WORD EMBEDDING FOR MALAY LANGUAGE USING WORD2VEC

Fung Chey (fungchey@student.usm.my)

INTRODUCTION

Word embedding is a NLP technique where the words from vocabulary are mapped to vectors of real numbers. The distance (cosine similarity) between the vectors helps in measuring words similarity and semantic relationship. It is widely used in document clustering, sentiment analysis, or even recommender system.

PROBLEM

Application of word embedding in Malay language is rare. Most of the Malaysia websites used Google as the default search engine in their hosted page. They rely on the Google indexing to return the search query among their website. The problem occurred on Malay language website when:

- The hosted webpage is limited to intranet purpose only, access to Google service is not possible.
- Customization on domain specific (e.g. traditional medicine, local tourism) query is required.
- Sentiment analysis (e.g. facebook comment, user feedback) in Malay language.

METHODOLOGY

1. Malay version of Wikipedia dump (~200MB) was loaded as a corpus.
2. Python's Gensim.Corpora package was used to normalize the corpus into 177 millions words.
3. Python's Gensim.Word2Vec package was applied to vectorized it into 400 dimensions.
4. Functions was created to retrieve the distance between words from the trained model:
   a. Test most similar: Find top 10 words that are closed to the input word.
   b. Test similarity: Find the relationship between two input words.
   c. Test doesn't match: Find the outliers in the list of the given words.
5. Perform PCA dimensionality reduction into 3D if wanted to visualize the results in Python's Matplotlib package.

```
test_most_similar("demam")
```

```
batuk 0.762915849685669
cirit 0.7364166975021362
birit 0.7257339954376221
muntah 0.7217694520950317
akut 0.719976544380188
denggi 0.7193765044212341
jangkitan 0.7163813710212708
bengkak 0.7148799896240234
radang 0.7126706838607788
simptom 0.7072034478187561
```

```
test_most_similar("langkawi")
```

```
tioman 0.6127931475639343
jerlun 0.5522773265838623
pangkor 0.5380395650863647
jitra 0.5357160568237305
awana 0.5328423976898193
labuan 0.5215494632720947
kulim 0.5120787620544434
kukup 0.5118043422698975
jemor 0.5095215439796448
bidong 0.5029813051223755
```

```
test_similarity ("kedah","perlis")
```

```
0.68158895
```

```
test_similarity ("kedah","sarawak")
```

```
0.37179914
```

RESULTS & DISCUSSION

1. Test most similar:
   a. Fever (demam) has high correlation with cough (batuk, 76%) and diarrhoea (cirit-birit, 73%).
   b. Langkawi has highest correlation with Tioman (61%) as both are recreational island, although they are located far apart geographically. The second highest is Jerlun (55%), which is small inland town located in the Kedah, same state as where Langkawi located in.
2. Test similarity:
   a. Kedah and Perlis has 68% similarity as both are northern region state in Penisular Malaysia.
   b. Kedah and Sarawak has only 37% similarity because Sarawak is located at East Malaysia, far apart geographically, and has no other significant similarity except that both are states in Malaysia.
3. Test doesn't match:
   a. Car (kereta), bicycle (basikal), motorcar (motor), chicken (ayam) were tested. Chicken (ayam) was identified as outliers.
   b. Banana (pisang), durian (durian), papaya (betik), bowl (mangkuk) were tested. Bowl (mangkuk) was identified as outliers.

CONCLUSION

Word2Vec is suitable in performing Malay word embedding task. The accuracy can be further improve by training the larger corpus. Note that the graph did not plotted according to absolute position as the PCA has reduced the dimensionality of vectors from 400 to 3 dimension for ease of visualization.

FUTURE WORK

Interested researchers can find the updates of this project from the author's public Github repository below:

https://github.com/fungchey/word2vec

```
test_doesnt_match ("kereta","basikal","motor","ayam")
```

```
ayam
```