Relatório final – mac0214 – atividade curricular em cultura e extensão

Eduardo Ribeiro Silva de Oliveira

05 de dezembro de 2024

Dados do aluno

Nome: Eduardo Ribeiro Silva de Oliveira
Endereço de email: eduardo.rso784@usp.br
NUSP: 11796920

Título do projeto

Construção de um dataframe de topônimos coletados da internet

Orientador

Nome: Patricia de Jesus Carvalhinhos
Email: patricia.carv@usp.br
Endereço profissional: Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (FFLCH-USP)

Resumo do projeto

Introdução: O projeto visa a construção de um dataframe contendo topônimos, ou seja, nomes próprios de lugares, coletados de diferentes fontes na internet. A motivação central é a organização da variedade e frequência de topônimos em diferentes contextos, que pode ser utilizada para estudos linguísticos, históricos e socioculturais.
Objetivos:

- Coletar dados de topônimos de diversas plataformas online, como sites de órgãos governamentais.
- Organizar os dados em um dataframe estruturado, facilitando a consulta e o acesso às informações coletadas.

Metodologia

O projeto foi desenvolvido em várias etapas:

1. **Coleta de dados:** A coleta de topônimos foi realizada em múltiplas etapas, utilizando ferramentas de scraping para extrair informações de fontes específicas, como o site da Câmara Municipal de São Paulo, Assembleia Legislativa, Senado Federal e Câmara dos Deputados. Foram utilizados termos-chaves relevantes, como "denomina", "altera", "acrescenta" e "substitui" para capturar os dados necessários.

2. **Construção do dataframe:** Os dados coletados foram organizados em um dataframe utilizando Python e bibliotecas específicas para tratamento de dados, como pandas. O dataframe foi estruturado para permitir fácil consulta e acesso posterior.

Cronograma

Agosto (20 horas)

- Pesquisa e definição de fontes de dados.
- Revisão de materiais e preparação inicial do conteúdo.

Setembro (50 horas)

- Desenvolvimento de scripts de coleta de dados.
- Implementação de scrapers para as principais fontes identificadas.

Outubro (25 horas)

- Coleta de dados e construção da primeira versão do dataframe.
- Ajustes nos scrapers para otimizar a coleta.

Novembro (55 horas)

- Revisão e otimização do dataframe.
- Conclusão e redação do relatório final do projeto.

Resultados obtidos

Os scrapers desenvolvidos para a Assembleia Legislativa de São Paulo (Alesp), Câmara Municipal e Prefeitura foram um dos principais resultados do projeto,

funcionando de forma eficiente e atendendo às expectativas. Além de realizar o scraping dos textos das páginas, o código implementado também foi capaz de realizar a lematização do texto extraído, enriquecendo o processo de organização dos dados.

Uma limitação observada para a construção de um dataframe ainda maior foi o armazenamento físico. Como meu notebook possui espaço limitado, os scripts foram executados continuamente por 7 dias, gerando uma base de dados com aproximadamente 30 GB. Esse tamanho ilustra a potencial riqueza e abrangência dos dados coletados, mas também evidencia a necessidade de recursos de armazenamento mais robustos para projetos futuros.

Os dataframes gerados estão disponíveis para consulta e download no Google Drive, no seguinte link: `https://drive.google.com/drive/folders/1CnxiH5gyZ1ElE2w-Oj_TxGRpkJWtuUym?usp=drive_link`.

O código completo e detalhado utilizado neste projeto pode ser acessado no repositório público do GitHub: `https://github.com/EduardoRSO/Toponimia`.

Relatório final

O projeto de construção de um dataframe de topônimos coletados da internet foi extremamente enriquecedor, proporcionando tanto o desenvolvimento de habilidades técnicas quanto a aplicação de conceitos de linguística e automação. O objetivo principal de criar um material estruturado e acessível para estudos sobre topônimos foi cumprido, com a produção de um dataframe robusto que facilita consultas diversas.

Durante o desenvolvimento, foram criados scrapers para diferentes fontes de dados, como órgãos governamentais, visando coletar topônimos de forma automatizada. Além disso, foram utilizadas técnicas de limpeza e organização dos dados para garantir a qualidade do dataframe final.

As atividades resultaram em um total de 150 horas dedicadas ao projeto, contemplando todas as fases de produção. A experiência proporcionou um aprendizado significativo na área de coleta de dados automatizada e no tratamento de dados para fins de organização. Os resultados obtidos incluem um material completo e de qualidade, que pode ser utilizado por pesquisadores interessados na organização e consulta de topônimos, contribuindo para sua formação acadêmica e desenvolvimento profissional.