FACOM- UFU
Professor: Wendel Melo

Terceiro trabalho de Organização e Recuperação da Informação 2023-2

Descrição

Este trabalho consiste no cálculo e plotagem do gráfico e precisão e revocação média para uma coleção de referência.

Deve ser entregue apenas um **único** programa desenvolvido em Python 3 que realize a tarefa solicitada. O programa deve usar apenas as bibliotecas padrão Python 3, isto é, as bibliotecas que já vem com a instalação padrão do interpretador da linguagem. Também há permissão de uso das bibliotecas *numpy* e *matplotlib*.

O trabalho deve ser feito **individualmente** e o código gerado deve ser anexado na respectiva tarefa do *MS Teams* no prazo indicado.

Aviso importante: se for detectado cópia ou qualquer tipo de trapaça entre trabalhos, todos os envolvidos serão punidos com a nota zero. Portanto, pense bem antes de pedir para copiar o trabalho do seu coleguinha, pois ele poderá ser punido também!

Antes de começar a desenvolver, certifique-se de que você compreendeu os slides sobre avaliação da recuperação. Após estudar os slides, volte aqui e leia a descrição novamente. Para fazer esse trabalho, confira o *notebook* e/ou o video sobre a aula de plotagem de gráfio com *Matplotlib* (a tarefa remota passada na aula de 15/02/2023)

A entrada do programa

Seu programa deverá receber um arquivo de entrada (cujo nome é passado pela linha de comando) especificando respostas ideais de um sistema fictício para consultas de referência. Para compreender o arquivo de entrada, vamos tomar como exemplo o exercício dos slides de avaliação de recuperação:

Exercício: Considere uma coleção de referência. Suponha que os conjuntos R1, R2 e R3 de documentos relevantes para as consultas q1, q2 e q3, respectivamente, tenham sido determinados por um grupo de especialistas. Os conjuntos R1, R2 e R3 são dados da seguinte forma:

R1 = {d3, d7, d12, d13, d26, d68}
R2 = {d1, d2, d9, d24, d51, d52, d70, d82}
R3 = {d2, d3, d6, d16, d20}

Considere que um novo algoritmo de recuperação chamado XYZ foi recém projetado. Suponha que esse algoritmo retorne, para as consultas q1, q2 e q3, os seguintes rankings de documentos:

Consulta q1 = {d1, d9, d26, d15, d2, d10, d74, d68, d32, d3, d53, d39, d56, d11, d4}.

Consulta q2 = {d3, d7, d8, d9, d19, d16, d37, d24, d20, d80, d67, d50, d46, d51, d29}.
Consulta q3 = {d2, d30, d25, d3, d9, d7d6, d39, d75, d19, d26 d16, d20, d51, d1}.

Construa o gráfico de precisão versus revocação para cada uma das consultas e o gráfico com a média de precisão por revocação do sistema XYZ

Assim, a primeira linha do arquivo de entrada contém o número *n* de consultas de referência (no exemplo anterior, *n* = 3). As *n* linhas seguintes especificam as saídas ideais para cada uma das consultas de referencia, onde a *i*-ésima linha especifica saída ideal para a consulta *i*. A resposta ideal de cada consulta estará inteiramente contida em uma linha, com os documentos separados por espaço. A seguir, as próximas *n* linhas especificam a resposta obtida pelo sistema para cada uma das consultas de referência, onde a *i*-ésima linha especifica saída do sistema para a consulta *i*. A resposta do sistema para cada consulta estará inteiramente contida em uma linha, com os documentos separados por espaço. Para o exemplo anterior, teríamos o seguinte arquivo de entrada:

```
3
3 7 12 13 26 68
1 2 9 24 51 52 70 82
2 3 6 16 20
1 9 26 15 2 10 74 68 32 3 53 39 56 11 4
3 7 8 9 19 16 37 24 20 80 67 50 46 51 29
2 30 25 3 9 76 39 75 19 26 16 20 51 1
```

exemplo de arquivo de entrada referencia.txt

Note que, em preto, na primeira linha, temos o número de consultas de referência (3). Em azul, as respostas ideais e, finalmente, em vermelho, as respostas do sistema. O nome do arquivo de entrada deverá ser recebido pela linha de comando. Assim, supondo que o arquivo de entrada se chame *referencia.txt* e que seu programa se chame *avaliacao.py*, chamaremos seu programa fazendo:

> python3 avaliacao.py referencia.txt

A saída do programa

Seu programa deverá gerar um aquivo de saída denominado ***media.txt*** com a precisão média do sistema em cada um dos 11 níveis padrão de revocação (0%, 10%, 20%, …, 90%, 100%). Basta armazenar no arquivo de saída os 11 valores de precisão (11 números no arquivo de saída e nada mais). Por exemplo, para a seguinte tabela arbitrária de precisão por revocação:

Revocação	Precisão
0%	64%
10%	64%
20%	50%
30%	48%
40%	40%
50%	40%
60%	32%
70%	30%
80%	0%
90%	0%
100%	0%

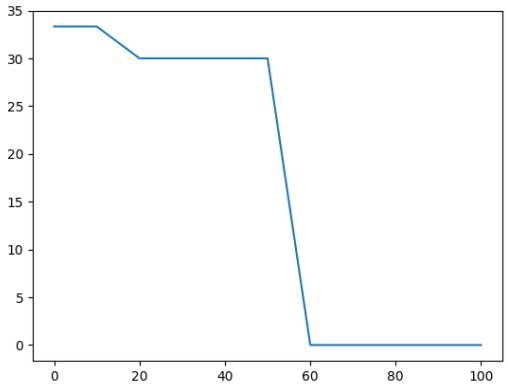
Teremos o seguinte arquivo de saída:

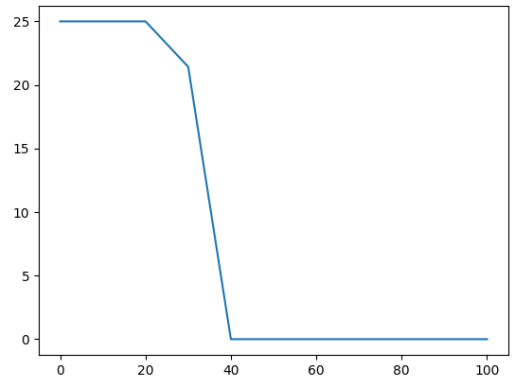
```
0.64  0.64  0.5  0.48  0.4  0.4  0.32  0.3  0  0  0
```

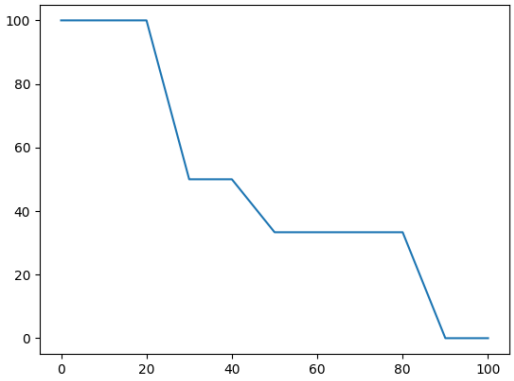
exemplo de arquivo de saída media.txt

Seu programa também deverá gerar um gráfico de precisão por revocação (nos níveis de revocação padrão) para cada uma das consultas de referência e um gráfico com a média do sistema. **Consulte o material da aula para entender como realizar os cálculos corretamente e para aprender a plotar gráficos com matplotlib**. Os gráficos podem ser plotados em tela (não é preciso salvar em arquivo). Note que a regra de interpolação deve ser usada para o cálculo das revocações padrão.

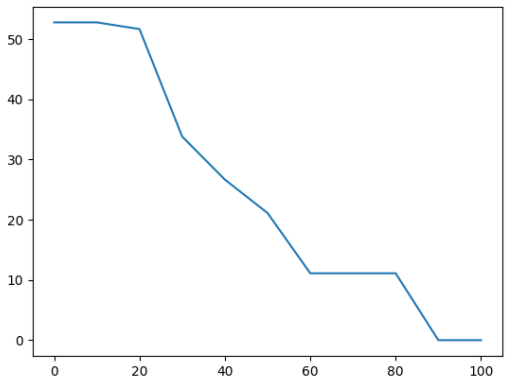
Para a entrada do exemplo aqui descrito, teríamos como saída:


35
30
25
20
15
10
5
0
0
20
40
60
80
100


25
20
15
10
5
0
0
20
40
60
80
100


100
80
60
40
20
0
0
20
40
60
80
100

Médias:


50
40
30
20
10
0
0
20
40
60
80
100