自然語言處理 HW2

組員 ：

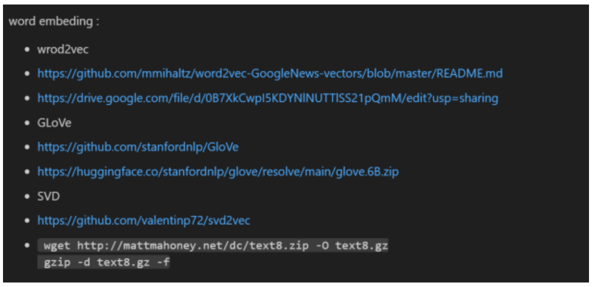
110590450 歐佳昀(70%) => 作業架構、流程、分析、改善

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



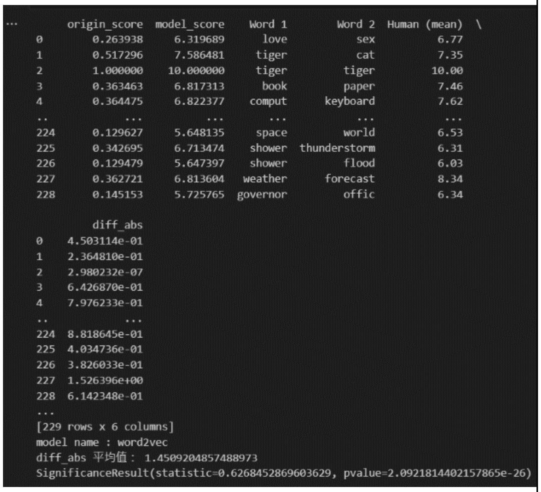
word embeding :

- wrod2vec
- https://github.com/mmihaltz/word2vec-GoogleNews-vectors/blob/master/README.md
- https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing
- GLoVe
- https://github.com/stanfordnlp/GloVe
- https://huggingface.co/stanfordnlp/glove/resolve/main/glove.6B.zip
- SVD
- https://github.com/valentinp72/svd2vec
- `wget http://mattmahoney.net/dc/text8.zip -O text8.gz`
  `gzip -d text8.gz -f`

Word similarity

WCorrelation on WordSim353

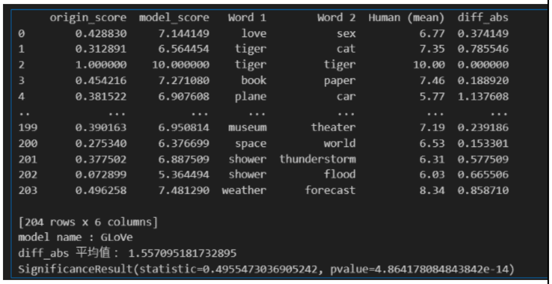
Word2vec



```
...      origin_score  model_score    Word 1        Word 2  Human (mean)  \
    0        0.263938     6.319689      love           sex          6.77
    1        0.517296     7.586481     tiger           cat          7.35
    2        1.000000    10.000000     tiger         tiger         10.00
    3        0.363463     6.817313      book         paper          7.46
    4        0.364475     6.822377    comput      keyboard          7.62
    ..            ...          ...       ...           ...           ...
    224      0.129627     5.648135     space         world          6.53
    225      0.342695     6.713474    shower  thunderstorm          6.31
    226      0.129479     5.647397    shower         flood          6.03
    227      0.362721     6.813604   weather      forecast          8.34
    228      0.145153     5.725765  governor         offic          6.34

             diff_abs
    0    4.503114e-01
    1    2.364810e-01
    2    2.980232e-07
    3    6.426870e-01
    4    7.976233e-01
    ..            ...
    224  8.818645e-01
    225  4.034736e-01
    226  3.826033e-01
    227  1.526396e+00
    228  6.142348e-01
    ...
    [229 rows x 6 columns]
    model name : word2vec
    diff_abs 平均值： 1.4509204857488973
    SignificanceResult(statistic=0.6268452869603629, pvalue=2.0921814402157865e-26)
```

GLoVe

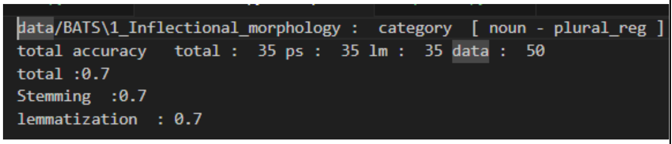
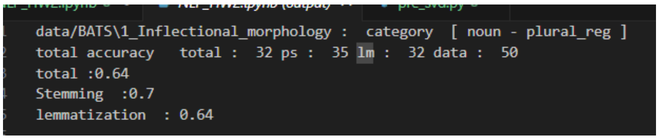
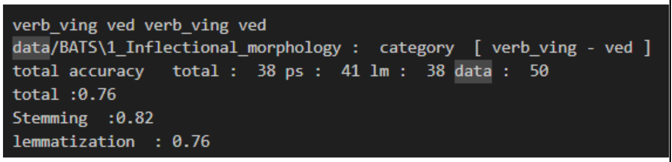
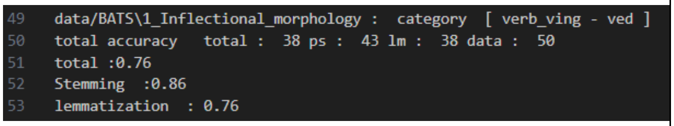
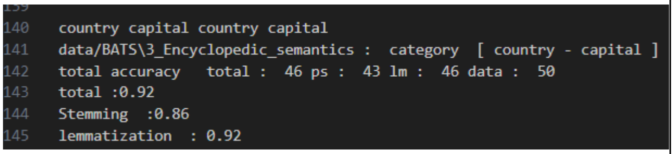
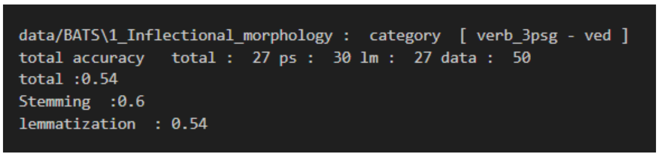
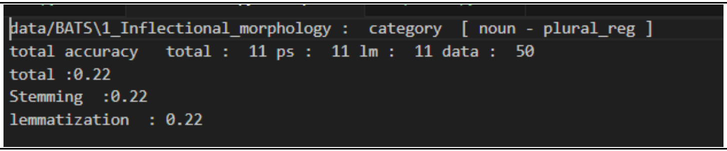
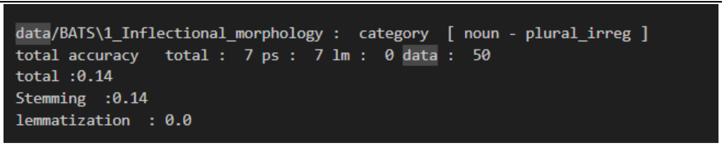
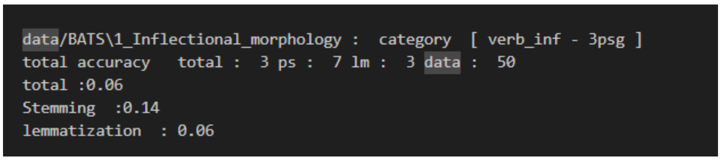


```
     origin_score  model_score   Word 1        Word 2  Human (mean)  diff_abs
0        0.428830     7.144149     love           sex          6.77  0.374149
1        0.312891     6.564454    tiger           cat          7.35  0.785546
2        1.000000    10.000000    tiger         tiger         10.00  0.000000
3        0.454216     7.271080     book         paper          7.46  0.188920
4        0.381522     6.907608    plane           car          5.77  1.137608
..            ...          ...      ...           ...           ...       ...
199      0.390163     6.950814   museum       theater          7.19  0.239186
200      0.275340     6.376699    space         world          6.53  0.153301
201      0.377502     6.887509   shower  thunderstorm          6.31  0.577509
202      0.072899     5.364494   shower         flood          6.03  0.665506
203      0.496258     7.481290  weather      forecast          8.34  0.858710

[204 rows x 6 columns]
model name : GLoVe
diff_abs 平均值： 1.557095181732895
SignificanceResult(statistic=0.4955473036905242, pvalue=4.864178084843842e-14)
```

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 total :0.7 Stemming :0.7 lemmatization : 0.7	 data/BATS\1_Inflectional_morphology : category [noun - plural_reg] total accuracy total : 32 ps : 35 lm : 32 data : 50 total :0.64 Stemming :0.7 lemmatization : 0.64
 verb_ving ved verb_ving ved data/BATS\1_Inflectional_morphology : category [verb_ving - ved] total accuracy total : 38 ps : 41 lm : 38 data : 50 total :0.76 Stemming :0.82 lemmatization : 0.76	 data/BATS\1_Inflectional_morphology : category [verb_ving - ved] total accuracy total : 38 ps : 43 lm : 38 data : 50 total :0.76 Stemming :0.86 lemmatization : 0.76
 country capital country capital data/BATS\3_Encyclopedic_semantics : category [country - capital] total accuracy total : 46 ps : 43 lm : 46 data : 50 total :0.92 Stemming :0.86 lemmatization : 0.92	 data/BATS\1_Inflectional_morphology : category [verb_3psg - ved] total accuracy total : 27 ps : 30 lm : 27 data : 50 total :0.54 Stemming :0.6 lemmatization : 0.54
*SVD - word2vec	
 data/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	
 data/BATS\1_Inflectional_morphology : category [noun - plural_irreg] total accuracy total : 7 ps : 7 lm : 0 data : 50 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	