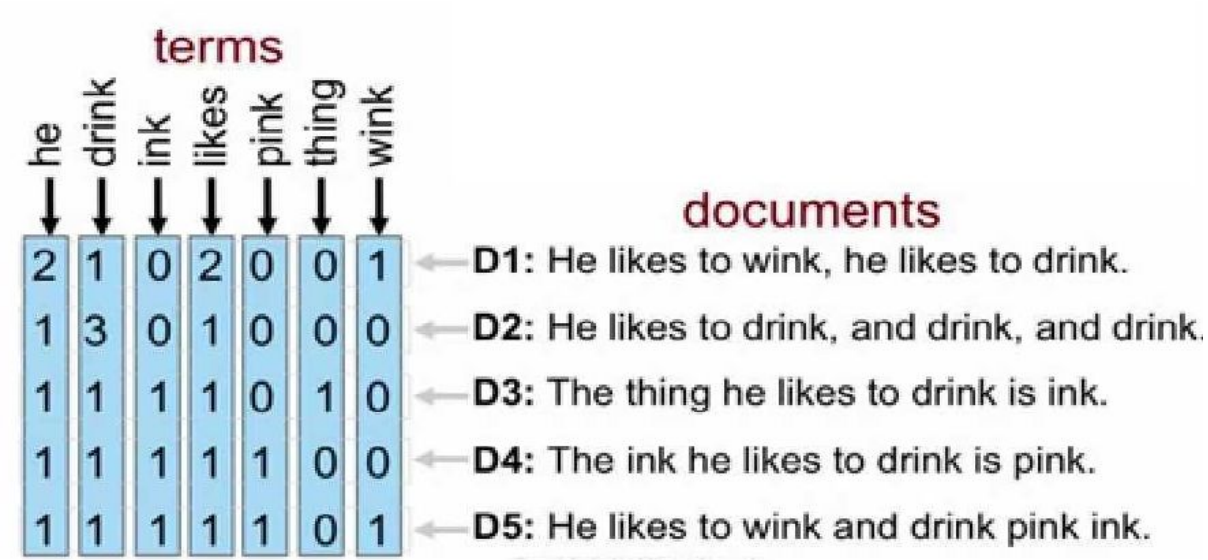
Coleta e Busca de Entidades Estruturadas em um Domínio - Parte 2

Violões
Lucas Melo, Mateus Gonçalves

Tarefa 1: Criar arquivo invertido



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], [3
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], [1
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], [
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], [
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

1	**Marca.giannini**	**[[1, 2], [1, 2], [1, 2], [3, 2], [1, 2], [10, 4], [7, 4], [40, 9], [4, 13], [4, 7], [1, 8], [7, 7]]**
2	Corda.aço	[[1, 2], [2, 2], [1, 3], [1, 3], [3, 2], [4, 1], [1, 1], [1, 2], [2, 1], [7, 3], [1, 2], [1, 3], [5, 1], [2, 3],
3	Categoria.acústico	[[1, 3], [1, 3], [2, 4], [1, 4], [2, 3], [6, 2], [4, 1], [5, 4], [4, 3], [1, 3], [19, 1], [5, 2], [7, 2], [8, 4],
4	Escala.rosewood	[[1, 1], [1, 1], [1, 1], [4, 2], [1, 1], [2, 1], [1, 2], [1, 1], [1, 1], [1, 1], [1, 3], [1, 3], [2, 2], [1, 1],
5	Tampo.spruce	[[1, 1], [1, 3], [5, 1], [3, 1], [2, 1], [1, 1], [1, 1], [1, 1], [3, 1], [3, 2], [2, 3], [2, 3], [3, 2], [1, 1],
6	Corda.nylon	[[2, 2], [4, 3], [1, 2], [2, 2], [1, 3], [9, 1], [2, 4], [1, 4], [4, 3], [1, 4], [1, 4], [1, 1], [2, 3], [9, 3],
7	Categoria.elétrico	[[3, 3], [3, 5], [2, 2], [1, 2], [1, 4], [1, 3], [1, 1], [6, 1], [1, 1], [1, 1], [1, 2], [2, 4], [1, 2], [1, 3],
8	Tampo.basswood	[[3, 1], [1, 1], [1, 1], [15, 2], [2, 2], [2, 2], [2, 2]]
9	Marca.andaluz	[[4, 2], [1, 2], [5, 2]]

- 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] += wf_{t,d} × 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[]
```

$(0.5 + 0.5\frac{f_{i,q}}{max_i \ f_{i,q}}) * \log \frac{N}{n_i}$

$f_{i,j} * \log \frac{N}{n_i}$

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:

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

DEMONSTRAÇÃO

Dificuldades

- Dificuldade de entendimento do algoritmo de Cosine Ranking:
  - Wtf, Wtq
- Até o final, não sabia se podia estar certo, tive que confiar