自然語言處理 HW2

組員:

110590450 歐佳昀(70%) => 作業架構、流程、分析、改善

110590452 莊于潔(30%) => 作業模型挑選、參數調整

環境: 使用 Jupyter Notebook 語言: python

- Results
- Word embeddings
- Word similarity
 - Correlation on WordSim353
- Analogy prediction
 - Accuracy for BATS categories

Word embeddings

```
word embeding:

• wrod2vec

• https://github.com/mmihaltz/word2vec-GoogleNews-vectors/blob/master/README.md

• https://drive.google.com/file/d/087XkCwpl5KDYNINUTTISS21pQmM/edit?usp=sharing

• GLoVe

• https://github.com/stanfordnlp/GloVe

• https://huggingface.co/stanfordnlp/glove/resolve/main/glove 68.zip

• SVD

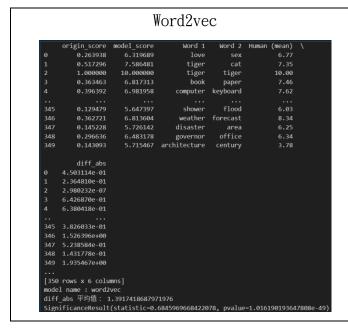
• https://github.com/valentinp72/svd2vec

• wget http://mattmahoney.net/dc/text8.zip -0 text8.gz

gzip -d text8.gz -f
```

Word similarity

WCorrelation on WordSim353



GLoVe origin score Word 2 Human (mean) diff abs 6.77 0.374149 7.35 0.785546 10.00 0.000000 6.564454 10.000000 tiger 7.46 0.188920 7.62 0.412694 0.454216 0.441461 keyboard 7.207306 computer .. 348 349 0.072899 flood 6.03 0.665506 5.364494 shower 350 351 0.261253 0.404522 6.306264 7.022609 6.25 0.056264 6.34 0.682609 disaster area 0.420720 7.103602 architecture 3.78 3.323602 [353 rows x 6 columns] model name : GLoVe diff_abs 平均值: 1.4636849824143916 SignificanceResult(statistic=0.5987723194963509, pvalue=1.0213953289911825e-35)

Analogy prediction Accuracy for BATS categories

```
GLoVe
                                                                                                                                                                                                                                                                                                            Word2vec
   data/BATS\1_Inflectional_morphology : category [ noun - plural_reg ]
total accuracy total : 35 ps : 35 lm : 35 data : 50
                                                                                                                                                                                                                                  data/BATS\1_Inflectional_morphology : category [ noun - plural_reg ] total accuracy total : 32 ps : 35 lm : 32 data : 50
    total :0.7
                                                                                                                                                                                                                                   total :0.64
    Stemming :0.7
                                                                                                                                                                                                                                   Stemming :0.7
                                                                                                                                                                                                                                   lemmatization : 0.64
    lemmatization : 0.7
   verb_ving ved verb_ving ved
                                                                                                                                                                                                                                        \label{lambdata/BATS_1_Inflectional_morphology: category [verb_ving - ved ] total accuracy total : 38 ps : 43 lm : 38 data : 50 }
   data/BATS\1_Inflectional_morphology : category [ verb_ving - ved ] total accuracy total : 38 ps : 41 lm : 38 data : 50
   total :0.76
                                                                                                                                                                                                                                        Stemming :0.86
   Stemming :0.82
                                                                                                                                                                                                                                        lemmatization : 0.76
   lemmatization : 0.76
                country capital country capital
                                                                                                                                                                                                                             data/BATS\1_Inflectional_morphology : category [ verb_3psg - ved ]
               data/BATS\3_Encyclopedic_semantics : category [ country - capital ] total accuracy total : 46 ps : 43 lm : 46 data : 50
                                                                                                                                                                                                                             total accuracy total : 27 ps : 30 lm : 27 data : 50
                                                                                                                                                                                                                             total :0.54
                                                                                                                                                                                                                             Stemming :0.6
                 Stemming :0.86
                                                                                                                                                                                                                              lemmatization : 0.54
                lemmatization : 0.92
*SVD - word2vec
bata/BATS\1_Inflectional_morphology : category [ noun - plural_reg ]
total accuracy total : 11 ps : 11 lm : 11 data : 50
 total :0.22
 Stemming :0.22
 lemmatization : 0.22
   \label{lem:data/BATS_label} $$  \data/BATS_1_Inflectional\_morphology: category [ noun - plural\_irreg ] $$  \data : 50 $$  \data: 50 $$  \dat
   total :0.14
   Stemming :0.14
   lemmatization : 0.0
     data/BATS\1_Inflectional_morphology : category [ verb_inf - 3psg ]
total accuracy total : 3 ps : 7 lm : 3 data : 50
      total :0.06
     Stemming :0.14
      lemmatization : 0.06
```