Ciência da Computação / Sistemas de Informação

Coleta de Dados para Mineração de Texto

Thiago Salles Santos - 6º módulo de Ciência da Computação, UFLA, bolsista PIBIC/FAPEMIG

Arthur Silveira Franco - 6º módulo de Ciência da Computação, UFLA, bolsista PIBIC/CNPq

Priscilla de Souza Silva - Coorientadora, Pós-Graduanda em Ciência da Computação, UFLA

Mozar José de Brito - Professor do Departamento de Gestão Agroindustrial, UFLA - mozarjdb@ufla.br Coorientador

Denilson Alves Pereira - Professor do Departamento de Ciência da Computação, UFLA - denilsonpereira@ufla.br Orientador - Orientador(a)

Resumo

Dados podem ser coletados automaticamente da Web e de redes sociais, e são muito úteis para diversas aplicações, como por exemplo, monitoramento de preço e análise de sentimentos. O objetivo deste trabalho é coletar dados de páginas Web e redes sociais acerca do assunto cachaça, para a realização das tarefas de reconhecimento das entidades nomeadas (da sigla em inglês NER) e de análise de sentimentos. Para a realização do projeto, foi feito um levantamento de páginas Web relacionadas à cachaça, como por exemplo, sites de venda e de avaliação, e outro levantamento de quais dados deveriam ser extraídos das páginas, como por exemplo, nome, preço, volume e descrição da cachaça. Posteriormente, foi realizado um estudo de técnicas e ferramentas de extração de dados Web, ficando decidido o uso da linguagem de programação Python, juntamente com o uso da biblioteca Requests para páginas Web estáticas e para a comunicação com API GraphAPI, da biblioteca Selenium para páginas Web dinâmicas, da Beautifulsoup para tratamento do HTML das páginas Web, da Doccano como ferramenta de rotulação dos dados para a tarefa de NER, além do uso de outras ferramentas e bibliotecas. Para a realização da tarefa de NER, foram coletados os dados das páginas Web respeitando as regras de extração de cada site. Após a coleta de dados, foi iniciada a rotulação manual das entidades nomeadas. Foram selecionadas mil instâncias dos dados extraídos para serem rotulados dentre 17 categorias de entidades, sendo 11 no domínio específico da bebida e 6 categorias genéricas. Ao término da rotulação, os membros participantes do projeto se reuniram para entrar em consenso sobre os dados rotulados, e assim estabelecer um dataset resultante da rotulação manual. Posteriormente, esse dataset foi avaliado experimentalmente usando um modelo de linguagem baseado em redes neurais. Já para a tarefa de análise de sentimentos, os dados extraídos são pertencentes às páginas da rede social Facebook, os quais foram coletados por meio da API GraphAPI fornecida pela Meta. Ao término da coleta, se obteve vários datasets separados por páginas da rede social Facebook. Portanto, conclui-se que a tarefa de produção de datasets é muito importante para comunidade acadêmica, por gerar os recursos que serão utilizados para futuros modelos de inteligência artificial. Como trabalho futuro, pretende-se utilizar os datasets resultantes da tarefa de análise de sentimentos em um modelo de rede neural.

Palavras-Chave: Coleta de dados, Dataset, Rotulação.
Instituição de Fomento: Universidade Federal de Lavras

Link do pitch: https://youtu.be/ajLyK5hP6nk