Recuperação de Artigos a Partir de uma Modelagem de Perguntas e Respostas

Thiago Coelho

Motivação

- Há uma base de dados com vários trechos de artigos
- Possuo um trecho de um artigo
- Como recuperar outros trechos do mesmo artigo?

Introdução

- Foram escolhidos artigos da Wikipedia em inglês
- Cada trecho corresponde a um parágrafo
- O primeiro parágrafo de cada artigo foi usado como consulta para recuperar outros trechos
- A ideia para o artigo foi baseada em outro artigo
    - *Sanity Check: A Strong Alignment and Information Retrieval Baseline for Question Answering.*

Sanity Check

- Utiliza dados de fóruns de perguntas e respostas, como Yahoo! Respostas
- A partir de uma pergunta, tenta ranquear as melhores respostas
- Propõe um método para escolher as respostas que mais estão em contexto com a pergunta

Modelagem

- Foi feita uma modelagem para utilizar a ideia do método do *Sanity Check*.
- Cada parágrafo utilizado como consulta foi considerado como uma pergunta
- Todos os demais parágrafos são respostas a todas as perguntas
- Esses parágrafos serão chamados de **pergunta** e **resposta**

GloVe: Global Vectors for Word Representation

- Algoritmo de aprendizado de máquina
- Representa palavras como vetores
- Foi utilizada uma base pré-treinada para a implementação do método proposto
  - Vetores de 300 dimensões
  - *Uncased*
  - Cerca de 1,9 milhão de termos

O Algoritmo

- Hiperparâmetros
- Pré-processamento
- Inverse Document Frequency (idf)
- Alinhamento das respostas
- Ranqueamento

Hiperparâmetros

- São definidos apenas três hiperparâmetros
  - K+, K- e λ

- Os valores usados foram, respectivamente, 5, 2, e 0.5.

Pré-processamento

- Feito utilizando expressão regular e a biblioteca do NLTK no python
- Remoção de *stopword*
- Remoção de todos os caracteres que não fossem letras
  - Números, parênteses, sinais de pontuação, etc
- Todas as palavras foram colocadas em caixa baixa
- Tokenização

Inverse Document Frequency

$$idf(q_i) = \log \frac{N - docfreq(q_i) + 0.5}{docfreq(q_i) + 0.5}$$

Em que N é o número de perguntas e $docfreq(q_i)$ é o número de perguntas que contêm q_i.

Alinhamento das Respostas

- Alinhamento de uma resposta com um termo da pergunta
- Uma resposta consiste de vários termos
- Alinhamento positivo (pos) e negativo (neg)

$$align(q_i, A) = pos(q_i, A) + \lambda \cdot neg(q_i, A)$$

Alinhamento Positivo

- Calculado com os K+ termos da resposta mais semelhantes a q_i
- Similaridade do cosseno dos vetores
- A intenção é encontrar respostas que possuam palavras mais em contexto com a pergunta

$$pos(q_i, A) = \sum_{k=1}^{K^+} \frac{1}{k} \cdot a^+_{q_i,k}$$

Alinhamento Negativo

- Calculado com os K- termos da resposta menos semelhantes a q_i
- A intenção é penalizar respostas que possuam palavras muito fora de contexto com a pergunta
- Este valor ainda será ponderado com o hiperparâmetro λ no cálculo do alinhamento geral

$$neg(q_i, A) = \sum_{k=1}^{K^-} \frac{1}{k} \cdot a^-_{q_i,k}$$

Ranqueamento

- As respostas foram ranqueadas baseadas em uma pontuação
- A pontuação considera o *idf* e o alinhamento de todos os termos de uma pergunta com cada resposta

$$s(Q, A) = \sum_{i=1}^{T_Q} idf(q_i) \cdot align(q_i, A)$$

Complexidade

- $align(q_i, A) \rightarrow O(T_A \cdot \log T_A)$
- $s(Q, A) \rightarrow O(T_Q \cdot T_A \cdot \log T_A)$
- Ranqueamento para uma pergunta $\rightarrow O(N_A \cdot T_Q \cdot T_A \cdot \log T_A)$

Desempenho

- Está em desvantagem quando comparado com a proposta do *Sanity Check* em questão de desempenho
- O valor de N_A tende a crescer muito mais do que na ideia original
- Pouco performático para uma base de dados com muitas respostas

Resultados

- Foram utilizadas 21 perguntas com aproximadamente 6.6 respostas relevantes para cada

- *Mean Average Precision*
  - 87,38% retornando 139 (todas) respostas por pergunta
  - 87,18% retornando 24 respostas por pergunta
  - 85,79% retornando 15 respostas por pergunta

- É possível retornar apenas 15 respostas por pergunta com quase nenhuma perda de qualidade

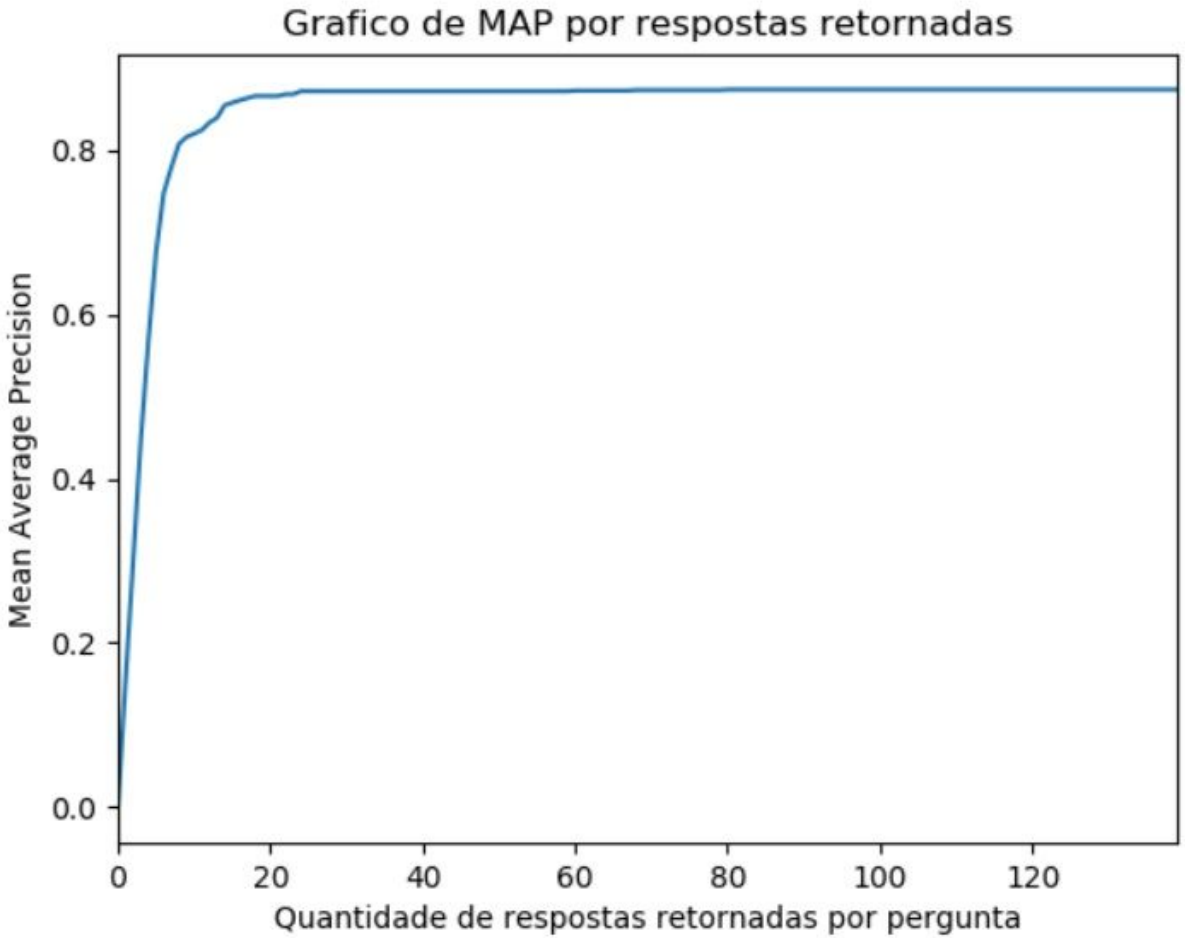
Grafico de MAP por respostas retornadas
Mean Average Precision
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
120
Quantidade de respostas retornadas por pergunta

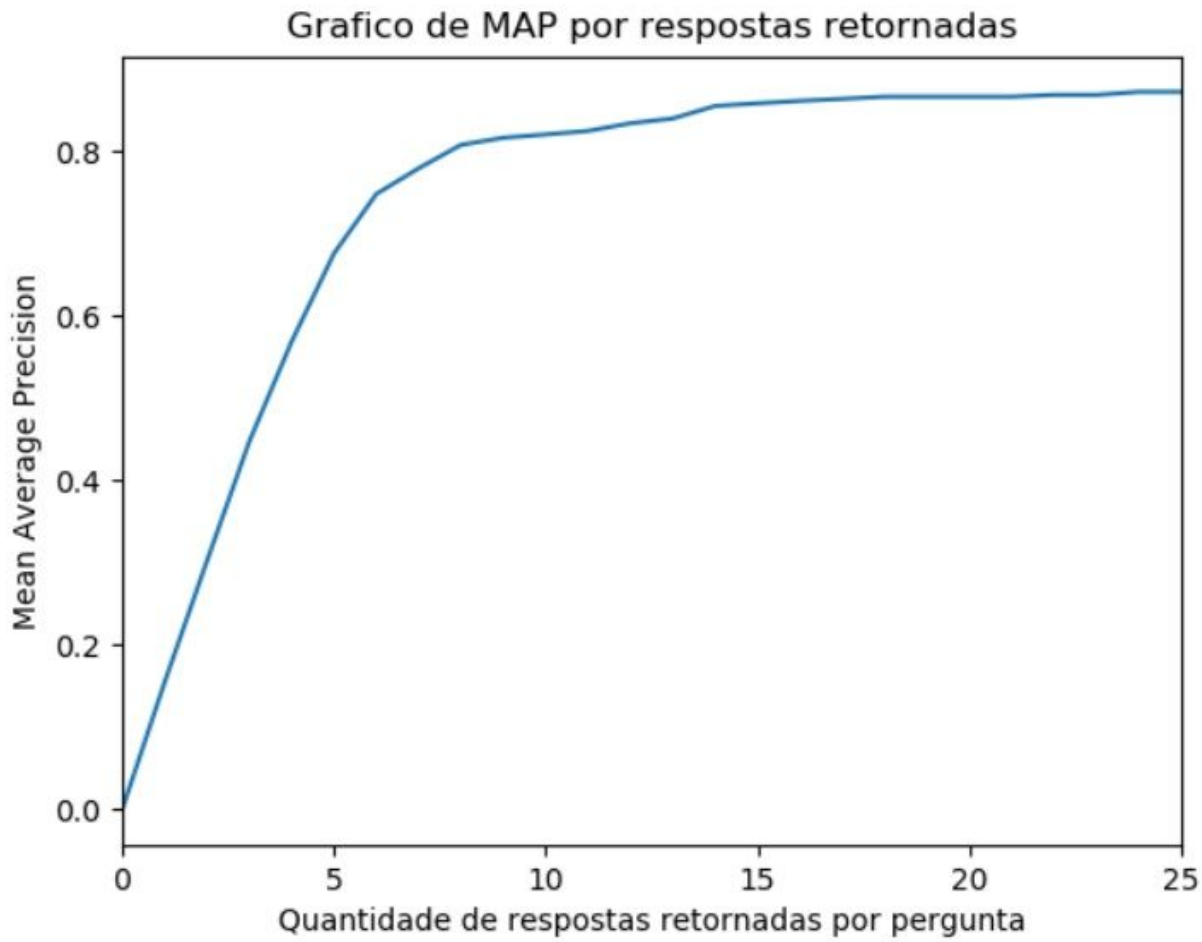
Grafico de MAP por respostas retornadas
Mean Average Precision
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
Quantidade de respostas retornadas por pergunta

Conclusão

- Os resultados demonstraram boa qualidade de recuperação dos dados

- Entretanto, o desempenho pode se tornar problemático em situações com uma grande quantidade de respostas possíveis

Referências

- *Sanity Check: A Strong Alignment and Information Retrieval Baseline for Question Answering*
- *Glove: Global vectors for word representation*
- https://nlp.stanford.edu/projects/glove/
- https://www.nltk.org

Dúvidas?