Trabalho de COS738 - 2022/1 - Busca e Mineração de Texto

Eduardo do Valle e Herbert Salazar

Relatório do algoritmo do Trabalho

O código foi desenvolvido em Python, para executá-lo basta processar o arquivo "compara.py", os logs da execução estão situados no arquivo "compara.log" situado na pasta "log" e os arquivos contendo as descrições dos grupos de cursos e das ocupações estão situados na pasta "dados" ("dadosOcupacoes.csv" teria as diretrizes das ocupações e "lista_grupos_cursos.csv" as dos grupos de cursos), que foram utilizados para descobrir a similaridade entre essas descrições. Enquanto os resultados estão no arquivo "distanciaVetores.csv", na pasta "Resultados", no qual temos em cada linha os resultados da similaridade pelo cosseno entre o vetor tf-idf da ocupação, representado pelo número na primeira coluna, e os vetores tf-idf dos grupos de cursos, apresentado na segunda coluna como vetor de tuplas que seguem a estrutura a seguir: Três valores em cada tupla, o primeiro seria o ranking de similaridade do grupo de curso determinado pelo valor do cosseno do ângulo entre os vetores , o segundo seria a identificação de qual grupo se refere esta tupla, e o terceiro seria o valor do cosseno do ângulo entre os vetores.

Bibliotecas utilizadas:

- punctuation de string
- distance de scipy.spatial
- RSLPStemmer de nltk.stem

- nltk
- numpy
- logging
- time
- csv
- math
- sys