FATEC RUBENS LARA
CIÊNCIA DE DADOS

GABRIELLE LARA MORICZ DE SOUZA

RELATÓRIO TÉCNICO
Trabalho de álgebra Linear

SANTOS
2025

1 INTRODUÇÃO

Relatório sobre o desenvolvimento de um código que analisa a similaridade de cosseno entre respostas para duas perguntas distintas a fim de adivinhar quais respostas veem da mesma pessoa

1.1 OBJETIVO

Que o código retorne as três respostas mais similares a uma busca e o ângulo dessa similaridade de maneira que uma destas respostas seja da mesma pessoa da qual provém a frase de busca.

2 METODOLOGIA

Para o desenvolvimento do código foi utilizado o ambiente de programação em linguagem python *Spyder* e as bibliotecas *nltk* e *math,* a primeira destas tendo sido utilizado somente para o teste de função. Foi feita a escolha de não remover stopwords para a análise principal uma vez que a frequência com que as pessoas utilizam este tipo de palavra implica sobre seus padrões de linguagem, que é o fator sendo utilizado para relacionar as diferentes respostas dadas por um mesmo individuo, uma vez que a temática é igual para todos.

2.1.1 CÓDIGO

O código pode ser divido entre quatro partes distintas. A leitura inicial dos dados contidos nos arquivos de texto acompanhantes e a criação de um *wordbank* contendo todas as palavras dos ‘documentos’. O processo de vetorização tanto dos termos de busca, que podem ser livres caso o usuário decida, tanto quanto de todos os ‘documentos’, os cálculos feitos a partir destes vetores, retornando o ângulo, e a escolha final dos três ângulos mais relevantes.

2.1.2 ARQUIVO SEPARADO DE FUNÇÕES

Com o intuito de deixar o código reutilizável para outros propósitos foi feito um arquivo separado com suas funcionalidades separadas em funções. Foi adicionada uma função não presente no código original feita para limpar conjuntos de dados massivos. Para o teste destas funções foi criado um arquivo de input livre.

2.2 DADOS

Os dados utilizados para o trabalho são respostas para as perguntas “por que você escolheu este curso” e “qual está sendo sua matéria favorita” dadas por alunos do segundo ciclo de ciência de dados. Foram obtidas dez respostas no total, no entanto três eram curtas demais e não possuíam palavras em comum entre as duas respostas, então foram descartadas. Todas as respostas utilizadas foram recebidas por áudio, colocando mais em evidência os padrões linguísticos de cada individuo por evitar a filtragem que ocorre quando se escreve um texto, e transcritas manualmente. Para popular os testes foram escritas cinco entradas de respostas fictícias escritas de maneira que uma resposta referencie a outra

3 RESULTADOS

Ao rodar o código com as respostas para a segunda pergunta como busca, Obtém-se uma resposta correta para 9 dos 12 casos de estudo. Com o vetor mais similar sendo o correto em 8 desses 9 casos, e o segundo mais similar no outro.

Uma busca feita com a segunda resposta escrita pela autora: *“a minha matéria favorita no momento entre álgebra e estatística eu gosto dos dois eu gostava bastante de algoritmos também quando a gente tava tendo uh mas eu gosto muito de quebrar fórmulas matemáticas o que é tipo o que eu mais ando me divertindo com depois das aulas de estatística eu acho legal então e álgebra é algo que eu não tinha visto muito sobre antes pra falar a verdade mas que eu eu realmente tô gostando pra caramba”* retorna:

NOME	RESPOSTA	ÂNGULO
EU	eu decidi fazer ciência de dados por que eu sempre gostei muito de estatística acho muito legal principalmente estatística probabilística mas não tinha uh perto então eu acabei decidindo fazer ciência de dados que era o mais próximo que tinha e eu tô gostando pra caramba do curso eu tô muito feliz com a escolha por que eu acho legal	44.67

ELISÂNGELA	eu escolhi esse curso por que quando eu fiz técnico de desenvolvimento de sistemas eu me interessei mais pela parte de banco de dados então e eu tenho uma visão mais analítica eu sempre gostei dessa parte assim de analisar as coisas de observar organizar então achei que era uma área que entraria mais pra mim do que é necessáriamente desenvolver sofware entendeu eu sei que tem programação no curso mas eu queria programar assim não é muito meu forte	51.29
PROGRAMADOR	eu sempre adorei programar principalmente algoritmos eu acho o desenvolvimento que essa área teve nos últimos anos simplesmente incrível então eu decidi que isso era o que eu realmente queria fazer com a minha vida	54.82

4 Considerações finais

Perguntas mais imediatamente similares teriam sido preferíveis, uma vez que muitos dos respondentes deram respostas dissimilares entre si e, no entanto, muito similares às outras respostas na segunda questão em especifico, também teria sido preferível utilizar mais dados para ter uma ciência melhor do funcionamento do código.