Report

PLSA: Text Document Clustering

Name: Keya Shukla

- **Problem Statement**

  Perform topic modelling using the 20 Newsgroup dataset (the dataset is also available in sklearn datasets sub-module). Perform the required data cleaning steps using NLP and then model the topics

  1. Using Latent Dirichlet Allocation (LDA).
  2. Using Probabilistic Latent Semantic Analysis (PLSA)

- **Prerequisites**

  - Software:
    - Python 3 (Use anaconda as your python distributor as well)

  - Tools:
    - Pandas
    - Sklearn

  - Dataset: Inbuilt Sklearn dataset 20newsgroup

- **Method Used**

  PLSA or Probabilistic Latent Semantic Analysis is a technique used to model information under a probabilistic framework. It is a statistical technique for the analysis of two-mode and co-occurrence data. PLSA characterizes each word in a document as a sample from a mixture model, where mixture components are conditionally independent multinomial distributions. Its main goal is to model cooccurrence information under a probabilistic framework in order to discover the underlying semantic structure of the data.

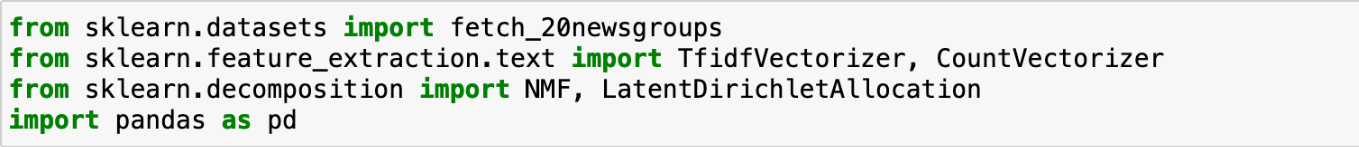
  Probabilistic latent semantic analysis (PLSA), also known as probabilistic latent semantic indexing (PLSI, especially in information retrieval circles) is a statistical technique for the analysis of two-mode and co-occurrence data. In effect, one can derive a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis, from which PLSA evolved.

  1. PLSA may be used in a discriminative setting, via Fisher kernels.

2. PLSA has applications in information retrieval and filtering, natural language processing, machine learning from text, and related areas.

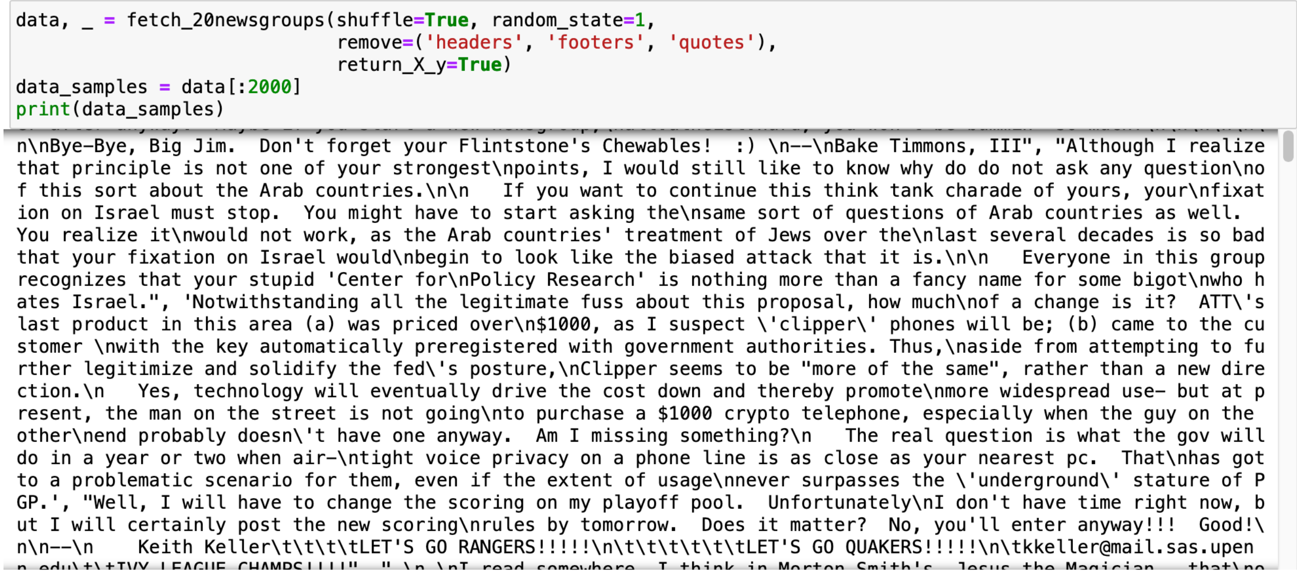
- **Implementation:**

1. Load all required libraries



```
from sklearn.datasets import fetch_20newsgroups
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.decomposition import NMF, LatentDirichletAllocation
import pandas as pd
```

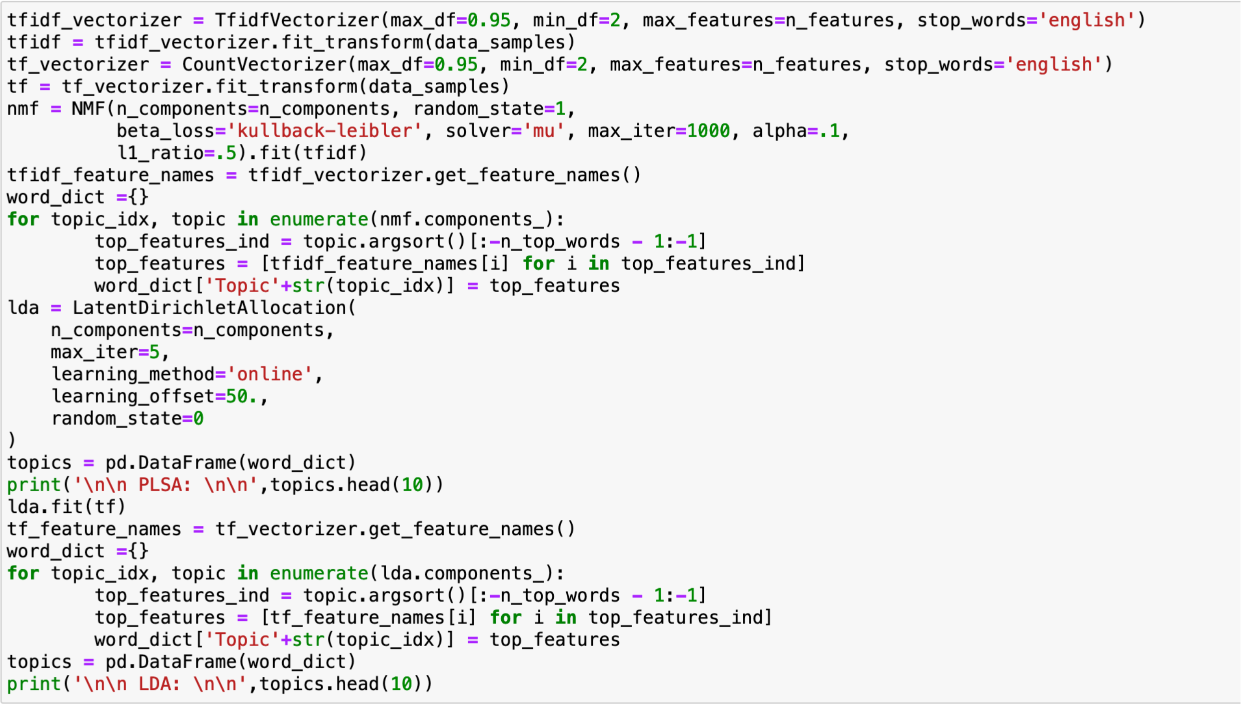
2. Calling dataset



```
data, _ = fetch_20newsgroups(shuffle=True, random_state=1,
                             remove=('headers', 'footers', 'quotes'),
                             return_X_y=True)
data_samples = data[:2000]
print(data_samples)
```

```
n\nBye-Bye, Big Jim.  Don't forget your Flintstone's Chewables!  :) \n--\nBake Timmons, III", "Although I realize
that principle is not one of your strongest\npoints, I would still like to know why do do not ask any question\no
f this sort about the Arab countries.\n\n   If you want to continue this think tank charade of yours, your\nfixat
ion on Israel must stop.  You might have to start asking the\nsame sort of questions of Arab countries as well.
You realize it\nwould not work, as the Arab countries' treatment of Jews over the\nlast several decades is so bad
that your fixation on Israel would\nbegin to look like the biased attack that it is.\n\n   Everyone in this group
recognizes that your stupid 'Center for\nPolicy Research' is nothing more than a fancy name for some bigot\nwho h
ates Israel.", 'Notwithstanding all the legitimate fuss about this proposal, how much\nof a change is it?  ATT\'s
last product in this area (a) was priced over\n$1000, as I suspect \'clipper\' phones will be; (b) came to the cu
stomer \nwith the key automatically preregistered with government authorities. Thus,\naside from attempting to fu
rther legitimize and solidify the fed\'s posture,\nClipper seems to be "more of the same", rather than a new dire
ction.\n   Yes, technology will eventually drive the cost down and thereby promote\nmore widespread use- but at p
resent, the man on the street is not going\nto purchase a $1000 crypto telephone, especially when the guy on the
other\nend probably doesn\'t have one anyway.  Am I missing something?\n   The real question is what the gov will
do in a year or two when air-\ntight voice privacy on a phone line is as close as your nearest pc.  That\nhas got
to a problematic scenario for them, even if the extent of usage\nnever surpasses the \'underground\' stature of P
GP.', "Well, I will have to change the scoring on my playoff pool.  Unfortunately\nI don't have time right now, b
ut I will certainly post the new scoring\nrules by tomorrow.  Does it matter?  No, you'll enter anyway!!!  Good!\
n\n--\n    Keith Keller\t\t\t\tLET'S GO RANGERS!!!!!\n\t\t\t\t\tLET'S GO QUAKERS!!!!!\n\tkkeller@mail.sas.upen
n.edu\t\tIVY LEAGUE CHAMPS!!!!", " \n \nI read somewhere, I think in Morton Smith's  Jesus the Magician , that\no
```

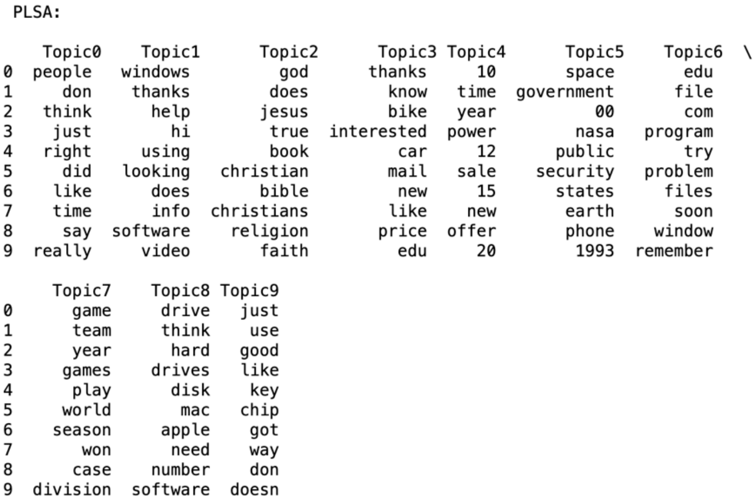
3. Implementing LDA and PLSA



```
tfidf_vectorizer = TfidfVectorizer(max_df=0.95, min_df=2, max_features=n_features, stop_words='english')
tfidf = tfidf_vectorizer.fit_transform(data_samples)
tf_vectorizer = CountVectorizer(max_df=0.95, min_df=2, max_features=n_features, stop_words='english')
tf = tf_vectorizer.fit_transform(data_samples)
nmf = NMF(n_components=n_components, random_state=1,
          beta_loss='kullback-leibler', solver='mu', max_iter=1000, alpha=.1,
          l1_ratio=.5).fit(tfidf)
tfidf_feature_names = tfidf_vectorizer.get_feature_names()
word_dict ={}
for topic_idx, topic in enumerate(nmf.components_):
        top_features_ind = topic.argsort()[:-n_top_words - 1:-1]
        top_features = [tfidf_feature_names[i] for i in top_features_ind]
        word_dict['Topic'+str(topic_idx)] = top_features
lda = LatentDirichletAllocation(
    n_components=n_components,
    max_iter=5,
    learning_method='online',
    learning_offset=50.,
    random_state=0
)
topics = pd.DataFrame(word_dict)
print('\n\n PLSA: \n\n',topics.head(10))
lda.fit(tf)
tf_feature_names = tf_vectorizer.get_feature_names()
word_dict ={}
for topic_idx, topic in enumerate(lda.components_):
        top_features_ind = topic.argsort()[:-n_top_words - 1:-1]
        top_features = [tf_feature_names[i] for i in top_features_ind]
        word_dict['Topic'+str(topic_idx)] = top_features
topics = pd.DataFrame(word_dict)
print('\n\n LDA: \n\n',topics.head(10))
```

- **Results:**

1. For PLSA



```
 PLSA: 

    Topic0    Topic1      Topic2      Topic3  Topic4      Topic5    Topic6  \
0  people   windows         god      thanks      10       space       edu   
1     don    thanks        does        know    time  government      file   
2   think      help       jesus        bike    year          00       com   
3    just        hi        true  interested   power        nasa   program   
4   right     using        book         car      12      public       try   
5     did   looking   christian        mail    sale    security   problem   
6    like      does       bible         new      15      states     files   
7    time      info  christians        like     new       earth      soon   
8     say  software    religion       price   offer       phone    window   
9  really     video       faith         edu      20        1993  remember   

     Topic7    Topic8  Topic9  
0      game     drive    just  
1      team     think     use  
2      year      hard    good  
3     games    drives    like  
4      play      disk     key  
5     world       mac    chip  
6    season     apple     got  
7       won      need     way  
8      case    number     don  
9  division  software   doesn  
```

2. For LDA

```
 LDA:

      Topic0  Topic1    Topic2      Topic3       Topic4   Topic5     Topic6  \
0        hiv   drive       edu          vs  performance    space     israel
1     health     car       com          gm       wanted     scsi        000
2       aids    disk      mail      thanks       robert    earth    section
3    disease    hard   windows         win        speed     moon    turkish
4    medical  drives      file  interested       couldn  surface   military
5       care    game      send      copies         math    probe   armenian
6      study   power  graphics        john           ok    lunar      greek
7   research   speed       use       email       change    orbit     killed
8       said    card   version        text      address  mission      state
9       1993    just       ftp          st      include     nasa  armenians

  Topic7      Topic8  Topic9
0     10         key    just
1     55  government     don
2     11      people  people
3     15         law    like
4     18      public   think
5     12        chip    know
6     20      church    time
7     00  encryption     say
8     13     clipper     god
9     93        used    good
```