**Fakulti Teknologi Maklumat dan Komunikasi**

**Universiti Teknikal Malaysia Melaka**

**BITI 3413 Natural Language Processing**

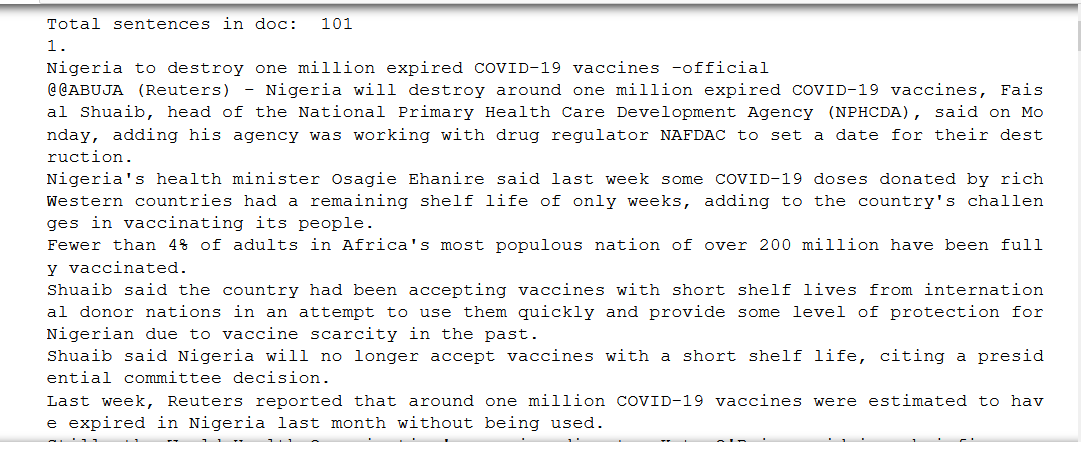
**Assignment 02 (Group of Two people)- 10%**

Group members:

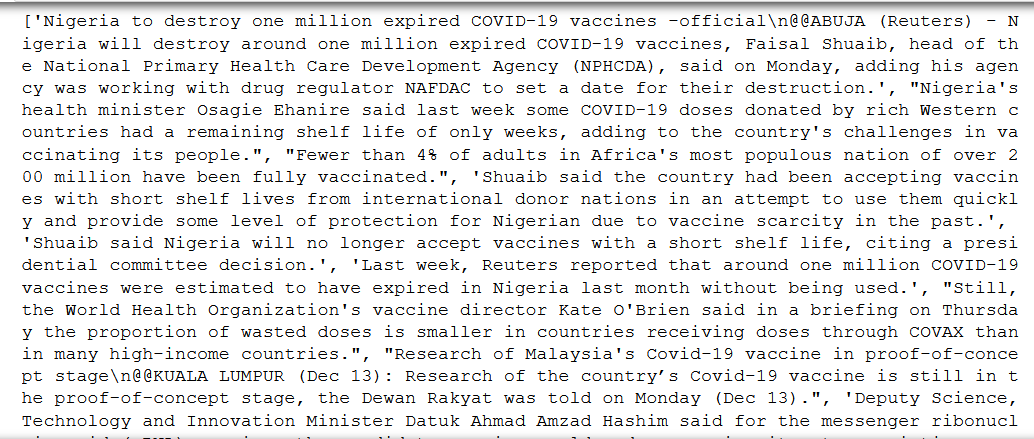
Lim Wen Ni B031910441

Reca Seng Binti Mohd Fadzil Seng B031910187

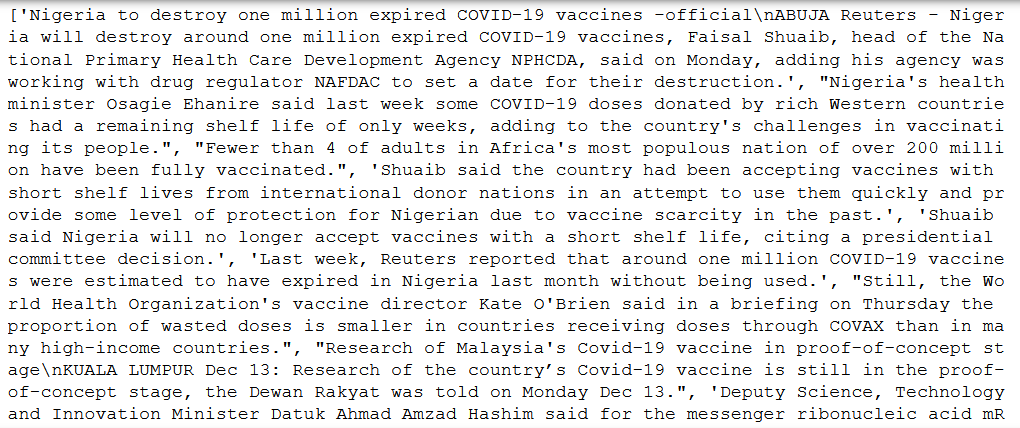
First of all, we have done selecting data from the Google by typing ‘COVID-19 vaccination news’ on the search bar. Then, we collected the written news displayed by the Google manually. Our text data before pre-processing has a total of 2448 words. For pre-processing part, firstly, we have done sentence segmentation on the text data using NLTK sent\_tokenize.



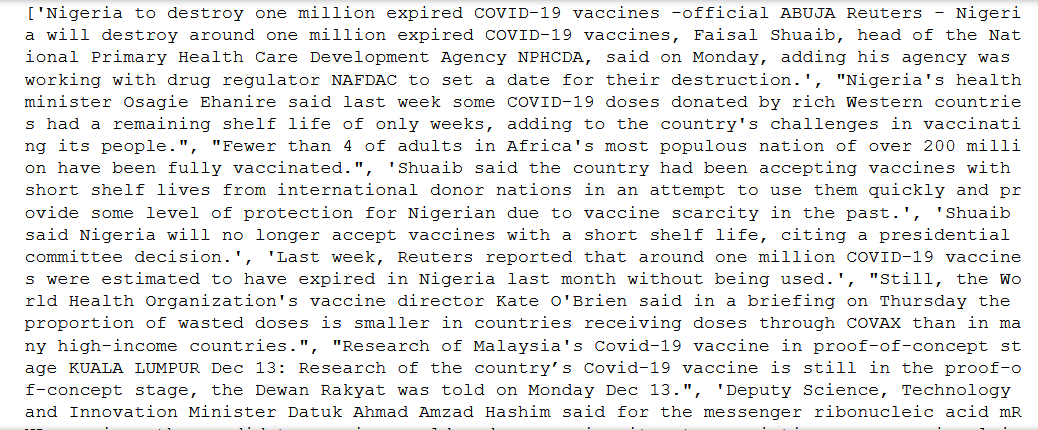
Secondly, as the numbering is also counted as one sentence, we decide to remove the numbering as part of the text data.

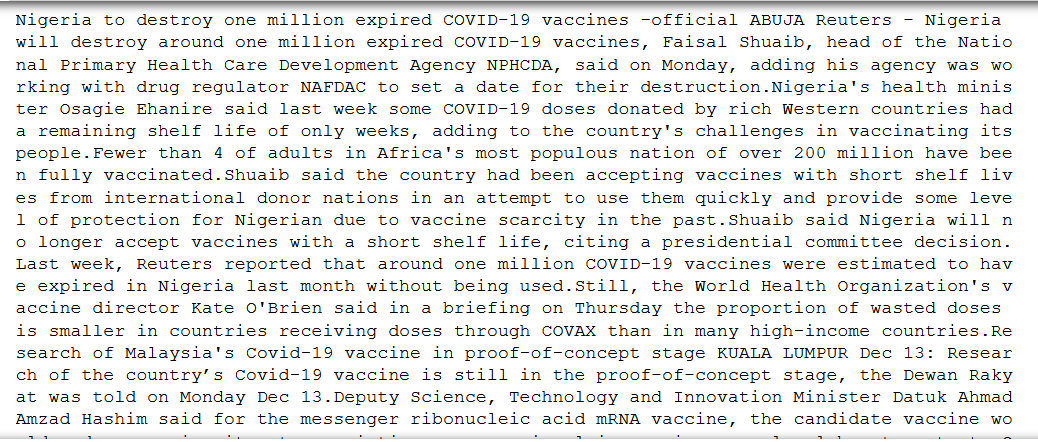


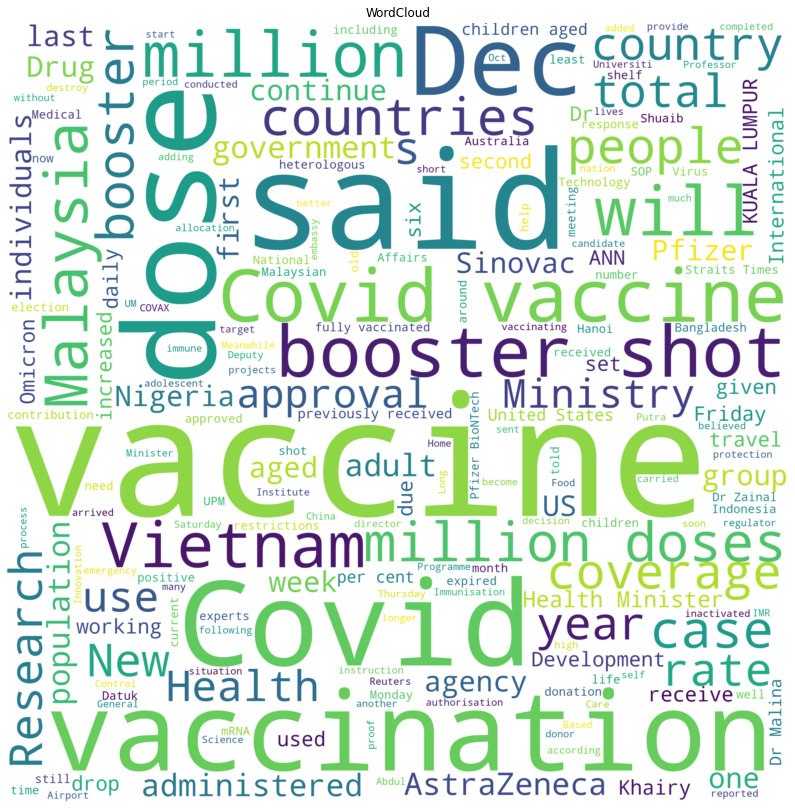
Thirdly, we remove all the special characters in the text data.

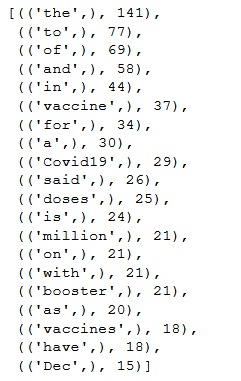
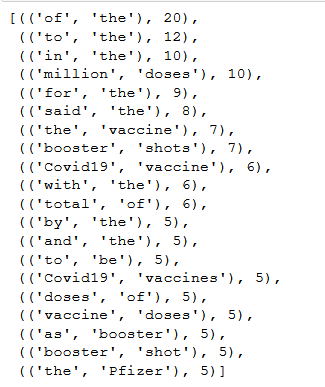
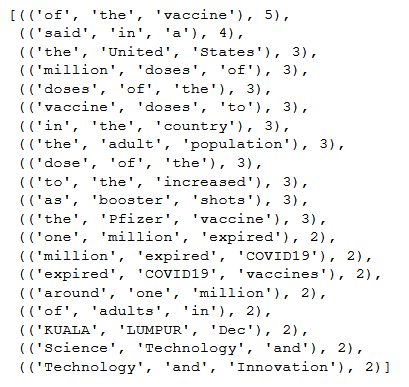


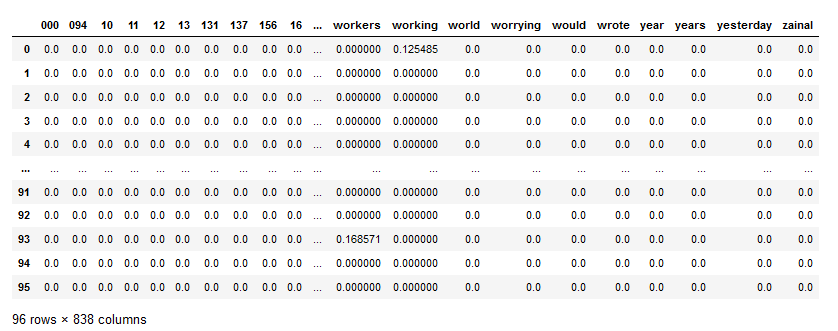
Fourth, since the ‘\n’ and ‘\’ are still present in the text data, we decide to remove them too.

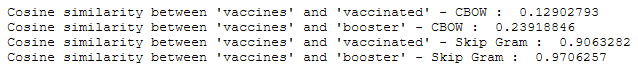


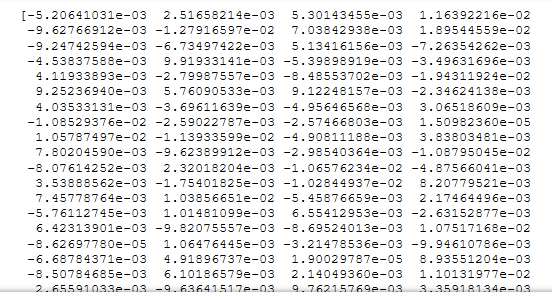
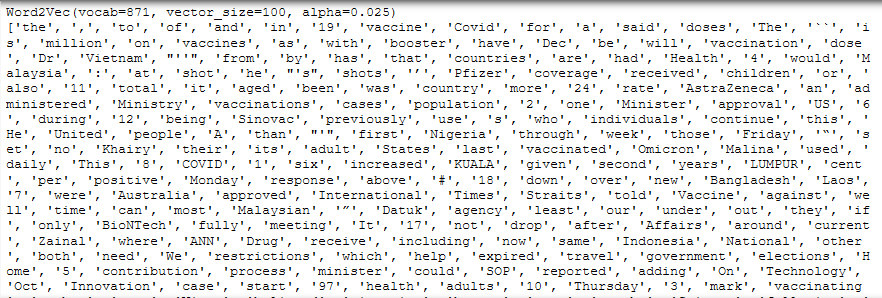
 Last but not least, for pre-processing, we combine all the sentences in the list into single text data with join() method. Thus, the corpus size is now 2351 (2351 words).

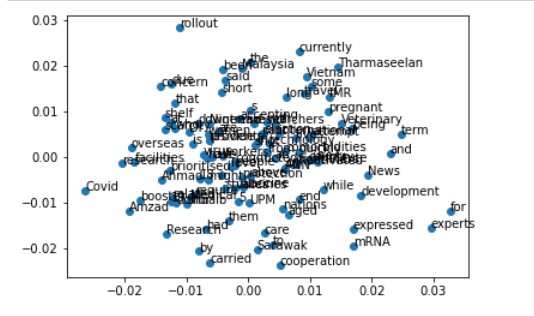
Moving on to the wordcloud (**2a**), this is the visualization of the frequent words appear in the text data. From the visualization above, the most frequent word found is ‘vaccine’ followed by ‘Covid’, ‘said, ‘Dec’, ‘dose’ and ‘vaccination. 

Meanwhile, the top 20 most frequent unigram, bigram and trigram are as follow:

The tfidf representation generated is as follows:

The generated word-embedding representation (either Skip-gram or CBOW) is as follows:

From the above results, we can see that the context of the word ‘vaccines’ is much closer to the word ‘booster’ rather than the word ‘vaccinated’ in both CBOW and Skip Gram method. In the English grammar context, the above results is considered true also. This is because both ‘vaccines’ and ‘booster’ are both noun while ‘vaccinated’ is a verb. For the visualization of embedding are as follows:



Resources: 1. Brownlee, J. (2020, September 3). *How to Develop Word Embeddings in Python with Gensim*. Machine Learning Mastery. <https://machinelearningmastery.com/develop-word-embeddings-python-gensim/> 2. Richter, M. (2021, December 10). *Comparing Word Embeddings - Towards Data Science*. Medium. <https://towardsdatascience.com/comparing-word-embeddings-c2efd2455fe3> 3. Johnson, D. (2022, January 1). *NLTK Tokenize: Words and Sentences Tokenizer with Example*. Guru99. <https://www.guru99.com/tokenize-words-sentences-nltk.html>