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.

```
Total sentences in doc:  101
1.
Nigeria to destroy one million expired COVID-19 vaccines -official
@@ABUJA (Reuters) - Nigeria will destroy around one million expired COVID-19 vaccines, Fais
al Shuaib, head of the National Primary Health Care Development Agency (NPHCDA), said on Mo
nday, adding his agency was working with drug regulator NAFDAC to set a date for their dest
ruction.
Nigeria's health minister Osagie Ehanire said last week some COVID-19 doses donated by rich
Western countries had a remaining shelf life of only weeks, adding to the country's challen
ges in vaccinating its people.
Fewer than 4% of adults in Africa's most populous nation of over 200 million have been full
y vaccinated.
Shuaib said the country had been accepting vaccines with short shelf lives from internation
al donor nations in an attempt to use them quickly and provide some level of protection for
Nigerian due to vaccine scarcity in the past.
Shuaib said Nigeria will no longer accept vaccines with a short shelf life, citing a presid
ential committee decision.
Last week, Reuters reported that around one million COVID-19 vaccines were estimated to hav
e expired in Nigeria last month without being used.
```

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

```
['Nigeria to destroy one million expired COVID-19 vaccines -official\n@@ABUJA (Reuters) - N
igeria will destroy around one million expired COVID-19 vaccines, Faisal Shuaib, head of th
e National Primary Health Care Development Agency (NPHCDA), said on Monday, adding his agen
cy was working with drug regulator NAFDAC to set a date for their destruction.', "Nigeria's
health minister Osagie Ehanire said last week some COVID-19 doses donated by rich Western c
ountries had a remaining shelf life of only weeks, adding to the country's challenges in va
ccinating its people.", "Fewer than 4% of adults in Africa's most populous nation of over 2
00 million have been fully vaccinated.", 'Shuaib said the country had been accepting vaccin
es with short shelf lives from international donor nations in an attempt to use them quickl
y and provide some level of protection for Nigerian due to vaccine scarcity in the past.',
'Shuaib said Nigeria will no longer accept vaccines with a short shelf life, citing a presi
dential committee decision.', 'Last week, Reuters reported that around one million COVID-19
vaccines were estimated to have expired in Nigeria last month without being used.', "Still,
the World Health Organization's vaccine director Kate O'Brien said in a briefing on Thursda
y the proportion of wasted doses is smaller in countries receiving doses through COVAX than
in many high-income countries.", "Research of Malaysia's Covid-19 vaccine in proof-of-conce
pt stage\n@@KUALA LUMPUR (Dec 13): Research of the country's Covid-19 vaccine is still in t
he proof-of-concept stage, the Dewan Rakyat was told on Monday (Dec 13).", 'Deputy Science,
Technology and Innovation Minister Datuk Ahmad Amzad Hashim said for the messenger ribonucl
```

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

```
['Nigeria to destroy one million expired COVID-19 vaccines -official\nABUJA Reuters - Niger
ia will destroy around one million expired COVID-19 vaccines, Faisal Shuaib, head of the Na
tional Primary Health Care Development Agency NPHCDA, said on Monday, adding his agency was
working with drug regulator NAFDAC to set a date for their destruction.', "Nigeria's health
minister Osagie Ehanire said last week some COVID-19 doses donated by rich Western countrie
s had a remaining shelf life of only weeks, adding to the country's challenges in vaccinati
ng its people.", "Fewer than 4 of adults in Africa's most populous nation of over 200 milli
on have been fully vaccinated.", 'Shuaib said the country had been accepting vaccines with
short shelf lives from international donor nations in an attempt to use them quickly and pr
ovide some level of protection for Nigerian due to vaccine scarcity in the past.', 'Shuaib
said Nigeria will no longer accept vaccines with a short shelf life, citing a presidential
committee decision.', 'Last week, Reuters reported that around one million COVID-19 vaccine
s were estimated to have expired in Nigeria last month without being used.', "Still, the Wo
rld Health Organization's vaccine director Kate O'Brien said in a briefing on Thursday the
proportion of wasted doses is smaller in countries receiving doses through COVAX than in ma
ny high-income countries.", "Research of Malaysia's Covid-19 vaccine in proof-of-concept st
age\nKUALA LUMPUR Dec 13: Research of the country’s Covid-19 vaccine is still in the proof-
of-concept stage, the Dewan Rakyat was told on Monday Dec 13.", 'Deputy Science, Technology
and Innovation Minister Datuk Ahmad Amzad Hashim said for the messenger ribonucleic acid mR
```

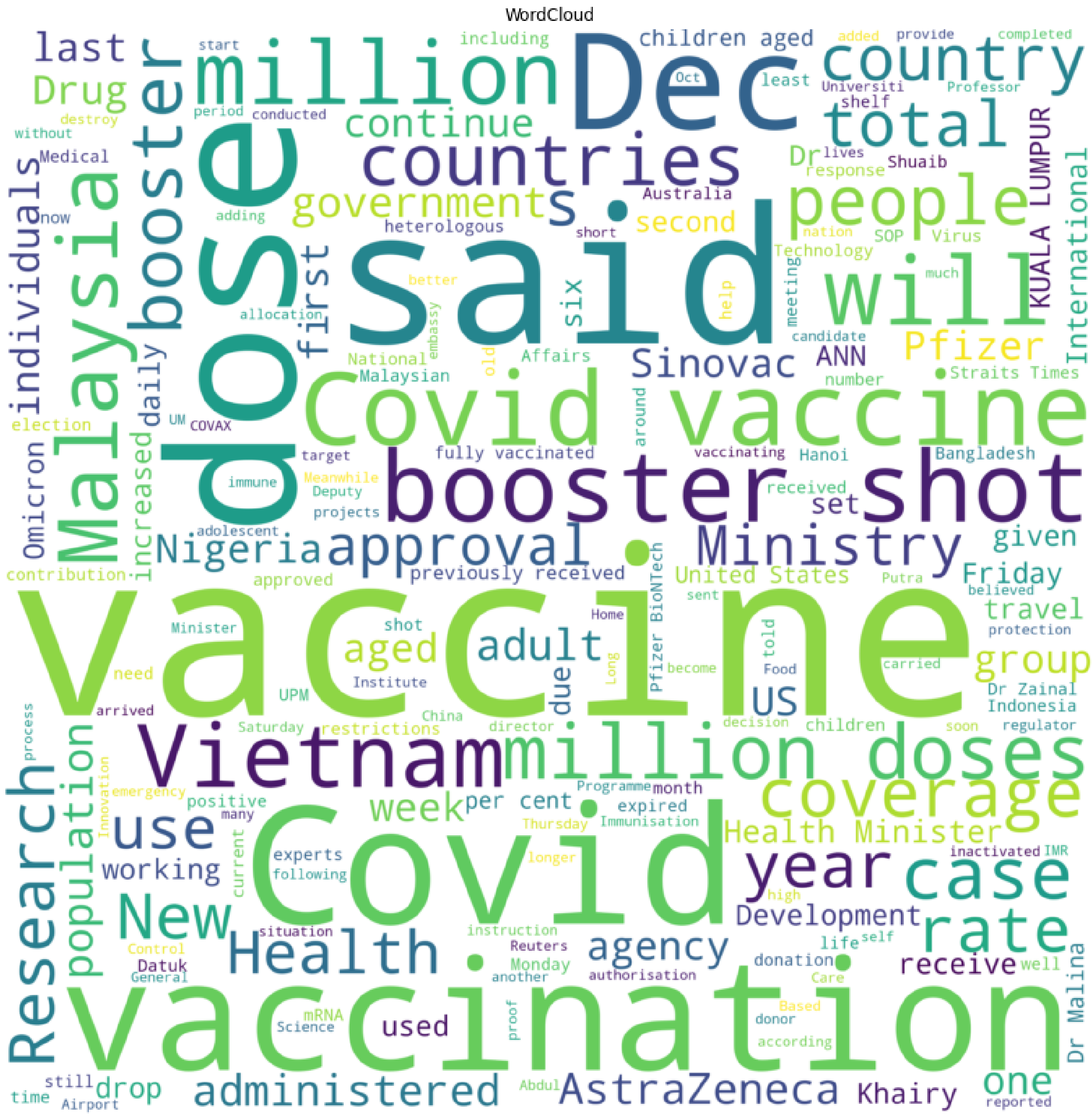
Fourth, since the ‘\n’ and ‘\’ are still present in the text data, we decide to remove them too.

```
['Nigeria to destroy one million expired COVID-19 vaccines -official ABUJA Reuters - Nigeri
a will destroy around one million expired COVID-19 vaccines, Faisal Shuaib, head of the Nat
ional Primary Health Care Development Agency NPHCDA, said on Monday, adding his agency was
working with drug regulator NAFDAC to set a date for their destruction.', "Nigeria's health
minister Osagie Ehanire said last week some COVID-19 doses donated by rich Western countrie
s had a remaining shelf life of only weeks, adding to the country's challenges in vaccinati
ng its people.", "Fewer than 4 of adults in Africa's most populous nation of over 200 milli
on have been fully vaccinated.", 'Shuaib said the country had been accepting vaccines with
short shelf lives from international donor nations in an attempt to use them quickly and pr
ovide some level of protection for Nigerian due to vaccine scarcity in the past.', 'Shuaib
said Nigeria will no longer accept vaccines with a short shelf life, citing a presidential
committee decision.', 'Last week, Reuters reported that around one million COVID-19 vaccine
s were estimated to have expired in Nigeria last month without being used.', "Still, the Wo
rld Health Organization's vaccine director Kate O'Brien said in a briefing on Thursday the
proportion of wasted doses is smaller in countries receiving doses through COVAX than in ma
ny high-income countries.", "Research of Malaysia's Covid-19 vaccine in proof-of-concept st
age KUALA LUMPUR Dec 13: Research of the country’s Covid-19 vaccine is still in the proof-o
f-concept stage, the Dewan Rakyat was told on Monday Dec 13.", 'Deputy Science, Technology
and Innovation Minister Datuk Ahmad Amzad Hashim said for the messenger ribonucleic acid mR
```

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).

```
Nigeria to destroy one million expired COVID-19 vaccines -official ABUJA Reuters - Nigeria
will destroy around one million expired COVID-19 vaccines, Faisal Shuaib, head of the Natio
nal Primary Health Care Development Agency NPHCDA, said on Monday, adding his agency was wo
rking with drug regulator NAFDAC to set a date for their destruction.Nigeria's health minis
ter Osagie Ehanire said last week some COVID-19 doses donated by rich Western countries had
a remaining shelf life of only weeks, adding to the country's challenges in vaccinating its
people.Fewer than 4 of adults in Africa's most populous nation of over 200 million have bee
n fully vaccinated.Shuaib said the country had been accepting vaccines with short shelf liv
es from international donor nations in an attempt to use them quickly and provide some leve
l of protection for Nigerian due to vaccine scarcity in the past.Shuaib said Nigeria will n
o longer accept vaccines with a short shelf life, citing a presidential committee decision.
Last week, Reuters reported that around one million COVID-19 vaccines were estimated to hav
e expired in Nigeria last month without being used.Still, the World Health Organization's v
accine director Kate O'Brien said in a briefing on Thursday the proportion of wasted doses
is smaller in countries receiving doses through COVAX than in many high-income countries.Re
search of Malaysia's Covid-19 vaccine in proof-of-concept stage KUALA LUMPUR Dec 13: Resear
ch of the country’s Covid-19 vaccine is still in the proof-of-concept stage, the Dewan Raky
at was told on Monday Dec 13.Deputy Science, Technology and Innovation Minister Datuk Ahmad
Amzad Hashim said for the messenger ribonucleic acid mRNA vaccine, the candidate vaccine wo
```

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',), 141),
 (('to',), 77),
 (('of',), 69),
 (('and',), 58),
 (('in',), 44),
 (('vaccine',), 37),
 (('for',), 34),
 (('a',), 30),
 (('Covid19',), 29),
 (('said',), 26),
 (('doses',), 25),
 (('is',), 24),
 (('million',), 21),
 (('on',), 21),
 (('with',), 21),
 (('booster',), 21),
 (('as',), 20),
 (('vaccines',), 18),
 (('have',), 18),
 (('Dec',), 15)]
```

```
[(('of', 'the'), 20),
 (('to', 'the'), 12),
 (('in', 'the'), 10),
 (('million', 'doses'), 10),
 (('for', 'the'), 9),
 (('said', 'the'), 8),
 (('the', 'vaccine'), 7),
 (('booster', 'shots'), 7),
 (('Covid19', 'vaccine'), 6),
 (('with', 'the'), 6),
 (('total', 'of'), 6),
 (('by', 'the'), 5),
 (('and', 'the'), 5),
 (('to', 'be'), 5),
 (('Covid19', 'vaccines'), 5),
 (('doses', 'of'), 5),
 (('vaccine', 'doses'), 5),
 (('as', 'booster'), 5),
 (('booster', 'shot'), 5),
 (('the', 'Pfizer'), 5)]
```

```
[(('of', 'the', 'vaccine'), 5),
 (('said', 'in', 'a'), 4),
 (('the', 'United', 'States'), 3),
 (('million', 'doses', 'of'), 3),
 (('doses', 'of', 'the'), 3),
 (('vaccine', 'doses', 'to'), 3),
 (('in', 'the', 'country'), 3),
 (('the', 'adult', 'population'), 3),
 (('dose', 'of', 'the'), 3),
 (('to', 'the', 'increased'), 3),
 (('as', 'booster', 'shots'), 3),
 (('the', 'Pfizer', 'vaccine'), 3),
 (('one', 'million', 'expired'), 2),
 (('million', 'expired', 'COVID19'), 2),
 (('expired', 'COVID19', 'vaccines'), 2),
 (('around', 'one', 'million'), 2),
 (('of', 'adults', 'in'), 2),
 (('KUALA', 'LUMPUR', 'Dec'), 2),
 (('Science', 'Technology', 'and'), 2),
 (('Technology', 'and', 'Innovation'), 2)]
```

The tfidf representation generated is as follows:

	000	094	10	11	12	13	131	137	156	16	...	workers	working	world	worrying	would	wrote	year	years	yesterday	zainal
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.125485	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
91	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
92	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
93	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.168571	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
94	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
95	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

96 rows × 838 columns

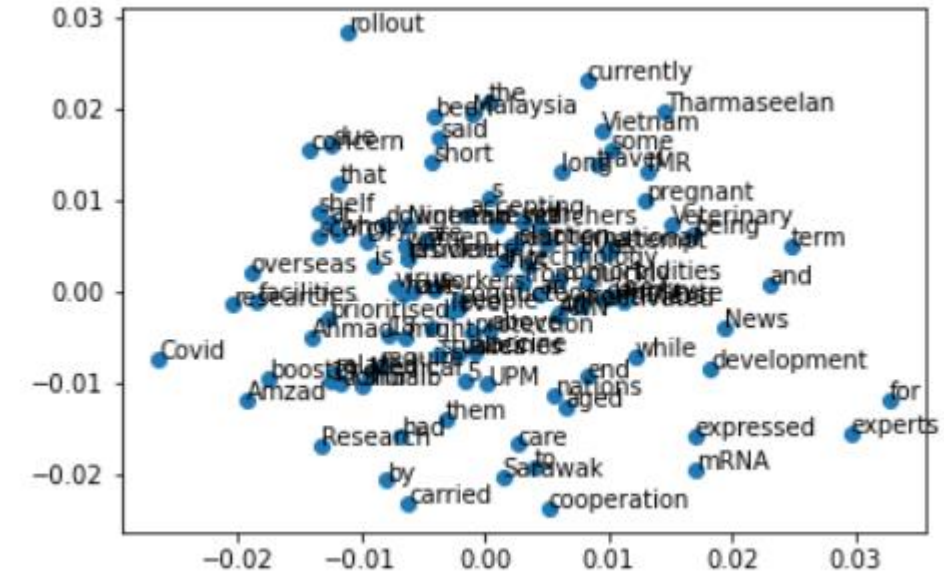
The generated word-embedding representation (either Skip-gram or CBOW) is as follows:

```
Cosine similarity between 'vaccines' and 'vaccinated' - CBOW :  0.12902793
Cosine similarity between 'vaccines' and 'booster' - CBOW :  0.23918846
Cosine similarity between 'vaccines' and 'vaccinated' - Skip Gram :  0.9063282
Cosine similarity between 'vaccines' and 'booster' - Skip Gram :  0.9706257
```

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:

```
Word2Vec(vocab=871, vector_size=100, alpha=0.025)
['the', ',', 'to', 'of', 'and', 'in', '19', 'vaccine', 'Covid', 'for', 'a', 'said', 'doses', 'The', '``', 'i
s', 'million', 'on', 'vaccines', 'as', 'with', 'booster', 'have', 'Dec', 'be', 'will', 'vaccination', 'dose
', 'Dr', 'Vietnam', "''", 'from', 'by', 'has', 'that', 'countries', 'are', 'had', 'Health', '4', 'would', 'M
alaysia', ':', 'at', 'shot', 'he', "'s", 'shots', '’', 'Pfizer', 'coverage', 'received', 'children', 'or', '
also', '11', 'total', 'it', 'aged', 'been', 'was', 'country', 'more', '24', 'rate', 'AstraZeneca', 'an', 'ad
ministered', 'Ministry', 'vaccinations', 'cases', 'population', '2', 'one', 'Minister', 'approval', 'US', '6
', 'during', '12', 'being', 'Sinovac', 'previously', 'use', 's', 'who', 'individuals', 'continue', 'this', '
He', 'United', 'people', 'A', 'than', "'", 'first', 'Nigeria', 'through', 'week', 'those', 'Friday', '‘', 's
et', 'no', 'Khairy', 'their', 'its', 'adult', 'States', 'last', 'vaccinated', 'Omicron', 'Malina', 'used', '
daily', 'This', '8', 'COVID', '1', 'six', 'increased', 'KUALA', 'given', 'second', 'years', 'LUMPUR', 'cent
', 'per', 'positive', 'Monday', 'response', 'above', '#', '18', 'down', 'over', 'new', 'Bangladesh', 'Laos',
'7', 'were', 'Australia', 'approved', 'International', 'Times', 'Straits', 'told', 'Vaccine', 'against', 'we
ll', 'time', 'can', 'most', 'Malaysian', '’', 'Datuk', 'agency', 'least', 'our', 'under', 'out', 'they', 'if
', 'only', 'BioNTech', 'fully', 'meeting', 'It', '17', 'not', 'drop', 'after', 'Affairs', 'around', 'current
', 'Zainal', 'where', 'ANN', 'Drug', 'receive', 'including', 'now', 'same', 'Indonesia', 'National', 'other
', 'both', 'need', 'We', 'restrictions', 'which', 'help', 'expired', 'travel', 'government', 'elections', 'H
ome', '5', 'contribution', 'process', 'minister', 'could', 'SOP', 'reported', 'adding', 'On', 'Technology',
'Oct', 'Innovation', 'case', 'start', '97', 'health', 'adults', '10', 'Thursday', '3', 'mark', 'vaccinating
```

```
[-5.20641031e-03  2.51658214e-03  5.30143455e-03  1.16392216e-02
 -9.62766912e-03 -1.27916597e-02  7.03842938e-03  1.89544559e-02
 -9.24742594e-03 -6.73497422e-03  5.13416156e-03 -7.26354262e-03
 -4.53837588e-03  9.91933141e-03 -5.39898919e-03 -3.49631696e-03
  4.11933893e-03 -2.79987557e-03 -8.48553702e-03 -1.94311924e-02
  9.25236940e-03  5.76090533e-03  9.12248157e-03 -2.34624138e-03
  4.03533131e-03 -3.69611639e-03 -4.95646568e-03  3.06518609e-03
 -1.08529376e-02 -2.59022787e-03 -2.57466803e-03  1.50982360e-05
  1.05787497e-02 -1.13933599e-02 -4.90811188e-03  3.83803481e-03
  7.80204590e-03 -9.62389912e-03 -2.98540364e-03 -1.08795045e-02
 -8.07614252e-03  2.32018204e-03 -1.06576234e-02 -4.87566041e-03
  3.53888562e-03 -1.75401825e-03 -1.02844937e-02  8.20779521e-03
  7.45778764e-03  1.03856651e-02 -5.45876659e-03  2.17464496e-03
 -5.76112745e-03  1.01481099e-03  6.55412953e-03 -2.63152877e-03
  6.42313901e-03 -9.82075557e-03 -8.69524013e-03  1.07517168e-02
 -8.62697780e-05  1.06476445e-03 -3.21478536e-03 -9.94610786e-03
 -6.68784371e-03  4.91896737e-03  1.90029787e-05  8.93551204e-03
 -8.50784685e-03  6.10186579e-03  2.14049360e-03  1.10131977e-02
  2.65591033e-03 -9.63641517e-03  9.76215769e-03  3.35918134e-03
```



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