UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS
CURSO: CIÊNCIA DA COMPUTAÇÃO
DISCIPLINA: RECUPERAÇÃO DA INFORMAÇÃO

JOÃO PAULO DE ARAÚJO
MARCUS VINÍCIUS MARTINS MELO
WERMESON ROCHA DA SILVA

MODELO VETORIAL EM SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO

Trabalho apresentado à Disciplina de Recuperação da Informação, ministrada pelo Prof. Lívio Antonio Melo Freire, para obtenção parcial de créditos e título de Bacharel em Ciência da Computação.

Maio / 2018
CRATEÚS-CE

Sumário

1 INTRODUÇÃO

A eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que ele utiliza, influenciando diretamente em seu modo de operação.

Segundo o site wikipedia, o **Modelo De Espaço Vetorial**, ou simplesmente modelo vetorial, representa documentos e consultas como vetores de termos. Termos são ocorrências únicas nos documentos. Os documentos devolvidos como resultado para uma consulta são representados similarmente, ou seja, o vetor resultado para uma consulta é montado através de um cálculo de similaridade.

2 OBJETIVO DO TRABALHO

A proposta deste trabalho se objetiva em desenvolver um pequeno sistema de recuperação da informação baseado no **Modelo De Espaço Vetorial,** juntamente com uma **avaliação corpora rotulada por humanos,** utilizando-se de cálculos de **precisão** e **cobertura**. A implementação deve, ainda, deve operar com a coleção Cystic Fibrosis e, de acordo com os resultados obtidos, gerar métricas capazes de mesurar a sua eficiência.

3 METODOLOGIA

Para a elaboração deste trabalho, as atividades necessárias podem ser divididas em termos de fases temporais, baseadas nas sequências propostas na descrição do trabalho.

A divisão em fases gera as etapas: planejamento, implementação dos algoritmos sugeridos em sala de aula, responsáveis pelos cálculos das métricas necessárias para o funcionamento da implementação.

A iniciação do trabalho constituiu-se no planejamento, onde a definição e a forma de organização dos principais componentes, responsáveis pela composição da implementação, foram definidos, assim como um estudo detalhado de como programar as funções necessárias para o funcionamento da implementação e um estudo mais aprofundado da linguagem de programação **Python**.

Vencida a etapa de planejamento, a etapa seguinte foi a implementação da aplicação. Nesta etapa, os conhecimentos adqueridos em sala de aula foram de extrema importância, dado que a compreensão dos termos e métricas utilizados na implementação requeriam um bom conhecimento de suas fundamentações. Nesta etapa a estrutura principal do sistema, **o índice invertido**, foi implementada, como sugerida no material de apoio, assim como as implementações das funções necessárias para a realização dos cálculos das métricas **Frequência do Termo (*tf*)**, **Frequência Inversa (*idf*)** e **Medidas de Similaridade**, responsável pelo cálculo do grau de similaridade entre dois vetores.

Concluída a etapa de implementação da aplicação, a etapa seguinte definiu-se na elaboração do sistema de avaliação responsável por mesurar a eficiência da aplicação. Nesta etapa, de início, os aquivos da coleção Cystic Fibrosis teve seus arquivos devidamente tratados, bem como a implementação das principais funções responsáveis pelos cálculos de **precisão** e **cobertura.** Estas implementadas com e sem **stem**, como demonstradas na seção referente aos resultado, além da **Medida F (F-measure)**, todas utilizando curva de interpolação na geração de seus gráficos, bem como os cálculos responsáveis necessários para a **precisão média** e **média da precisão médi**a sendo considerados os 100 primeiros documentos retornados pela aplicação.

4 DIFICULDADES ENCONTRADAS

Durante a elaboração do trabalho, duas dificuldades demonstraram-se bastante relevantes. A primeira delas deu-se pelo fato do não conhecimentos suficiente da linguagem de programação Python que possibilitasse a implementação, sendo esta contornada por meio de estudos da documentação da linguagem.

A segunda dificuldade deu-se pelo fato de que o nível de abstração nos pseudocódigos sugeridos para elaborar as principais funções da implementação ser bastante alto e confuso. Esta dificuldade difundiu-se, causando graves falhas na implementação que influenciaram bastante, e de forma negativa, nos resultados da parte de avaliação do sistema, o que ocasionou a união dos três alunos, responsáveis pela elaboração deste trabalho, **João Paulo De Araújo, Marcus Vinícius Martins Melo e Wermeson Rocha da Silva,** que compartilharam as implementações e os conhecimentos adqueridos em sala de aula para a obtenção de êxito, como mostra a Figura 4.1**.**

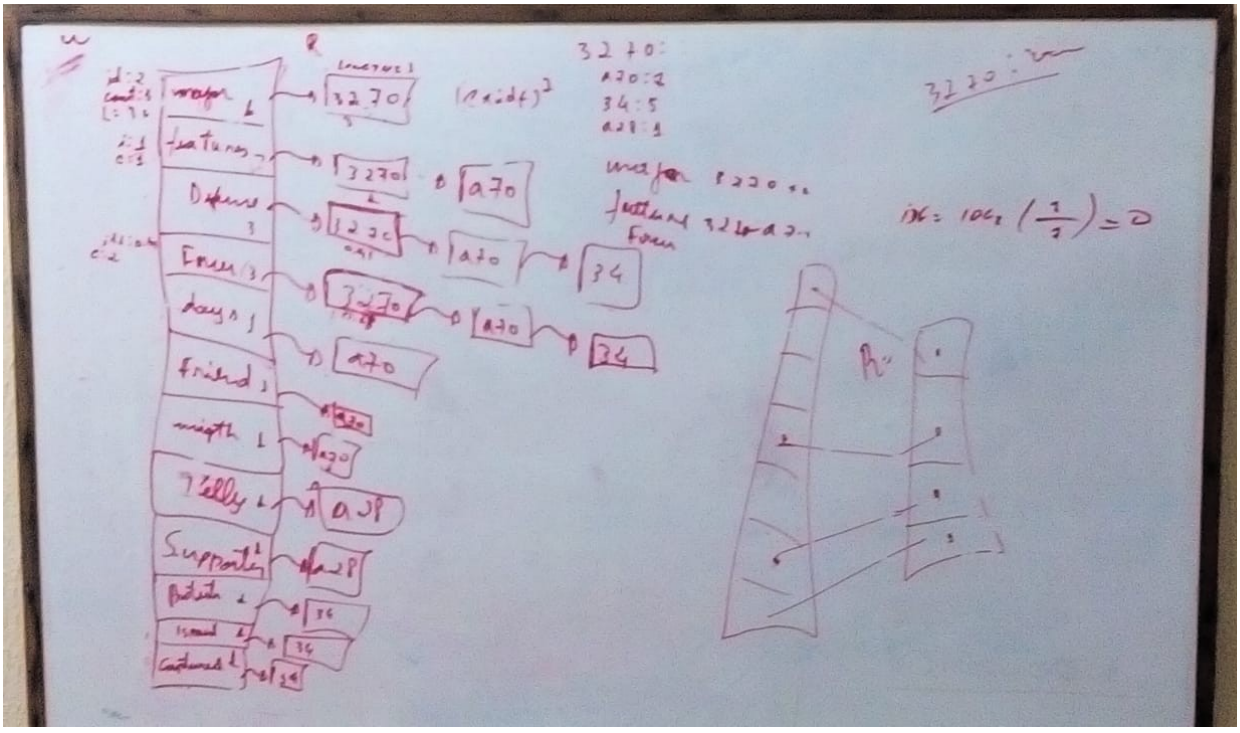


Figura 4.1 - Simulação do sistema de avaliação

5 RESULTADOS OBTIDOS

Os resultados obtidos com a implementação foram:

- As figuras abaixo mostram os resultados obtidos para a consulta ***"How are salivary glycoproteins from CF patients different from those of normal subjects?"***, o documento **633** como o de maior relevância, este retornado pela aplicação como o de maior relevância, quando realizada a respectiva consulta:

```
QN 00003
QU How are salivary glycoproteins from CF patients different from those of
   normal subjects?
NR 00043
RD   23 1000   40 0010  139 2122  190 0001  221 0001  246 0001  309 1000
    311 0011  325 0001  345 0010  347 0010  356 0010  370 1001  374 0001
    375 2221  439 0001  440 1120  454 1220  515 1000  520 0010  524 0010
    526 1010  527 1220  533 0001  535 0010  560 0010  561 0001  571 0010
    584 0010  604 0002  623 0010  633 2222  733 1000  742 1010  854 0100
    856 2112  950 0010  967 0001 1144 1111 1161 0001 1172 1000 1175 0001
   1196 0002
```

- Figura 5.1 - Arquivo referente a consulta 3 (*"How are salivary glycoproteins from CF patients different from those of normal subjects?")*.

```
do $ python3 rev.py data.pickle
insira a consulta: How are salivary glycoproteins from CF patients different fro
m those of normal subjects

1 - 633

2 - 77

3 - 1206

4 - 856

5 - 439

PAGINA 1 de 20


100 Arquivos de resultado da consulta


        1 - Mostrar as proximas 5 recuperações.
        2 - Mostrar o Mth documento recuperado.
```

Figura 5.2 – Resultados obtidos da consulta 3 (*"How are salivary glycoproteins from CF patients different from those of normal subjects?")*.

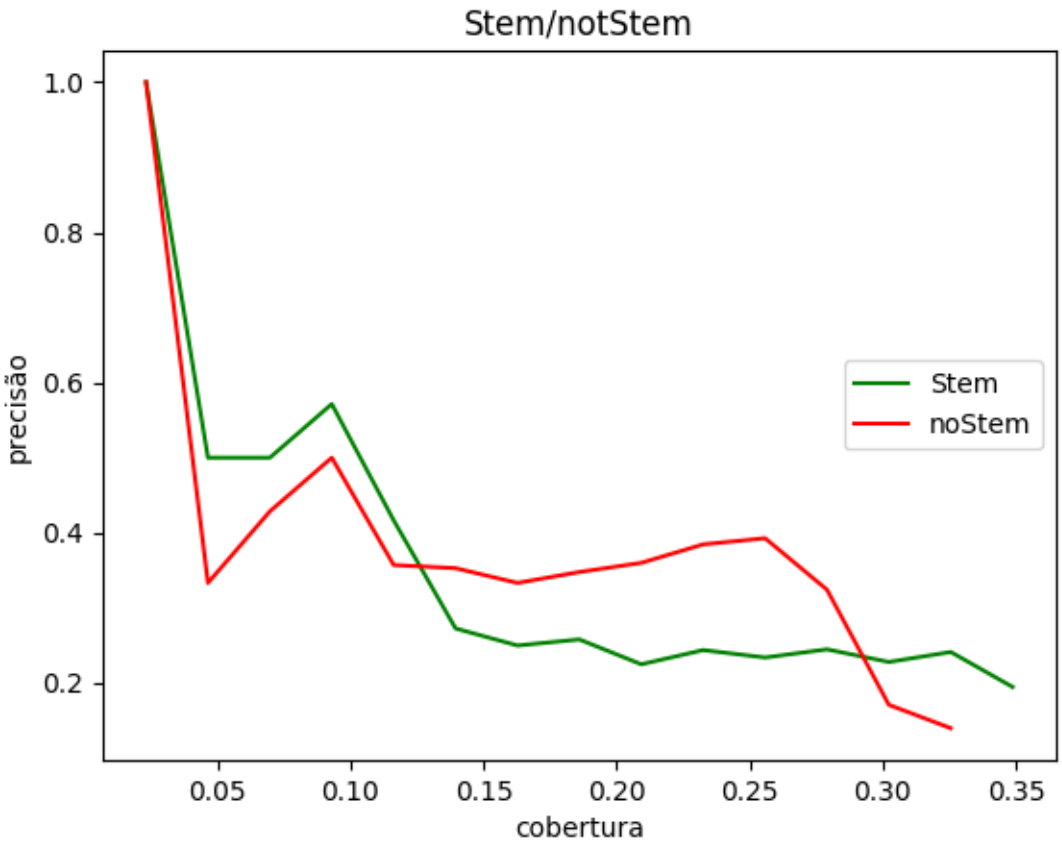


Figura 5.3 – Gráfico referente à consulta 3 (*"How are salivary glycoproteins from CF patients different from those of normal subjects?")*, com e sem ***STEM***.

- A média da **Medida F (F-measure)** obtida pela aplicação, atingiu, aproximadamente, 0.37, como mostrado nas figuras a seguir:

```
Busca: What is the pathology of the reproductive system (male or female) in    CF?
qtd origin: 44
qtd return: 99
F-Measure: 0.19928023895636174
Precisao Media: 0.18181818181818182

Busca: What structural or enzymatic differences are there between fibroblasts    from CF patients and non-CF patients?
qtd origin: 140
qtd return: 99
F-Measure: 0.3929547597054508
Precisao Media: 0.2607142857142856

Busca: What histochemical differences have been described between normal and    CF respiratory epithelia?
qtd origin: 22
qtd return: 99
F-Measure: 0.15441004584171938
Precisao Media: 0.18181818181818182

Busca: What is the best treatment for nasal polyps in CF patients?
qtd origin: 13
qtd return: 99
F-Measure: 0.4703400130212098
Precisao Media: 0.4230769230769231
media precisao media =  0.2680104246230246
media medida F = 0.365342
marcus@Marcus ~/Dropbox/Semestre 2018.1/Recuperação da informação/Indice_invertido $
```

Figura 5.4 – Resultado da **Medida F (F-measure).**

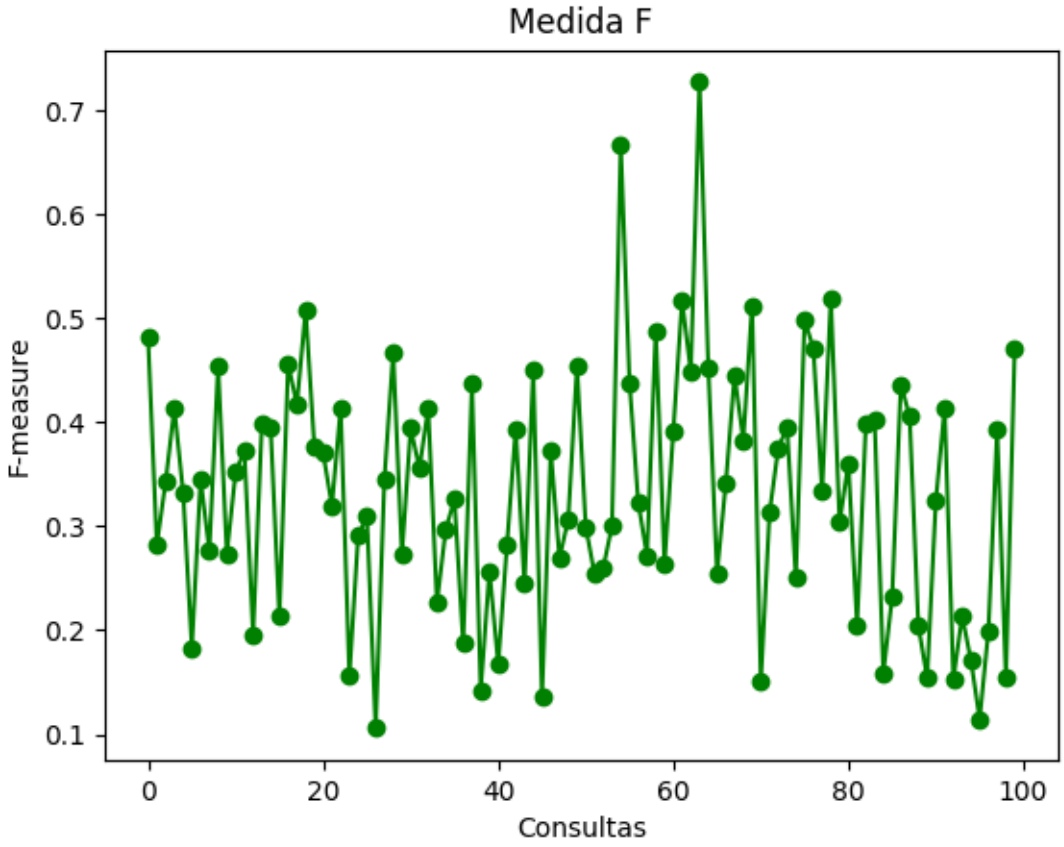
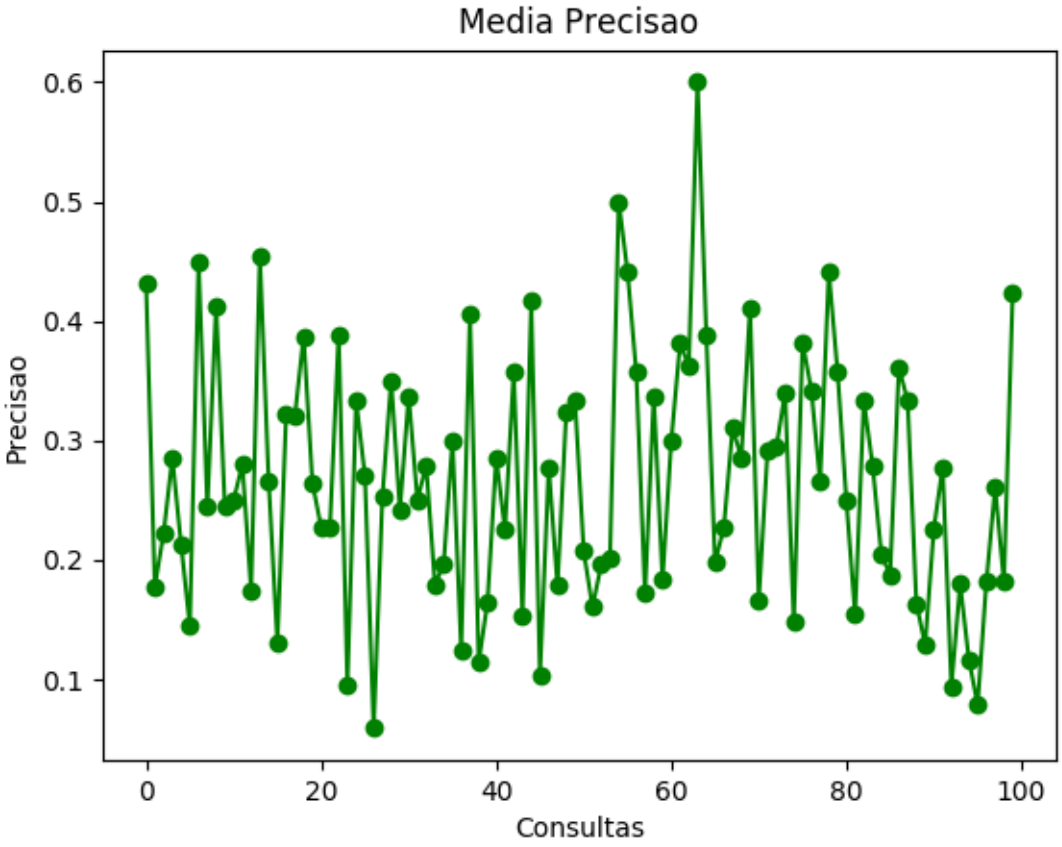


Figura 5.5 – Gráfico referente à Medida F (F-measure).

- A **média da precisão média** obtida pela aplicação, atingiu, aproximadamente, 0.27, como mostrado nas figuras a seguir:

```
PROBLEMS 7   OUTPUT   DEBUG CONSOLE   TERMINAL                    1: Python

Precisao Media: 0.07876712328767123

Busca: What is the pathology of the reproductive system (male or female) in    CF?
qtd origin: 44
qtd return: 99
F-Measure: 0.19928023895636174
Precisao Media: 0.18181818181818182

Busca: What structural or enzymatic differences are there between fibroblasts    from CF patients and non-CF patients?
qtd origin: 140
qtd return: 99
F-Measure: 0.3929547597054508
Precisao Media: 0.2607142857142856

Busca: What histochemical differences have been described between normal and    CF respiratory epithelia?
qtd origin: 22
qtd return: 99
F-Measure: 0.15441004584171938
Precisao Media: 0.18181818181818182

Busca: What is the best treatment for nasal polyps in CF patients?
qtd origin: 13
qtd return: 99
F-Measure: 0.4703400130212098
Precisao Media: 0.4230769230769231
media precisao media =  0.2680104246230246
marcus@Marcus ~/Dropbox/Semestre 2018.1/Recuperação da informação/Indice invertido $
```

Figura 5.6 – Resultado da Média da Precisão Média.



5.7 – Gráfico da Média da Precisão Média.

6 REFERÊNCIAS

- **Natural Language Toolkit;** Disponível em: https://www.nltk.org/;
- **DocumentacaoPython;** Disponível em: https://wiki.python.org.br/DocumentacaoPython;
- **How To Plot Data in Python 3 Using matplotlib;** Disponível em: https://www.digitalocean.com/community/tutorials/how-to-plot-data-in-python-3-using-matplotlib;
- **Matplotlib legends — position and arrangement;** Disponível em: https://www.codementor.io/tips/7214263758/matplotlib-legends-position-and-arrangement;