Approach: Machine Learning

The greatest evaluation forum on Question Classification

Instituto Superior Técnico
Língua Natural, Mini-Projeto 1

João Parreiro, 89483 / João Vieira, 90739

Outubro de 2020

1 Descrição do Modelo Utilizado

A abordagem usada para a classificação de questões foi a utilização de count vectorizer e cosine similarity.
O count vectorizer permite transformar um conjunto de documentos (questões, no contexto deste projeto), numa representação esparsa do número de vezes que cada token aparece em cada documento.
Foi utilizada a implementação do Count Vectorizer da biblioteca scikit-learn, que oferece uma forma simples de realizar a operação acima descrita.

A cosine similarity consiste numa medida de semelhança entre dois vetores não nulos (os vetores criados através do count vectorizer, neste caso). O seu valor é o do cosseno do ângulo definido pelos dois vetores.

$$\text{similarity} = \cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\,\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Portanto, dois vetores com a mesma orientação têm uma cosine similarity de 1, e dois vetores perpendiculares têm uma cosine similarity de 0.
Para classificar cada questão do conjunto de teste, é determinada a questão do conjunto de treino mais semelhante com esta, é obtida categoria correspondente, e esta é atribuida à questão do conjunto de teste.
Foi utilizada a implementação do Cosine Similarity da biblioteca scikit-learn, que oferece uma forma simples de realizar a operação acima descrita.

Não foi efetuado um pré-processamento de texto extensivo, uma vez que as funções das bibliotecas utilizadas já realizam muito desse trabalho. O pré-processamento realizado consistiu em reduzir todas as palavras de cada documento ao seu radical (*stemming*), e substituir todas as expressões entre aspas pela expressão `QUOTED_EXPRESSION`.

2 Accuracy

Com o modelo acima descrito, utilizando o conjunto de desenvolvimento, foi obtido o seguinte resultado:

- **Coarse accuracy:** 70.66%
- **Fine accuracy:** 62.57%

3 Análise de Erros

O método utlizado falha quando há questões que fogem ao padrão das outras da mesma categoria, ao não utilizar as palavras/expressões mais frequentes na mesma.
O resultado acima referido poderia talvez também ter sido melhor se tivessem sido utilizados outros procedimentos no pré-processamento das questões.

4 Bibliografia

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
https://en.wikipedia.org/wiki/Cosine_similarity