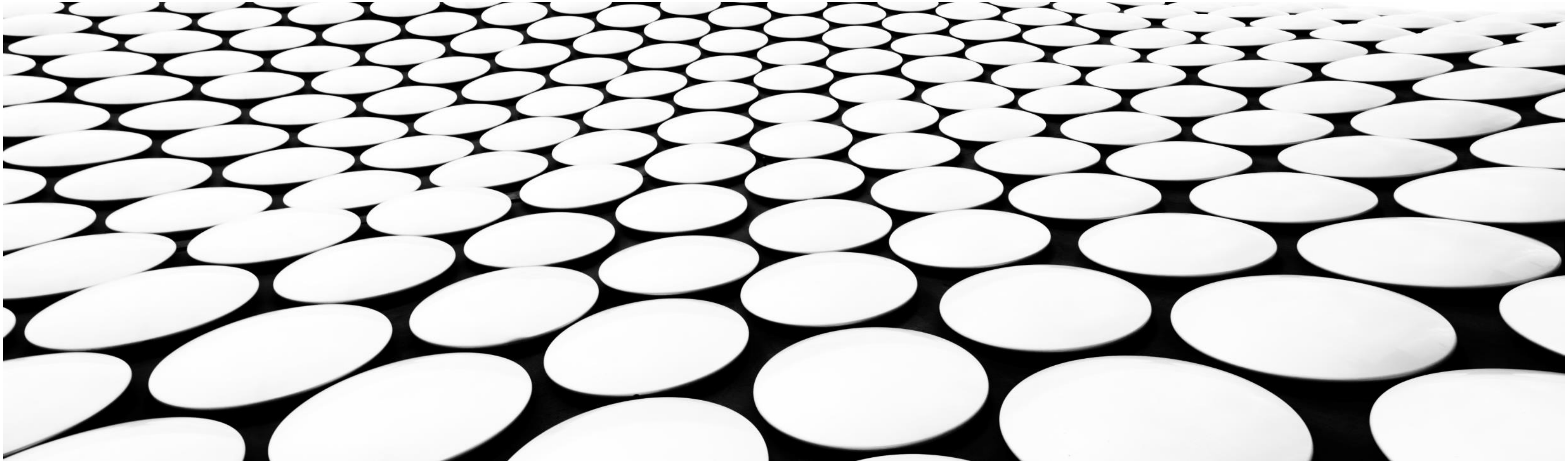
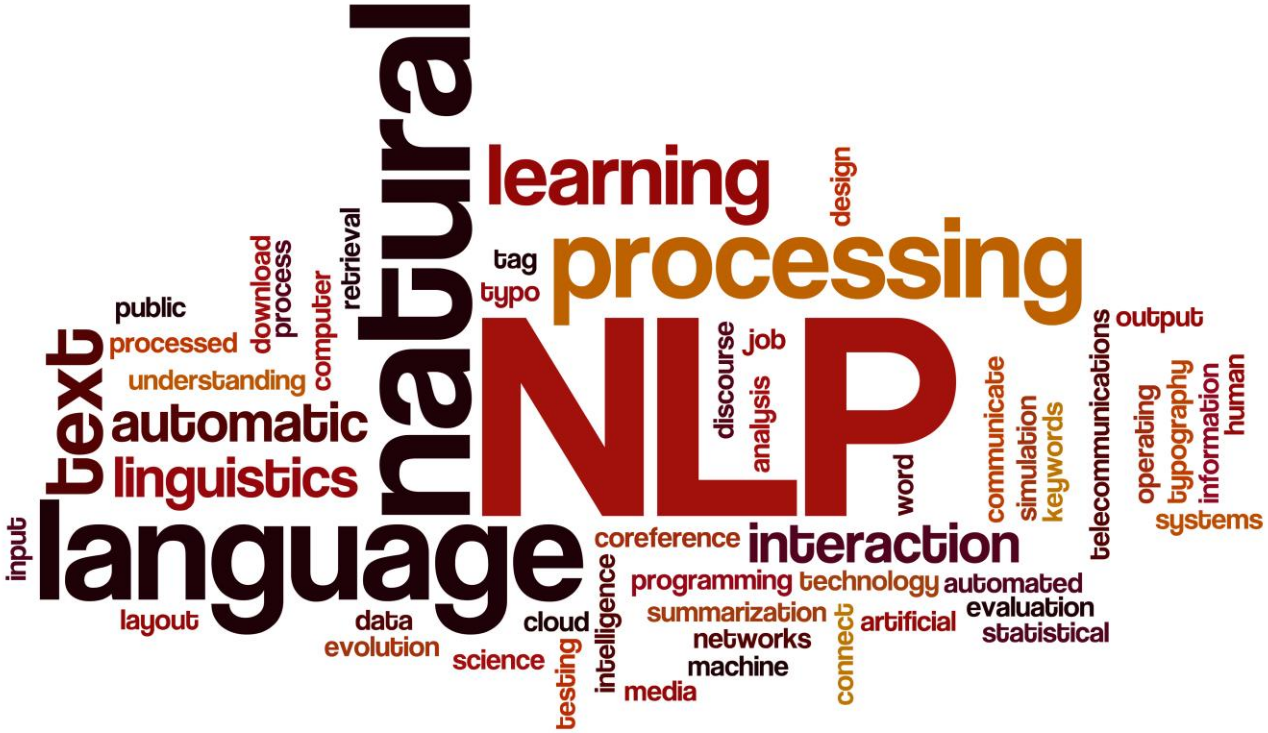
TEXT SUMMARIZATION

YANGYUE WANG



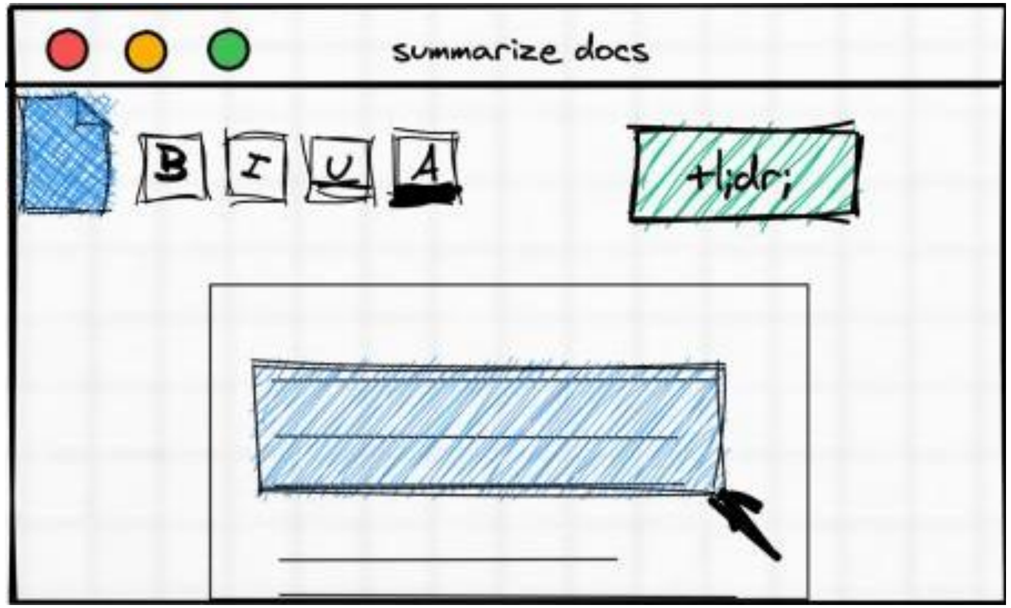
SECTIONS

- Summarization
- Word Cloud Visualization
- KMeans Clustering Sentences



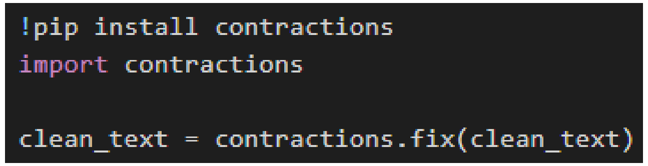
SUMMARIZATION PREPROCESSING

1. Remove HTML tags
2. Remove extra whitespaces
3. Convert accented characters to ASCII characters
4. Expand contractions
5. Remove special characters
6. Lowercase all texts
7. Remove numbers
8. Remove stop words
9. Lemmatization



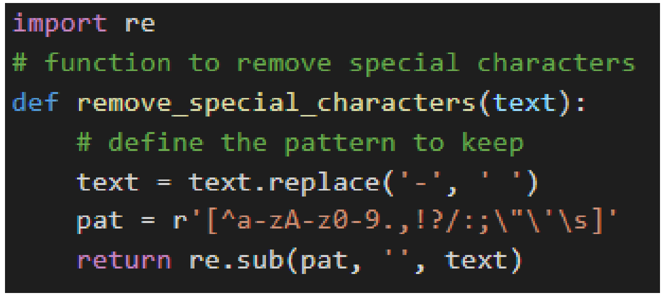
SUMMARIZATION PREPROCESSING

Remove Contractions



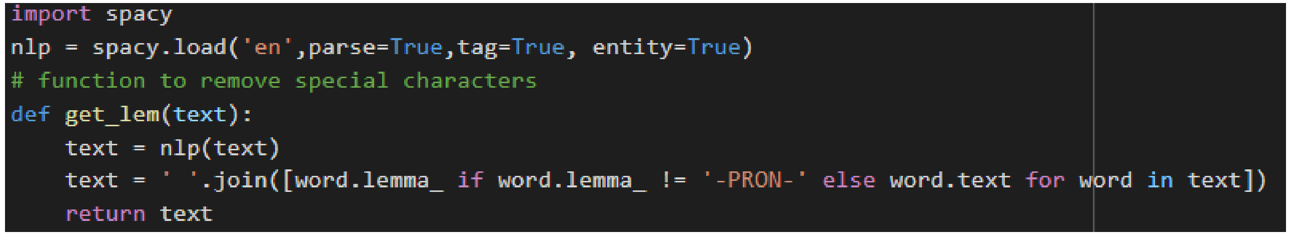
```
!pip install contractions
import contractions

clean_text = contractions.fix(clean_text)
```

Remove Special Characters



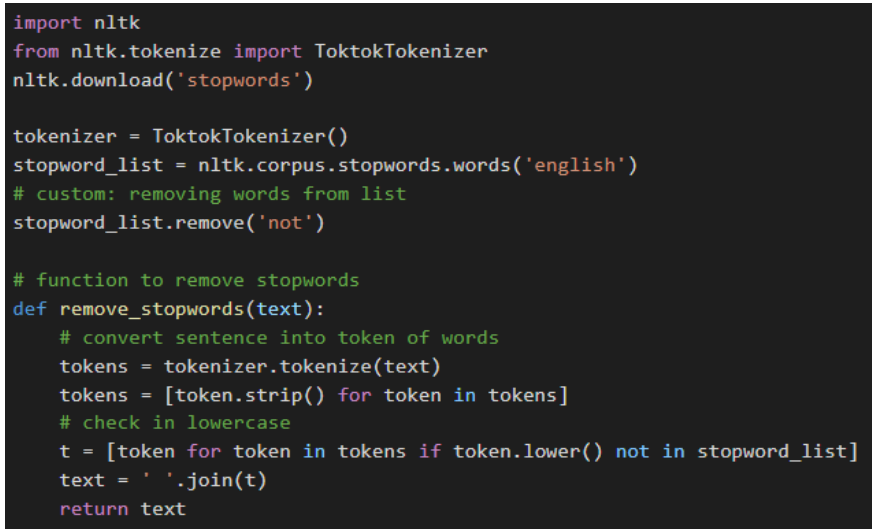
```
import re
# function to remove special characters
def remove_special_characters(text):
    # define the pattern to keep
    text = text.replace('-', ' ')
    pat = r'[^a-zA-z0-9.,!?/:;\"\'\s]'
    return re.sub(pat, '', text)
```

Lemmatization



```
import spacy
nlp = spacy.load('en',parse=True,tag=True, entity=True)
# function to remove special characters
def get_lem(text):
    text = nlp(text)
    text = ' '.join([word.lemma_ if word.lemma_ != '-PRON-' else word.text for word in text])
    return text
```

Remove Stop words

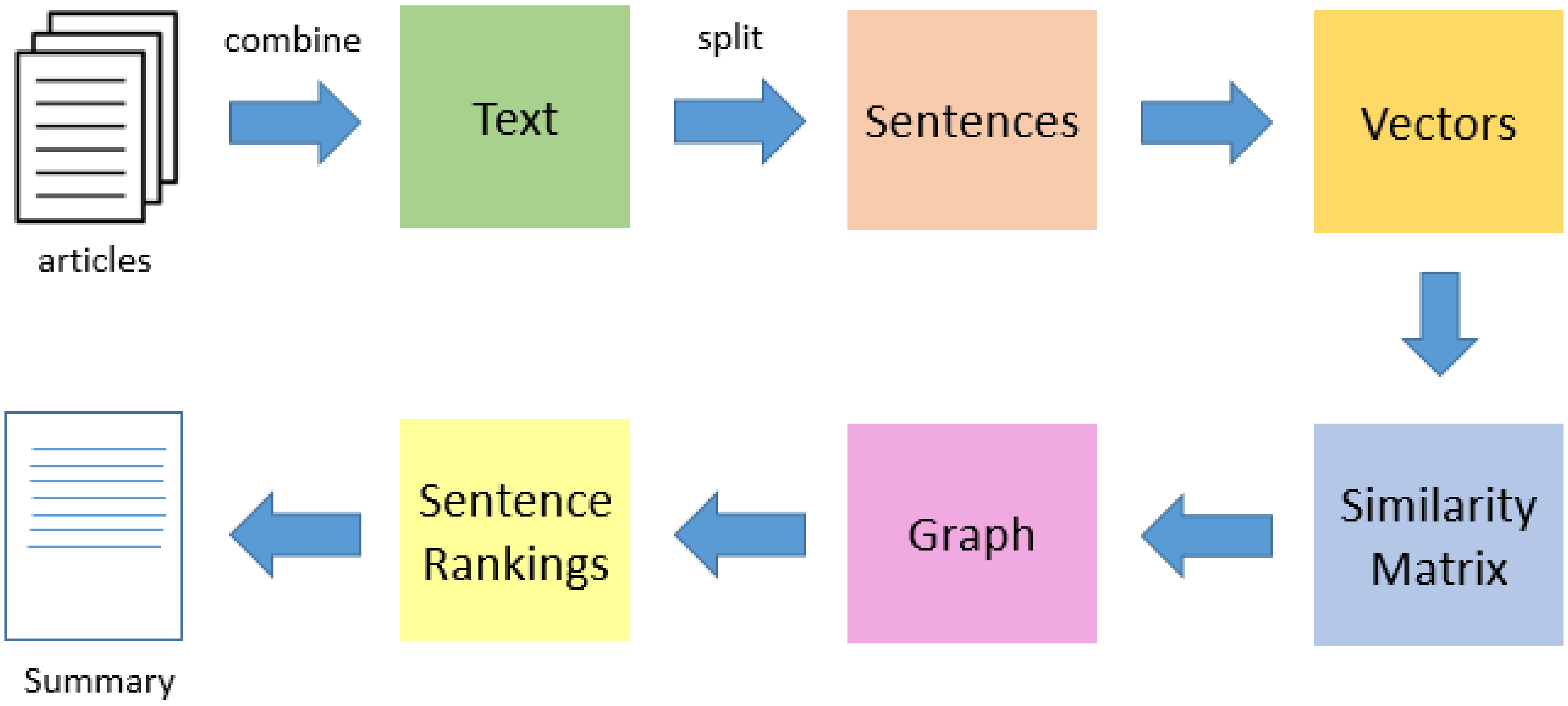


```
import nltk
from nltk.tokenize import ToktokTokenizer
nltk.download('stopwords')

tokenizer = ToktokTokenizer()
stopword_list = nltk.corpus.stopwords.words('english')
# custom: removing words from list
stopword_list.remove('not')

# function to remove stopwords
def remove_stopwords(text):
    # convert sentence into token of words
    tokens = tokenizer.tokenize(text)
    tokens = [token.strip() for token in tokens]
    # check in lowercase
    t = [token for token in tokens if token.lower() not in stopword_list]
    text = ' '.join(t)
    return text
```

SUMMARIZATION

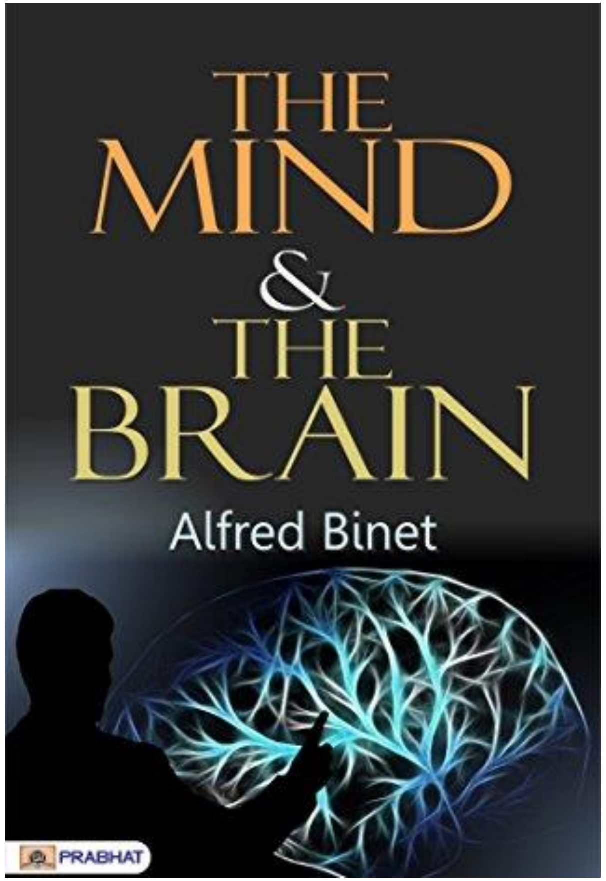
- **Step 1 - Read text and tokenize**
- **Step 2 - Generate similarity matrix across sentences**
- **Step 3 - Rank sentences in similarity matrix**
- **Step 4 - Sort the rank and pick top sentences**
- **Step 5 - Output the summarize text**

SUMMARIZATION ALGORITHM

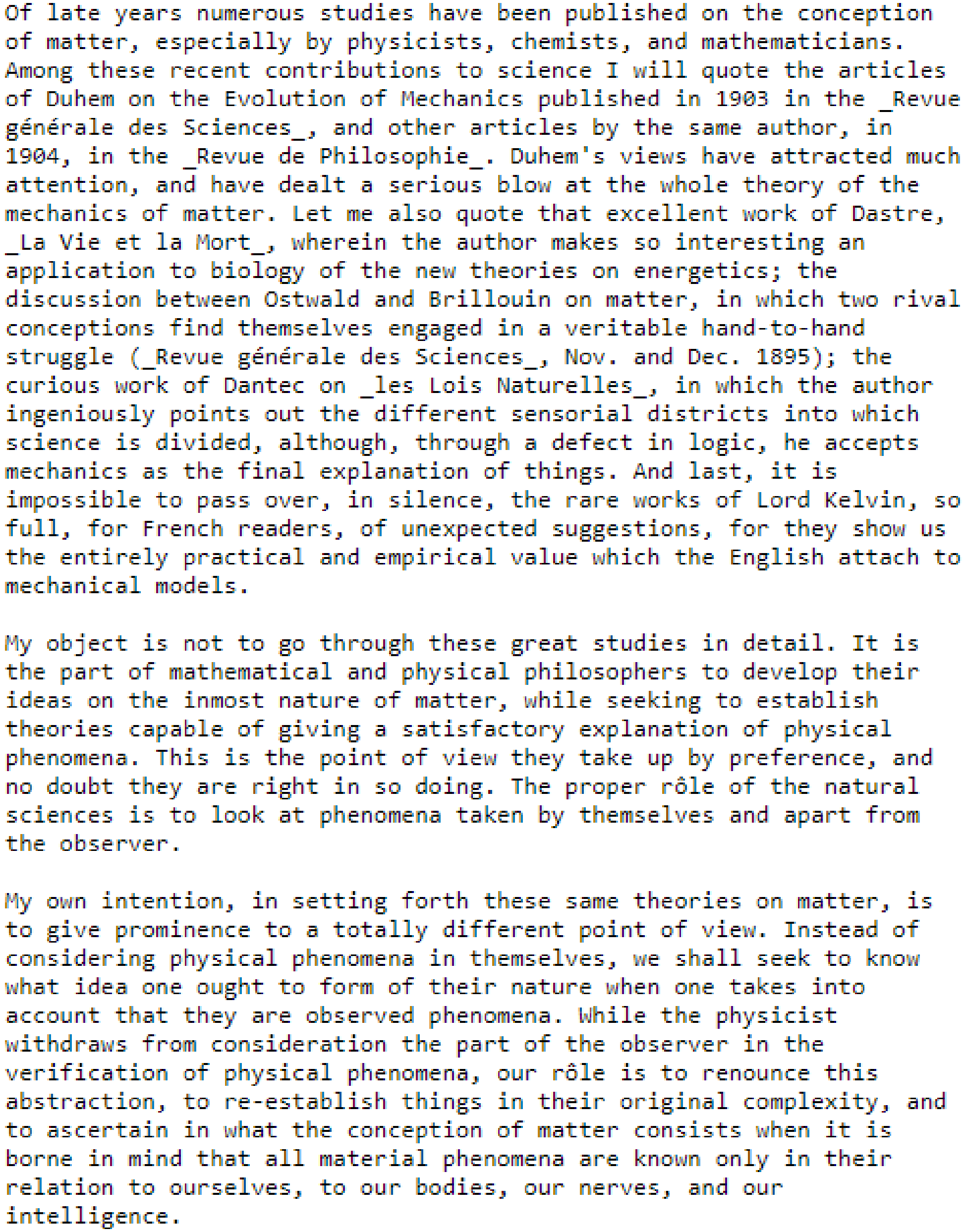


[1] https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

SUMMARIZATION



https://www.gutenberg.org/

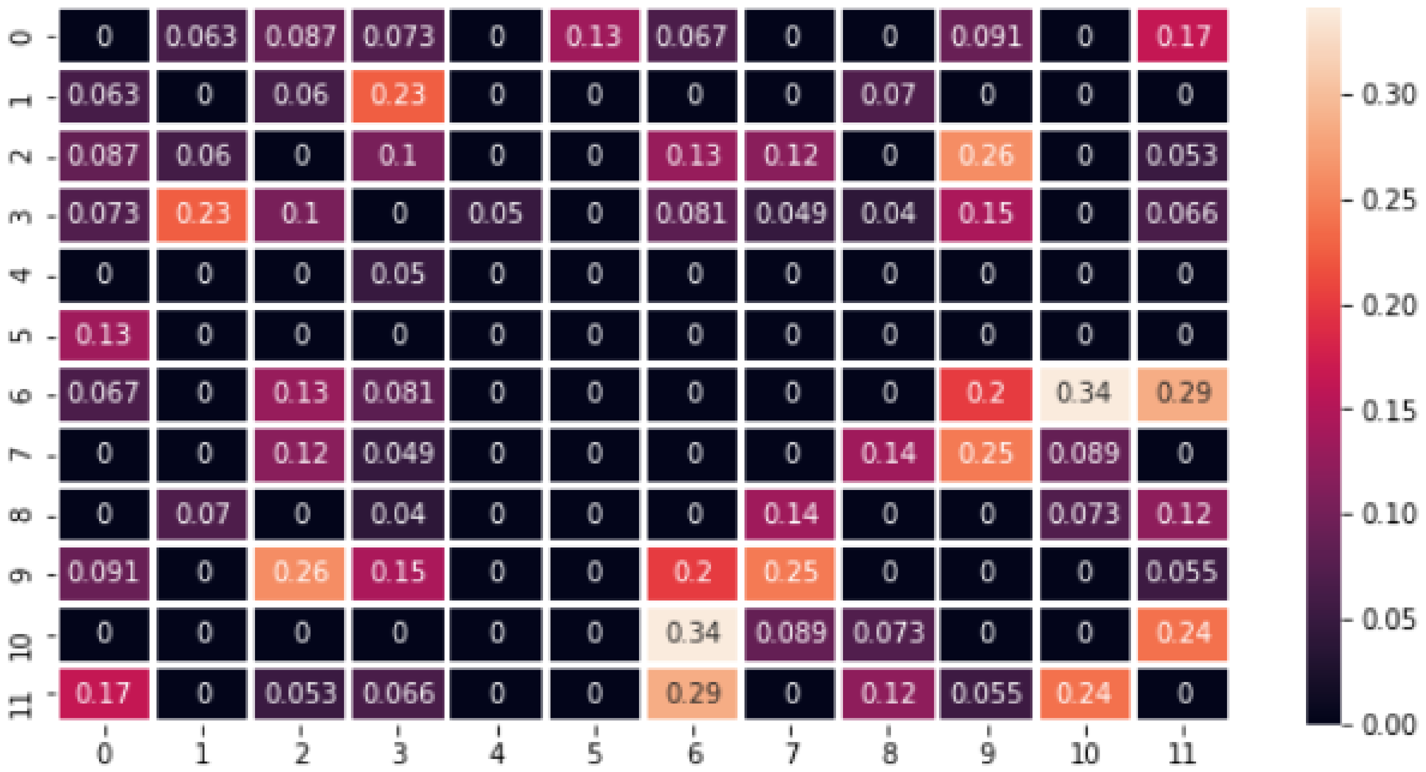


Of late years numerous studies have been published on the conception of matter, especially by physicists, chemists, and mathematicians. Among these recent contributions to science I will quote the articles of Duhem on the Evolution of Mechanics published in 1903 in the _Revue générale des Sciences_, and other articles by the same author, in 1904, in the _Revue de Philosophie_. Duhem's views have attracted much attention, and have dealt a serious blow at the whole theory of the mechanics of matter. Let me also quote that excellent work of Dastre, _La Vie et la Mort_, wherein the author makes so interesting an application to biology of the new theories on energetics; the discussion between Ostwald and Brillouin on matter, in which two rival conceptions find themselves engaged in a veritable hand-to-hand struggle (_Revue générale des Sciences_, Nov. and Dec. 1895); the curious work of Dantec on _les Lois Naturelles_, in which the author ingeniously points out the different sensorial districts into which science is divided, although, through a defect in logic, he accepts mechanics as the final explanation of things. And last, it is impossible to pass over, in silence, the rare works of Lord Kelvin, so full, for French readers, of unexpected suggestions, for they show us the entirely practical and empirical value which the English attach to mechanical models.

My object is not to go through these great studies in detail. It is the part of mathematical and physical philosophers to develop their ideas on the inmost nature of matter, while seeking to establish theories capable of giving a satisfactory explanation of physical phenomena. This is the point of view they take up by preference, and no doubt they are right in so doing. The proper rôle of the natural sciences is to look at phenomena taken by themselves and apart from the observer.

My own intention, in setting forth these same theories on matter, is to give prominence to a totally different point of view. Instead of considering physical phenomena in themselves, we shall seek to know what idea one ought to form of their nature when one takes into account that they are observed phenomena. While the physicist withdraws from consideration the part of the observer in the verification of physical phenomena, our rôle is to renounce this abstraction, to re-establish things in their original complexity, and to ascertain in what the conception of matter consists when it is borne in mind that all material phenomena are known only in their relation to ourselves, to our bodies, our nerves, and our intelligence.

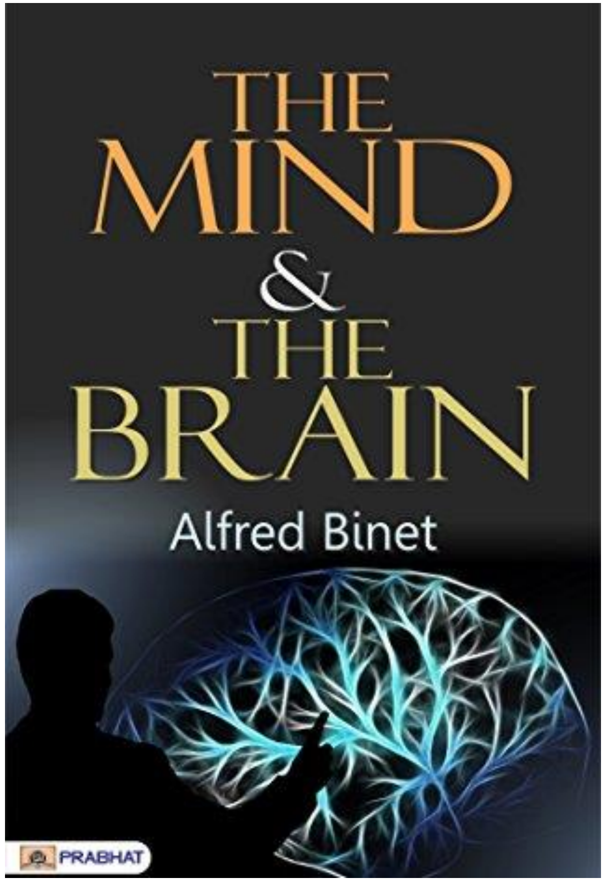
SIMILARITY MATRIX



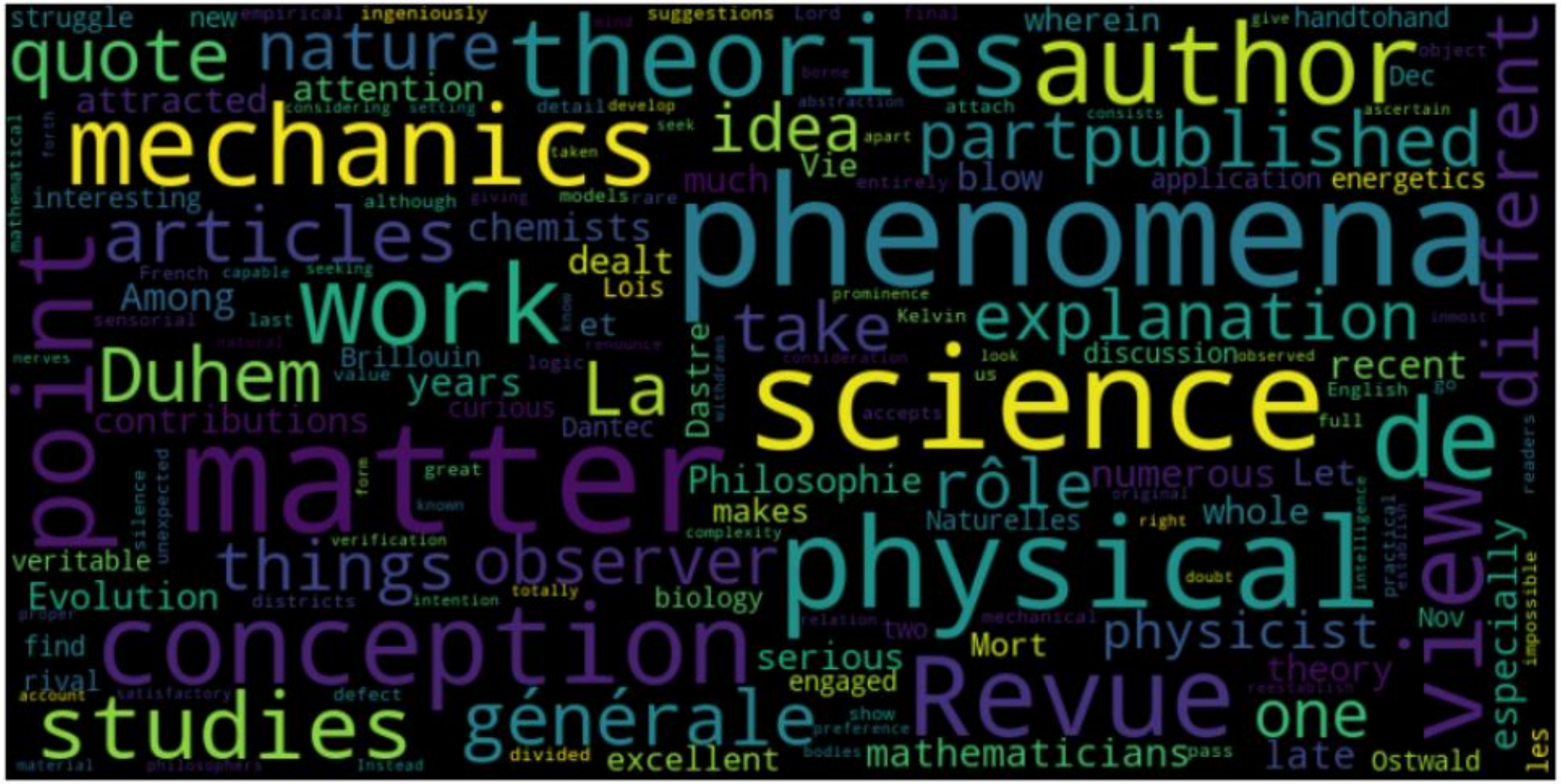
	0	1	2	3	4	5	6	7	8	9	10	11
0	0	0.063	0.087	0.073	0	0.13	0.067	0	0	0.091	0	0.17
1	0.063	0	0.06	0.23	0	0	0	0	0.07	0	0	0
2	0.087	0.06	0	0.1	0	0	0.13	0.12	0	0.26	0	0.053
3	0.073	0.23	0.1	0	0.05	0	0.081	0.049	0.04	0.15	0	0.066
4	0	0	0	0.05	0	0	0	0	0	0	0	0
5	0.13	0	0	0	0	0	0	0	0	0	0	0
6	0.067	0	0.13	0.081	0	0	0	0	0	0.2	0.34	0.29
7	0	0	0.12	0.049	0	0	0	0	0.14	0.25	0.089	0
8	0	0.07	0	0.04	0	0	0	0.14	0	0	0.073	0.12
9	0.091	0	0.26	0.15	0	0	0.2	0.25	0	0	0	0.055
10	0	0	0	0	0	0	0.34	0.089	0.073	0	0	0.24
11	0.17	0	0.053	0.066	0	0	0.29	0	0.12	0.055	0.24	0

SUMMARIZATION RESULT

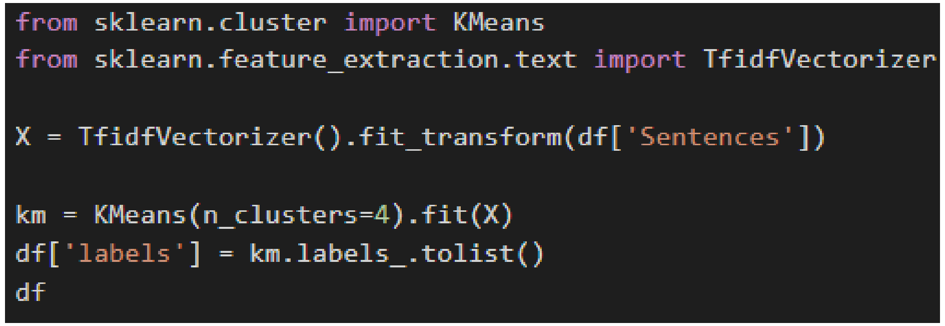
- it be the part of mathematical and physical philosopher to develop their idea on the inmost nature of matter while seek to establish theory capable of give a satisfactory explanation of physical phenomenon. my own intention in set forth these same theory on matter be to give prominence to a totally different point of view. while the physicist withdraw from consideration the part of the observer in the verification of physical phenomenon our role be to renounce this abstraction to re establish thing in their original complexity and to ascertain in what the conception of matter consist when it be bear in mind that all material phenomenon be know only in their relation to ourselves to our body our nerve and our intelligence



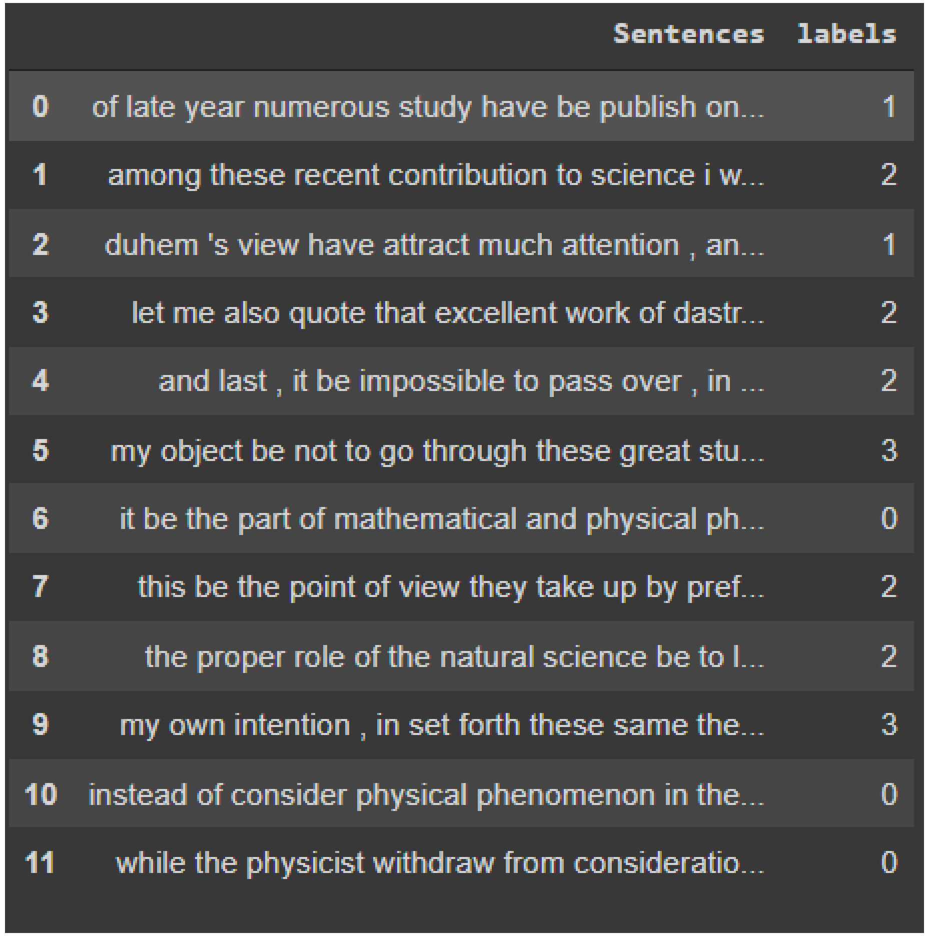
WORD CLOUD


struggle
new
empirical
ingeniously
mind
suggestions
Lord
final
wherein
give
handtohand
quote
nature
theories
author
object
borne
Dec
attracted
attention
considering
setting
detail
develop
abstraction
attach
consists
ascertain
mechanics
seek
idea
apart
part
published
different
mathematical
forth
taken
Vie
much
entirely
blow
application
energetics
interesting
although
giving
models
rare
phenomena
articles
chemists
point
dealt
French
capable
seeking
Lois
prominence
Among
work
know
et
take
Kelvin
explanation
inmost
sensorial
last
natural
renounce
nerves
Brillouin
logic
Dastre
withdrawn
consideration
look
discussion
observed
recent
value
Duhem
years
La
us
science
English
go
curious
accepts
full
contributions
Dantec
de
readers
matter
great
form
Philosophie
rôle
numerous
Let
original
known
silence
unexpected
makes
Naturelles
right
whole
intelligence
practical
establish
verification
complexity
veritable
things
observer
physical
doubt
view
especially
impossible
Evolution
districts
intention
totally
biology
proper
relation
two
mechanical
Nov
find
conception
physicist
Mort
serious
theory
rival
engaged
reestablish
account
satisfactory
defect
studies
générale
show
Revue
one
preference
divided
excellent
bodies
mathematicians
pass
late
Ostwald
les
material
philosophers
Instead

KMEANS CLUSTERING

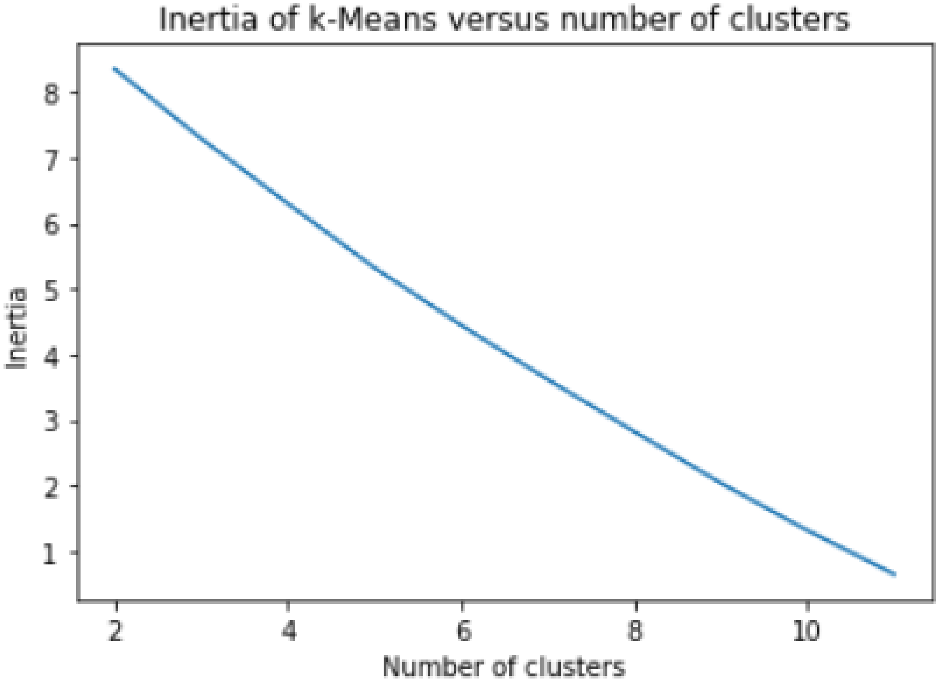


```
from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import TfidfVectorizer

X = TfidfVectorizer().fit_transform(df['Sentences'])

km = KMeans(n_clusters=4).fit(X)
df['labels'] = km.labels_.tolist()
df
```



	Sentences	labels
0	of late year numerous study have be publish on...	1
1	among these recent contribution to science i w...	2
2	duhem 's view have attract much attention , an...	1
3	let me also quote that excellent work of dastr...	2
4	and last , it be impossible to pass over , in ...	2
5	my object be not to go through these great stu...	3
6	it be the part of mathematical and physical ph...	0
7	this be the point of view they take up by pref...	2
8	the proper role of the natural science be to l...	2
9	my own intention , in set forth these same the...	3
10	instead of consider physical phenomenon in the...	0
11	while the physicist withdraw from consideratio...	0

ELBOW METHOD



TAKEAWAYS

- Use abstractive instead of extractive summarization
- Better preprocessing
- De-lemmatize the summarized text
- Speed up the algorithm for large text
- Better similarity matrix algorithm