Processamento de Linguagem Natural

Trabalho Final

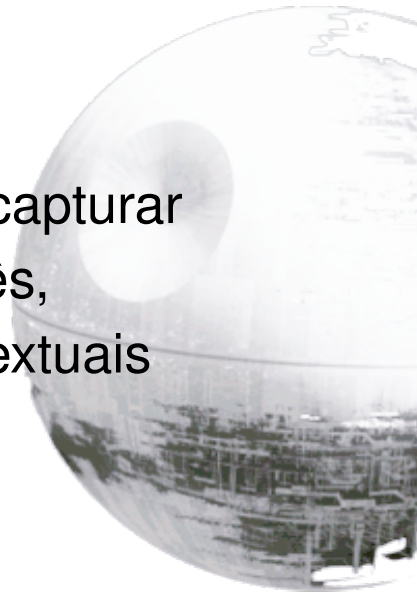
Alex Marino

14 de novembro de 2024

Conteúdo

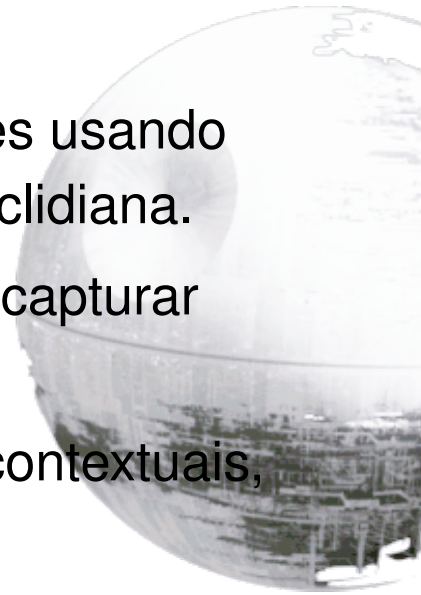
Trabalho Final de NLP: Análise Semântica de Embeddings em Português

Objetivo Geral:
Explorar e avaliar diferentes modelos de embeddings para capturar similaridade semântica entre palavras e frases em português, utilizando embeddings clássicos (Word2Vec, GloVe) e contextuais (BERT, LLM).



Objetivos Específicos

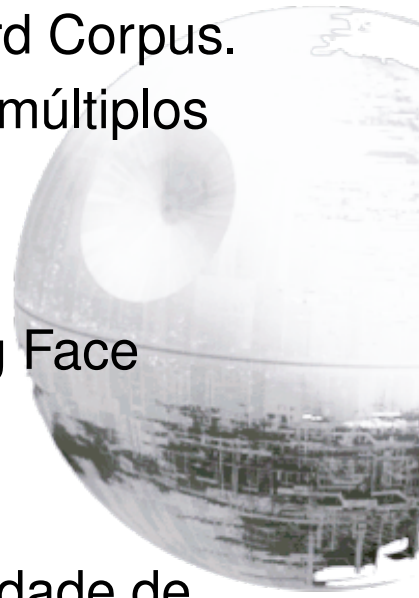
- Implementar modelos Word2Vec, GloVe e BERT para gerar embeddings de palavras e frases em português.
- Calcular a similaridade entre pares de palavras e frases usando métricas de similaridade como cosseno e distância euclidiana.
- Avaliar a capacidade dos embeddings contextuais em capturar múltiplos significados e contextos específicos.
- Comparar os resultados dos embeddings clássicos e contextuais, destacando suas limitações e pontos fortes.



Metodologia

Etapas:

1. **Coleta e Pré-processamento de Dados:**
   - Usar o Corpus Brasileiro ou o Portuguese Billion Word Corpus.
   - Realizar tokenização e preparar frases de teste com múltiplos contextos.
2. **Implementação dos Modelos de Embeddings:**
   - Word2Vec e GloVe com Gensim; BERT com Hugging Face Transformers.
3. **Cálculo de Similaridade Semântica:**
   - Comparar pares de palavras e frases usando similaridade de cosseno e distância euclidiana.



Metodologia (continuação)

4. **Análise Comparativa:**
   - Avaliar precisão dos embeddings em capturar múltiplos significados e contextos. - Comparar desempenho dos modelos clássicos e contextuais.
5. **Relatório e Discussão dos Resultados:**
   - Documentar as limitações dos modelos e os ganhos com embeddings contextuais. - Sugerir aplicações práticas baseadas nos resultados obtidos.

Entregáveis

- Código comentado com as implementações dos modelos e cálculo de similaridade.
- Relatório detalhado com gráficos, tabelas e análise qualitativa dos resultados.
- Apresentação dos principais resultados e discussões.

Dataset Sugerido

- **Corpus Brasileiro:** Disponível no site do NILC-USP.
- **Portuguese Billion Word Corpus:** Disponível no Google Drive do projeto.

Observação: Ambos os datasets são apropriados para NLP em português, fornecendo uma variedade de contextos para análise semântica.

MAY THE

SOURCE

BE WITH

YOU