Relatório final – MAC0215 – atividade curricular em pesquisa

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

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

Não houve modelo de linguagem natural, apenas utilizamos a seguinte abordagem que suporta uma parametrização:

- Verificar se a palavra possui um sufixo toponímico;
- Verificar se existem palavras similares no texto;
- Verificar se existe similaridade com os sinônimos;
- Verificar se existe similaridade com os topônimos do IBGE;
- Verificar se algum dos similares possui sufixo toponímico;
- Garantir que a pontuação esteja entre 0 e 1.

Assim, retornamos uma probabilidade da palavra ser um topônimo ou não.

A dificuldade para construir o modelo de linguagem natural é a base de treinamento: seriam necessários textos e os topônimos contidos no texto. No entanto, não encontrei essa base de dados pronta na internet, e construir uma excederia o escopo de 100 horas do projeto.

Para fazer o cálculo da confiança dos topônimos, utilizei uma base de dados com textos extraídos da Alesp, Prefeitura e a Câmara de Deputados. Os detalhes da construção desse dataframe constam no relatório de outro projeto desenvolvido neste semestre.

Código desenvolvido

Optei por uma modularização seguindo a estrutura abaixo:

- `classification_workflow_logic`: Coordena os demais módulos;
- `suffix_logic`: Verifica se uma palavra possui sufixos toponímicos;
- `synonym_logic`: Recupera os sinônimos de uma dada palavra;
- `lexical_similarity_logic`: Verifica a similaridade entre palavras;
- `ibge_toponyms_logic`: Base de dados com topônimos do IBGE (cidades e municípios).

O projeto

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.

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.

O código desenvolvido e os detalhes adicionais do projeto estão disponíveis no repositório público do GitHub: `https://github.com/EduardoRSO/Toponimia`.

Resultados obtidos

A classificação utilizando uma abordagem semântica e morfológica revelou-se limitada para a definição precisa de um topônimo em um texto. A principal dificuldade encontrada foi a subjetividade inerente à toponímia, que varia conforme o contexto linguístico, histórico e cultural de cada palavra ou expressão.

Embora o sistema desenvolvido tenha conseguido identificar padrões morfológicos, como sufixos toponímicos, e relacioná-los semanticamente a outras palavras no texto, os resultados não foram suficientemente robustos para categorizar de forma confiável uma palavra como topônimo ou não. Essa limitação está diretamente associada à complexidade e à subjetividade do campo de estudo, onde fatores contextuais e interpretações subjetivas desempenham um papel crucial.

Apesar disso, o projeto proporcionou insights importantes sobre o uso de técnicas de processamento de linguagem natural aplicadas à análise de topônimos, destacando a necessidade de bases de dados mais abrangentes e específicas, bem como a importância de incorporar conhecimentos linguísticos mais profundos para melhorar a precisão e a eficácia do sistema.