Artigo sobre o Projeto Semestral - Segundo Semestre - ECM514 - Ciência dos Dados

Alunos:
Marcel Marques Caceres - RA 17.00648-0
Kaique de Andrade Almeida - RA 17.01113-2
Filipe dos Santos Pugliesi - RA 18.02608-7
Johannes Mattheus Krouwel - RA 20.01248-9

Link para o vídeo da apresentação disponível no YouTube:
https://youtu.be/oPtiWn8m9Jo

O projeto do nosso grupo utiliza a API da Wikipédia para extrair dados de biografias de pessoas consideradas as maiores personalidades de seus países por votações em concursos televisivos.

O objetivo é realizar o processamento e análise de textos extraídos das biografias disponíveis na Wikipédia, estabelecendo associações entre as pessoas, suas áreas de atuação e criando modelos preditivos baseados nesses dados.

O projeto utiliza a API do Dropbox para salvar os arquivos contendo os textos extraídos das páginas da Wikipédia. Antes de realizar o processamento, há uma etapa de teste para verificar se a conexão com o Dropbox está funcionando corretamente.

Uma função automatiza a extração do conteúdo das páginas da Wikipédia. A extração é feita com base em uma lista de títulos de páginas previamente definida, que inclui personalidades de diversas nacionalidades, como britânicos, alemães, franceses, americanos, argentinos, portugueses, italianos e brasileiros. O conteúdo extraído é salvo no Dropbox para uso posterior.

Após a extração, os textos são carregados para o ambiente de análise, onde: são exibidos exemplos, como o início e o final de alguns arquivos, é feita a correspondência entre os títulos das páginas e os textos extraídos e dicionários são criados para relacionar as personalidades às suas respectivas áreas de atuação.

Com base nos dicionários gerados, o sistema cria rótulos para cada personalidade. Esses rótulos são usados em modelos preditivos para associar novos textos às áreas correspondentes.

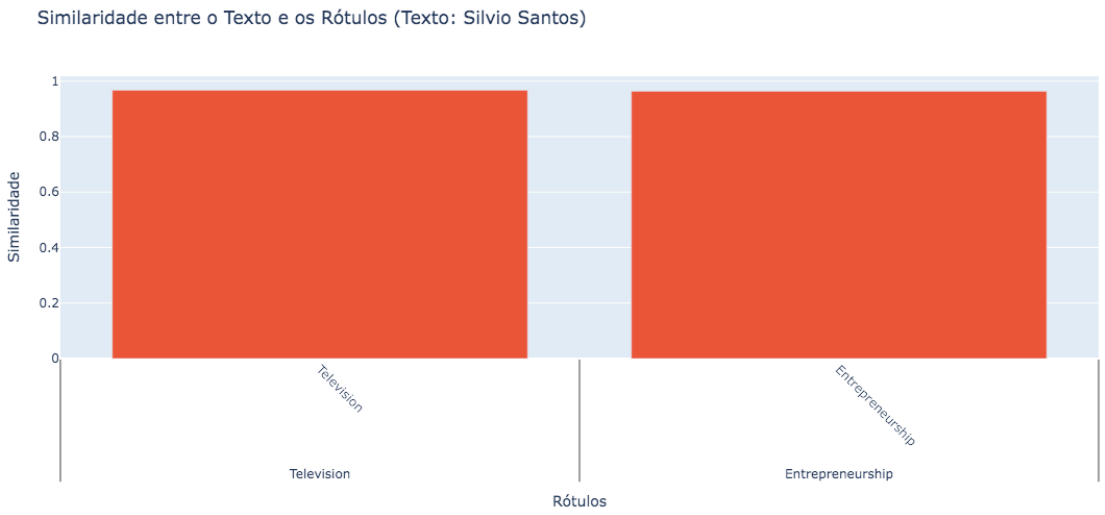
O projeto emprega modelos de aprendizado supervisionado, que são treinados para reconhecer padrões nos textos. Os parâmetros do modelo podem ser salvos para

serem reutilizados em análises futuras. Há também um modelo de aprendizado não supervisionado incluído para demonstração, embora não seja o foco principal.

Uma funcionalidade permite ao usuário inserir o título de uma página da Wikipédia. O sistema analisa o texto associado, testa a similaridade com os padrões aprendidos e retorna o rótulo mais provável. Caso o texto seja de Silvio Santos, por exemplo, o sistema pode identificar que ele pertence às áreas de "televisão" e "empreendedorismo", com alta similaridade.

O projeto também inclui uma ferramenta para gerar gráficos que mostram as similaridades entre textos e rótulos, auxiliando na interpretação dos resultados.

Portanto, podemos concluir que, este trabalho combina o uso de APIs, processamento de textos e aprendizado de máquina para criar um sistema que analisa e categoriza biografias de forma automática. Ele demonstra o potencial de integrar ferramentas e técnicas de ciência de dados para explorar informações de fontes amplas, como a Wikipédia.



Exemplo de gráfico que pode ser gerado pelo programa.