Trabalho 2 – ENTREGA DIA 19/06/2017

Considerando os conjuntos de dados de resenhas de filmes curto.zip e longo.zip, faça o pedido. Nesse trabalho, nenhuma biblioteca de mineração de texto pode ser utilizada. Todos os cálculos devem ser implementados. Pode fazer em qualquer linguagem, vocês me apresentarão quando estiver pronto.

1 – Com o conjunto de dados curto.zip,

→ crie o código para produzir a matriz de document frequency (DF) conforme a fórmula com log:

$$idf_t = \log(\frac{N}{df_t})$$

→Em seguida, gere um arquivo pdf com o resultado da matriz DF x Termo.

→ crie o código para produzir a matriz de Term frequency (TF) conforme a fórmula com log:

$$w_{t,d} = \begin{cases} 1 + \log \mathrm{tf}_{t,d}, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

→Em seguida, insira no arquivo pdf com o resultado da matriz TF x Documento.

→ crie o código para produzir a matriz de TF*IDF

→Em seguida, insira no arquivo pdf com o resultado da matriz TF*IDF.

2- Com o conjunto de dados longo.zip,

→ Utilizando os métodos ou funções criados acima, crie um sistema para receber uma query e buscar os 10 documentos mais relevantes com relação à query. Deve aparecer o nome do documento e a similaridade com a query. Vou pedir para ver essa parte funcionando no sistema de vocês.

Ex: cv999_13106.txt - 0.71
cv999_13199.txt - 068

→ Coloque uma query (escolhida por você) no pdf e os nomes dos documentos e a similaridade.

→Por fim, coloque no pdf a precisão, abrangência e F1 do sistema.