RELATÓRIO FINAL – MAC0215 – Atividade Curricular em Cultura e Extensão

Eduardo Ribeiro Silva de Oliveira

07 de Outubro de 2024

Dados do Aluno

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

Título do Projeto

Classificação Automatizada de Topônimos: Abordagens Semânticas e Morfológicas

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

Este projeto tem como objetivo desenvolver um sistema de classificação automatizada de topônimos utilizando abordagens semânticas e morfológicas. A classificação será baseada na semelhança entre palavras, com foco na análise da presença de sufixos e no uso de sinônimos. A introdução do projeto abordará a importância de métodos eficazes para a categorização de topônimos em estudos linguísticos e geográficos. Os objetivos principais incluem a implementação de algoritmos que avaliem a semelhança lexical e a criação de regras específicas para a detecção e categorização de sufixos e sinônimos, visando melhorar a precisão na classificação dos topônimos.

Metodologia

O projeto foi realizado em várias etapas, conforme descrito abaixo:

1. **Revisão Bibliográfica**: Pesquisa de estudos e metodologias existentes na classificação de topônimos, com foco em abordagens semânticas e morfológicas. Foram utilizados artigos acadêmicos e livros especializados para estabelecer uma base teórica sólida.

2. **Coleta e Preparação de Dados**: Reunião de um conjunto de dados de topônimos de fontes diversas, como documentos governamentais e redes sociais, seguido pela limpeza e preparação dos dados para garantir a qualidade do input para o modelo.

3. **Desenvolvimento do Modelo**: Implementação de algoritmos de classificação baseados em semelhança lexical, identificação de sufixos e sinônimos. Utilizamos bibliotecas como spaCy, Word2Vec e WordNet para a construção do modelo.

4. **Treinamento do Modelo**: Aplicação dos algoritmos aos dados coletados e treinamento do modelo para avaliar a precisão e eficácia da classificação, utilizando técnicas de embeddings semânticos e regras específicas de categorização.

5. **Documentação e Apresentação**: Documentação detalhada de todas as etapas do projeto, além da preparação para a apresentação final dos resultados.

Cronograma

Agosto (25 horas)

Revisão Bibliográfica

Setembro (30 horas)

Definição do Escopo do Projeto
    Coleta de Dados

Outubro (40 horas)

Preparação de Dados
    Desenvolvimento do Modelo (primeira parte)

Novembro (45 horas)

Desenvolvimento do Modelo (continuação)
    Treinamento do Modelo
    Documentação Parcial

Dezembro (10 horas)

Finalização da Documentação

Relatório Final

O projeto de classificação automatizada de topônimos foi extremamente enriquecedor, proporcionando tanto o desenvolvimento de habilidades técnicas em processamento de linguagem natural quanto o entendimento mais profundo sobre a categorização de topônimos. A aplicação de abordagens semânticas e morfológicas possibilitou a criação de um sistema eficaz de classificação, que se mostrou relevante para estudos linguísticos e geográficos.

Durante a execução do projeto, realizamos uma revisão bibliográfica para entender as metodologias existentes e definir as melhores abordagens a serem aplicadas. Em seguida, foi realizada a coleta de dados de diferentes fontes, seguida pela preparação dos mesmos, o que envolveu técnicas de limpeza e normalização. Na fase de desenvolvimento, utilizamos ferramentas como spaCy, Word2Vec e WordNet para implementar o modelo de classificação.

O treinamento do modelo foi uma etapa crucial, onde aplicamos os algoritmos desenvolvidos para categorizar os topônimos com base na semelhança lexical e na presença de sufixos e sinônimos. A documentação foi realizada utilizando a metodologia de auto-documentação proposta por Valdemar W. Setzer, garantindo que cada etapa do desenvolvimento fosse bem descrita e pudesse ser facilmente consultada.

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 processamento de linguagem natural e análise linguística, e os resultados obtidos incluem um sistema de classificação automatizado que pode ser utilizado por pesquisadores interessados na análise de topônimos, contribuindo para sua formação acadêmica e desenvolvimento profissional.