

# Project Report

## Text Document Clustering

Name: Keshav Anand

Course: AI and ML

(Batch-4)

Duration: 12 months

Problem Statement: Text document clustering using PLSA.

### Prerequisites

---

What things you need to install the software and how to install them:

Python 3.6 This setup requires that your machine has latest version of python. The following url <https://www.python.org/downloads/> can be referred to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: <https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/>. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.

Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url <https://www.anaconda.com/download/> You will also need to download and install below 3 packages after you install either python or anaconda from the steps above Sklearn (scikit-learn) numpy scipy if you have chosen to install python 3.6 then run below commands in command prompt/terminal to install these packages `pip install -U scikit-learn` `pip install numpy` `pip install scipy` if you have chosen to install anaconda then run below commands in anaconda prompt to install these packages `conda install -c scikit-learn` `conda install -c anaconda numpy` `conda install -c anaconda scipy`

### Dataset used:

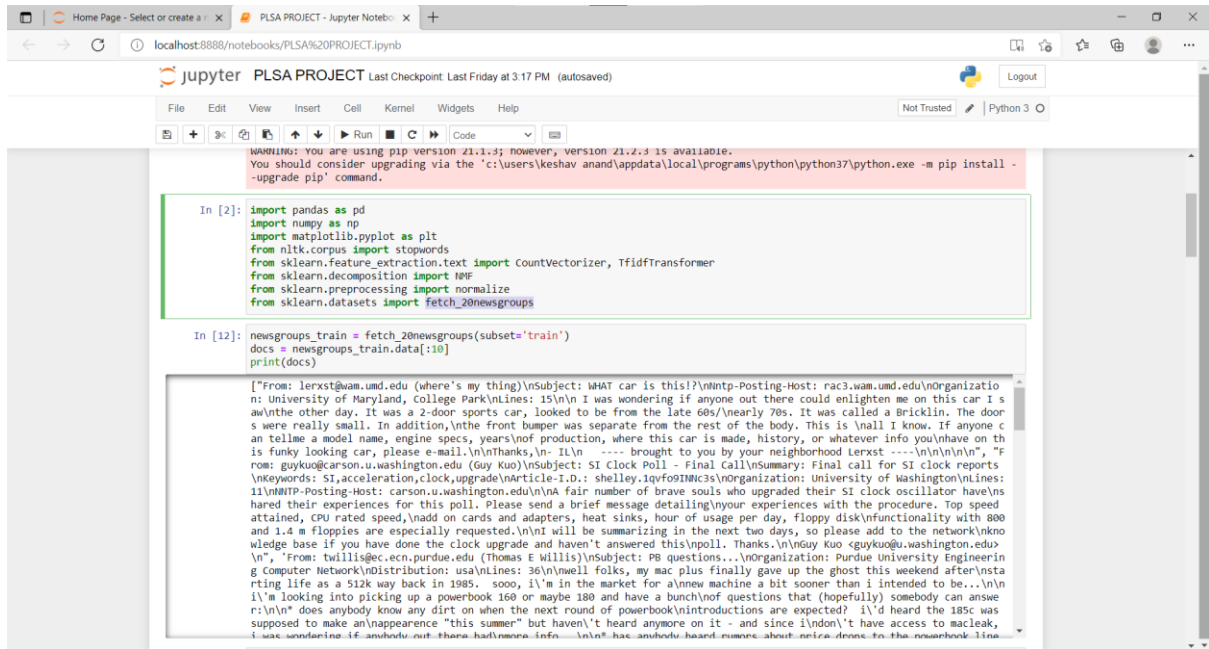
---

The dataset used is fetch\_20newsgroups dataset which is an in-built dataset available in scikit-learn library.

## Method used for Clustering:

### PLSA (Probabilistic Latent Semantic Analysis)

#### Screenshots of Source Code and Output:

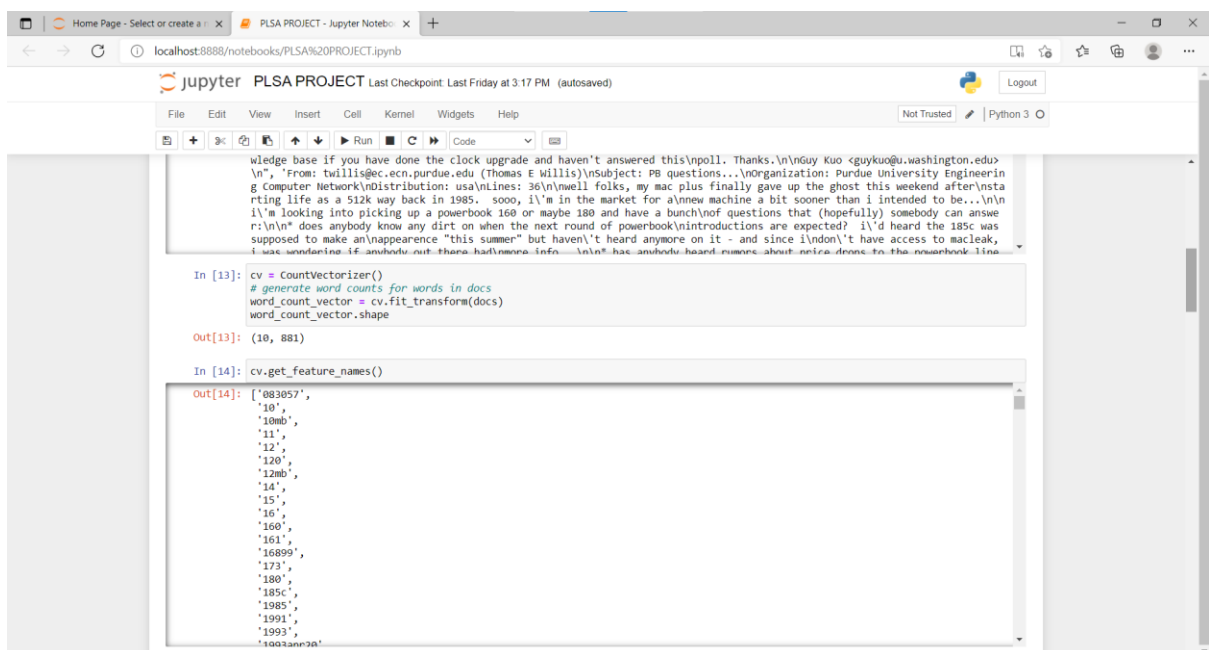


```
WARNING: You are using pip version 21.1.3; however, version 21.2.3 is available.
You should consider upgrading via the 'c:\users\keshav anand\appdata\local\programs\python\python37\python.exe -m pip install --upgrade pip' command.

In [2]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.decomposition import NMF
from sklearn.preprocessing import normalize
from sklearn.datasets import fetch_20newsgroups

In [12]: newsgroups_train = fetch_20newsgroups(subset='train')
docs = newsgroups_train.data[:10]
print(docs)

["From: lerxst@wam.umd.edu (where's my thing)\nSubject: WHAT car is this?\nntp-Posting-Host: rac3.wam.umd.edu\nOrganization: University of Maryland, College Park\nLines: 15\n\nI was wondering if anyone out there could enlighten me on this car I saw the other day. It was a 2-door sports car, looked to be from the late 60s/nearly 70s. It was called a Bricklin. The doors were really small. In addition, the front bumper was separate from the rest of the body. This is all I know. If anyone can tell me a model name, engine specs, years of production, where this car is made, history, or whatever info you have on this funky looking car, please e-mail.\n\nThanks,\n\nIL\n\n---- brought to you by your neighborhood Lerxst ----\n\n\nFrom: guykuo@carson.u.washington.edu (Guy Kuo)\nSubject: SI Clock Poll - Final Call\nSummary: Final call for SI clock reports\nKeywords: SI, acceleration, clock, upgrade\nArticle-ID.: shelly.1qvf09IMWC3\nOrganization: University of Washington\nLines: 11\nntp-Posting-Host: carson.u.washington.edu\nA fair number of brave souls who upgraded their SI clock oscillator have shared their experiences for this poll. Please send a brief message detailing your experiences with the procedure. Top speed attained, CPU rated speed, nadd on cards and adapters, heat sinks, hour of usage per day, floppy disk functionality with 800 and 1.4 m floppies are especially requested.\n\nI will be summarizing in the next two days, so please add to the network knowledge base if you have done the clock upgrade and haven't answered this poll. Thanks.\n\nGuy Kuo <guykuo@u.washington.edu>\n\n\nFrom: twillis@ec.ecn.purdue.edu (Thomas E Willis)\nSubject: P8 questions...\nOrganization: Purdue University Engineering Computer Network\nDistribution: usa\nLines: 36\n\nwell folks, my mac plus finally gave up the ghost this weekend after starting life as a 512k way back in 1985. sooo, i'm in the market for a new machine a bit sooner than i intended to be...\n\ni'm looking into picking up a powerbook 160 or maybe 180 and have a bunch of questions that (hopefully) somebody can answer:\n\n1) does anybody know any dirt on when the next round of powerbook introductions are expected? i'd heard the 185c was supposed to make an appearance "this summer" but haven't heard anymore on it - and since i don't have access to maclean, i was wondering if anybody out there had more info...\n\n2) has anybody heard rumors about new designs for the powerbook line...
```



```
wledge base if you have done the clock upgrade and haven't answered this poll. Thanks.\n\nGuy Kuo <guykuo@u.washington.edu>\n\n\nFrom: twillis@ec.ecn.purdue.edu (Thomas E Willis)\nSubject: P8 questions...\nOrganization: Purdue University Engineering Computer Network\nDistribution: usa\nLines: 36\n\nwell folks, my mac plus finally gave up the ghost this weekend after starting life as a 512k way back in 1985. sooo, i'm in the market for a new machine a bit sooner than i intended to be...\n\ni'm looking into picking up a powerbook 160 or maybe 180 and have a bunch of questions that (hopefully) somebody can answer:\n\n1) does anybody know any dirt on when the next round of powerbook introductions are expected? i'd heard the 185c was supposed to make an appearance "this summer" but haven't heard anymore on it - and since i don't have access to maclean, i was wondering if anybody out there had more info...\n\n2) has anybody heard rumors about new designs for the powerbook line...
```

```
In [13]: cv = CountVectorizer()
# generate word counts for words in docs
word_count_vector = cv.fit_transform(docs)
word_count_vector.shape

Out[13]: (10, 881)

In [14]: cv.get_feature_names()

Out[14]: ['083057',
'10',
'10mb',
'11',
'12',
'120',
'12mb',
'14',
'15',
'16',
'160',
'161',
'16899',
'173',
'180',
'185c',
'185',
'1991',
'1993',
'1993an',
'20']
```

```
localhost:8888/notebooks/PLSA%20PROJECT.ipynb

jupyter PLSA PROJECT Last Checkpoint: Last Friday at 3:17 PM (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [19]: tfidf_transformer = TfidfTransformer()
tfidf_vector = tfidf_transformer.fit_transform(word_count_vector)
feature_names = cv.get_feature_names()
# get tfidf vector for first document
first_document_vector = tfidf_vector[2]
# print the vector
df = pd.DataFrame(first_document_vector.T.todense(), index = feature_names, columns = ["tfidf"])
df.sort_values(by = ["tfidf"], ascending = False)
df = pd.read_excel("abcnews-date-text.xlsx")
data_text = df[["headline_text"]].astype("str")
data_text.shape

Out[19]: (1048575, 1)

In [34]: data_text = data_text.loc[1:100000, :]

In [35]: import nltk
nltk.download("stopwords")

[nltk_data] Downloading package stopwords to C:\Users\KESHAV
[nltk_data] ANAND\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!

Out[35]: True
```

```
localhost:8888/notebooks/PLSA%20PROJECT.ipynb

jupyter PLSA PROJECT Last Checkpoint: Last Friday at 3:17 PM (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [42]: stopw = stopwords.words("english")
def stopwords_remove(x):
    terms = x.split(' ')
    terms = [w for w in terms if w not in stopw]
    sentence = ' '.join(terms)
    return sentence
data_text["Refined_headlines"] = data_text["headline_text"].apply(lambda x: stopwords_remove(x))

In [43]: data_text.head()
#stopw

Out[43]:
  headline_text  Refined_headlines
1 act fire witnesses must be aware of defamation act fire witnesses must aware defamation
2 a g calls for infrastructure protection summit g calls infrastructure protection summit
3 air nz staff in aust strike for pay rise air nz staff aust strike pay rise
4 air nz strike to affect australian travellers air nz strike affect australian travellers
5 ambitious olsson wins triple jump ambitious olsson wins triple jump

In [44]: def word_count(x):
    terms = x.split()
    return len(terms)
data_text["word_count"] = data_text["Refined_headlines"].apply(lambda x: word_count(x))

In [45]: data_text.head()

Out[45]:
  headline_text  Refined_headlines  word_count
1 act fire witnesses must be aware of defamation act fire witnesses must aware defamation 6
2 a g calls for infrastructure protection summit g calls infrastructure protection summit 5
3 air nz staff in aust strike for pay rise air nz staff aust strike pay rise 7
```

