Ciência da Computação / Sistemas de Informação

Criação de Bases de Dados para as Tarefas de Reconhecimento de Entidades Nomeadas e de Mineração de Opinião

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

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

Priscilla de Souza Silva - Pós-Graduanda do Departamento de Ciência da Computação, UFLA.

Mozar José de Brito - Coorientador DGA, UFLA.

Denilson Alves Perreira - Orientador DCC, UFLA. - Orientador(a)

Resumo

A coleta de dados da Web, ou Web Scraping, é uma tarefa utilizada para extração de dados de páginas Web e sua conversão para um formato estruturado para posterior análise. Consiste no download de uma página e navegação em seu código HTML para identificar os dados de interesse. Outro tipo de coleta de dados é aquela feita a partir de redes sociais, como Twitter, por meio de suas API\\\'s. A coleta tem várias utilidade, tais como monitorar o preço de produtos, coletar notícias de jornais, entre outros. Neste trabalho, foram coletados dados sobre a cachaça, para a criação de bases de dados para avaliação de projetos sobre reconhecimento de entidades nomeadas (NER) e mineração de opinião. Para a execução do projeto, foi realizado um levantamento de 24 sites de vendas e suas páginas no Twitter. Para a coleta e extração dos dados dos sítes, foram desenvolvidos scrips em Python utilizando as bibliotecas BeautifulSoup e Selenium. Para a coleta das publicações no Twitter, foi utilizada a sua API. Foram coletados, estruturados e armazenados 20.248 postagens no Twitter, para a tarefa de mineração de opinião, e dados de 3.404 páginas de vendas, para a tarefa de NER. Em seguida, foi feita a preparação dos dados, removendo atributos irrelevantes para geração das bases de dados, como espaçamento dos textos, e dividindo os textos em sentenças. Por fim, foram selecionados 1.000 documentos para serem rotulados manualmente para a tarefa de NER. A rotulação foi feita manualmente por três alunos, seguindo uma diretriz elaborada para guiar o processo e utilizando a ferramenta de rotulação Doccano. Em seguida, os resultados foram comparados automaticamente, aquelas entidades em que pelo menos dois rotuladores concordaram foram para o dataset final sem necessidade de uma revisão manual, como foi feita com 403 entidades em que houve discordância. O grau de concordância geral entre as rotulações, obtido através do coeficiente Fleiss' Kappa foi de 0.857, a qual é considerada quase perfeita. Como resultado, o dataset nomeado CachacaNER está disponível em formato IOB2, com 183.019 tokens rotulados manualmente e com resultados experimentais de 92,79% na métrica F1, utilizando o modelo de linguagem pré-treinado BERTimbau. Concluí-se que a adição de mais um dataset para a tarefa NER contribui com a comunidade acadêmica em avaliações experimentais de modelos de linguagem na língua portuguesa. Como trabalho futuro, recomenda-se a mineração de opinião dos dados coletados do Twitter.

Palavras-Chave: Coleta de dados, Dataset, NER.
Instituição de Fomento: CNPq; FAPEMIG

Link do pitch: https://youtu.be/2toOT3Cz88Q