Plano de Trabalho - Projeto Final IA376E/1s2020

Busca de documentos usando vetores densos

Rafael Gonçalves (186062)
Thomas Portugal (187646)

Resumo

Assim como o uso de *word embeddings* (MIKOLOV et al., 2013) possibilitou a incorporação de relações semânticas no tratamento de palavras, o projeto visa verificar a viabilidade da utilização de vetores densos, produzidos por modelos baseados em transformers – como o BERT (DEVLIN et al., 2018) –, para a tarefa de busca de documentos. A ideia é que vetores produzidos por um modelo neural de linguagem poderiam capturar dimensões semânticas em expressões de busca – evitando o problema de uma busca não incluir documentos com palavras similares, mas não iguais, às da busca –, o que normalmente foge do escopo dos mecanismos mais tradicionais.

1 Objetivos

A partir de uma literatura relevante estudada no curso, verificar empiricamente quais contribuições o uso de vetores densos (*embeddings*) pode trazer para o problema de busca de documentos (*document retrieval*).

2 Análise Bibliográfica

Em (NOGUEIRA et al., 2019a) os autores propõem a abordagem Doc2Query, visando ampliar documentos com termos que sejam representativos do conteúdo destes. Com modelos neurais sequence to sequence, a proposta é criar possíveis perguntas que seriam respondidas por aquele documento. O artigo compara diversas técnicas, tanto tradicionais, quanto utilizando redes neurais, para avaliar a eficácia da abordagem.

Já em (CHANG et al., 2020), foram comparadas duas estratégias baseadas em transformer para o problema de *document retrieval*. Ambas se basearam no pre-treinamento de transformers BERT para este problema em específico.

Nossa abordagem será utilizar a estratégia contida em (NOGUEIRA et al., 2019a) para transformar documentos em querys, e então utilizar uma das abordagens contida em (CHANG et al., 2020) – especificamente o uso de dois transformers (um para processar *queries* e outro para processar os documentos (ou no nosso caso, *queries* gerados de documentos) e posterior avaliação da similaridade usando produto interno – para resolver o problema de *document retrieval*.

3 Datasets a serem utilizados

O dataset utilizado será o MsMARCO disponível em: https://microsoft.github.io/msmarco/. Este dataset é composto por perguntas (queries) retiradas do banco de dados do buscador BING. Junto dessas perguntas, há 10 trechos contidos nas páginas retornadas pelo buscador. Os trechos, por sua vez, possuem uma classificação binária que indica se neles há, ou não, a resposta para a pergunta. Além desses trechos, ainda há uma resposta redigida por um ser humano com base nos documentos disponíveis, o tipo de resposta e a URL referentes aos documentos.

4 Metodologia

O desenvolvimento do projeto final se dará nas seguintes etapas:

A. Levantamento e leitura da bibliografia, escolha de um conjunto de dados adequado para o projeto e redação deste plano de trabalho.

B. Familiarização e adequação do dataset ao formato exigido pelo problema (e pelos modelos).

C. Implementação dos modelos de busca baseados em transformer (BERT).

D. Redação de um relatório final que visa sintetizar os resultados obtidos na execução do projeto.

5 Cronograma

etapa / semana	1	2	3	4	5
A	X				
B		X	X		
C			X	X	X
D					X

Referências

CHANG, W.-C. et al. *Pre-training Tasks for Embedding-based Large-scale Retrieval*. 2020. arXiv: 2002.03932 [cs.LG].

DEVLIN, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805 [cs.CL].

MIKOLOV, T. et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

NOGUEIRA, R.; CHO, K. *Passage Re-ranking with BERT*. 2019. arXiv: 1901.04085 [cs.IR].

NOGUEIRA, R. et al. *Document Expansion by Query Prediction*. 2019a. arXiv: 1904.08375 [cs.IR].

NOGUEIRA, R. et al. *Multi-Stage Document Ranking with BERT*. 2019b. arXiv: 1910.14424 [cs.IR].