NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Synthetic Word Embeddings

For downstream NLP applications

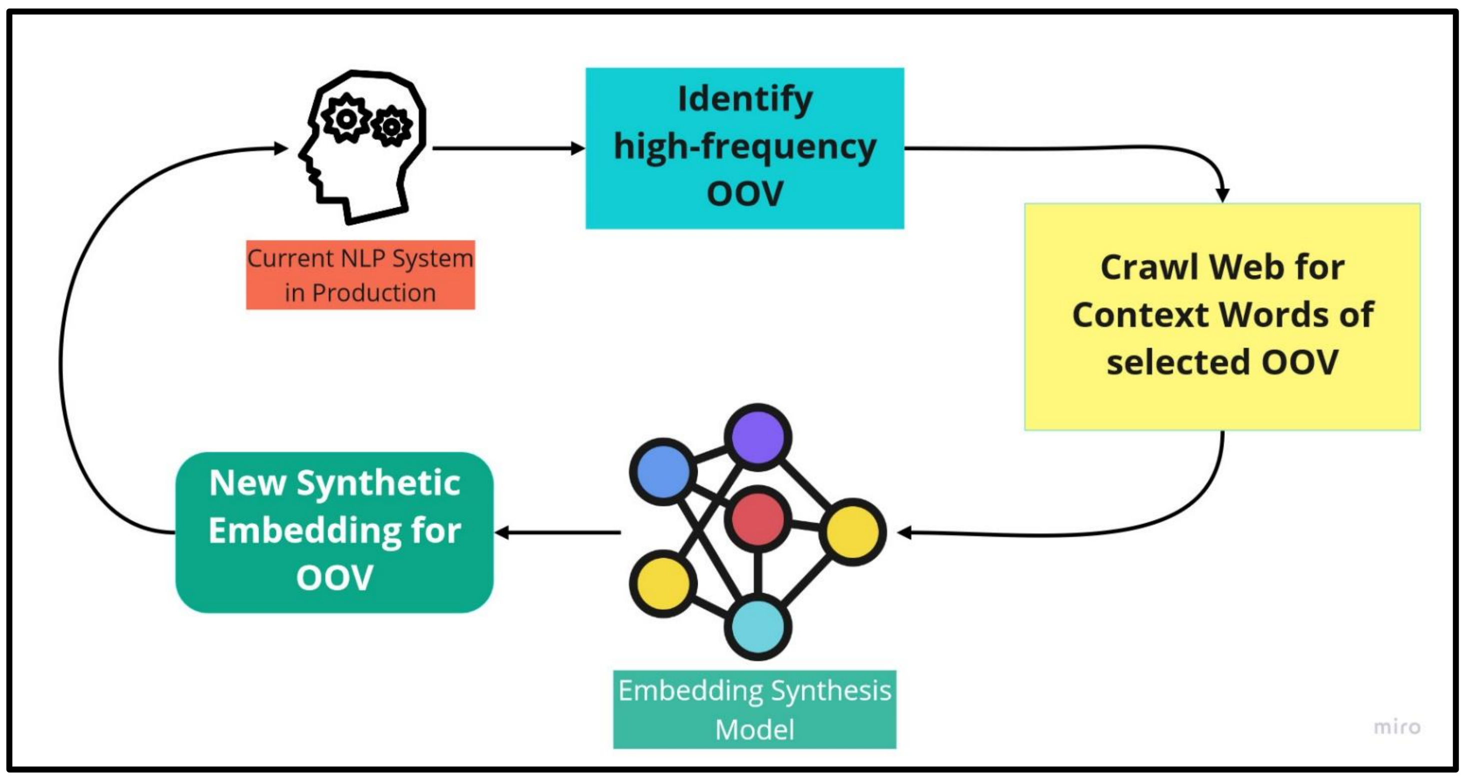
Student: Hoang Viet

Supervisor: A/P Chng Eng Siong

Project Objectives:

This Project aims to create synthetic embedding for unknown words using a novel approach, to enable machine learning applications to understand unprecedented phenomena and entities such as COVID-19 and PCR test. Our experiments show that the proposed network can generate new embedding containing relevant semantics using existing knowledge in latent space. Downstream tasks can then incorporate these synthetic embedding to understand new entities better

Workflow:


Current NLP System in Production
Identify high-frequency OOV
Crawl Web for Context Words of selected OOV
Embedding Synthesis Model
New Synthetic Embedding for OOV
miro

Best Architecture:

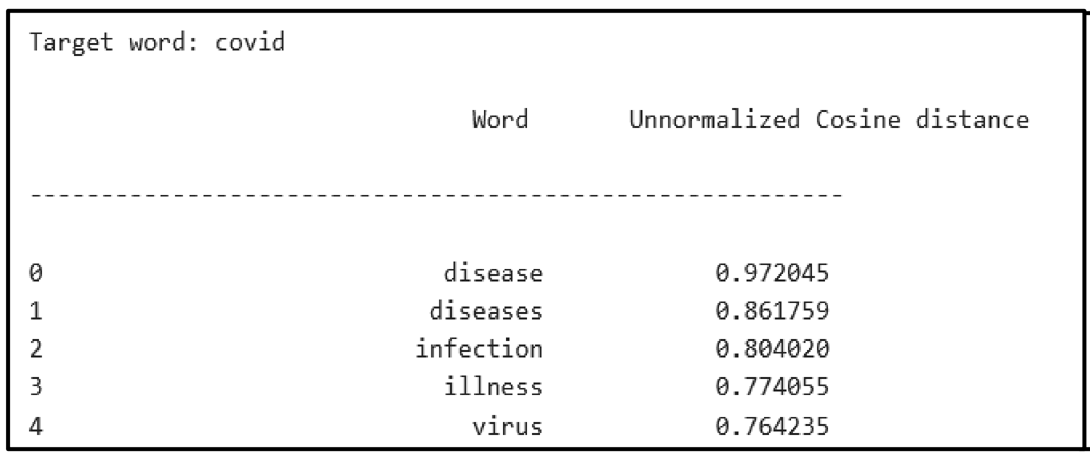
Multi- Layer Perceptrons

Embedding Tested:

GloVe Embeddings	RoBERTa Embeddings
100 dimensions	768 dimensions

Results:

The proposed workflow is able to synthesize embeddings for unknown words with great semantic relevance to the entities. These synthetic embeddings can then be easily inserted into the existing applications

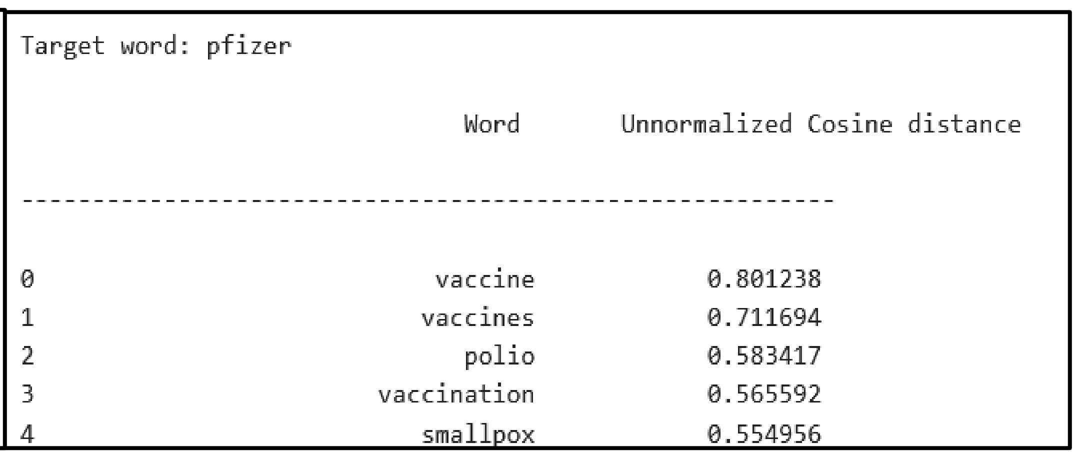


```
Target word: covid

                          Word      Unnormalized Cosine distance

---------------------------------------------------------

0                      disease             0.972045
1                     diseases             0.861759
2                    infection             0.804020
3                      illness             0.774055
4                        virus             0.764235
```



```
Target word: pfizer

                          Word      Unnormalized Cosine distance

---------------------------------------------------------

0                      vaccine             0.801238
1                     vaccines             0.711694
2                        polio             0.583417
3                  vaccination             0.565592
4                     smallpox             0.554956
```

GitHub Demo

www.ntu.edu.sg/scse