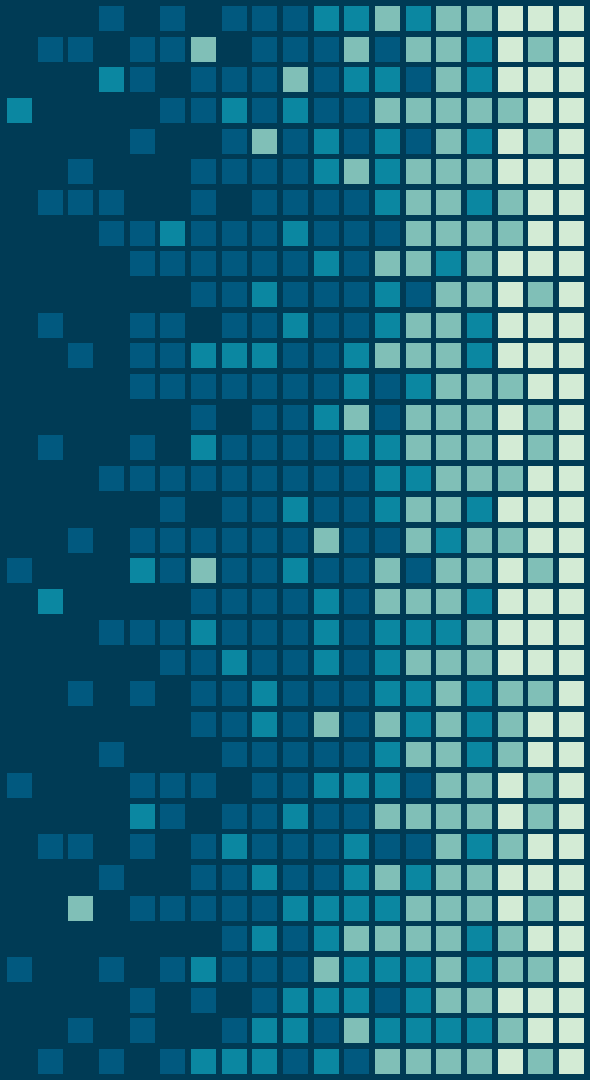
# Recuperação de Informação

Integrantes:

- Bruno Vitorino (bvcl)
- Leonardo Galdino (lcgm)
- Lucas Santana (lss5)

21/11/2019



# Ideia

- Domínio: Jogadores de Futebol (em inglês).
- Engenho de busca vertical sobre esse domínio.
- Principais atributos: nome, posição, nacionalidade, número (da camisa), time, perna principal.

# Parte 1: Crawler, classificador, extrator.

# Responsabilidade dos integrantes

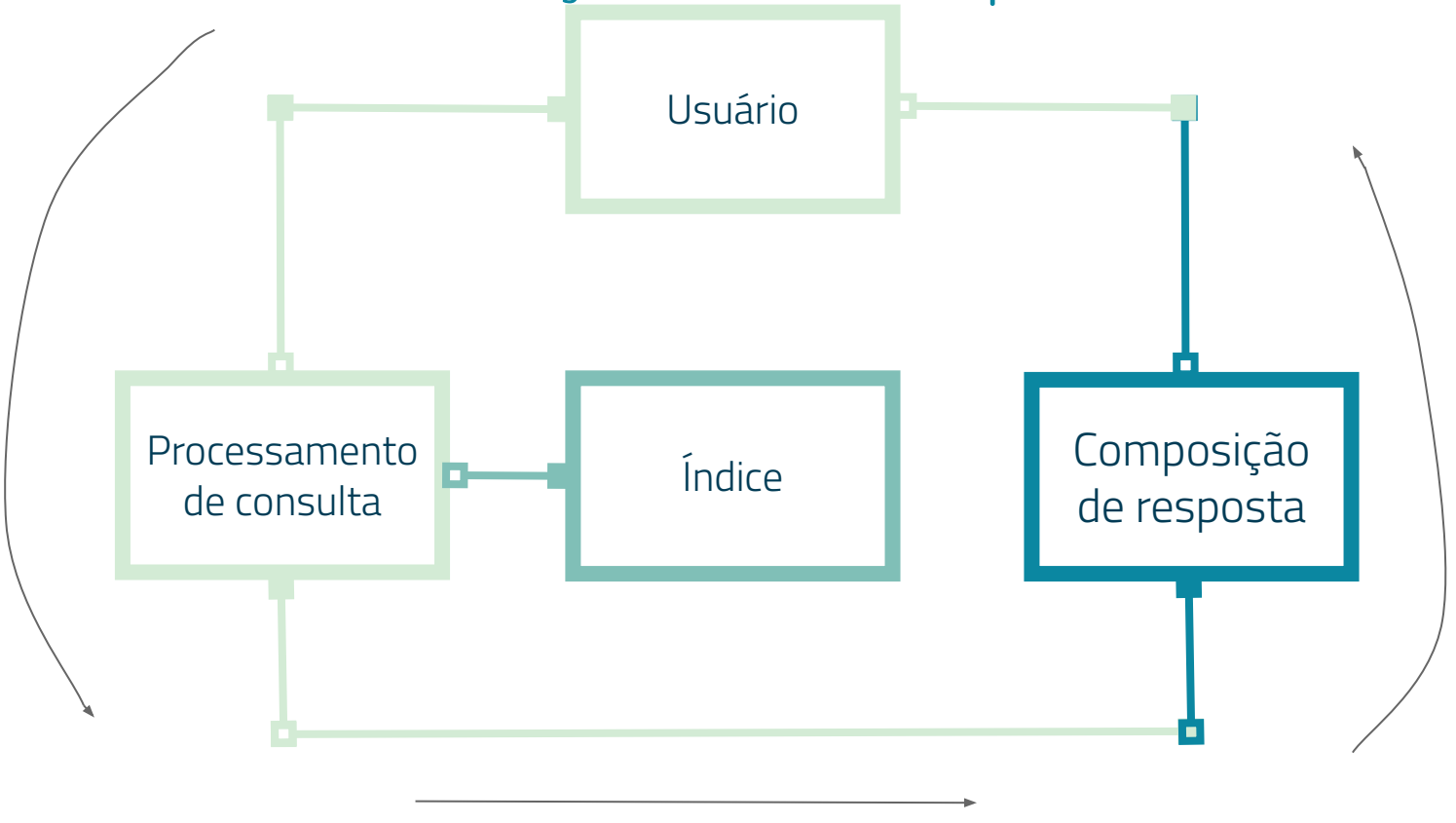
- Bruno Vitorino: Crawler (navegar pelo mundo de páginas existentes para achar dados)
- Leonardo Galdino: Classificador (detectar e selecionar páginas que contêm dados sobre jogadores)
- Lucas Santana: Extrator (coletar dados das páginas selecionadas)

# Parte 2: Índice, processamento de consulta, composição de resposta.

# 5533

Páginas em nosso engenho de busca.

# Interação entre as partes



# Responsabilidade dos integrantes

- Lucas Santana: Índice (mapeamento pré-processado sobre quais documentos do engenho contêm cada termo do vocabulário).
- Leonardo Galdino: Processamento de consulta (transformar a necessidade de informação do usuário em forma de texto para uma lista ordenada de documentos).
- Bruno Vitorino: Composição de respostas (transformar a lista ordenada de documentos em uma resposta HTML para que o usuário possa consumir em seu browser).

# Índice Invertido

- Linguagem usada: Python;
- Inicialmente: Serializer.py extrai todas as entidades (5533, no total, agora chamadas de documentos) e adiciona num aquivo JSON;
- Freq_index.py acessa o arquivo JSON e cria um arquivo txt do index invertido com frequência:
  - **field//term freqN ID1:freq1 ID2:freq2 ... IDn:freqn**
- Gera também uma versão com compressão nos postings, só adicionando os gaps:
  - **field//term freqN ID1:freq1 (ID2-ID1):freq2 ... (IDn - IDn-1):freqn**
- Bit_index.py transforma os postings do index em um arquivo binário usando a codificação gamma (γ).

# Índice

| Armazenamento | Tamanho |
|---|---|
| Pasta c/ arquivos .html | 1.1 GB |
| Arquivo .json | 2.5 MB |
| Freq_index.txt | 1.7 MB |
| Freq_index.txt c/ gap nos postings | 1.2 MB |
| Arquivo .bin com codificação gamma (γ) | 211.2 kB (só posting) // 574.9 kB (c/ termos em .txt) |

# Processamento de consulta

- Backend em Flask (Python).
- Tokenização e remoção de stopwords (uso da biblioteca NLTK).
- Ranqueamento de documentos: baseado no modelo de espaço de vetores. As componentes do vetor de um documento são dadas pelo TF-IDF (vetor no espaço da consulta). Também é possível usar presença ou ausência do termo no documento como componente do vetor.
- Mapeamento do atributo número para o devido quartil.
- Coleta de métricas: correlação de ranking (Kendall Tau) e medição do tempo de resposta.

# Correlação de Ranking: TF-IDF vs Presença/Ausência de termo

```
lcgm@lcgm-ss51:~/workspace/PlayersInformationRetrieval/src$ pipenv run python compute_metrics.py ranking_correlation
[nltk_data] Downloading package stopwords to /home/lcgm/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
For query [neymar messi] on field [term]: 0.23809523809523814 Kendal Tau correlation. Rankings with 10 documents.
For query [cristiano fifa ronaldo] on field [term]: 0.9908045977011495 Kendal Tau correlation. Rankings with 30 documents.
For query [cristiano brazil attacker] on field [term]: -0.27356321839080455 Kendal Tau correlation. Rankings with 30 documents
For query [Leonel Messi] on field [name]: 1.0 Kendal Tau correlation. Rankings with 10 documents.
For query [Robinho] on field [name]: 1.0 Kendal Tau correlation. Rankings with 10 documents.
For query [defender] on field [position]: 1.0 Kendal Tau correlation. Rankings with 50 documents.
For query [brazil] on field [nationality]: 1.0 Kendal Tau correlation. Rankings with 30 documents.
For query [33] on field [number]: 1.0 Kendal Tau correlation. Rankings with 20 documents.
For query [juventus] on field [team]: 1.0 Kendal Tau correlation. Rankings with 15 documents.
For query [both] on field [foot]: 0.8666666666666667 Kendal Tau correlation. Rankings with 30 documents.
lcgm@lcgm-ss51:~/workspace/PlayersInformationRetrieval/src$
```

Teste outras consultas manualmente!

# Tempo de resposta: TF-IDF vs Presença/Ausência de termo

```
lcgm@lcgm-ss51:~/workspace/PlayersInformationRetrieval/src$ pipenv run python compute_metrics.py timing
[nltk_data] Downloading package stopwords to /home/lcgm/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
For query [neymar messi] on field [term], 10 documents with tf_idf: 8.3937ms avg | 0.14934303945729588ms standard deviation.
For query [neymar messi] on field [term], 10 documents without tf_idf: 7.09424ms avg | 0.2780783669791992ms standard deviation.
For query [cristiano fifa ronaldo] on field [term], 30 documents with tf_idf: 292.12190000000004ms avg | 17.752114984253044ms standard deviation.
For query [cristiano fifa ronaldo] on field [term], 30 documents without tf_idf: 241.34509ms avg | 17.227290827476164ms standard deviation.
For query [cristiano brazil attacker] on field [term], 30 documents with tf_idf: 810.74011ms avg | 19.11155148541575ms standard deviation.
For query [cristiano brazil attacker] on field [term], 30 documents without tf_idf: 564.1474499999999ms avg | 43.07953704659545ms standard deviation.
For query [Leonel Messi] on field [name], 10 documents with tf_idf: 4.0254ms avg | 0.07522222968234747ms standard deviation.
For query [Leonel Messi] on field [name], 10 documents without tf_idf: 3.27749ms avg | 0.05374622279422493ms standard deviation.
For query [Robinho] on field [name], 10 documents with tf_idf: 2.1549699999999996ms avg | 0.04830635715321681ms standard deviation.
For query [Robinho] on field [name], 10 documents without tf_idf: 1.97415ms avg | 0.13402806693850 63ms standard deviation.
For query [defender] on field [position], 50 documents with tf_idf: 547.77643ms avg | 385.28582996749634ms standard deviation.
For query [defender] on field [position], 50 documents without tf_idf: 307.42664ms avg | 142.2597518609028ms standard deviation.
For query [brazil] on field [nationality], 30 documents with tf_idf: 298.58296ms avg | 2.263045252876233ms standard deviation.
For query [brazil] on field [nationality], 30 documents without tf_idf: 259.80054ms avg | 0.6438458536267228ms standard deviation.
For query [33] on field [number], 20 documents with tf_idf: 481.91171999999995ms avg | 1.7506856554625745ms standard deviation.
For query [33] on field [number], 20 documents without tf_idf: 419.42172999999997ms avg | 4.219561765399364ms standard deviation.
For query [juventus] on field [team], 15 documents with tf_idf: 22.30632ms avg | 0.14450353041718497ms standard deviation.
For query [juventus] on field [team], 15 documents without tf_idf: 19.86114ms avg | 0.08654910674748688ms standard deviation.
For query [both] on field [foot], 30 documents with tf_idf: 159.45255ms avg | 1.8996737804744848ms standard deviation.
For query [both] on field [foot], 30 documents without tf_idf: 141.47129ms avg | 0.6876957909636287ms standard deviation.
lcgm@lcgm-ss51:~/workspace/PlayersInformationRetrieval/src$ gpl
```

# Teste outras consultas manualmente!

# Composição de resposta

- Backend
  - Ranking dos documentos como entrada
  - Recupera as informações para os documentos do ranking
  - Constrói o HTML no backend com as informações recuperadas
  - Retorna para o front o HTML já pronto

- Frontend
  - Angular 6
  - Interface para o usuário realizar a consulta
  - Envia a requisição com os parâmetros para o backend
  - Recebe o retorno e exibe para o usuário

# Football Player Search

**Generic search**

**Name**

**Position**

**Nationality**

**Number**

**Team**

**Kicking Leg**

**Max. number of documents**

10

**TF-IDF on vector model**

Submit

back

TF-IDF on vector model

# Results

Response time: 0.301 seconds.

Neymar da Silva Santos Junior
https://us.soccerway.com
| **Name:** Neymar da Silva Santos Junior | **Nationality:** Brazil | **Position:** Attacker |

Neymar
https://www.uefa.com
| **Name:** Neymar | **Nationality:** BRA | **Position:** Forward | **Team:** Paris Saint-Germain |

Neymar
https://www.eurosport.com
| **Name:** Neymar | **Nationality:** Brazil | **Position:** Attack |

Jose Hernandez
https://www.mlssoccer.com
| **Name:** Jose Hernandez | **Nationality:** Venezuela | **Number:** 13 | **Position:** Defender | **Team:** Atlanta United FC |

Krisztian Nemeth
https://www.mlssoccer.com
| **Name:** Krisztian Nemeth | **Nationality:** Hungary | **Number:** 9 | **Position:** Forward | **Team:** Sporting Kansas City |

Harrison Afful
https://www.mlssoccer.com
| **Name:** Harrison Afful | **Nationality:** Ghana | **Number:** 25 | **Position:** Defender | **Team:** Columbus Crew SC |

Will Johnson
https://www.mlssoccer.com
| **Name:** Will Johnson | **Nationality:** USA | **Number:** 4 | **Position:** Midfielder | **Team:** Orlando City SC |

# THANKS!

## Any questions?

bvcl@cin.ufpe.br
lcgm@cin.ufpe.br
lss5@cin.ufpe.br