",
    ...
  ]
}
```

New boosts:

```
"multi_match": {
  "query": "iphone 8",
  "type": "cross_fields",
  "operator": "and",
  "fields": [
    → "title^90.0",
    "description^0.0",
    "model^5.0",
    "category^85.0",
    "brand^30.0",
    "attribute^0.0",
    "operation^10.0",
    "adjective^50.0",
    "state^50.0",
    "city^25.0",
    "neighbourhood^70.0",
    "address^0.0",
    "neighbor_cities^35.0",
    "neighbor_neighbourhoods^0.0",
    "poi^5.0",
    ...
  ]
}
```

Problems with BM25

👍 **Better precision**

+8.6% in offline DCG

Although...

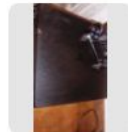
1 - 20 de 59.822 resultados



Ps4

R\$1950,00

Fortaleza



Ps4

R\$1750,00

Porto Alegre



PS4

R\$2200,00

Brasília



PS4

R\$1600,00

Araguari



Ps4

R\$2090,00

Cuiabá

Problems with BM25

👍 **Better precision**

+8.6% in offline DCG

Although...

👎 **Poor title quality and diversity**

-21.2% in title words count

-37.3% in title dissimilarity

👎 **Poor democracy in terms of impressions between ads with different titles**

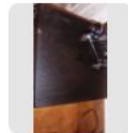
1 - 20 de 59.822 resultados



Ps4

R\$1950,00

Fortaleza



Ps4

R\$1750,00

Porto Alegre



PS4

R\$2200,00

Brasília



PS4

R\$1600,00

Araguari

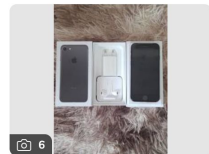


Ps4

R\$2090,00

Cuiabá

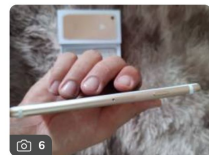
More problems with BM25!



iPhone iPhone iPhone iPhone
iPhone iPhone iPhone iPhone
iPhone iPhone iPhone iPhone...

R\$ 1.800

Fortaleza - CE



iPhone iPhone iPhone iPhone
iPhone iPhone iPhone iPhone
iPhone iPhone iPhone iPhone

R\$ 1.800

Fortaleza - CE



Invicta Invicta Invicta
Invicta Invicta Invicta Invi...
R\$599

Madureira - 19/06 às 18:50



Invicta Invicta Invicta
Invicta Invicta Invicta Invi...
R\$549

Irajá - 19/06 às 18:43



2) 99985-4343

Sofá Sofá Sofá Sofá Sofá Sofá
Sofá Sofá Sofá Sofá Sofá Sof...

R\$ 850

Ontem às 17:30



Sofá Sofá Sofá Sofá Sofá Sofá
Sofá Sofá Sofá Sofá Sofá Sof...

R\$ 950

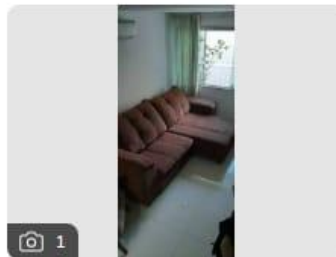
Ontem às 17:20

The last straw with BM25!

Translation:

"Retractable sofa. I'm also going to write 'sofa sofa sofa' hahahaha"

This user was mocking us! 🤪 🗑️



Sofá retrátil. Também vou colocar sofá, sofá, sofá, hahahaha

R\$ 1.650
~~R\$ 1.700~~

Rio de Janeiro, Recreio dos Bandeirantes - DDD 21

Learnings until then

Title is a very important field

Users can "hack" BM25 in a C2C use case

Term frequency in short texts
is not that important

Field length leads to better diversity,
but depends on term frequency

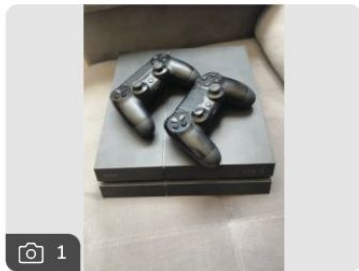


$$\sum_i^n IDF(q_i) \frac{f(q_i, D) * (k1 + 1)}{f(q_i, D) + k1 * (1 - b + b * \frac{fieldLen}{avgFieldLen})}$$

and the ultimate Insight...

word **ORDER** and **POSITION** matter!

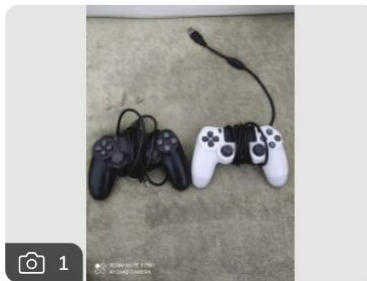
(at least in Portuguese)



PS4 com dois controles

Santos - SP

Video game console



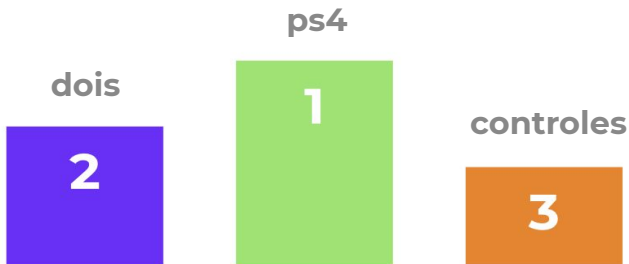
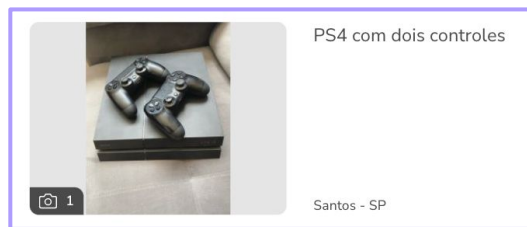
CONTROLE DE PS4 ORIGINAL

Rio de Janeiro, Vila da Penha - DDD 21

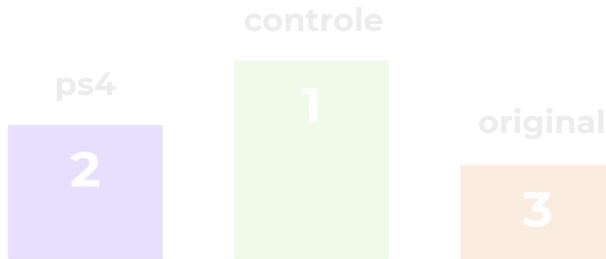
Video game controller

"Term podium" match

Indexing item 1

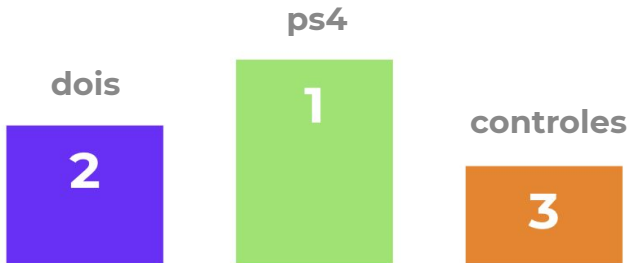


Indexing item 2

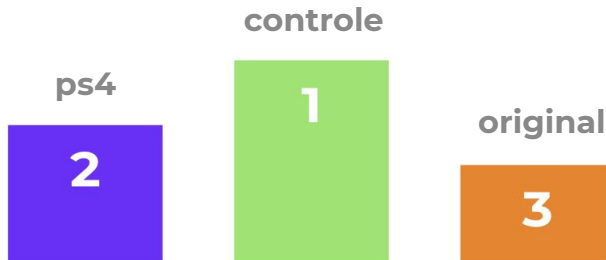


"Term podium" match

Indexing item 1



Indexing item 2



"Term podium" match



"Term podium" match

$$Score(PQ, PI) = \sum_{i=0}^{N_{max}^Q-1} \sum_{j=0}^{N_{max}^I-1} match(PQ_i^{rev}, PI_j^{rev}) \cdot PW(i * N_{max}^Q + j)$$

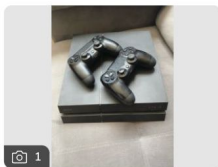
$$match(a, b) = \begin{cases} 1, & \text{if } a = b. \\ 0, & \text{otherwise.} \end{cases}$$

$$PW(n) = 2^n$$

$$N_{max}^Q = \min(3, |PQ|)$$

$$N_{max}^I = \min(3, |PI|)$$

"Term podium" match



PS4 com dois controles

Santos - SP

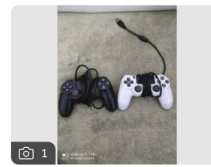
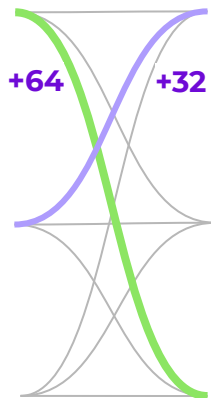
Podium Query

controle



Podium Item 1

ps4



CONTROLE DE PS4 ORIGINAL

Rio de Janeiro, Vila da Penha - DDD 21

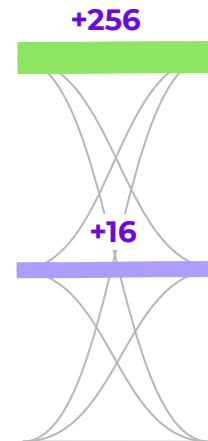
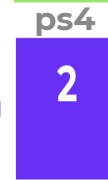
Podium Query

controle




Podium Item 2

controle

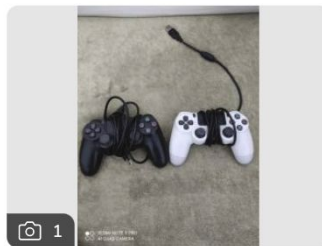


"Term podium" match

Search Result:

$\text{Score}(\text{query}, \text{item2}) = 16 + 256 = 272$

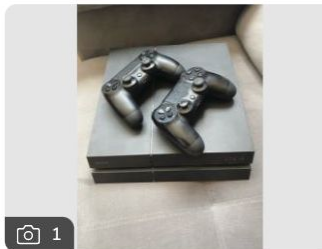


CONTROLE DE PS4 ORIGINAL

Rio de Janeiro, Vila da Penha - DDD 21

...

$\text{Score}(\text{query}, \text{item1}) = 32 + 64 = 96$



PS4 com dois controles

Santos - SP

Benefits of "Term Podium"

Known maximum score

More "Equally relevant" documents have the same score, allowing to break a tie using other fields (ex: recency)

Easier to interpret compared to BM25

Recap: Past x Term Podium

ps4



2017

1 - 20 de 58.608 resultados



Xbox One X 2100 à vista ou
12 X 200,00 AC Trocas !!!!!!!

R\$2200,00

Brasília



Hoverboard

R\$550,00

Rio de Janeiro



Hoverboard

R\$550,00

Rio de Janeiro



Street Fighter V PS4 Game
Original P HD

R\$49,00

Vila Velha



Xbox series X Novo lacrado
com Nota Fiscal 5500 AC
cartao 12 x 500

R\$,00

Brasília

2021

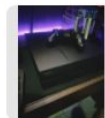
1 - 20 de 58.608 resultados



Ps4 fat 1 terá

R\$1599,00

São Paulo



PS4 fat + jogos

R\$1600,00

Viana



Ps4 Slim 1tb + 2 Controles e 1R\$1850,00
jogo

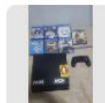
Santo André



Aluga-se PS4

R\$50,00

Boa Vista



PS4 Fat 1 TB.

R\$1700,00

Marinópolis

Term podium results

Better relevance

- + **3.8%** in offline DCG
- + **2.9%** in clicks @5
- + **3.6%** in conversion rate

Better diversity

- +**43.6%** in title words count in the search results
- +**64.1%** in title dissimilarity in the search results

Better democracy

No more BM25 hacks from our users

Conclusions

Begin with metrics (if possible)

Know your data

Understand the methods and algorithms used

Start with the simple stuff

Read this thread: <https://relevancy.slack.com/archives/C47DXEJUA/p1627368535279900>



Thank you!

Haystack Conference - Sept 29th, 2021

Contact information

Email: leonardo.wajnsztok@olxbr.com

Linkedin: [@leonardowajnsztok](#)