NLP_HW1

NTU b10705013 陳彥廷

Questions

Q1: Which embedding model do you use? What are the pre-processing steps? What are the hyperparameter settings?

A1:

我使用的模型是 word2vec。

在 preprocess 時，我使用了 nltk 的 stopword 來移除沒有辨識度的字，並且使用了 porter stemmer 來還原詞性。

訓練時，我設定的參數是 200 維的 vector、windows size 設定為 5，並移除出現 1 次以下的罕見字，訓練的 epoch 則為 5。

Q2: What is the performance for different categories or sub-categories?

A2:

Category: Semantic, Accuracy: 73.0634795354606%

Category: Syntatic, Accuracy: 60.590163934426236%

Sub-Category: capital-common-countries, Accuracy: 91.10671936758892%

Sub-Category: capital-world, Accuracy: 84.9027409372237%

Sub-Category: currency, Accuracy: 16.05080831408776%

Sub-Category: city-in-state, Accuracy: 67.04499391974058%

Sub-Category: family, Accuracy: 76.08695652173914%

Sub-Category: gram1-adjective-to-adverb, Accuracy: 22.883064516129032%

Sub-Category: gram2-opposite, Accuracy: 22.906403940886698%

Sub-Category: gram3-comparative, Accuracy: 82.05705705705707%

Sub-Category: gram4-superlative, Accuracy: 53.5650623885918%

Sub-Category: gram5-present-participle, Accuracy: 53.78787878787878%

Sub-Category: gram6-nationality-adjective, Accuracy: 86.67917448405254%

Sub-Category: gram7-past-tense, Accuracy: 61.21794871794872%

Sub-Category: gram8-plural, Accuracy: 72.67267267267268%

Sub-Category: gram9-plural-verbs, Accuracy: 55.632183908045974% Q3: What do you believe is the primary factor causing the accuracy differences for your approach?

A3:

我認為主要影響結果的是 vector size、windows size、min count 以及 preprocess 的處理。

vector	windows	min count	preprocess	accuracy
150	5	1	--	70/57
150	5	1	stopword	71/59
150	5	1	stopword + stem	72/59
200	5	1	stopword + stem	73/61
200	5	0	stopword + stem	73/60
200	3	1	stopword + stem	71/62
200	7	1	stopword + stem	70/62

Q4: What’s your discovery from your t-SNE visualization plots?

A4:

同一類型的詞會在靠近的地方，如 (he、she) 跟 (her、his) 而且類似的對比會呈現平行的樣子，證明向量可以說明詞的關聯程度。

Q5: What’s the difference in word representations if you increase the amount of training data?

A5:

增加訓練資料有助於了解低頻詞的表現，同時也會更準確地得到詞的關聯性，有助於準確率的提升。

Settings

All code run in win11、CPU i5-12400、Python 3.10.11

References

TODO#3、TODO#6、TODO#7 are written by Chatgpt.