

自然語言處理 HW2

組員：

110590450 歐佳昀(60%) => 作業架構、流程、分析、改善

110590452 莊于潔(40%) => 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 embedding :

- word2vec
  - https://github.com/mnihaltz/word2vec-GoogleNews-vectors/blob/master/README.md
  - https://drive.google.com/file/d/0B7XkCwpl5KDYNINUTISS21pQmM/edit?usp=sharing
- GloVe
  - https://github.com/stanfordnlp/GloVe
  - https://huggingface.co/stanfordnlp/glove/resolve/main/glove.6B.zip
- SVD
  - https://github.com/valentin72/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      0.396392    6.981958    computer keyboard  7.62
...      ...      ...      ...      ...      ...
345     0.129479    5.647397    shower  flood    6.03
346     0.362721    6.813604    weather forecast  8.34
347     0.145228    5.726142    disaster area      6.25
348     0.296636    6.483178    governor office    6.34
349     0.143093    5.715467    architecture century  3.78

diff_abs
0      4.503114e-01
1      2.364810e-01
2      2.980232e-07
3      6.426870e-01
4      6.380418e-01
...      ...
345     3.826033e-01
346     1.526396e+00
347     5.238584e-01
348     1.431778e-01
349     1.935467e+00
...
[350 rows x 6 columns]
model name : word2vec
diff_abs 平均值: 1.3917418687971976
SignificanceResult(statistic=0.6845969668422078, pvalue=1.016190193647808e-49)
```

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      0.441461    7.207306    computer keyboard  7.62  0.412694
...      ...      ...      ...      ...      ...
348     0.072899    5.364494    shower  flood    6.03  0.665506
349     0.496258    7.481290    weather forecast  8.34  0.858710
350     0.261253    6.306264    disaster area      6.25  0.056264
351     0.404522    7.022609    governor office    6.34  0.682609
352     0.420720    7.103602    architecture century  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
<pre>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</pre>	<pre>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</pre>
<pre>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</pre>	<pre>49 data/BATS\1_Inflectional_morphology : category [verb_ving - ved] 50 total accuracy total : 38 ps : 43 lm : 38 data : 50 51 total :0.76 52 Stemming :0.86 53 lemmatization : 0.76</pre>
<pre>139 country capital country capital 140 data/BATS\3_Encyclopedic_semantics : category [country - capital] 141 total accuracy total : 46 ps : 43 lm : 46 data : 50 142 total :0.92 143 Stemming :0.86 144 lemmatization : 0.92</pre>	<pre>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</pre>
*SVD - word2vec	
<pre>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</pre>	
<pre>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</pre>	
<pre>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</pre>	