

Quarto trabalho de Organização e Recuperação da Informação 2023-1

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 gráfico com *Matplotlib* (a tarefa remota passada na aula de reposição do dia 03/11/2023)

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 R_1 , R_2 e R_3 de documentos relevantes para as consultas q_1 , q_2 e q_3 , respectivamente, tenham sido determinados por um grupo de especialistas. Os conjuntos R_1 , R_2 e R_3 são dados da seguinte forma:

$R_1 = \{d_3, d_7, d_{12}, d_{13}, d_{26}, d_{68}\}$

$R_2 = \{d_1, d_2, d_9, d_{24}, d_{51}, d_{52}, d_{70}, d_{82}\}$

$R_3 = \{d_2, d_3, d_6, d_{16}, d_{20}\}$

Considere que um novo algoritmo de recuperação chamado XYZ foi recém projetado. Suponha que esse algoritmo retorne, para as consultas q_1 , q_2 e q_3 , os seguintes rankings de documentos:

Consulta $q_1 = \{d_1, d_9, d_{26}, d_{15}, d_2, d_{10}, d_{74}, d_{68}, d_{32}, d_3, d_{53}, d_{39}, d_{56}, d_{11}, d_4\}$.

Consulta $q_2 = \{d_3, d_7, d_8, d_9, d_{19}, d_{16}, d_{37}, d_{24}, d_{20}, d_{80}, d_{67}, d_{50}, d_{46}, d_{51}, d_{29}\}$.
Consulta $q_3 = \{d_2, d_{30}, d_{25}, d_3, d_9, d_{7d6}, d_{39}, d_{75}, d_{19}, d_{26}, d_{16}, d_{20}, d_{51}, d_1\}$.

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 referência, 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 arquivo 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 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%

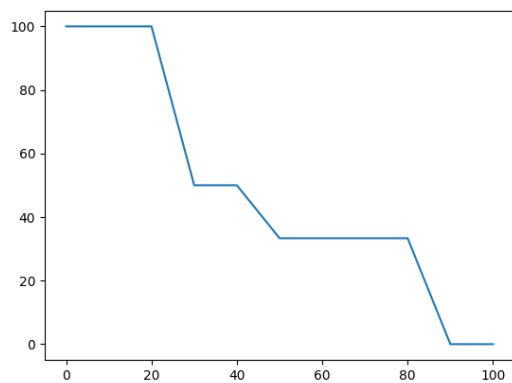
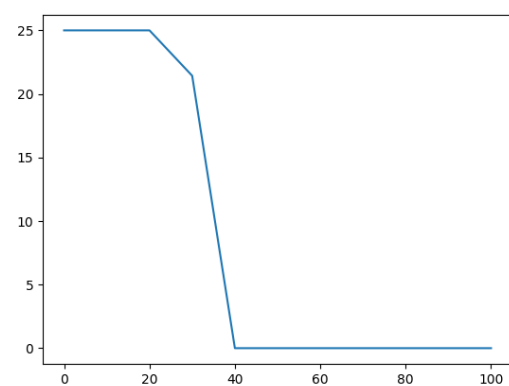
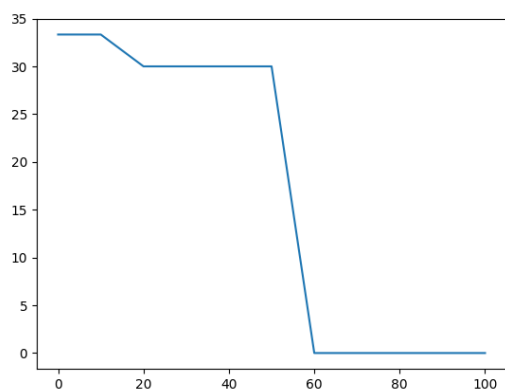
Teremos o seguinte arquivo de saída:

```
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:



Médias:

