Assignment- In20-S2-CS5616

Word Embedding Assignment

Name	**Jayasuriya D.P.**
Index number	**209337M**

Question 1

How Genism was used to create word vectors.

Genism is an opensource python library for natural language processing. This library will enable us to develop word embeddings by training our own model on a custom corpus whether with CBOW or skip-grams algorithms.

Genism word2vec required that a format of list of lists for training where every document is contained in a list and every list contains lists of tokens of that documents. Genism wod2vec model class definition as follows, important ones are highlighted.

class gensim.models.word2vec.Word2Vec(***sentences***=*None*, *corpus_file=None*, ***size***=*100*, *alpha=0.025*, ***window***=*5*, ***min_count***=*5*, *max_vocab_size=None*, *sample=0.001*, *seed=1*, ***workers***=*3*, *min_alpha=0.0001*, ***sg***=*0*, *hs=0*, *negative=5*, *ns_exponent=0.75*, *cbow_mean=1*, *hashfxn=<built-in function hash>*, *iter=5*, *null_word=0*, *trim_rule=None*, *sorted_vocab=1*, *batch_words=10000*, *compute_loss=False*, *callbacks=()*, *max_final_vocab=None*)

Argument	Value
sentences	The sentences iterable can be simply a list of lists of tokens
size	Dimensionality of the word vectors.
window	Maximum distance between the current and predicted word within a sentence.
min_count	Ignores all words with total frequency lower than this.
workers	Use these many worker threads to train the model (=faster training with multicore machines).
sg	Training algorithm: 1 for skip-gram; otherwise CBOW.

Table 1 – Arguments for word2vec class

Example:

```
model = Word2Vec(word_list, size=100, window=5, min_count=1, workers=4, sg=0)
```

Question 2

How similarity between words was measured

Similarity between words were measured using

*most_similar(**positive**=None, **negative**=None, **topn**=10, restrict_vocab=None, indexer=None)*

- **positive** (list of str, optional) – List of words that contribute positively.
- **negative** (list of str, optional) – List of words that contribute negatively.
- **topn** (int or None, optional) – Number of top-N similar words to return, when topn is int. When topn is None, then similarities for all words are returned.

Ex:

vector1 = model_cbow.wv["කටයුතු"]

model_cbow.wv.most_similar([vector1])

Below are the 10 most common words in corpus, here I took the words having more than three letters when calculating the top words.

'සඳහා', 'යුතු', 'කිරීම', 'කටයුතු', 'විසින්', 'සඳහන්', 'ලේකම්', 'වැනි', 'කවරේද', 'යටතේ'

For each word top ten similar words were predicted from each model. The results are as follows.

Word	CBOW Model	Skip Gram Model
සඳහා	සඳහා, සඳහා, වලින්, වෙනුවෙන්, පිණිස, යාවත්කාලීන, කිරීමටත්, අවශ්‍යතා, පදනම්, කරගෙන	සඳහා, කිරීම, යොදා, පියවා, ආරම්භ, සැපයුම, බඳවා, ක්‍රියාමාර්ගයක්, අපේක්ෂිත, ගෙන්වා
යුතු	යුතු, යුතුය, හැකි, යුත්තේ, කෙරෙන, අදාළ, දැක්වෙන, කල්, නොහැකි, ලියා	යුතු, ඉල්ලුම්පත්‍රයේ, ඉල්ලුම්පත්‍රය, ලිපියකින්, දිය, නොගෙවිය, යුතුය, ගම්‍ය, විය, සැලකිය
කිරීම	කිරීම, කිරීම්, කිරීමට, කිරීමේ, කිරීමේදී, කිරීමටත්, කිරිම, කරමින්, කෙරෙන, කොට	කිරීම, ගිවිසුම, පිළියෙල, ආරක්ෂා, අනිසි, ඇගැයීම, වාර්ෂිකය, දැනුවත්, අපේක්ෂා, කරගත්තාවූ
කටයුතු	කටයුතු, සැලසුම්, අලුත්වැඩියා, කඩිනම්න්, සිදු, සංවිධානය, විධිමත්, සාකච්ඡා, දෝෂ, සම්පාදනය	කටයුතු, මගපෙන්වීම, අවශ්‍ය, දීමට, කරන්නේද, වැදගත්වන, දෙයි, මුද්‍රාහැරිය, සූත්‍ර, නොපමාව
විසින්	වවිසින්, **සාදන**, අදාලව, විගණකාධිපති, මණ්ඩලය, දිනැතිව, එවා, ලිපියට, වෙත, කරුණට	විසින්, ලදුව, **සාදන**, සාදනු, අති, ලිපියෙන්, අමාත්‍යවරයා, උක්ත, දිනැති, දිනැතිව
සඳහන්	සඳහන්, දැක්වෙන, සඳහන්, වේ, යන්න, **යන්නත්**, කරන්නෙහිද, **වගන්තිවල**, කරුණු, ලබන්නේකෙසේද	සඳහන්, **වගන්තිවල**, පරිදි, වාර්තාවෙහි, ලිපියෙහි, **යන්නත්**, කොන්දේසිවලට, විධිවිධානයන්, මෙම, හිග
ලේකම්	ලේකම්, ප්‍රාදේශීය, දිස්ත්‍රික්, මාතලේ, **දිසාපති**, කෝරළේ, **කාර්යාලය**, පස්බාගේ, **කාර්යාලයේ**, ප්‍රාදේශිය	ලේකම්, **දිසාපති**, කාර්යාලවල, දිස්ත්‍රික්, ප්‍රාදේශීය, මැදගම, කොට්ඨාස, **කාර්යාලය**, **කාර්යාලයේ**, මූලටියන
වැනි	වැනි, ආඥාපනතේ, සුරාබදු, අධිකාරය, වෘක්ෂලතා, වනසත්ව, ඡේදය, බලන්න, පරිච්ඡේදයේ, වගන්තියේ	වැනි, අධිකාරය, වගන්තිය, ආඥාපනතේ, පරිශිෂ්ටය, කාණ්ඩයේ, ඡේදය, වගන්ති, කියවිය, පරිච්ඡේදයේ

කවරේද	කවරේද, **යන්න**, ප්‍රතිඵල, ලිපිනයන්, ඒවායින්, එතුමා, වශඵතුමා, ඒවායේ, බලපෑ, අපේක්ෂිත	කවරේද, ක්‍රියාමාර්ග, **යන්න**, තිබේද, පියවර, පෙරදාතම, වැඩපිළිවෙලක්, උසාවිය, හේතු, ගෙන
යටතේ	යටතේ, සංචාරක, වගන්තියේ, **සුරාබදු, වගන්තිය, ආඥාපනතේ**, ආරක්ෂක, සංග්‍රහයේ, ආඥාපනත, දැක්වීමේ	යටතේ, **ආඥාපනතේ**, පනතේ, වගන්ති, **වගන්තිය**, කාණ්ඩයේ, **සුරාබදු**, අධිකාරය, ඡේදය, උපවගන්තිය

Table 2 – Comparison between similar words by two models

Then FastText pretrained model for sinhala is downloaded and checked the similar words for the same top ten common words.

Similar words predict by fasttext model trained by Facebook research group.

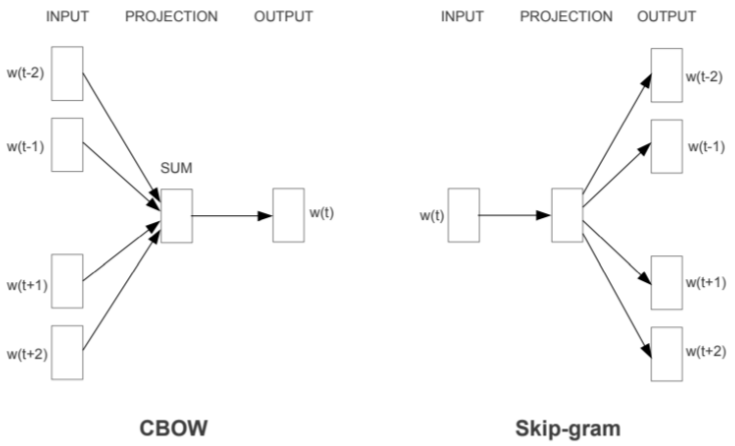
Word	FastText model top 10 similar words
සඳහා	සඳහා, සදහා, සඳහා, කිරීම, සඳහාඳ, සඳහාවු, සඳහාඑජ, සඳහා1.54ක්, සඳහාමද, සඳහාඑම
යුතු	යුතු, හැකි, යුතුය, යුතුමුත්, යුතුම, යුතුම්ව, යුතුයි.අවශ්‍ය, යුතුමද, විය, නොහැකි
කිරීම	කිරීම, කිරීමට, කිරීම.ඒ, කිරීම, ගැනීම, කිරීමල, කිරීමේ, කිරීම3, වීම, කිරීමී
කටයුතු	කටයුතු, කටයුතුව, කටයුතුවලදීත්, කලාකටයුතු, කටයුතුවලද, කටයුතුවලදීද, කටයුතුකල, කටයුතුකලා, සේවාකටයුතු, සම්ගකටයුතු
විසින්	විසින්, මගින්, විසින්ම, වෙත, විසින, විසිනි, මහතාවිසින්, මඟින්, හර්ෂල්විසින්, තමාවිසින්ම
සඳහන්	සඳහන්, සඳහන්ඉනි, සදහන්, සඳහන්කර, සඳහන්ද, සඳහන්කලා, සඳහන්ය.ඒ, සඳහන්වූ, සඳහනේ, සඳහන්වන
ලේකම්	ලේකම්, ලේකම්ය, ලේකම්ල, ලේකම්ගෙ, ලේකම්ද, ලේකම්කාර්යාලය, ලේකම්තුමිය, කාර්යාලයලේකම්, මහලේකම්, ලේකම්කම
වැනි	වැනි, වැනිදෑද, සියවසේ, වැනිදේට, හවැනි, මෙවැනි, 16වැනි, වැනිදයින්, වැනිදේද, 01වැනි
කවරේද	කවරේද, කවරේදැ, කියනුම, කවරේදැයි, හමාස, ආර්‍යසත්‍යය, කිමෙක්ද, කවරේට, කෙසේද, දුක්ඛආර්‍යසත්‍යය
යටතේ	යටතේ, යටතේී, යටතේයි, 3244, යටතේමය, යටතේත්, යටතේම, යටතේය, යටතේය.එම, යටතේද

Table 3 – Top 10 similar words for FastText by facebook https://github.com/facebookresearch/fastText

Question 3

Your analysis on the results

Word Embeddings is a technique that is used to represent words as vector forms. In Word2Vec, we train a model with a hidden layer that can predict the target word based on its context words. The assumption is that the meaning of the word can be inferred by the neighboring words. In their paper they mention about two models, continuous bag of words (CBOW) and continuous skip-gram model.



In simple words CBOW model, the distributed representation of context used to predict the word in the middle as contrast to in skip-gram, the distributed representation of the input word is used to predict the context.

In this assignment we train the Sinhala corpus for these two methods and took the top 10 most used words (with more than 3 letters to avoid small joining words like හෝ, හා, විසින්. The predicted the most similar words for those 10-word using both previous models we trained. Table 2 shows the comparison. There it's clear that only some of the words are matched, but both words context is similar. Due to two training methods may be the prediction is different but it learnt the context correctly.

Then we took the FastText model trained by face book for Sinhala language and try to predict the same words using that model. Table 3 shows the results. FastText is essentially an extension of wor2vec morel This treats each word as composed of characters. FastText use n-grams and learns the vectors only for complete words found in the training. (learns for the n-grams that found within each word). This generate better word embeddings for rare words and handle out-of-vocab words. Also because of that, this only trained for complete words.

Those results are not comparable with the results we got. Also, this is trained using much more large word corpus and different context than the given word file for word2vec models. Because of this reason and also the improvements over word2vec, FastText gives better results than the Word2Vec method.

Code

GitHub Link to the complete code Jupiter notebook – same file attached with this the submission

https://github.com/DulanGit/Word-Embeddings/blob/master/gensim/Word%20Embeddings.ipynb

References

https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

https://medium.com/analytics-vidhya/word2vec-cbow-skip-gram-algorithmic-optimizations-921d6f62d739

https://towardsdatascience.com/fasttext-under-the-hood-11efc57b2b3

https://fasttext.cc/

https://medium.com/@bruceyanghy/nlp-and-deep-learning-all-in-one-part-ii-word2vec-glove-and-fasttext-184bd03a7ba

https://github.com/facebookresearch/fastText

https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

https://en.wikipedia.org/wiki/Sinhala_language

https://radimrehurek.com/gensim/models/fasttext.html

https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors.WordEmbeddingsKeyedVectors.most_similar

https://radimrehurek.com/gensim/models/word2vec.html