

```
In [ ]: !pip install pyLDAvis

Requirement already satisfied: pyLDAvis in /usr/local/lib/python3.7/dist-packages (3.2.1)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (0.22.2.post1)
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (57.4.0)
Requirement already satisfied: pandas>=1.2.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.3.3)
Requirement already satisfied: gensim in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (3.6.0)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (2.11.3)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.4.1)
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.21.2)
Requirement already satisfied: funcy in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.16)
Requirement already satisfied: future in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (0.16.0)
Requirement already satisfied: sklearn in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (0.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (1.0.1)
Requirement already satisfied: numexpr in /usr/local/lib/python3.7/dist-packages (from pyLDAvis) (2.7.3)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=1.2.0->pyLDAvis) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas>=1.2.0->pyLDAvis) (2018.9)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.7.3->pandas>=1.2.0->pyLDAvis) (1.15.0)
Requirement already satisfied: smart-open>=1.2.1 in /usr/local/lib/python3.7/dist-packages (from gensim->pyLDAvis) (5.1.0)
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/dist-packages (from Jinja2->pyLDAvis) (2.0.1)
```

```
In [ ]: from google.colab import drive
drive.mount('/content/gdrive')

#changing the working directory
# %cd /content/gdrive/My Drive/Dataset (Colab)
#Check the present working directory using pwd command
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
```

```
In [ ]: # Importing module
import pandas as pd
import os
# Read data into papers
papers = pd.read_csv('/content/gdrive/MyDrive/Dataset (Colab)/papers1.csv')
# Print head
papers.head()
# papers.info()
```

```
Out[ ]: 
```

	id	year	title	event_type	pdf_name	abstract	paper_text
0	1	1987	Self-Organization of Associative Database and ...	NaN	1-self-organization-of-associative-database-an...	Abstract Missing	767\n\nSELF-ORGANIZATION OF ASSOCIATIVE DATABA...
1	10	1987	A Mean Field Theory of Layer IV of Visual Cort...	NaN	10-a-mean-field-theory-of-layer-iv-of-visual-c...	Abstract Missing	683\n\nA MEAN FIELD THEORY OF LAYER IV OF VISU...
2	100	1988	Storing Covariance by the Associative Long-Ter...	NaN	100-storing-covariance-by-the-associative-long...	Abstract Missing	394\n\nSTORING COVARIANCE BY THE ASSOCIATIVE EV...
3	1000	1994	Bayesian Query Construction for Neural Network...	NaN	1000-bayesian-query-construction-for-neural-ne...	Abstract Missing	Bayesian Query Construction for NeuralNetwork...
4	1001	1994	Neural Network Ensembles, Cross Validation, an...	NaN	1001-neural-network-ensembles-cross-validation...	Abstract Missing	Neural Network Ensembles, Cross/Validation, &

```
In [ ]: # Remove the columns
papers = papers.drop(columns=['id', 'event_type', 'pdf_name'], axis=1)
# Print out the first rows of papers
papers.head()
```

```
Out[ ]: 
```

	year	title	abstract	paper_text
0	1987	Self-Organization of Associative Database and ...	Abstract Missing	767\n\nSELF-ORGANIZATION OF ASSOCIATIVE DATABA...
1	1987	A Mean Field Theory of Layer IV of Visual Cort...	Abstract Missing	683\n\nA MEAN FIELD THEORY OF LAYER IV OF VISU...
2	1988	Storing Covariance by the Associative Long-Ter...	Abstract Missing	394\n\nSTORING COVARIANCE BY THE ASSOCIATIVE EV...
3	1994	Bayesian Query Construction for Neural Network...	Abstract Missing	Bayesian Query Construction for NeuralNetwork...
4	1994	Neural Network Ensembles, Cross Validation, an...	Abstract Missing	Neural Network Ensembles, Cross/Validation, a...

```
In [ ]: # Load the regular expression library
import re
# Remove punctuation
papers['paper_text_processed'] = \
papers['paper_text'].map(lambda x: re.sub('[\.\!\?]', '', x))

# print(papers['paper_text_processed'])

# # Convert the titles to lowercase
papers['paper_text_processed'] = \
papers['paper_text_processed'].map(lambda x: x.lower())
# # Print out the first rows of papers
papers['paper_text_processed'].head()
```

```
Out[ ]: 
```

```
0    767\n\nself-organization of associative databa...
1    683\n\nA mean field theory of layer iv of visu...
2    394\n\nstoring covariance by the associative v...
3    bayesian query construction for neuralnetwor...
4    neural network ensembles cross/validation and...
Name: paper_text_processed, dtype: object
```

```
In [ ]: # Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = '\n'.join(list(papers['paper_text_processed'].values))
# Print(long_string(0:30))
# Create a WordCloud object
wordcloud = WordCloud(background_color="black", max_words=200, contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Out[ ]: 
```



```
In [ ]: import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use'])
def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))

def remove_stopwords(texts):
    docs=[]
    # for doc in texts:
    #     docs=[]
    #     for word in simple_preprocess(str(doc)):
    #         if word not in stop_words:
    #             docs.append(word)
    #     final=""
    #     docs.append(final)
    #     return docs
    return [word for word in simple_preprocess(str(doc))
            if word not in stop_words for doc in texts]
```

```
data = papers.paper_text_processed.values.tolist()
# print(data[1]) # contains all the rows in a list as an individual element
# print(papers['paper_text_processed'])
data_words = list(sent_to_words(data)) #it saves a sentence as a list by treating every word as an element of

# print((data_words[1]))
# remove stop words
# data_words = remove_stopwords(data_words)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

```
In [ ]: import gensim.corpora as corpora
# Create Dictionary
id2word = corpora.Dictionary(data_words)
print(id2word)
# Create Corpus
texts = data_words
print(texts)
# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
# View
print(corpus[1:1][0][:30])
```


[illegible]

[illegible]

[illegible]

