Comp 472 Mini 3

Kimchheng heng

word2vec-google-news-300

- Pre train data
- Return ready to used data
- We can used
  - wv[word] , wv is the dataset return from gensim.download.load model and question is word we want to look for
  - If it does not exist in the model dataset, keyError would be through
  - Use wv.similarity(w1,w2) to check the similarity between two word, the higher value it is the closer it is

Compare between model

word2vec-google-news-300,3000000,70,79,0.8860759493670886

Glove-wiki-gigaword-50,400000,57,80,0.7125

glove-wiki-gigaword-100,400000,65,80,0.8125

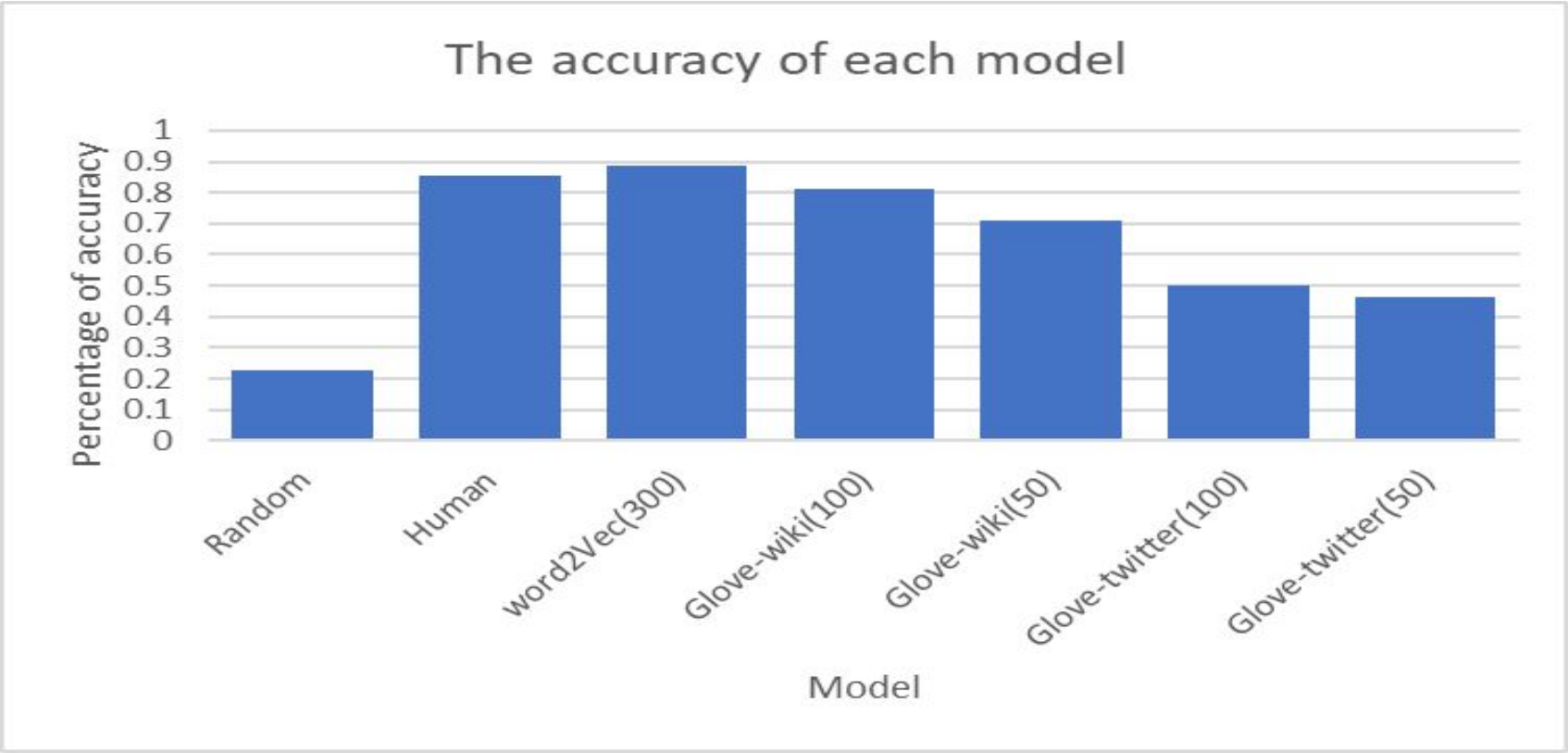
glove-twitter-50,1193514,36,78,0.4615384615384616

glove-twitter-100,1193514,39,78,0.5

random,0,23,80,0.225

Human average performance of 85.57% , which is higher than most of the model but lower than Word2vec

Compare all the model



Some analysis

- The size of corpus could not determine the accuracy
- Compare between two different dimension parameter
  - The higher value have higher accuracy
  - The lower value have lower accuracy
- Maybe the higher similarity value does not guarantee the correct answer