NLP 447 Project 2 Summary

Yuan Wang / ywang340

According to the instruction, there are four kinds of measurement methods implemented in this project:

- **the cosine similarity for word2vec vectors**
  Use pre-trained model "glove-wiki-gigaword-100" from genism package, where 100 indicates the number of dimensions between word vectors.

- **Wu-Palmer similarity on the TRIPS ontology**
  Load the lex-ont.json file to get lf_parents word. It is possible to have more than one lf-parents, so collect all the previous words in a list, then find the intersection of two lists as LCS. Choose the LCS with higher depth and the parents include the LCS, finally compute the similarity.

- **the cosine similarity for vectors computed using the Brown corpus**
  Build N by N metrics based on the length of words number from trips-brown_NV_overlap.txt, update the metrics based on Brown corpus, then calculate the cosine similarity from vector of input words

- **Forth novel method**
  Use the word2vec based on Brown corpus to predict similarity score

The output odd word is based on the similarity score, for each group (three words), if two of words get max score, the odd word is the rest of word

Discussion:

I labeled the assumed correct odd words for each group; it looks that the first and the second method have higher accuracy than others. However, my novel 4th method may correct for the last group of word which depends on how you interpretate the words. Waste and save can be opposite words, so help is the odd. The third method depends on the training data and looks the more the better.

Output:

tripleid	word2vec_score_choice	Wu_Palmer_score_choice	brown-vector_score_choice	4th_novel_score_choice
0	house	house	cat	house
1	bottle	bottle	bottle	bottle
2	drum	health	drum	health

3	doc	friend	doc	friend
4	waste	waste	waste	help

Score:

tripleid	word1	word2	word2vec_score	Wu_Palmer_score	brown-vector_score	4th_novel_score
0	dog	cat	0.8798075	1.0	0.8139726301062110	0.92693484
0	dog	house	0.43759328	0.36363636363636400	0.88328404000978	0.81000423
0	cat	house	0.37882093	0.36363636363636400	0.7008910505999740	0.5469933
1	bottle	house	0.27423105	0.42105263157894700	0.8158530834541380	0.8384596
1	bottle	run	0.19380157	0.35294117647058800	0.8425647690436400	0.78050363
1	house	run	0.49696347	0.8333333333333330	0.9034234796245880	0.87831223
2	drum	health	-0.0077056717	0.3157894736842110	0.5563925773978070	0.9067832
2	drum	milk	0.08721569	0.7777777777777780	0.5975480176686130	0.9736985
2	health	milk	0.35118803	0.3157894736842110	0.6634753713833010	0.91598296
3	doc	queen	0.092088275	0.9166666666666670	0.562078996167126	0.9460243
3	doc	friend	0.28238988	0.8333333333333330	0.7200374126405000	0.6631671
3	queen	friend	0.48006862	0.8333333333333330	0.7307981474642620	0.7541387
4	save	help	0.7064365	0.9473684210526320	0.8547973240768060	0.8912734
4	save	waste	0.36981145	0.6666666666666670	0.7645042165335950	0.903338
4	help	waste	0.37867028	0.631578947368421	0.7202843082428950	0.6582974

Commands:

python3 p2_ywang340.py input.csv output.csv score.csv

```
(base) dhcp-10-5-39-146:project2 wayoo$ python3 p2_ywang340.py input.csv output.csv score.csv
    tripleid   word1   word2  word2vec_score  Wu_Palmer_score  brown-vector_score  4th_novel_score
0          0     dog     cat        0.879807         1.000000            0.813973         0.924504
1          0     dog   house        0.437593         0.363636            0.883284         0.776587
2          0     cat   house        0.378821         0.363636            0.700891         0.586813
3          1  bottle   house        0.274231         0.421053            0.815853         0.866143
4          1  bottle     run        0.193802         0.352941            0.842565         0.801114
5          1   house     run        0.496963         0.833333            0.903423         0.878986
6          2    drum  health       -0.007706         0.315789            0.556393         0.906599
7          2    drum    milk        0.087216         0.777778            0.597548         0.950923
8          2  health    milk        0.351188         0.315789            0.663475         0.904502
9          3     doc   queen        0.092088         0.916667            0.562079         0.917080
10         3     doc  friend        0.282390         0.833333            0.720037         0.679939
11         3   queen  friend        0.480069         0.833333            0.730798         0.754086
12         4    save    help        0.706437         0.947368            0.854797         0.849234
13         4    save   waste        0.369811         0.666667            0.764504         0.905361
14         4    help   waste        0.378670         0.631579            0.720284         0.737258
   tripleid word2vec_score_choice Wu_Palmer_score_choice brown-vector_score_choice 4th_novel_score_choice
0         0                 house                 house                       cat                 house
1         1                bottle                bottle                    bottle                bottle
2         2                  drum                health                      drum                health
3         3                   doc                friend                       doc                friend
4         4                 waste                 waste                     waste                  help
(base) dhcp-10-5-39-146:project2 wayoo$
```