



Coleta e Busca de Entidades Estruturadas em um Domínio - Parte 2

Violões

Lucas Melo, Mateus Gonçalves

Tarefa 1: Criar arquivo invertido

| terms | | | | | | | |
|-------|-------|-----|-------|------|-------|------|--|
| he | drink | ink | likes | pink | thing | wink | |
| 2 | 1 | 0 | 2 | 0 | 0 | 1 | ← D1: He likes to wink, he likes to drink. |
| 1 | 3 | 0 | 1 | 0 | 0 | 0 | ← D2: He likes to drink, and drink, and drink. |
| 1 | 1 | 1 | 1 | 0 | 1 | 0 | ← D3: The thing he likes to drink is ink. |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | ← D4: The ink he likes to drink is pink. |
| 1 | 1 | 1 | 1 | 1 | 0 | 1 | ← D5: He likes to wink and drink pink ink. |



Atributos

- Marca
- Cordas (aço/nylon)
- Categoria (acústico/elétrico)
- Escala
- Tampo

Índice Invertido sem compressão

| | | |
|---|--------------------|--|
| 1 | Marca.giannini | [[1, 2], [2, 2], [3, 2], [6, 2], [7, 2], [17, 4], [24, 4], [64, 9], [68, 13], [72, 7], [73, 8], [80, 7] |
| 2 | Corda.aço | [[1, 2], [3, 2], [4, 3], [5, 3], [8, 2], [12, 1], [13, 1], [14, 2], [16, 1], [23, 3], [24, 2], [25, 3], [30, |
| 3 | Categoria.acústico | [[1, 3], [2, 3], [4, 4], [5, 4], [7, 3], [13, 2], [17, 1], [22, 4], [26, 3], [27, 3], [46, 1], [51, 2], [5] |
| 4 | Escala.rosewood | [[1, 1], [2, 1], [3, 1], [7, 2], [8, 1], [10, 1], [11, 2], [12, 1], [13, 1], [14, 1], [15, 3], [16, 3], [18, |
| 5 | Tampo.spruce | [[1, 1], [2, 3], [7, 1], [10, 1], [12, 1], [13, 1], [14, 1], [15, 1], [18, 1], [21, 2], [23, 3], [25, 3], |
| 6 | Corda.nylon | [[2, 2], [6, 3], [7, 2], [9, 2], [10, 3], [19, 1], [21, 4], [22, 4], [26, 3], [27, 4], [28, 4], [29, 1], [3 |
| 7 | Categoria.elétrico | [[3, 3], [6, 5], [8, 2], [9, 2], [10, 4], [11, 3], [12, 1], [18, 1], [19, 1], [20, 1], [21, 2], [23, 4], [2 |
| 8 | Tampo.basswood | [[3, 1], [4, 1], [5, 1], [20, 2], [22, 2], [24, 2], [26, 2]] |
| 9 | Marca.andaluz | [[4, 2], [5, 2], [10, 2]] |

- Quantidade de pares atributo-valor
 - **36 entradas**
- Tamanho do índice:
 - **3.8 KB**

Índice Invertido sem compressão

[illegible]

- Quantidade de pares atributo-valor
 - **36 entradas**
- Tamanho do índice:
 - **3.49 KB**



Dificuldades

- Em muitas páginas, as informações dos atributos não são estruturadas.
- Vários sites apresentam informações sobre outros violões que não o da página em questão, o que confunde as características do violão em questão com as características de violões de outras páginas, dificultando a análise do par Atributo-Valor real daquela página.
- Alguns sites não apresentam todos os atributos do violão.
- Em cada site, determinar strings particulares que delimitassem áreas com informações de interesse para a extração do texto.



Tarefa 2: Processamento da Query

- Document at a time
- Cosine Ranking
 - Sem TF-IDF da query (Baseado apenas na frequência)
 - Com TF-IDF da query
- Correlação de Spearman
- Correlação de Kendau Tal



Cosine Ranking sem TF-IDF

- Considerei:
 - Frequência de cada termo da query
 - Quantidade de documentos que cada termo aparecia

Cosine Ranking com TF-IDF

COSINESCORE(q)

```
1  float Scores[N] = 0
2  Initialize Length[N]
3  for each query term  $t$ 
4  do calculate  $w_{t,q}$  and fetch postings list for  $t$ 
5    for each pair( $d, tf_{t,d}$ ) in postings list
6      do Scores[d] +=  $w_{t,d} \times w_{t,q}$ 
7  Read the array Length[d]
8  for each  $d$ 
9    do Scores[d] = Scores[d] / Length[d]
10 return Top K components of Scores[]
```

Diagram illustrating the calculation of the cosine score for a document d based on the TF-IDF weights:

- The weight $w_{t,d}$ is calculated as $(0.5 + 0.5 \frac{f_{i,q}}{\max_i f_{i,q}}) * \log \frac{N}{n_i}$.
- The weight $w_{t,q}$ is calculated as $f_{i,j} * \log \frac{N}{n_i}$.



Correlação de Spearman e Kendau Tal

- Spearman:

$$S(\mathcal{R}_1, \mathcal{R}_2) = 1 - \frac{6 \times \sum_{j=1}^K (s_{1,j} - s_{2,j})^2}{K \times (K^2 - 1)}$$

- Kendau tal:

$$\tau(\mathcal{R}_1, \mathcal{R}_2) = 1 - \frac{2 \times \Delta(\mathcal{R}_1, \mathcal{R}_2)}{K(K-1)}$$



Cinco Pesquisas

- “giannini aço rosewood acústico spruce”:

```
giannini aço rosewood acústico spruce
Top 5 escolhas sem TF-IDF:
https://www.americanas.com.br/produto/8084563
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-folk-giannini-gf1r-eletrico-aco-natural-cutway-11117101.html
https://www.milsons.com.br/violao-aco-acustico-gs15-bk-classico-giannini-p1762/
https://www.milsons.com.br/violao-nylon-acustico-gn15-nt-classico-giannini-p1136/
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-estudo-acustico-nylon-preto-gn14-start-giannini-4011361.html
-----
Top 5 escolhas com TF-IDF:
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-estudo-acustico-nylon-preto-gn14-start-giannini-4011361.html
https://www.novamusic.com.br/produto/violao-giannini-infantil-nr-14-30/
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-folk-giannini-gf1r-eletrico-aco-natural-cutway-11117101.html
https://www.milsons.com.br/violao-aco-acustico-gs15-bk-classico-giannini-p1762/
https://www.milsons.com.br/violao-aco-acustico-x11-nt-andaluz-p1706/
-----
Correlacao de Spearman:
0.5227841540170307
-----
Correlacao de Kendal Tau:
0.6838208071084784
-----
```

Cinco Pesquisas

- “spruce acústico giannini”:

```
spruce acústico giannini
Top 5 escolhas sem TF-IDF:
https://www.americanas.com.br/produto/8084563
https://www.milsons.com.br/violao-nylon-acustico-gn15-nt-classico-giannini-p1136/
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-estudo-acustico-nylon-preto-gn14-start-giannini-4011361.html
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-folk-giannini-gf1r-eletrico-aco-natural-cutway-11117101.html
https://www.milsons.com.br/violao-aco-acustico-gs15-bk-classico-giannini-p1762/
-----
Top 5 escolhas com TF-IDF:
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-estudo-acustico-nylon-preto-gn14-start-giannini-4011361.html
https://www.novamusic.com.br/produto/violao-giannini-infantil-nr-14-30/
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-folk-giannini-gf1r-eletrico-aco-natural-cutway-11117101.html
https://www.milsons.com.br/violao-aco-acustico-gs15-bk-classico-giannini-p1762/
https://www.americanas.com.br/produto/8084563
-----
Correlacao de Spearman:
0.8943479388125486
-----
Correlacao de Kendal Tau:
0.8515037593984962
-----
```



Cinco Pesquisas

- “ibanez rosewood elétrico”:

```
ibanez rosewood elétrico
Top 5 escolhas sem TF-IDF:
https://www.novamusic.com.br/produto/violao-ibanez-tcy-8/
https://www.americanas.com.br/produto/8019215
https://www.novamusic.com.br/produto/violao-ibanez-aeg-6tn/
https://www.americanas.com.br/produto/8088174
https://www.americanas.com.br/produto/15090738
-----
Top 5 escolhas com TF-IDF:
https://www.novamusic.com.br/produto/violao-ibanez-tcy-8/
https://www.novamusic.com.br/produto/violao-ibanez-aeg-6tn/
https://www.novamusic.com.br/produto/violao-ibanez-ga3-ece/
https://www.milsons.com.br/violao-nylon-eletrico-s3904-andaluz-p1707/
https://www.milsons.com.br/violao-nylon-eletrico-nf14-flat-cutway-giannini-p1826/
-----
Correlacao de Spearman:
0.748118377785897
-----
Correlacao de Kendal Tau:
0.7855924978687128
-----
```



Cinco Pesquisas

- “aço acústico andaluz”:

```
aço acústico andaluz
Top 5 escolhas sem TF-IDF:
https://www.milsons.com.br/violao-aco-acustico-x11-nt-andaluz-p1706/
https://www.milsons.com.br/violao-aco-acustico-x11-bk-andaluz-p1705/
https://www.milsons.com.br/violao-aco-acustico-gs15-bk-classico-giannini-p1762/
https://www.americanas.com.br/produto/8084563
https://www.novamusic.com.br/produto/violao-epiphone-dr-100/
-----
Top 5 escolhas com TF-IDF:
https://www.milsons.com.br/violao-aco-acustico-x11-nt-andaluz-p1706/
https://www.milsons.com.br/violao-aco-acustico-x11-bk-andaluz-p1705/
https://www.milsons.com.br/violao-nylon-eletrico-s3904-andaluz-p1707/
https://www.milsons.com.br/violao-aco-acustico-gs15-bk-classico-giannini-p1762/
https://www.milsons.com.br/violao-nylon-acustico-gnc3spc-classico-giannini-p1787/
-----
Correlacao de Spearman:
0.5758031442241969
-----
Correlacao de Kendal Tau:
0.7188311688311688
-----
```



Cinco Pesquisas

- “di giorgio acústico nylon:

```
di giorgio acustico nylon
Top 5 escolhas sem TF-IDF:
https://www.multisom.com.br/produto/violao-acustico-di-giorgio-estudante-18-natural-com-cordas-de-nylon-296
https://www.americanas.com.br/produto/8084563
https://www.multisom.com.br/produto/violao-classico-di-giorgio-38e-natural-eletrico-cordas-de-nylon-com-afinador-p10-xlr-9006
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-classico-acustico-nylon-gna-111nt-alinhado-harmonics-13449442.html
https://www.casasbahia.com.br/instrumentosmusicais/cordas/violao/violao-classico-acustico-nylon-gna-111nt-alinhado-harmonics-13449442.html
-----
Top 5 escolhas com TF-IDF:
https://www.multisom.com.br/produto/violao-mini-jumbo-di-giorgio-irbl-iron-azul-eletrico-cordas-de-aco-e-com-afinador-16900
https://www.multisom.com.br/produto/violao-mini-jumbo-di-giorgio-irgd-iron-gold-eletrico-cordas-de-aco-e-com-afinador-16901
https://www.multisom.com.br/produto/violao-di-giorgio-talent-i-natural-cordas-de-nylon-eletrico-e-com-afinador-1501
https://www.multisom.com.br/produto/violao-classico-di-giorgio-38e-natural-eletrico-cordas-de-nylon-com-afinador-p10-xlr-9006
https://www.multisom.com.br/produto/violao-acustico-di-giorgio-estudante-18-natural-com-cordas-de-nylon-296
-----
Correlacao de Spearman:
0.7210526315789474
-----
Correlacao de Kendal Tau:
0.7678812415654521
-----
```



DEMONSTRAÇÃO



Dificuldades

- Dificuldade de entendimento do algoritmo de Cosine Ranking:
 - W_{tf} , W_{tq}
- Até o final, não sabia se podia estar certo, tive que confiar