Student Information

Name: 謝安亭

Student ID: 110065516

GitHub ID:antinghsieh555

Instructions

1. First: do the **take home** exercises in the DM2022-Lab1-Master. You may need to copy some cells from the Lab notebook to this notebook. **This part is worth 20% of your grade.**

1. Second: follow the same process from the DM2022-Lab1-Master on **the new dataset**. You don't need to explain all details as we did (some **minimal comments** explaining your code are useful though). **This part is worth 30% of your grade.**
    - Download the the new dataset. The dataset contains a `sentence` and `score` label. Read the specificiations of the dataset for details.
    - You are allowed to use and modify the `helper` functions in the folder of the first lab session (notice they may need modification) or create your own.

1. Third: please attempt the following tasks on **the new dataset**. **This part is worth 30% of your grade.**
    - Generate meaningful **new data visualizations**. Refer to online resources and the Data Mining textbook for inspiration and ideas.
    - Generate **TF-IDF features** from the tokens of each text. This will generating a document matrix, however, the weights will be computed differently (using the TF-IDF value of each word per document as opposed to the word frequency). Refer to this Sciki-learn guide .
    - Implement a simple **Naive Bayes classifier** that automatically classifies the records into their categories. Use both the TF-IDF features and word frequency features to build two seperate classifiers. Comment on the differences. Refer to this article.

1. Fourth: In the lab, we applied each step really quickly just to illustrate how to work with your dataset. There are somethings that are not ideal or the most efficient/meaningful. Each dataset can be habdled differently as well. What are those inefficent parts you noticed? How can you improve the Data preprocessing for these specific datasets? **This part is worth 10% of your grade.**

1. Fifth: It's hard for us to follow if your code is messy, so please **tidy up your notebook** and **add minimal comments where needed**. **This part is worth 10% of your grade.**

You can submit your homework following these guidelines: Git Intro & How to hand your homework. Make sure to commit and save your changes to your repository **BEFORE the deadline (October 20th 11:59 pm, Thursday)**.

In [158…
```
### Begin Assignment Here
```

*** Section 1 ***

Data Mining Lab 1

In this lab session we will focus on the use of scientific computing libraries to efficiently process, transform, and manage data. Furthermore, we will provide best practices and introduce visualization tools for effectively conducting big data analysis and visualization. In this lab session we will focus on the use of scientific computing libraries to efficiently process, transform, and manage data. Furthermore, we will provide best practices and introduce visualization tools for effectively conducting big data analysis and visualization.

Table of Contents

Introduction

In this notebook I will explore a text-based, document-based dataset using scientific computing tools such as Pandas and Numpy. In addition, several fundamental Data Mining concepts will be explored and explained in details, ranging from calculating distance measures to computing term frequency vectors. Coding examples, visualizations and demonstrations will be provided where necessary. Furthermore, additional exercises are provided after special topics. These exercises are geared towards testing the proficiency of students and motivate students to explore beyond the techniques covered in the notebook.

Requirements

Here are the computing and software requirements

Computing Resources

- Operating system: Preferably Linux or MacOS

- RAM: 8 GB
- Disk space: Mininium 8 GB

Software Requirements

Here is a list of the required programs and libraries necessary for this lab session:

Language:

- Python 3+ (Note: coding will be done strictly on Python 3)
  - We are using Python 3.9.6.
  - You can use newer version, but use at your own risk.

Environment:

Using an environment is to avoid some library conflict problems. You can refer this Setup Instructions to install and setup.

- Anaconda (recommended but not required)
  - Install anaconda environment
- Python virtualenv (recommended to Linux/MacOS user)
  - Install virtual environment
- Kaggle Kernel
  - Run on the cloud (with some limitations)
  - Reference: Kaggle Kernels Instructions

Necessary Libraries:

- Jupyter (Strongly recommended but not required)
  - Install `jupyter` and Use `$jupyter notebook` in terminal to run
- Scikit Learn
  - Install `sklearn` latest python library
- Pandas
  - Install `pandas` python library
- Numpy
  - Install `numpy` python library
- Matplotlib
  - Install `maplotlib` for python
- Plotly
  - Install and signup for `plotly`
- Seaborn
  - Install and signup for `seaborn`
- NLTK
  - Install `nltk` library

In [159…
```
# TEST necessary for when working with external scripts
%load_ext autoreload
%autoreload 2
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

1. The Data

In this notebook we will explore the popular 20 newsgroup dataset, originally provided here. The dataset is called "Twenty Newsgroups", which means there are 20 categories of news articles available in the entire dataset. A short description of the dataset, provided by the authors, is provided below:

- *The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of our knowledge, it was originally collected by Ken Lang, probably for his paper "Newsweeder: Learning to filter netnews," though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.*

If you need more information about the dataset please refer to the reference provided above. Below is a snapshot of the dataset already converted into a table. Keep in mind that the original dataset is not in this nice pretty format. That work is left to us. That is one of the tasks that will be covered in this notebook: how to convert raw data into convenient tabular formats using Pandas.

	text	category_name
0	From: sd345@city.ac.uk (Michael Collier) Subje...	comp.graphics
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...	comp.graphics
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	soc.religion.christian
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	soc.religion.christian
4	From: stanly@grok11.columbiasc.ncr.com (stanly...	soc.religion.christian
5	From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...	soc.religion.christian
6	From: jodfishe@silver.ucs.indiana.edu (joseph ...	soc.religion.christian
7	From: aldridge@netcom.com (Jacquelin Aldridge)...	sci.med
8	From: geb@cs.pitt.edu (Gordon Banks) Subject: ...	sci.med
9	From: libman@hsc.usc.edu (Marlena Libman) Subj...	sci.med

---

2. Data Preparation

In the following we will use the built-in dataset loader for 20 newsgroups from scikit-learn. Alternatively, it is possible to download the dataset manually from the website and use the sklearn.datasets.load_files function by pointing it to the 20news-bydate-train sub-folder of the uncompressed archive folder.

In order to get faster execution times for this first example we will work on a partial dataset with only 4 categories out of the 20 available in the dataset:

In [160…
```
# categories
categories = ['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']
```

In [161…
```
# obtain the documents containing the categories provided
from sklearn.datasets import fetch_20newsgroups
```

```
twenty_train = fetch_20newsgroups(subset='train', categories=categories,
                                  shuffle=True, random_state=42)
```

Let's take at look some of the records that are contained in our subset of the data

In [162…
```
twenty_train.data[0:2]
```

Out[162]:
```
['From: sd345@city.ac.uk (Michael Collier)\nSubject: Converting images to HP LaserJe
t III?\nNntp-Posting-Host: hampton\nOrganization: The City University\nLines: 14\n\n
Does anyone know of a good way (standard PC application/PD utility) to\nconvert tif/
img/tga files into LaserJet III format.  We would also like to\ndo the same, convert
ing to HPGL (HP plotter) files.\n\nPlease email any response.\n\nIs this the correct
group?\n\nThanks in advance.  Michael.\n-- \nMichael Collier (Programmer)
The Computer Unit,\nEmail: M.P.Collier@uk.ac.city                The City Universit
y,\nTel: 071 477-8000 x3769                     London,\nFax: 071 477-8565
EC1V 0HB.\n',
 "From: ani@ms.uky.edu (Aniruddha B. Deglurkar)\nSubject: help: Splitting a trimming
region along a mesh \nOrganization: University Of Kentucky, Dept. of Math Sciences\n
Lines: 28\n\n\n\n\tHi,\n\n\tI have a problem, I hope some of the 'gurus' can help me
solve.\n\n\tBackground of the problem:\n\tI have a rectangular mesh in the uv domai
n, i.e  the mesh is a \n\tmapping of a 3d Bezier patch into 2d. The area in this dom
ain\n\twhich is inside a trimming loop had to be rendered. The trimming\n\tloop is a
set of 2d Bezier curve segments.\n\tFor the sake of notation: the mesh is made up of
cells.\n\n\tMy problem is this :\n\tThe trimming area has to be split up into indivi
dual smaller\n\tcells bounded by the trimming curve segments. If a cell\n\tis wholly
inside the area...then it is output as a whole ,\n\telse it is trivially rejected.
\n\n\tDoes any body know how thiss can be done, or is there any algo. \n\tsomewhere
for doing this.\n\n\tAny help would be appreciated.\n\n\tThanks, \n\tAni.\n-- \nTo g
et irritated is human, to stay cool, divine.\n"]
```

Note the `twenty_train` is just a bunch of objects that can be accessed as python dictionaries; so, you can do the following operations on `twenty_train`

In [163…
```
twenty_train.target_names
```

Out[163]:
```
['alt.atheism', 'comp.graphics', 'sci.med', 'soc.religion.christian']
```

In [164…
```
len(twenty_train.data)
```

Out[164]:
```
2257
```

In [165…
```
len(twenty_train.filenames)
```

Out[165]:
```
2257
```

We can also print an example from the subset

In [166…
```
# An example of what the subset contains
print("\n".join(twenty_train.data[0].split("\n")))
```

```
From: sd345@city.ac.uk (Michael Collier)
Subject: Converting images to HP LaserJet III?
Nntp-Posting-Host: hampton
Organization: The City University
Lines: 14

Does anyone know of a good way (standard PC application/PD utility) to
convert tif/img/tga files into LaserJet III format.  We would also like to
do the same, converting to HPGL (HP plotter) files.

Please email any response.

Is this the correct group?

Thanks in advance.  Michael.
--
Michael Collier (Programmer)                The Computer Unit,
Email: M.P.Collier@uk.ac.city               The City University,
Tel: 071 477-8000 x3769                     London,
Fax: 071 477-8565                           EC1V 0HB.
```

... and determine the label of the example via `target_names` key value

In [167…
```
print(twenty_train.target_names[twenty_train.target[0]])
```

```
comp.graphics
```

In [168…
```
twenty_train.target[0]
```

Out[168]:
```
1
```

... we can also get the category of 10 documents via `target` key value

In [169…
```
# category of first 10 documents.
twenty_train.target[:10]
```

Out[169]:
```
array([1, 1, 3, 3, 3, 3, 3, 2, 2, 2])
```

Note: As you can observe, both approaches above provide two different ways of obtaining the `category` value for the dataset. Ideally, we want to have access to both types -- numerical and nominal -- in the event some particular library favors a particular type.

As you may have already noticed as well, there is no **tabular format** for the current version of the data. As data miners, we are interested in having our dataset in the most convenient format as possible; something we can manipulate easily and is compatible with our algorithms, and so forth.

Here is one way to get access to the *text* version of the label of a subset of our training data:

In [170…
```
for t in twenty_train.target[:10]:
    print(twenty_train.target_names[t])
```

```
comp.graphics
comp.graphics
soc.religion.christian
soc.religion.christian
soc.religion.christian
soc.religion.christian
soc.religion.christian
sci.med
sci.med
sci.med
```

>>> Exercise 1 (5 min):

In this exercise, please print out the *text* data for the first three samples in the dataset. (See the above code for help)

In [171…

```
# Answer here
for i in twenty_train.data[:3]:
    print(i)
```

From: sd345@city.ac.uk (Michael Collier)
Subject: Converting images to HP LaserJet III?
Nntp-Posting-Host: hampton
Organization: The City University
Lines: 14

Does anyone know of a good way (standard PC application/PD utility) to
convert tif/img/tga files into LaserJet III format. We would also like to
do the same, converting to HPGL (HP plotter) files.

Please email any response.

Is this the correct group?

Thanks in advance. Michael.
--
Michael Collier (Programmer)
Email: M.P.Collier@uk.ac.city
Tel: 071 477-8000 x3769
Fax: 071 477-8565

The Computer Unit,
The City University,
London,
EC1V 0HB.

From: ani@ms.uky.edu (Aniruddha B. Deglurkar)
Subject: help: Splitting a trimming region along a mesh
Organization: University Of Kentucky, Dept. of Math Sciences
Lines: 28

Hi,

I have a problem, I hope some of the 'gurus' can help me solve.

Background of the problem:
I have a rectangular mesh in the uv domain, i.e the mesh is a
mapping of a 3d Bezier patch into 2d. The area in this domain
which is inside a trimming loop had to be rendered. The trimming
loop is a set of 2d Bezier curve segments.
For the sake of notation: the mesh is made up of cells.

My problem is this :
The trimming area has to be split up into individual smaller
cells bounded by the trimming curve segments. If a cell
is wholly inside the area...then it is output as a whole ,
else it is trivially rejected.

Does any body know how thiss can be done, or is there any algo.
somewhere for doing this.

Any help would be appreciated.

Thanks,
Ani.
--
To get irritated is human, to stay cool, divine.

From: djohnson@cs.ucsd.edu (Darin Johnson)
Subject: Re: harrassed at work, could use some prayers
Organization: =CSE Dept., U.C. San Diego
Lines: 63

(Well, I'll email also, but this may apply to other people, so
I'll post also.)

>I've been working at this company for eight years in various
>engineering jobs. I'm female. Yesterday I counted and realized that
>on seven different occasions I've been sexually harrassed at this
>company.

>I dreaded coming back to work today. What if my boss comes in to ask
>me some kind of question...

```
Your boss should be the person bring these problems to.  If he/she
does not seem to take any action, keep going up higher and higher.
Sexual harrassment does not need to be tolerated, and it can be an
enormous emotional support to discuss this with someone and know that
they are trying to do something about it.  If you feel you can not
discuss this with your boss, perhaps your company has a personnel
department that can work for you while preserving your privacy.  Most
companies will want to deal with this problem because constant anxiety
does seriously affect how effectively employees do their jobs.

It is unclear from your letter if you have done this or not.  It is
not inconceivable that management remains ignorant of employee
problems/strife even after eight years (it's a miracle if they do
notice).  Perhaps your manager did not bring to the attention of
higher ups?  If the company indeed does seem to want to ignore the
entire problem, there may be a state agency willing to fight with
you.  (check with a lawyer, a women's resource center, etc to find out)

You may also want to discuss this with your paster, priest, husband,
etc.  That is, someone you know will not be judgemental and that is
supportive, comforting, etc.  This will bring a lot of healing.

>So I returned at 11:25, only to find that ever single
>person had already left for lunch.  They left at 11:15 or so.  No one
>could be bothered to call me at the other building, even though my
>number was posted.

This happens to a lot of people.  Honest.  I believe it may seem
to be due to gross insensitivity because of the feelings you are
going through.  People in offices tend to be more insensitive while
working than they normally are (maybe it's the hustle or stress or...)
I've had this happen to me a lot, often because they didn't realize
my car was broken, etc.  Then they will come back and wonder why I
didn't want to go (this would tend to make me stop being angry at
being ignored and make me laugh).  Once, we went off without our
boss, who was paying for the lunch :-)

>For this
>reason I hope good Mr. Moderator allows me this latest indulgence.

Well, if you can't turn to the computer for support, what would
we do?  (signs of the computer age :-)

In closing, please don't let the hateful actions of a single person
harm you.  They are doing it because they are still the playground
bully and enjoy seeing the hurt they cause.  And you should not
accept the opinions of an imbecile that you are worthless - much
wiser people hold you in great esteem.
--
Darin Johnson
djohnson@ucsd.edu
  - Luxury!  In MY day, we had to make do with 5 bytes of swap...
```

---

3. Data Transformation

So we want to explore and understand our data a little bit better. Before we do that we definitely need to apply some transformations just so we can have our dataset in a nice format to be able to explore it freely and more efficient. Lucky for us, there are powerful scientific tools to transform our data into that tabular format we are so farmiliar with. So that is what we will do in the next section--transform our data into a nice table format.

3.1 Converting Dictionary into Pandas Dataframe

Here we will show you how to convert dictionary objects into a pandas dataframe. And by the way, a pandas dataframe is nothing more than a table magically stored for efficient information retrieval.

In [172…
```
import pandas as pd

# my functions
import helpers.data_mining_helpers as dmh

# construct dataframe from a list
X = pd.DataFrame.from_records(dmh.format_rows(twenty_train), columns= ['text'])
```

In [173…
```
len(X)
```

Out[173]:
```
2257
```

In [174…
```
X[0:2]
```

Out[174]:

	text
0	From: sd345@city.ac.uk (Michael Collier) Subje...
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...

In [175…
```
for t in X["text"][:3]:
    print(t)
```

```
From: sd345@city.ac.uk (Michael Collier) Subject: Converting images to HP LaserJet I
II? Nntp-Posting-Host: hampton Organization: The City University Lines: 14  Does any
one know of a good way (standard PC application/PD utility) to convert tif/img/tga f
iles into LaserJet III format.  We would also like to do the same, converting to HPG
L (HP plotter) files.  Please email any response.  Is this the correct group?  Thank
s in advance.  Michael. --  Michael Collier (Programmer)                The Compute
r Unit, Email: M.P.Collier@uk.ac.city                The City University, Tel: 071 4
77-8000 x3769                        London, Fax: 071 477-8565
EC1V 0HB.
From: ani@ms.uky.edu (Aniruddha B. Deglurkar) Subject: help: Splitting a trimming re
gion along a mesh  Organization: University Of Kentucky, Dept. of Math Sciences Line
s: 28           Hi,     I have a problem, I hope some of the 'gurus' can help me sol
ve.     Background of the problem:      I have a rectangular mesh in the uv domain,
i.e  the mesh is a      mapping of a 3d Bezier patch into 2d. The area in this domai
n       which is inside a trimming loop had to be rendered. The trimming        loop
is a set of 2d Bezier curve segments.   For the sake of notation: the mesh is made u
p of cells.     My problem is this :    The trimming area has to be split up into in
dividual smaller        cells bounded by the trimming curve segments. If a cell
is wholly inside the area...then it is output as a whole ,      else it is trivially
rejected.       Does any body know how thiss can be done, or is there any algo.
somewhere for doing this.       Any help would be appreciated.          Thanks,
Ani. --  To get irritated is human, to stay cool, divine.
From: djohnson@cs.ucsd.edu (Darin Johnson) Subject: Re: harrassed at work, could use
some prayers Organization: =CSE Dept., U.C. San Diego Lines: 63  (Well, I'll email a
lso, but this may apply to other people, so I'll post also.)  >I've been working at
this company for eight years in various >engineering jobs.  I'm female.  Yesterday I
counted and realized that >on seven different occasions I've been sexually harrassed
at this >company.  >I dreaded coming back to work today.  What if my boss comes in t
o ask >me some kind of question...  Your boss should be the person bring these probl
ems to.  If he/she does not seem to take any action, keep going up higher and highe
r. Sexual harrassment does not need to be tolerated, and it can be an enormous emoti
onal support to discuss this with someone and know that they are trying to do someth
ing about it.  If you feel you can not discuss this with your boss, perhaps your com
pany has a personnel department that can work for you while preserving your privacy.
Most companies will want to deal with this problem because constant anxiety does ser
iously affect how effectively employees do their jobs.  It is unclear from your lett
er if you have done this or not.  It is not inconceivable that management remains ig
norant of employee problems/strife even after eight years (it's a miracle if they do
notice).  Perhaps your manager did not bring to the attention of higher ups?  If the
company indeed does seem to want to ignore the entire problem, there may be a state
agency willing to fight with you.  (check with a lawyer, a women's resource center,
etc to find out)  You may also want to discuss this with your paster, priest, husban
d, etc.  That is, someone you know will not be judgemental and that is supportive, c
omforting, etc.  This will bring a lot of healing.  >So I returned at 11:25, only to
find that ever single >person had already left for lunch.  They left at 11:15 or so.
No one >could be bothered to call me at the other building, even though my >number w
as posted.  This happens to a lot of people.  Honest.  I believe it may seem to be d
ue to gross insensitivity because of the feelings you are going through.  People in
offices tend to be more insensitive while working than they normally are (maybe it's
the hustle or stress or...) I've had this happen to me a lot, often because they did
n't realize my car was broken, etc.  Then they will come back and wonder why I did
n't want to go (this would tend to make me stop being angry at being ignored and mak
e me laugh).  Once, we went off without our boss, who was paying for the lunch :-)
>For this >reason I hope good Mr. Moderator allows me this latest indulgence.  Well,
if you can't turn to the computer for support, what would we do?  (signs of the comp
uter age :-)  In closing, please don't let the hateful actions of a single person ha
rm you.  They are doing it because they are still the playground bully and enjoy see
ing the hurt they cause.  And you should not accept the opinions of an imbecile that
you are worthless - much wiser people hold you in great esteem. --  Darin Johnson dj
ohnson@ucsd.edu   - Luxury!  In MY day, we had to make do with 5 bytes of swap...
```

Adding Columns

One of the great advantages of a pandas dataframe is its flexibility. We can add columns to the current dataset programmatically with very little effort.

In [176…

```
# add category to the dataframe
X['category'] = twenty_train.target
```

In [177…
```
# add category label also
X['category_name'] = X.category.apply(lambda t: dmh.format_labels(t, twenty_train))
```

Now we can print and see what our table looks like.

In [178…
```
X[0:10]
```

Out[178]:

	text	category	category_name
0	From: sd345@city.ac.uk (Michael Collier) Subje...	1	comp.graphics
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...	1	comp.graphics
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	3	soc.religion.christian
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	3	soc.religion.christian
4	From: stanly@grok11.columbiasc.ncr.com (stanly...	3	soc.religion.christian
5	From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...	3	soc.religion.christian
6	From: jodfishe@silver.ucs.indiana.edu (joseph ...	3	soc.religion.christian
7	From: aldridge@netcom.com (Jacquelin Aldridge)...	2	sci.med
8	From: geb@cs.pitt.edu (Gordon Banks) Subject: ...	2	sci.med
9	From: libman@hsc.usc.edu (Marlena Libman) Subj...	2	sci.med

Nice! Isn't it? With this format we can conduct many operations easily and efficiently since Pandas dataframes provide us with a wide range of built-in features/functionalities. These features are operations which can directly and quickly be applied to the dataset. These operations may include standard operations like **removing records with missing values** and **aggregating new fields** to the current table (hereinafter referred to as a dataframe), which is desirable in almost every data mining project. Go Pandas!

3.2 Familiarizing yourself with the Data

To begin to show you the awesomeness of Pandas dataframes, let us look at how to run a simple query on our dataset. We want to query for the first 10 rows (documents), and we only want to keep the `text` and `category_name` attributes or fields.

In [179…
```
# a simple query
X[:10][["text","category_name"]]
```

Out[179]:

	text	category_name
0	From: sd345@city.ac.uk (Michael Collier) Subje...	comp.graphics
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...	comp.graphics
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	soc.religion.christian
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	soc.religion.christian
4	From: stanly@grok11.columbiasc.ncr.com (stanly...	soc.religion.christian
5	From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...	soc.religion.christian
6	From: jodfishe@silver.ucs.indiana.edu (joseph ...	soc.religion.christian
7	From: aldridge@netcom.com (Jacquelin Aldridge)...	sci.med
8	From: geb@cs.pitt.edu (Gordon Banks) Subject: ...	sci.med
9	From: libman@hsc.usc.edu (Marlena Libman) Subj...	sci.med

Let us look at a few more interesting queries to familiarize ourselves with the efficiency and

conveniency of Pandas dataframes.

Let's query the last 10 records

In [180…
```
X[-10:]
```

Out[180]:

	text	category	category_name
2247	From: daniels@math.ufl.edu (TV's Big Dealer) S...	3	soc.religion.christian
2248	From: "danny hawrysio" <danny.hawrysio@canrem....	1	comp.graphics
2249	From: shellgate!llo@uu4.psi.com (Larry L. Over...	3	soc.religion.christian
2250	From: ingles@engin.umich.edu (Ray Ingles) Subj...	0	alt.atheism
2251	From: Mark-Tarbell@suite.com Subject: Amniocen...	2	sci.med
2252	From: roos@Operoni.Helsinki.FI (Christophe Roo...	2	sci.med
2253	From: mhollowa@ic.sunysb.edu (Michael Holloway...	2	sci.med
2254	From: sasghm@theseus.unx.sas.com (Gary Merrill...	2	sci.med
2255	From: Dan Wallach <dwallach@cs.berkeley.edu> S...	2	sci.med
2256	From: dyer@spdcc.com (Steve Dyer) Subject: Re:...	2	sci.med

Ready for some sourcery? Brace yourselves! Let us see if we can query the first 10th record in our dataframe. For this we will use the build-in function called `loc`. This allows us to explicity define the columns you want to query.

In [181…
```
# using loc (by label)
X.loc[:10, 'text']
```

Out[181]:
```
0     From: sd345@city.ac.uk (Michael Collier) Subje...
1     From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...
2     From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...
3     From: s0612596@let.rug.nl (M.M. Zwart) Subject...
4     From: stanly@grok11.columbiasc.ncr.com (stanly...
5     From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...
6     From: jodfishe@silver.ucs.indiana.edu (joseph ...
7     From: aldridge@netcom.com (Jacquelin Aldridge)...
8     From: geb@cs.pitt.edu (Gordon Banks) Subject: ...
9     From: libman@hsc.usc.edu (Marlena Libman) Subj...
10    From: anasaz!karl@anasazi.com (Karl Dussik) Su...
Name: text, dtype: object
```

You can also use the `iloc` function to query a selection of our dataset by position. Take a look at this great discussion on the differences between the `iloc` and `loc` functions.

In [182…
```
# using iloc (by position)
X.iloc[:10, 0]
```

Out[182]:
```
0    From: sd345@city.ac.uk (Michael Collier) Subje...
1    From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...
2    From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...
3    From: s0612596@let.rug.nl (M.M. Zwart) Subject...
4    From: stanly@grok11.columbiasc.ncr.com (stanly...
5    From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...
6    From: jodfishe@silver.ucs.indiana.edu (joseph ...
7    From: aldridge@netcom.com (Jacquelin Aldridge)...
8    From: geb@cs.pitt.edu (Gordon Banks) Subject: ...
9    From: libman@hsc.usc.edu (Marlena Libman) Subj...
Name: text, dtype: object
```

>>> Exercise 2 (take home):

Experiment with other querying techniques using pandas dataframes. Refer to their documentation for more information.

In [183…
```
#Answer here
X.query('category > 2')
```

Out[183]:

	text	category	category_name
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	3	soc.religion.christian
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	3	soc.religion.christian
4	From: stanly@grok11.columbiasc.ncr.com (stanly...	3	soc.religion.christian
5	From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...	3	soc.religion.christian
6	From: jodfishe@silver.ucs.indiana.edu (joseph ...	3	soc.religion.christian
...	...	...	...
2229	From: jcj@tellabs.com (jcj) Subject: Re: proof...	3	soc.religion.christian
2230	From: news@cbnewsk.att.com Subject: Re: Bible ...	3	soc.religion.christian
2246	From: lmvec@westminster.ac.uk (William Hargrea...	3	soc.religion.christian
2247	From: daniels@math.ufl.edu (TV's Big Dealer) S...	3	soc.religion.christian
2249	From: shellgate!llo@uu4.psi.com (Larry L. Over...	3	soc.religion.christian

599 rows × 3 columns

>>> Exercise 3 (5 min):

Try to fetch records belonging to the `sci.med` category, and query every 10th record. Only show the first 5 records.

In [184…
```
# Answer here
X.loc[lambda f: f.category_name == 'sci.med'].iloc[::10, :][0:5]
```

Out[184]:

	text	category	category_name
7	From: aldridge@netcom.com (Jacquelin Aldridge)...	2	sci.med
49	From: jimj@contractor.EBay.Sun.COM (Jim Jones)...	2	sci.med
82	From: jason@ab20.larc.nasa.gov (Jason Austin) ...	2	sci.med
118	From: rogers@calamari.hi.com (Andrew Rogers) S...	2	sci.med
142	From: lady@uhunix.uhcc.Hawaii.Edu (Lee Lady) S...	2	sci.med

---

4. Data Mining using Pandas

Let's do some serious work now. Let's learn to program some of the ideas and concepts learned so far in the data mining course. This is the only way we can be convince ourselves of the true power of Pandas dataframes.

4.1 Missing Values

First, let us consider that our dataset has some *missing values* and we want to remove those values. In its current state our dataset has no missing values, but for practice sake we will add some

records with missing values and then write some code to deal with these objects that contain missing values. You will see for yourself how easy it is to deal with missing values once you have your data transformed into a Pandas dataframe.

Before we jump into coding, let us do a quick review of what we have learned in the Data Mining course. Specifically, let's review the methods used to deal with missing values.

The most common reasons for having missing values in datasets has to do with how the data was initially collected. A good example of this is when a patient comes into the ER room, the data is collected as quickly as possible and depending on the conditions of the patients, the personal data being collected is either incomplete or partially complete. In the former and latter cases, we are presented with a case of "missing values". Knowing that patients data is particularly critical and can be used by the health authorities to conduct some interesting analysis, we as the data miners are left with the tough task of deciding what to do with these missing and incomplete records. We need to deal with these records because they are definitely going to affect our analysis or learning algorithms. So what do we do? There are several ways to handle missing values, and some of the more effective ways are presented below (Note: You can reference the slides - Session 1 Handout for the additional information).

- **Eliminate Data Objects** - Here we completely discard records once they contain some missing values. This is the easiest approach and the one we will be using in this notebook. The immediate drawback of going with this approach is that you lose some information, and in some cases too much of it. Now imagine that half of the records have at least one or more missing values. Here you are presented with the tough decision of quantity vs quality. In any event, this decision must be made carefully, hence the reason for emphasizing it here in this notebook.

- **Estimate Missing Values** - Here we try to estimate the missing values based on some criteria. Although this approach may be proven to be effective, it is not always the case, especially when we are dealing with sensitive data, like **Gender** or **Names**. For fields like **Address**, there could be ways to obtain these missing addresses using some data aggregation technique or obtain the information directly from other databases or public data sources.

- **Ignore the missing value during analysis** - Here we basically ignore the missing values and proceed with our analysis. Although this is the most naive way to handle missing values it may proof effective, especially when the missing values includes information that is not important to the analysis being conducted. But think about it for a while. Would you ignore missing values, especially when in this day and age it is difficult to obtain high quality datasets? Again, there are some tradeoffs, which we will talk about later in the notebook.

- **Replace with all possible values** - As an efficient and responsible data miner, we sometimes just need to put in the hard hours of work and find ways to makes up for these missing values. This last option is a very wise option for cases where data is scarce (which is almost always) or when dealing with sensitive data. Imagine that our dataset has an **Age** field, which contains many missing values. Since **Age** is a continuous variable, it means that we can build a separate model for calculating the age for the incomplete records based on some rule-based approach or probabilistic approach.

As mentioned earlier, we are going to go with the first option but you may be asked to compute missing values, using a different approach, as an exercise. Let's get to it!

First we want to add the dummy records with missing values since the dataset we have is perfectly composed and cleaned that it contains no missing values. First let us check for ourselves that

indeed the dataset doesn't contain any missing values. We can do that easily by using the following built-in function provided by Pandas.

In [185…
```
# check missing values
X.isnull()
```

Out[185]:

	text	category	category_name
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	False	False
…	…	…	…
2252	False	False	False
2253	False	False	False
2254	False	False	False
2255	False	False	False
2256	False	False	False

2257 rows × 3 columns

The `isnull` function looks through the entire dataset for null values and returns `True` wherever it finds any missing field or record. As you will see above, and as we anticipated, our dataset looks clean and all values are present, since `isnull` returns **False** for all fields and records. But let us start to get our hands dirty and build a nice little function to check each of the records, column by column, and return a nice little message telling us the amount of missing records found. This excerice will also encourage us to explore other capabilities of pandas dataframes. In most cases, the build-in functions are good enough, but as you saw above when the entire table was printed, it is impossible to tell if there are missing records just by looking at preview of records manually, especially in cases where the dataset is huge. We want a more reliable way to achieve this. Let's get to it!

In [186…
```
X.isnull().apply(lambda x: dmh.check_missing_values(x))
```

Out[186]:

	text	category	category_name
0	The amoung of missing records is:	The amoung of missing records is:	The amoung of missing records is:
1	0	0	0

Okay, a lot happened there in that one line of code, so let's break it down. First, with the `isnull` we tranformed our table into the **True/False** table you see above, where **True** in this case means that the data is missing and **False** means that the data is present. We then take the transformed table and apply a function to each row that essentially counts to see if there are missing values in each record and print out how much missing values we found. In other words the `check_missing_values` function looks through each field (attribute or column) in the dataset and counts how many missing values were found.

There are many other clever ways to check for missing data, and that is what makes Pandas so beautiful to work with. You get the control you need as a data scientist or just a person working in data mining projects. Indeed, Pandas makes your life easy!

>>> Exercise 4 (5 min):

Let's try something different. Instead of calculating missing values by column let's try to calculate the missing values in every record instead of every column.
$Hint$: `axis` parameter. Check the documentation for more information.

In [187…]
```
# Answer here
X.isnull().apply(lambda x: dmh.check_missing_values(x), axis=1)
```

Out[187]:
```
0       (The amoung of missing records is: , 0)
1       (The amoung of missing records is: , 0)
2       (The amoung of missing records is: , 0)
3       (The amoung of missing records is: , 0)
4       (The amoung of missing records is: , 0)
                        ...
2252    (The amoung of missing records is: , 0)
2253    (The amoung of missing records is: , 0)
2254    (The amoung of missing records is: , 0)
2255    (The amoung of missing records is: , 0)
2256    (The amoung of missing records is: , 0)
Length: 2257, dtype: object
```

We have our function to check for missing records, now let us do something mischievous and insert some dummy data into the dataframe and test the reliability of our function. This dummy data is intended to corrupt the dataset. I mean this happens a lot today, especially when hackers want to hijack or corrupt a database.

We will insert a `Series`, which is basically a "one-dimensional labeled array capable of holding data of any type (integer, string, float, python objects, etc.). The axis labels are collectively called index.", into our current dataframe.

In [188…]
```
dummy_series = pd.Series(["dummy_record", 1], index=["text", "category"])
```

In [189…]
```
dummy_series
```

Out[189]:
```
text        dummy_record
category               1
dtype: object
```

In [190…]
```
dummy_series.to_frame().T
# .to_frame() -> Convert Series to DataFrame
# .T          -> Transpose
```

Out[190]:

	text	category
0	dummy_record	1

In [191…]
```
result_with_series = pd.concat([X, dummy_series.to_frame().T], ignore_index=True)
```

In [192…]
```
# check if the records was commited into result
len(result_with_series)
```

Out[192]:
```
2258
```

Now we that we have added the record with some missing values. Let try our function and see if it can detect that there is a missing value on the resulting dataframe.

In [193…]
```
result_with_series.isnull().apply(lambda x: dmh.check_missing_values(x))
```

Out[193]:

	text	category	category_name
0	The amoung of missing records is:	The amoung of missing records is:	The amoung of missing records is:
1	0	0	1

Indeed there is a missing value in this new dataframe. Specifically, the missing value comes from the `category_name` attribute. As I mentioned before, there are many ways to conduct specific operations on the dataframes. In this case let us use a simple dictionary and try to insert it into our original dataframe `X`. Notice that above we are not changing the `X` dataframe as results are directly applied to the assignment variable provided. But in the event that we just want to keep things simple, we can just directly apply the changes to `X` and assign it to itself as we will do below. This modification will create a need to remove this dummy record later on, which means that we need to learn more about Pandas dataframes. This is getting intense! But just relax, everything will be fine!

In [194…
```
# dummy record as dictionary format
dummy_dict = [{'text': 'dummy_record',
               'category': 1
              }]
```

In [195…
```
X = pd.concat([X, pd.DataFrame(dummy_dict)], ignore_index=True)
```

In [196…
```
len(X)
```

Out[196]:
```
2258
```

In [197…
```
X.isnull().apply(lambda x: dmh.check_missing_values(x))
```

Out[197]:

	text	category	category_name
0	The amoung of missing records is:	The amoung of missing records is:	The amoung of missing records is:
1	0	0	1

So now that we can see that our data has missing values, we want to remove the records with missing values. The code to drop the record with missing that we just added, is the following:

In [198…
```
X.dropna(inplace=True)
```

... and now let us test to see if we gotten rid of the records with missing values.

In [199…
```
X.isnull().apply(lambda x: dmh.check_missing_values(x))
```

Out[199]:

	text	category	category_name
0	The amoung of missing records is:	The amoung of missing records is:	The amoung of missing records is:
1	0	0	0

In [200…
```
len(X)
```

Out[200]:
```
2257
```

And we are back with our original dataset, clean and tidy as we want it. That's enough on how to deal with missing values, let us now move unto something more fun.

But just in case you want to learn more about how to deal with missing data, refer to the official Pandas documentation.

>>> Exercise 5 (take home)

There is an old saying that goes, "The devil is in the details." When we are working with extremely large data, it's difficult to check records one by one (as we have been doing so far). And also, we don't even know what kind of missing values we are facing. Thus, "debugging" skills get sharper as we spend more time solving bugs. Let's focus on a different method to check for missing values and the kinds of missing values you may encounter. It's not easy to check for missing values as you will find out in a minute.

Please check the data and the process below, describe what you observe and why it happened.
$Hint$: why `.isnull()` didn't work?

In [201…
```
import numpy as np

NA_dict = [{ 'id': 'A', 'missing_example': np.nan },
           { 'id': 'B'                    },
           { 'id': 'C', 'missing_example': 'NaN'  },
           { 'id': 'D', 'missing_example': 'None' },
           { 'id': 'E', 'missing_example':  None  },
           { 'id': 'F', 'missing_example': ''     }]

NA_df = pd.DataFrame(NA_dict, columns = ['id','missing_example'])
NA_df
```

Out[201]:

	id	missing_example
0	A	NaN
1	B	NaN
2	C	NaN
3	D	None
4	E	None
5	F	

In [202…
```
NA_df['missing_example'].isnull()
```

Out[202]:
```
0     True
1     True
2    False
3    False
4     True
5    False
Name: missing_example, dtype: bool
```

In [203…
```
# Answer here
"""
For function of isnull(), there are only np.nan, None, or completely blank considered
In the other words, the column where id is "C" and id is "D", their data of 'NaN' and
Also, the column where id is "F,'' its data of '' will be regarded as a value with a
Therefore, they won't be seen as missing data and isnull() function won't work.
"""
```

Out[203]:
```
'\nFor function of isnull(), there are only np.nan, None, or completely blank consid
ered missing data.\nIn the other words, the column where id is "C" and id is "D", th
eir data of \'NaN\' and \'None\' will be regarded as string by the computer. \nAlso,
the column where id is "F,\'\' its data of \'\' will be regarded as a value with a b
lank key. \nTherefore, they won\'t be seen as missing data and isnull() function won
\'t work.\n'
```

4.2 Dealing with Duplicate Data

Dealing with duplicate data is just as painful as dealing with missing data. The worst case is that you have duplicate data that has missing values. But let us not get carried away. Let us stick with the basics. As we have learned in our Data Mining course, duplicate data can occur because of many reasons. The majority of the times it has to do with how we store data or how we collect and merge data. For instance, we may have collected and stored a tweet, and a retweet of that same tweet as two different records; this results in a case of data duplication; the only difference being that one is the original tweet and the other the retweeted one. Here you will learn that dealing with duplicate data is not as challenging as missing values. But this also all depends on what you consider as duplicate data, i.e., this all depends on your criteria for what is considered as a duplicate record and also what type of data you are dealing with. For textual data, it may not be so trivial as it is for numerical values or images. Anyhow, let us look at some code on how to deal with duplicate records in our `X` dataframe.

First, let us check how many duplicates we have in our current dataset. Here is the line of code that checks for duplicates; it is very similar to the `isnull` function that we used to check for missing values.

In [204…
```
X.duplicated()
```

Out[204]:
```
0       False
1       False
2       False
3       False
4       False
        ...
2252    False
2253    False
2254    False
2255    False
2256    False
Length: 2257, dtype: bool
```

We can also check the sum of duplicate records by simply doing:

In [205…
```
sum(X.duplicated())
```

Out[205]:
```
0
```

Based on that output, you may be asking why did the `duplicated` operation only returned one single column that indicates whether there is a duplicate record or not. So yes, all the `duplicated()` operation does is to check per records instead of per column. That is why the operation only returns one value instead of three values for each column. It appears that we don't have any duplicates since none of our records resulted in `True`. If we want to check for duplicates as we did above for some particular column, instead of all columns, we do something as shown below. As you may have noticed, in the case where we select some columns instead of checking by all columns, we are kind of lowering the criteria of what is considered as a duplicate record. So let us only check for duplicates by only checking the `text` attribute.

In [206…
```
sum(X.duplicated('text'))
```

Out[206]:
```
0
```

Now let us create some duplicated dummy records and append it to the main dataframe `X`. Subsequenlty, let us try to get rid of the duplicates.

In [207…
```
dummy_duplicate_dict = [{
                                'text': 'dummy record',
```

```
                                'category': 1,
                                'category_name': "dummy category"
                        },
                        {
                                'text': 'dummy record',
                                'category': 1,
                                'category_name': "dummy category"
                        }]
```

In [208…
```
X = pd.concat([X, pd.DataFrame(dummy_duplicate_dict)], ignore_index=True)
```

In [209…
```
len(X)
```

Out[209]:
```
2259
```

In [210…
```
sum(X.duplicated('text'))
```

Out[210]:
```
1
```

We have added the dummy duplicates to `X`. Now we are faced with the decision as to what to do with the duplicated records after we have found it. In our case, we want to get rid of all the duplicated records without preserving a copy. We can simply do that with the following line of code:

In [211…
```
X.drop_duplicates(keep=False, inplace=True) # inplace applies changes directly on our
```

In [212…
```
len(X)
```

Out[212]:
```
2257
```

Check out the Pandas documentation for more information on dealing with duplicate data.

---

5. Data Preprocessing

In the Data Mining course we learned about the many ways of performing data preprocessing. In reality, the list is quiet general as the specifics of what data preprocessing involves is too much to cover in one course. This is especially true when you are dealing with unstructured data, as we are dealing with in this particular notebook. But let us look at some examples for each data preprocessing technique that we learned in the class. We will cover each item one by one, and provide example code for each category. You will learn how to perform each of the operations, using Pandas, that cover the essentials to Preprocessing in Data Mining. We are not going to follow any strict order, but the items we will cover in the preprocessing section of this notebook are as follows:

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature Subset Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformation

5.1 Sampling

The first concept that we are going to cover from the above list is sampling. Sampling refers to the technique used for selecting data. The functionalities that we use to selected data through queries provided by Pandas are actually basic methods for sampling. The reasons for sampling are sometimes due to the size of data -- we want a smaller subset of the data that is still representatitive enough as compared to the original dataset.

We don't have a problem of size in our current dataset since it is just a couple thousand records long. But if we pay attention to how much content is included in the `text` field of each of those records, you will realize that sampling may not be a bad idea after all. In fact, we have already done some sampling by just reducing the records we are using here in this notebook; remember that we are only using four categories from the all the 20 categories available. Let us get an idea on how to sample using pandas operations.

In [213…
```
# Answer here
# Use for answer Exercise 6
# Initialize X

X_initial = X
```

In [214…
```
X_sample = X.sample(n=1000) #random state
```

In [215…
```
len(X_sample)
```

Out[215]:
```
1000
```

In [216…
```
X_sample[0:4]
```

Out[216]:

	text	category	category_name
998	From: labson@borneo.corp.sgi.com (Joel Labson)...	3	soc.religion.christian
1779	From: wjhovi01@ulkyvx.louisville.edu Subject: ...	3	soc.religion.christian
856	From: halsall@murray.fordham.edu (Paul Halsall...	3	soc.religion.christian
890	From: jim.zisfein@factory.com (Jim Zisfein) S...	2	sci.med

---

>>> Exercise 6 (take home):

Notice any changes to the `X` dataframe? What are they? Report every change you noticed as compared to the previous state of `X`. Feel free to query and look more closely at the dataframe for these changes.

In [217…
```
# Answer here

"""
From cell above, we can notice there are not any differences between initial X and X
Then, X_sample is randomly chosen from X, so index is not sorting.
And X_sample will not change X value.

This result can be proved by the following code.
"""
```

Out[217]:
```
'\nFrom cell above, we can notice there are not any differences between initial X an
d X after sample.\nThen, X_sample is randomly chosen from X, so index is not sortin
g.\nAnd X_sample will not change X value.\n\nThis result can be proved by the follow
ing code.\n'
```

In [218…
```
# Answer here
# Compare X_initial and X after .sample
```

```
diff = False


for i in range(len(X)):
    if(X_initial["text"][i] != X["text"][i]):
        print("text", i)
        diff = True
    if(X_initial["category"][i] != X["category"][i]):
        print("category", i)
        diff = True
    if(X_initial["category_name"][i] != X["category_name"][i]):
        print("category_name", i)
        diff = True

if (diff == True):
    print("There are some different")
else:
    print("There isn't any different")
```

```
There isn't any different
```

Let's do something cool here while we are working with sampling! Let us look at the distribution of categories in both the sample and original dataset. Let us visualize and analyze the disparity between the two datasets. To generate some visualizations, we are going to use `matplotlib` python library. With matplotlib, things are faster and compatability-wise it may just be the best visualization library for visualizing content extracted from dataframes and when using Jupyter notebooks. Let's take a loot at the magic of `matplotlib` below.

In [219…
```
import matplotlib.pyplot as plt
%matplotlib inline
```

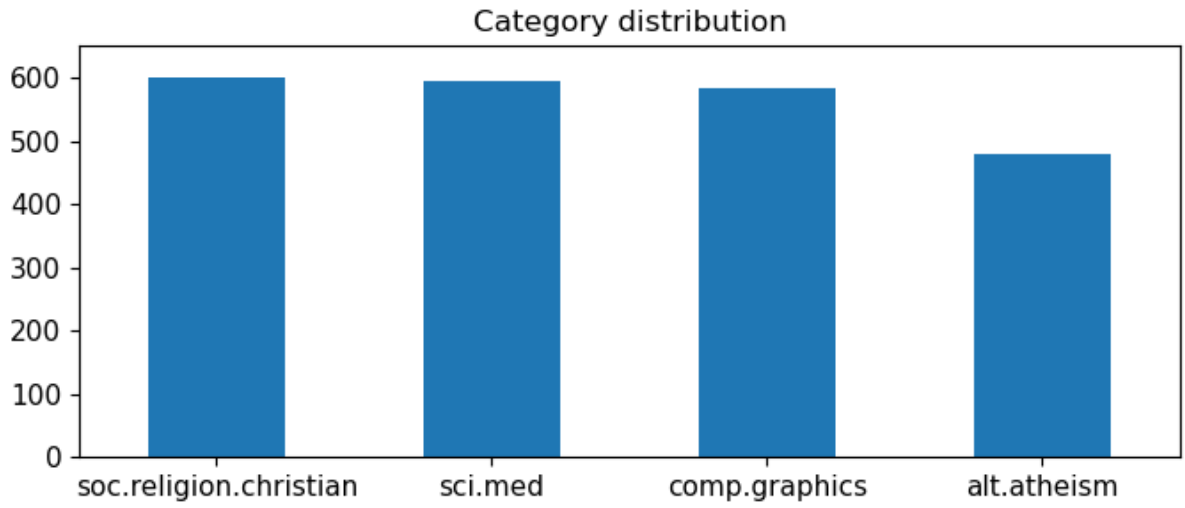
In [220…
```
categories
```

Out[220]:
```
['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']
```

In [221…
```
print(X.category_name.value_counts())

# plot barchart for X
X.category_name.value_counts().plot(kind = 'bar',
                                    title = 'Category distribution',
                                    ylim = [0, 650],
                                    rot = 0, fontsize = 11, figsize = (8,3))
```

```
soc.religion.christian    599
sci.med                   594
comp.graphics             584
alt.atheism               480
Name: category_name, dtype: int64
```

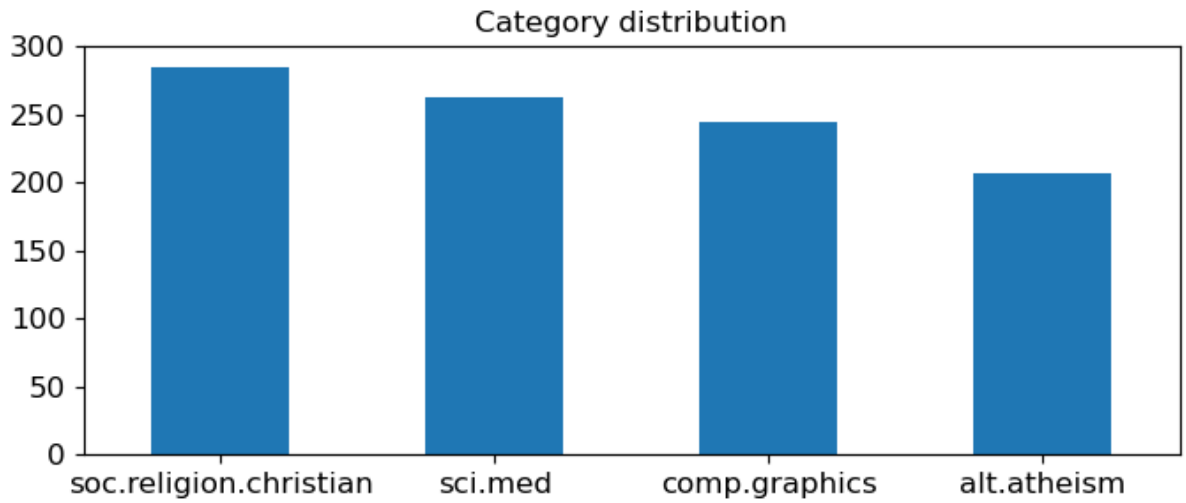
Out[221]:
```
<AxesSubplot:title={'center':'Category distribution'}>
```



In [222…
```
print(X_sample.category_name.value_counts())

# plot barchart for X_sample
X_sample.category_name.value_counts().plot(kind = 'bar',
                                           title = 'Category distribution',
                                           ylim = [0, 300],
                                           rot = 0, fontsize = 12, figsize = (8,3))
```

```
soc.religion.christian    285
sci.med                   263
comp.graphics             245
alt.atheism               207
Name: category_name, dtype: int64
```

Out[222]: <AxesSubplot:title={'center':'Category distribution'}>



You can use following command to see other available styles to prettify your charts.

```
print(plt.style.available)
```

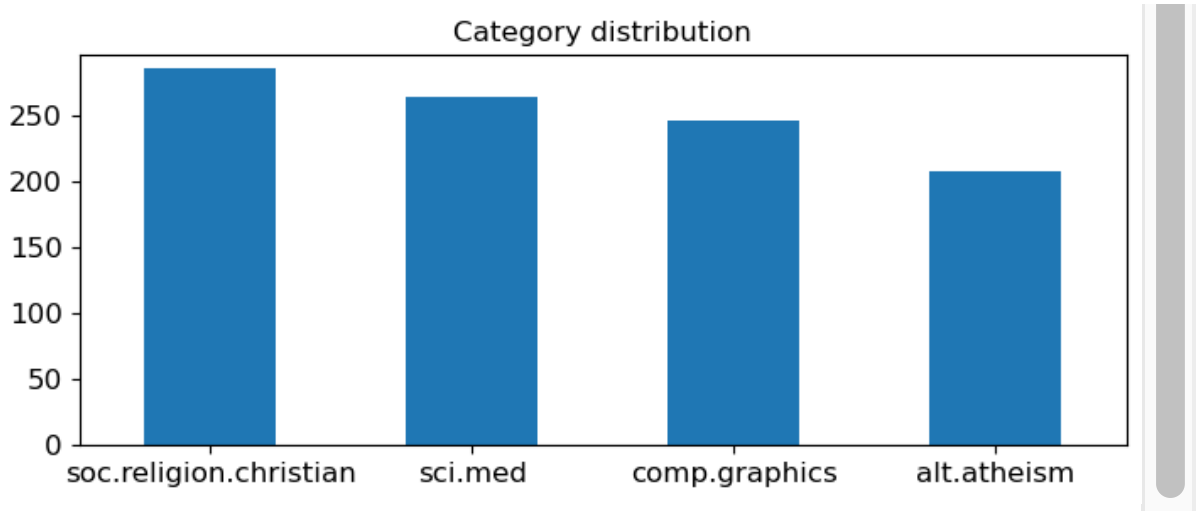
>>> Exercise 7 (5 min):

Notice that for the `ylim` parameters we hardcoded the maximum value for y. Is it possible to automate this instead of hard-coding it? How would you go about doing that? (Hint: look at code above for clues)

In [223…
```
# Answer here
upper_bound = max(X_sample.category_name.value_counts()) + 10

print(X_sample.category_name.value_counts())
# plot barchart for X_sample
X_sample.category_name.value_counts().plot(kind = 'bar',
                                           title = 'Category distribution',
                                           ylim = [0, upper_bound],
                                           rot = 0, fontsize = 12, figsize = (8,3))
```

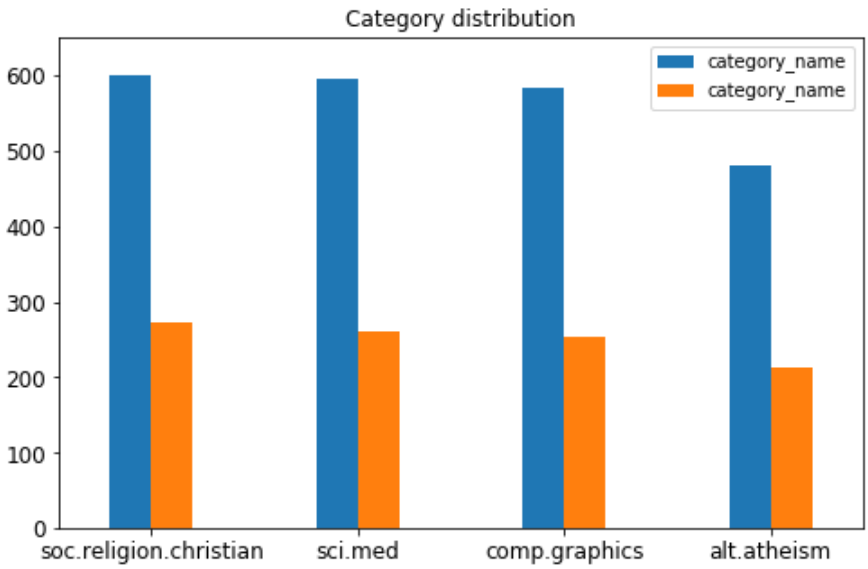
```
soc.religion.christian    285
sci.med                   263
comp.graphics             245
alt.atheism               207
Name: category_name, dtype: int64
```

Out[223]: <AxesSubplot:title={'center':'Category distribution'}>



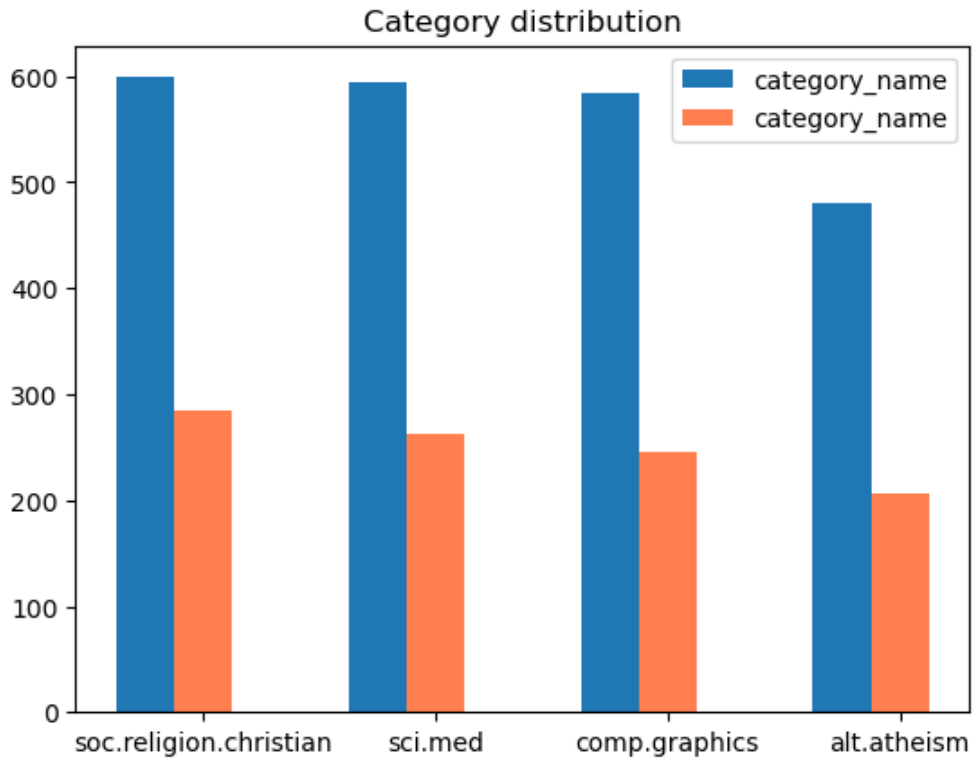
>>> Exercise 8 (take home):

We can also do a side-by-side comparison of the distribution between the two datasets, but maybe you can try that as an excerise. Below we show you an snapshot of the type of chart we are looking for.



In [224…

```
# Answer here

bar1 = X.category_name.value_counts()
bar2 = X_sample.category_name.value_counts()
index = np.arange(4)

plt.bar(index,bar1.tolist(), label='category_name', width=0.25)
plt.bar(index+0.25, bar2.tolist(), label='category_name', width=0.25, color= 'coral')
plt.xticks(index+0.25, ['soc.religion.christian', 'sci.med', 'comp.graphics','alt.ath
plt.title("Category distribution")
plt.legend()

plt.show()
```



One thing that stood out from the both datasets, is that the distribution of the categories remain relatively the same, which is a good sign for us data scientist. There are many ways to conduct sampling on the dataset and still obtain a representative enough dataset. That is not the main focus in this notebook, but if you would like to know more about sampling and how the `sample` feature works, just reference the Pandas documentation and you will find interesting ways to conduct more advanced sampling.

5.2 Feature Creation

The other operation from the list above that we are going to practise on is the so-called feature creation. As the name suggests, in feature creation we are looking at creating new interesting and useful features from the original dataset; a feature which captures the most important information from the raw information we already have access to. In our `X` table, we would like to create some features from the `text` field, but we are still not sure what kind of features we want to create. We can think of an interesting problem we want to solve, or something we want to analyze from the data, or some questions we want to answer. This is one process to come up with features -- this process is usually called `feature engineering` in the data science community.

We know what feature creation is so let us get real involved with our dataset and make it more interesting by adding some special features or attributes if you will. First, we are going to obtain the **unigrams** for each text. (Unigram is just a fancy word we use in Text Mining which stands for 'tokens' or 'individual words'.) Yes, we want to extract all the words found in each text and append it as a new feature to the pandas dataframe. The reason for extracting unigrams is not so clear yet, but we can start to think of obtaining some statistics about the articles we have: something like **word distribution** or **word frequency**.

Before going into any further coding, we will also introduce a useful text mining library called NLTK. The NLTK library is a natural language processing tool used for text mining tasks, so might as well we start to familiarize ourselves with it from now (It may come in handy for the final project!). In partcular, we are going to use the NLTK library to conduct tokenization because we are interested in

splitting a sentence into its individual components, which we refer to as words, emojis, emails, etc. So let us go for it! We can call the `nltk` library as follows:

```
import nltk
```

In [225…
```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/antinghsieh/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[225]:
```
True
```

In [226…
```
# takes a like a minute or two to process
X['unigrams'] = X['text'].apply(lambda x: dmh.tokenize_text(x))
```

In [227…
```
X[0:4]["unigrams"]
```

Out[227]:
```
0    [From, :, sd345, @, city.ac.uk, (, Michael, Co...
1    [From, :, ani, @, ms.uky.edu, (, Aniruddha, B....
2    [From, :, djohnson, @, cs.ucsd.edu, (, Darin, ...
3    [From, :, s0612596, @, let.rug.nl, (, M.M, ., ...
Name: unigrams, dtype: object
```

If you take a closer look at the `X` table now, you will see the new columns `unigrams` that we have added. You will notice that it contains an array of tokens, which were extracted from the original `text` field. At first glance, you will notice that the tokenizer is not doing a great job, let us take a closer at a single record and see what was the exact result of the tokenization using the `nltk` library.

In [228…
```
X[0:4]
```

Out[228]:

	text	category	category_name	unigrams
0	From: sd345@city.ac.uk (Michael Collier) Subje...	1	comp.graphics	[From, :, sd345, @, city.ac.uk, (, Michael, Co...
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...	1	comp.graphics	[From, :, ani, @, ms.uky.edu, (, Aniruddha, B....
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	3	soc.religion.christian	[From, :, djohnson, @, cs.ucsd.edu, (, Darin, ...
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	3	soc.religion.christian	[From, :, s0612596, @, let.rug.nl, (, M.M, ., ...

In [229…
```
list(X[0:1]['unigrams'])
```

Out[229]:
```
[['From',
  ':',
  'sd345',
  '@',
  'city.ac.uk',
  '(',
  'Michael',
  'Collier',
  ')',
  'Subject',
  ':',
  'Converting',
  'images',
  'to',
  'HP',
  'LaserJet',
  'III',
  '?',
  'Nntp-Posting-Host',
  ':',
  'hampton',
  'Organization',
  ':',
  'The',
  'City',
  'University',
  'Lines',
  ':',
  '14',
  'Does',
  'anyone',
  'know',
  'of',
  'a',
  'good',
  'way',
  '(',
  'standard',
  'PC',
  'application/PD',
  'utility',
  ')',
  'to',
  'convert',
  'tif/img/tga',
  'files',
  'into',
  'LaserJet',
  'III',
  'format',
  '.',
  'We',
  'would',
  'also',
  'like',
  'to',
  'do',
  'the',
  'same',
  ',',
  'converting',
  'to',
  'HPGL',
  '(',
  'HP',
  'plotter',
  ')',
  'files',
  '.',
  'Please',
  'email',
```

```
 'any',
 'response',
 '.',
 'Is',
 'this',
 'the',
 'correct',
 'group',
 '?',
 'Thanks',
 'in',
 'advance',
 '.',
 'Michael',
 '.',
 '--',
 'Michael',
 'Collier',
 '(',
 'Programmer',
 ')',
 'The',
 'Computer',
 'Unit',
 ',',
 'Email',
 ':',
 'M.P.Collier',
 '@',
 'uk.ac.city',
 'The',
 'City',
 'University',
 ',',
 'Tel',
 ':',
 '071',
 '477-8000',
 'x3769',
 'London',
 ',',
 'Fax',
 ':',
 '071',
 '477-8565',
 'EC1V',
 '0HB',
 '.']]
```

The `nltk` library does a pretty decent job of tokenizing our text. There are many other tokenizers online, such as spaCy, and the built in libraries provided by scikit-learn. We are making use of the NLTK library because it is open source and because it does a good job of segmentating text-based data.

5.3 Feature subset selection

Okay, so we are making some headway here. Let us now make things a bit more interesting. We are going to do something different from what we have been doing thus far. We are going use a bit of everything that we have learned so far. Briefly speaking, we are going to move away from our main dataset (one form of feature subset selection), and we are going to generate a document-term matrix from the original dataset. In other words we are going to be creating something like this.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Initially, it won't have the same shape as the table above, but we will get into that later. For now, let us use scikit learn built in functionalities to generate this document. You will see for yourself how easy it is to generate this table without much coding.

In [230…

```
from sklearn.feature_extraction.text import CountVectorizer

count_vect = CountVectorizer()
X_counts = count_vect.fit_transform(X.text) #learn the vocabulary and return document
print(X_counts[0])
```

```
  (0, 14887)    1
  (0, 29022)    1
  (0, 8696)     4
  (0, 4017)     2
  (0, 33256)    2
  (0, 21661)    3
  (0, 9031)     3
  (0, 31077)    1
  (0, 9805)     2
  (0, 17366)    1
  (0, 32493)    4
  (0, 16916)    2
  (0, 19780)    2
  (0, 17302)    2
  (0, 23122)    1
  (0, 25663)    1
  (0, 16881)    1
  (0, 16082)    1
  (0, 23915)    1
  (0, 32142)    5
  (0, 33597)    2
  (0, 20253)    1
  (0, 587)      1
  (0, 12051)    1
  (0, 5201)     1
  :     :
  (0, 25361)    1
  (0, 25337)    1
  (0, 12833)    2
  (0, 5195)     1
  (0, 27836)    1
  (0, 18474)    1
  (0, 32270)    1
  (0, 9932)     1
  (0, 15837)    1
  (0, 32135)    1
  (0, 17556)    1
  (0, 4378)     1
  (0, 26175)    1
  (0, 9338)     1
  (0, 33572)    1
  (0, 31915)    1
  (0, 177)      2
  (0, 2326)     2
  (0, 3062)     1
  (0, 35416)    1
  (0, 20459)    1
  (0, 14085)    1
  (0, 3166)     1
  (0, 12541)    1
  (0, 230)      1
```

What we did with those two lines of code is that we transformed the articles into a **term-document matrix**. Those lines of code tokenize each article using a built-in, default tokenizer (often referred to as an `analzyer`) and then produces the word frequency vector for each document. We can create our own analyzers or even use the nltk analyzer that we previously built. To keep things tidy and minimal we are going to use the default analyzer provided by `CountVectorizer`. Let us look closely at this analyzer.

In [231…
```
analyze = count_vect.build_analyzer()
analyze("I am craving for a hawaiian pizza right now")
```

Out[231]:
```
['am', 'craving', 'for', 'hawaiian', 'pizza', 'right', 'now']
```

>>> Exercise 9 (5 min):

Let's analyze the first record of our X dataframe with the new analyzer we have just built. Go ahead try it!

In [232…

```
# Answer here
analyze(" ".join(list(X[:1].text)))
```

Out[232]:
```
['from',
 'sd345',
 'city',
 'ac',
 'uk',
 'michael',
 'collier',
 'subject',
 'converting',
 'images',
 'to',
 'hp',
 'laserjet',
 'iii',
 'nntp',
 'posting',
 'host',
 'hampton',
 'organization',
 'the',
 'city',
 'university',
 'lines',
 '14',
 'does',
 'anyone',
 'know',
 'of',
 'good',
 'way',
 'standard',
 'pc',
 'application',
 'pd',
 'utility',
 'to',
 'convert',
 'tif',
 'img',
 'tga',
 'files',
 'into',
 'laserjet',
 'iii',
 'format',
 'we',
 'would',
 'also',
 'like',
 'to',
 'do',
 'the',
 'same',
 'converting',
 'to',
 'hpgl',
 'hp',
 'plotter',
 'files',
 'please',
 'email',
 'any',
 'response',
 'is',
 'this',
 'the',
 'correct',
 'group',
 'thanks',
 'in',
 'advance',
```

```
'michael',
'michael',
'collier',
'programmer',
'the',
'computer',
'unit',
'email',
'collier',
'uk',
'ac',
'city',
'the',
'city',
'university',
'tel',
'071',
'477',
'8000',
'x3769',
'london',
'fax',
'071',
'477',
'8565',
'ec1v',
'0hb']
```

Now let us look at the term-document matrix we built above.

In [233…
```
# We can check the shape of this matrix by:
X_counts.shape
```

Out[233]:
```
(2257, 35788)
```

In [234…
```
# We can obtain the feature names of the vectorizer, i.e., the terms
# usually on the horizontal axis
count_vect.get_feature_names_out()[0:10]
```

Out[234]:
```
array(['00', '000', '0000', '0000001200', '000005102000', '0001',
       '000100255pixel', '00014', '000406', '0007'], dtype=object)
```

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Above we can see the features found in the all the documents `X`, which are basically all the terms found in all the documents. As I said earlier, the transformation is not in the pretty format (table) we saw above -- the term-document matrix. We can do many things with the `count_vect` vectorizer and its transformation `X_counts`. You can find more information on other cool stuff you can do with the CountVectorizer.

Now let us try to obtain something that is as close to the pretty table I provided above. Before jumping into the code for doing just that, it is important to mention that the reason for choosing the `fit_transform` for the `CountVectorizer` is that it efficiently learns the vocabulary dictionary and returns a term-document matrix.

In the next bit of code, we want to extract the first five articles and transform them into document-term matrix, or in this case a 2-dimensional array. Here it goes.

In [235…

```
X_counts.shape
```

Out[235]:

```
(2257, 35788)
```

In [236…

```
# we convert from sparse array to normal array
X_counts[0:5, 0:100].toarray()
```

Out[236]:

```
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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,
        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, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])
```

In [237…

```
count_vect.get_feature_names_out()[0:1]
```

Out[237]:

```
array(['00'], dtype=object)
```

As you can see the result is just this huge sparse matrix, which is computationally intensive to generate and difficult to visualize. But we can see that the fifth record, specifically, contains a `1` in the beginning, which from our feature names we can deduce that this article contains exactly one `00` term.

>>> Exercise 10 (take home):

We said that the `1` at the beginning of the fifth record represents the `00` term. Notice that there is another 1 in the same record. Can you provide code that can verify what word this 1 represents from the vocabulary. Try to do this as efficient as possible.

In [238…

```
# Answer here

# Check the second one
order = 0
# choose the fifth record
fifRec = X_counts[4:5,0:100].toarray()
```

```
for i in range(100):
    if(fifRec[0][i] == 1):
        order = order + 1
    if(fifRec[0][i] == 1 and order ==2):
        print("The word is", count_vect.get_feature_names()[i])
        break
```

```
The word is 01
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

To get you started in thinking about how to better analyze your data or transformation, let us look at this nice little heat map of our term-document matrix. It may come as a surpise to see the gems you can mine when you start to look at the data from a different perspective. Visualization are good for this reason.

In [239…

```
# first twenty features only
plot_x = ["term_"+str(i) for i in count_vect.get_feature_names_out()[0:20]]
```

In [240…

```
# obtain document index
plot_y = ["doc_"+ str(i) for i in list(X.index)[0:20]]
```

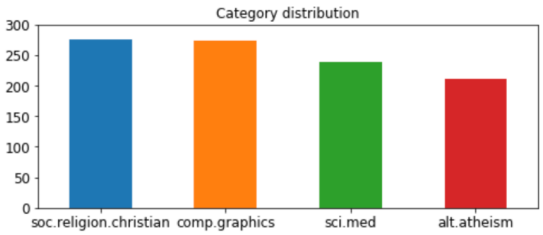
In [241…

```
plot_z = X_counts[0:20, 0:20].toarray()
plot_z
```

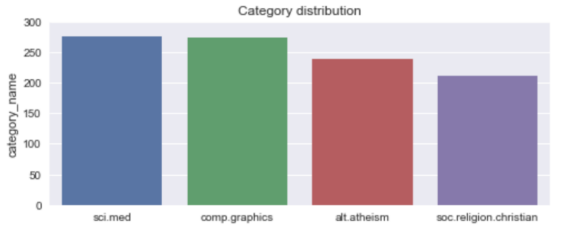
Out[241]:

```
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 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],
       [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])
```

For the heat map, we are going to use another visualization library called `seaborn`. It's built on top of matplotlib and closely integrated with pandas data structures. One of the biggest advantages of seaborn is that its default aesthetics are much more visually appealing than matplotlib. See comparison below.



By Matplotlib



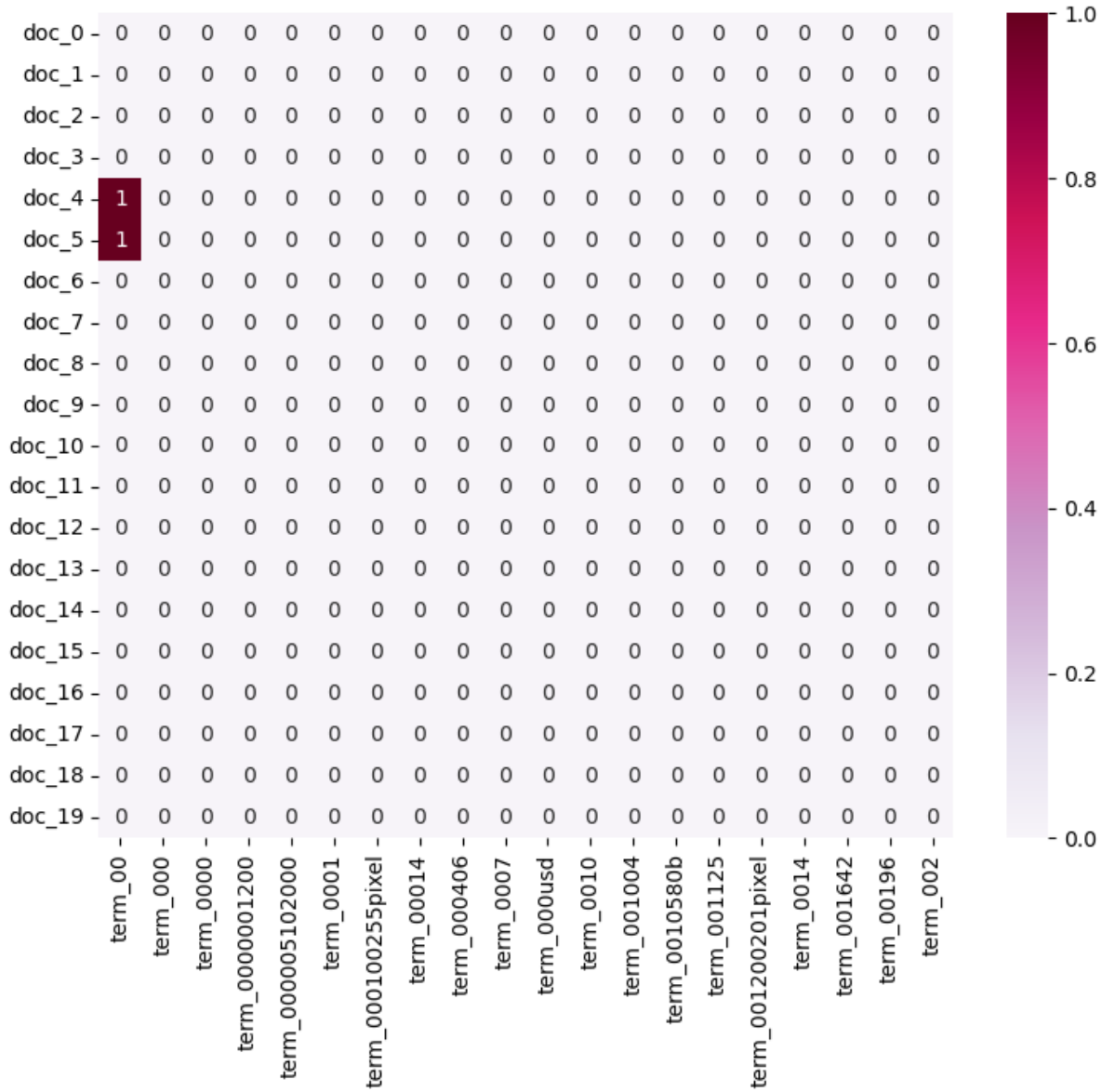
By Seaborn

The other big advantage of seaborn is that seaborn has some built-in plots that matplotlib does not support. Most of these can eventually be replicated by hacking away at matplotlib, but they're not built in and require much more effort to build.

So without further ado, let us try it now!

In [242…

```
import seaborn as sns

df_todraw = pd.DataFrame(plot_z, columns = plot_x, index = plot_y)
plt.subplots(figsize=(9, 7))
ax = sns.heatmap(df_todraw,
                 cmap="PuRd",
                 vmin=0, vmax=1, annot=True)
```



Check out more beautiful color palettes here: https://python-graph-gallery.com/197-available-

color-palettes-with-matplotlib/

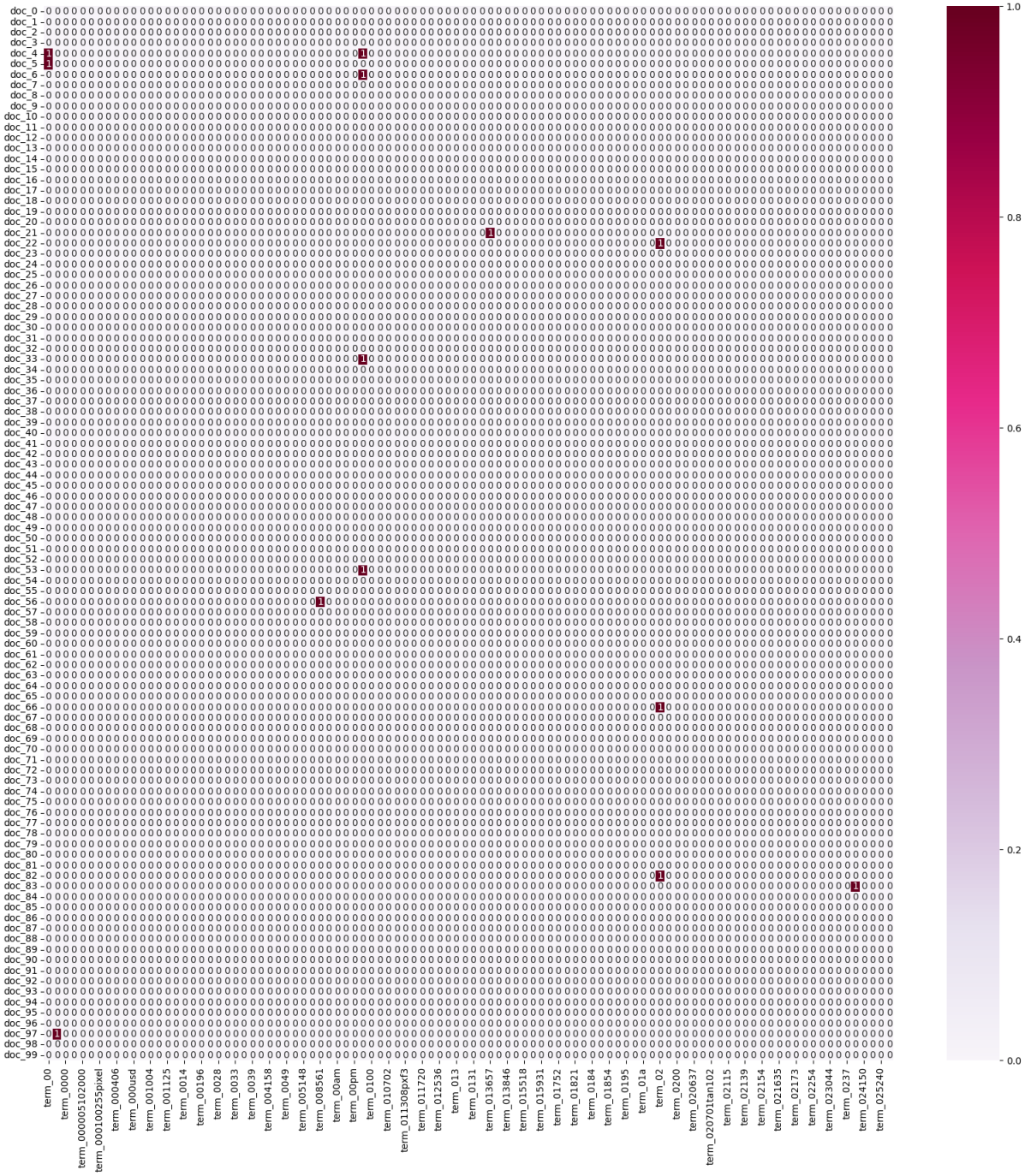
>>> Exercise 11 (take home):

From the chart above, we can see how sparse the term-document matrix is; i.e., there is only one terms with frequency of `1` in the subselection of the matrix. By the way, you may have noticed that we only selected 20 articles and 20 terms to plot the histrogram. As an excersise you can try to modify the code above to plot the entire term-document matrix or just a sample of it. How would you do this efficiently? Remember there is a lot of words in the vocab. Report below what methods you would use to get a nice and useful visualization

In [243…

```
# Answer here
print("GO!")
# x,y,z
plot_x_all = ["term_" + str(i) for i in count_vect.get_feature_names()[0:100]]
plot_y_all = ["doc_" + str(i) for i in list(X.index)[0:100]]
plot_z_all = X_counts[0:100, 0:100].toarray()

# plot
df_all_todraw = pd.DataFrame(plot_z_all, columns = plot_x_all, index = plot_y_all)
plt.subplots(figsize=(20, 20))
ax = sns.heatmap(df_all_todraw,
                 cmap="PuRd",
                 vmin=0, vmax=1, annot=True)

# plt.show()
print('END')
```

```
GO!
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

END
```



The great thing about what we have done so far is that we now open doors to new problems. Let us be optimistic. Even though we have the problem of sparsity and a very high dimensional data, we are now closer to uncovering wonders from the data. You see, the price you pay for the hard work is worth it because now you are gaining a lot of knowledge from what was just a list of what appeared to be irrelevant articles. Just the fact that you can blow up the data and find out interesting characteristics about the dataset in just a couple lines of code, is something that truly inspires me to practise Data Science. That's the motivation right there!

5.4 Dimensionality Reduction

Since we have just touched on the concept of sparsity most naturally the problem of "curse of dimentionality" comes up. I am not going to get into the full details of what dimensionality reduction is and what it is good for just the fact that is an excellent technique for visualizing data efficiently (please refer to notes for more information). All I can say is that we are going to deal with the issue

of sparsity with a few lines of code. And we are going to try to visualize our data more efficiently with the results.

We are going to make use of Principal Component Analysis to efficiently reduce the dimensions of our data, with the main goal of "finding a projection that captures the largest amount of variation in the data." This concept is important as it is very useful for visualizing and observing the characteristics of our dataset.

PCA Algorithm

Input: Raw term-vector matrix

Output: Projections

In [244…
```
from sklearn.decomposition import PCA
```

In [245…
```
X_reduced = PCA(n_components = 2).fit_transform(X_counts.toarray())
```

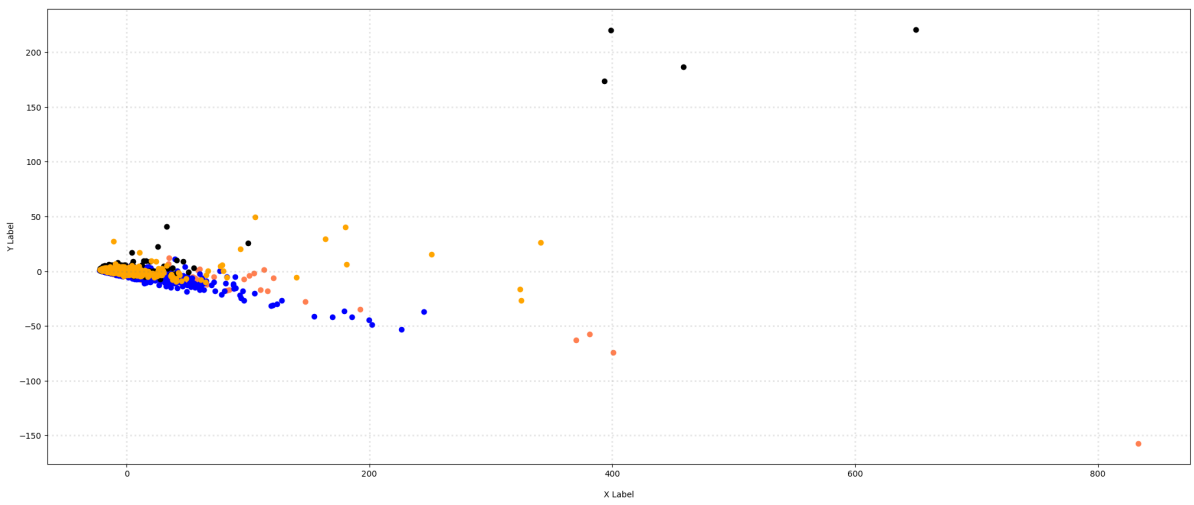
In [246…
```
X_reduced.shape
```

Out[246]:
```
(2257, 2)
```

In [247…
```
categories
```

Out[247]:
```
['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']
```

In [248…
```
col = ['coral', 'blue', 'black', 'orange']

# plot
fig = plt.figure(figsize = (25,10))
ax = fig.subplots()

for c, category in zip(col, categories):
    xs = X_reduced[X['category_name'] == category].T[0]
    ys = X_reduced[X['category_name'] == category].T[1]

    ax.scatter(xs, ys, c = c, marker='o')

ax.grid(color='gray', linestyle=':', linewidth=2, alpha=0.2)
ax.set_xlabel('\nX Label')
ax.set_ylabel('\nY Label')

plt.show()
```



From the 2D visualization above, we can see a slight "hint of separation in the data"; i.e., they might have some special grouping by category, but it is not immediately clear. The PCA was applied to the raw frequencies and this is considered a very naive approach as some words are not really unique

to a document. Only categorizing by word frequency is considered a "bag of words" approach. Later on in the course you will learn about different approaches on how to create better features from the term-vector matrix, such as term-frequency inverse document frequency so-called TF-IDF.

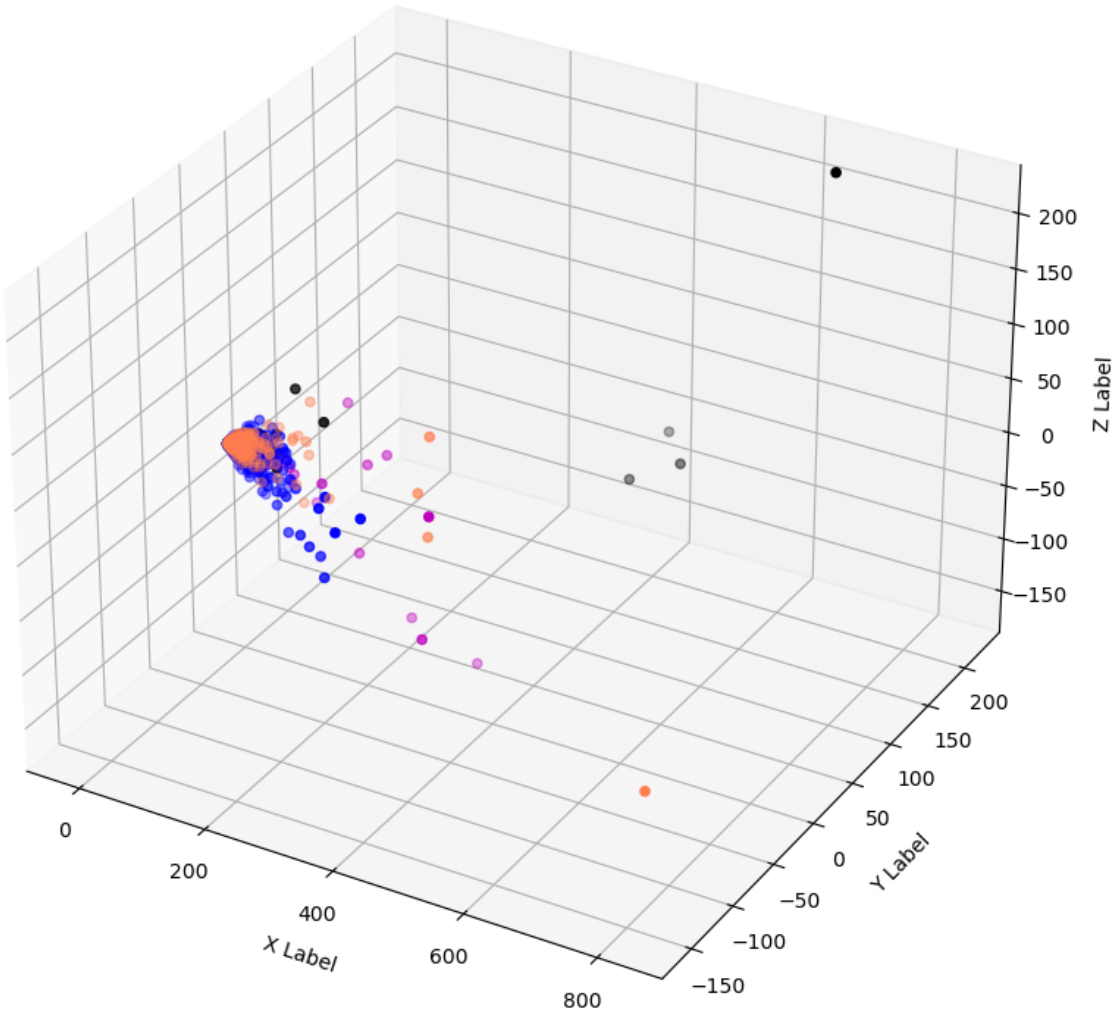
>>> Exercise 12 (take home):

Please try to reduce the dimension to 3, and plot the result use 3-D plot. Use at least 3 different angle (camera position) to check your result and describe what you found.

Hint: you can refer to Axes3D in the documentation.

In [249…

```
# Answer here
from mpl_toolkits.mplot3d import Axes3D

X_3dim = PCA(n_components = 3).fit_transform(X_counts.toarray())

# plot
col = ['coral', 'blue', 'black', 'm']

fig = plt.figure(figsize = (25,10))
ax = fig.add_subplot(111, projection='3d')


for c, category in zip(col, categories):
    xs = X_3dim[X['category_name'] == category].T[0]
    ys = X_3dim[X['category_name'] == category].T[1]
    zs = X_3dim[X['category_name'] == category].T[2]

    ax.scatter(xs, ys, zs, c = c, marker='o')

ax.grid(color='gray', linestyle=':', linewidth=2, alpha=0.2)
ax.set_xlabel('\nX Label')
ax.set_ylabel('\nY Label')
ax.set_zlabel('\nZ Label')

plt.show()
```



5.5 Attribute Transformation / Aggregation

We can do other things with the term-vector matrix besides applying dimensionality reduction technique to deal with sparsity problem. Here we are going to generate a simple distribution of the words found in all the entire set of articles. Intuitively, this may not make any sense, but in data science sometimes we take some things for granted, and we just have to explore the data first before making any premature conclusions. On the topic of attribute transformation, we will take the word distribution and put the distribution in a scale that makes it easy to analyze patterns in the distrubution of words. Let us get into it!

First, we need to compute these frequencies for each term in all documents. Visually speaking, we are seeking to add values of the 2D matrix, vertically; i.e., sum of each column. You can also refer to this process as aggregation, which we won't explore further in this notebook because of the type of data we are dealing with. But I believe you get the idea of what that includes.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0
	3	8	5	2	5	8	2	5	3	2

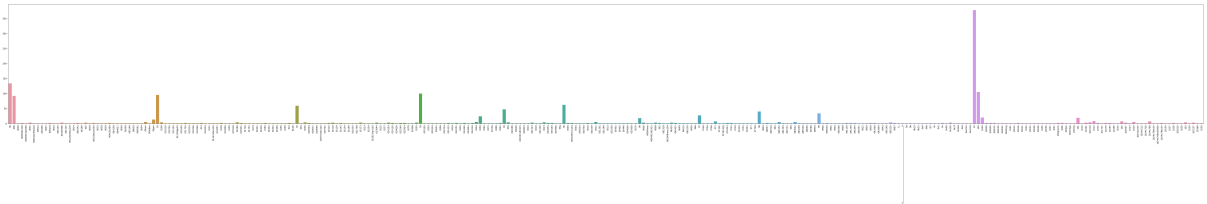
In [250…
```
# note this takes time to compute. You may want to reduce the amount of terms you wan
term_frequencies = []
for j in range(0,X_counts.shape[1]):
    term_frequencies.append(sum(X_counts[:,j].toarray()))
```

In [251…
```
term_frequencies = np.asarray(X_counts.sum(axis=0))[0]
```

In [252…
```
term_frequencies[0] #sum of first term
```

Out[252]:
```
134
```

In [253…
```
plt.subplots(figsize=(100, 10))
g = sns.barplot(x=count_vect.get_feature_names_out()[:300],
            y=term_frequencies[:300])
g.set_xticklabels(count_vect.get_feature_names_out()[:300], rotation = 90);
```



>>> Exercise 13 (take home):

If you want a nicer interactive visualization here, I would encourage you try to install and use plotly to achieve this.

In [254…
```
# Answer here

!pip install plotly
!pip install plotly --upgrade
!pip install chart_studio
```

```
Requirement already satisfied: plotly in /Users/antinghsieh/opt/anaconda3/envs/dataM
ining/lib/python3.9/site-packages (5.10.0)
Requirement already satisfied: tenacity>=6.2.0 in /Users/antinghsieh/opt/anaconda3/e
nvs/dataMining/lib/python3.9/site-packages (from plotly) (8.0.1)
Requirement already satisfied: plotly in /Users/antinghsieh/opt/anaconda3/envs/dataM
ining/lib/python3.9/site-packages (5.10.0)
Requirement already satisfied: tenacity>=6.2.0 in /Users/antinghsieh/opt/anaconda3/e
nvs/dataMining/lib/python3.9/site-packages (from plotly) (8.0.1)
Requirement already satisfied: chart_studio in /Users/antinghsieh/opt/anaconda3/env
s/dataMining/lib/python3.9/site-packages (1.1.0)
Requirement already satisfied: retrying>=1.3.3 in /Users/antinghsieh/opt/anaconda3/e
nvs/dataMining/lib/python3.9/site-packages (from chart_studio) (1.3.3)
Requirement already satisfied: plotly in /Users/antinghsieh/opt/anaconda3/envs/dataM
ining/lib/python3.9/site-packages (from chart_studio) (5.10.0)
Requirement already satisfied: six in /Users/antinghsieh/opt/anaconda3/envs/dataMini
ng/lib/python3.9/site-packages (from chart_studio) (1.16.0)
Requirement already satisfied: requests in /Users/antinghsieh/opt/anaconda3/envs/dat
aMining/lib/python3.9/site-packages (from chart_studio) (2.28.1)
Requirement already satisfied: tenacity>=6.2.0 in /Users/antinghsieh/opt/anaconda3/e
nvs/dataMining/lib/python3.9/site-packages (from plotly->chart_studio) (8.0.1)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /Users/antinghsieh/opt/anaco
nda3/envs/dataMining/lib/python3.9/site-packages (from requests->chart_studio) (1.2
6.11)
Requirement already satisfied: certifi>=2017.4.17 in /Users/antinghsieh/opt/anaconda
3/envs/dataMining/lib/python3.9/site-packages (from requests->chart_studio) (2022.9.
14)
Requirement already satisfied: charset-normalizer<3,>=2 in /Users/antinghsieh/opt/an
aconda3/envs/dataMining/lib/python3.9/site-packages (from requests->chart_studio)
(2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /Users/antinghsieh/opt/anaconda3/env
s/dataMining/lib/python3.9/site-packages (from requests->chart_studio) (3.3)
```

In [255…

```python
# Answer here
import plotly
import plotly as py
import plotly.graph_objs as go
import chart_studio
import plotly.offline as pyo
from plotly.offline import iplot
import plotly.figure_factory as ff

# Cconstruct the df with term and frequencies
df_term_frequencies = pd.DataFrame(columns = ["term", "frequencies"])

for i in range(300):
    df_term_frequencies.loc[i, "term"] = str(count_vect.get_feature_names()[i])
    df_term_frequencies.loc[i, "frequencies"] = int(term_frequencies[i])

# plotly interactive visualization
term = df_term_frequencies.term
plotly_data = [go.Bar(x=df_term_frequencies.term, y=df_term_frequencies.frequencies)]
layout = go.Layout(title='Interactive Visualization Figure by plotly', font={'size':3
fig = go.Figure(data=plotly_data, layout=layout)
plotly.offline.init_notebook_mode()
plotly.offline.iplot(fig,filename='basic-scatter')
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

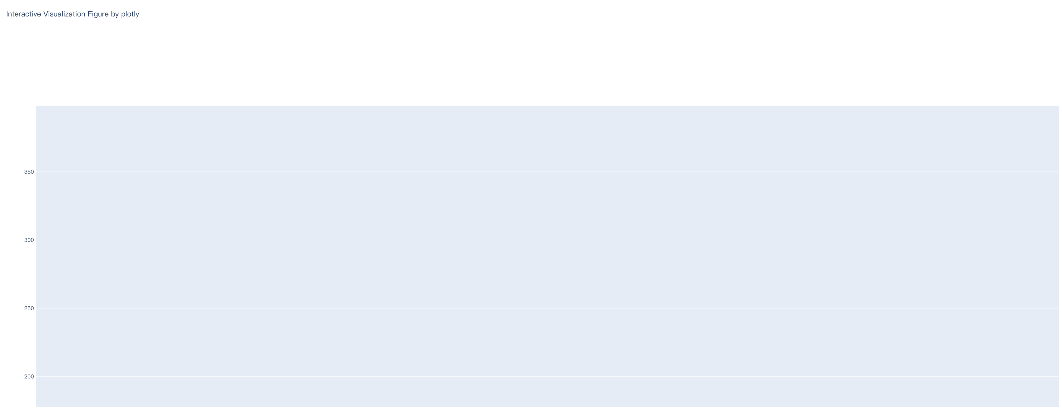
```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```



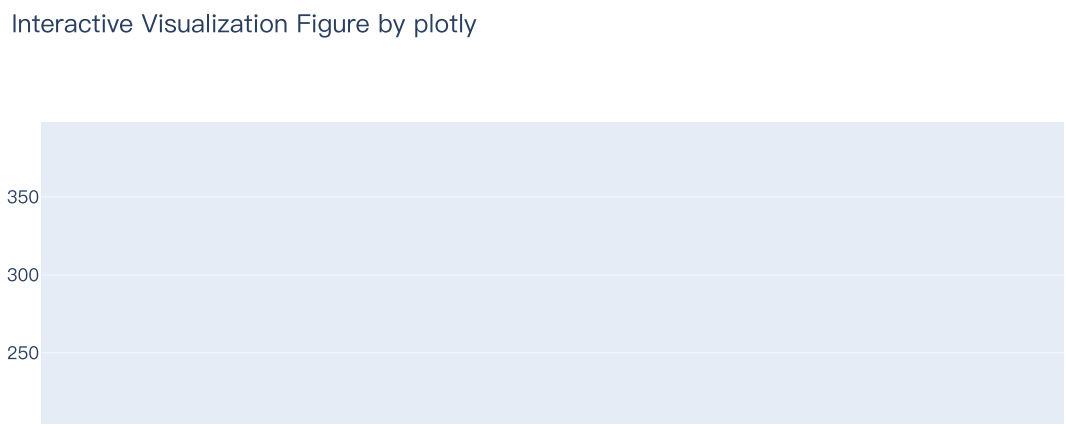
>>> Exercise 14 (take home):

The chart above contains all the vocabulary, and it's computationally intensive to both compute and visualize. Can you efficiently reduce the number of terms you want to visualize as an exercise.

In [256…

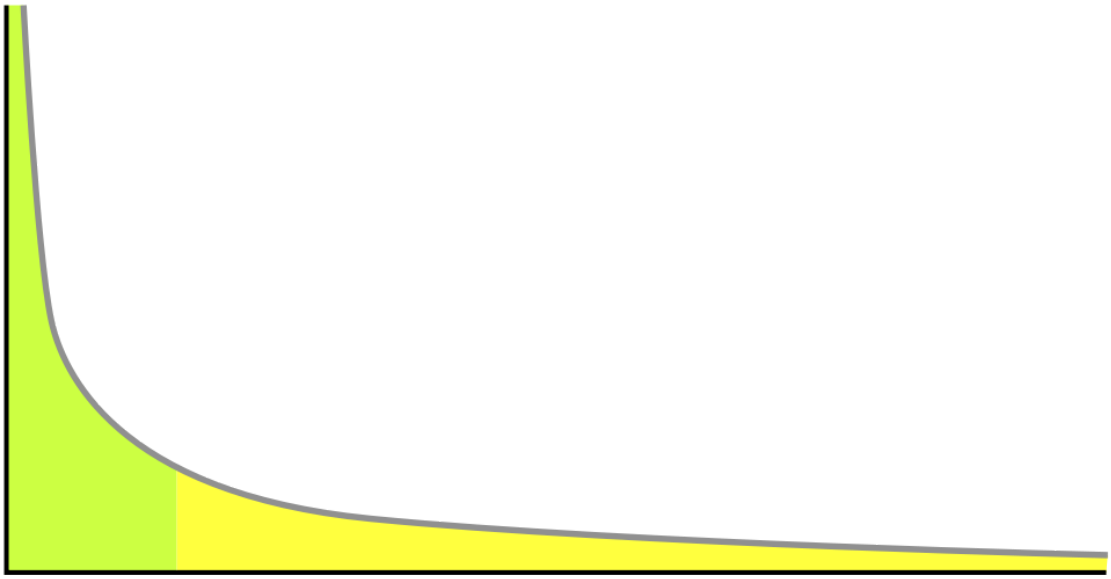
```python
# Answer here

# Screening by the frequency of the term > 15
# Time Complexity of Screening by this function only O(1)
df_reduce = df_term_frequencies[df_term_frequencies["frequencies"]>15]
```

```
# plotly interactive visualization
reduce_data = [go.Bar(x=df_reduce.term, y=df_reduce.frequencies)]
reduce_layout = go.Layout(title='Interactive Visualization Figure by plotly', font={'
reduce_fig = go.Figure(data=reduce_data, layout=reduce_layout)
plotly.offline.init_notebook_mode()
plotly.offline.iplot(reduce_fig,filename='basic-scatter')
```



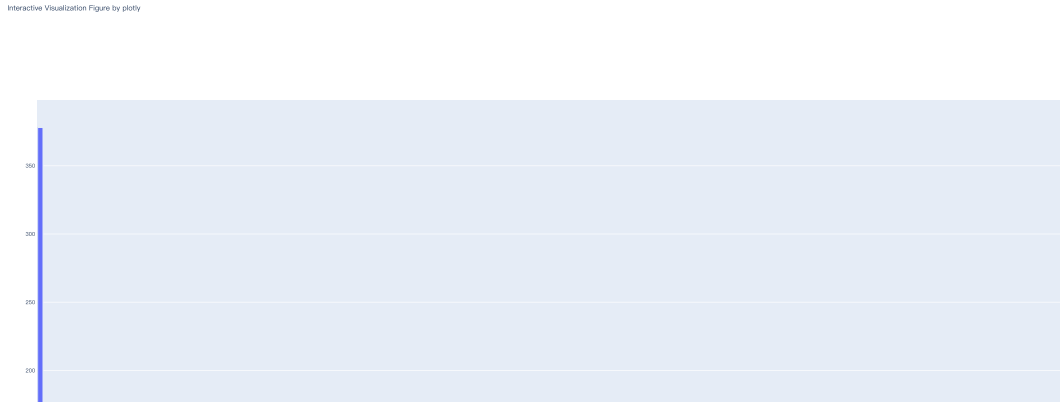
>>> Exercise 15 (take home):

Additionally, you can attempt to sort the terms on the `x-axis` by frequency instead of in alphabetical order. This way the visualization is more meaninfgul and you will be able to observe the so called long tail (get familiar with this term since it will appear a lot in data mining and other statistics courses). see picture below



In [257…

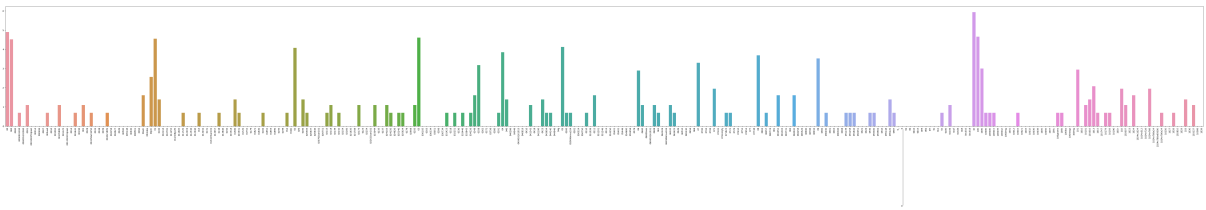
```
# Answer here

# Sort the terms on the x-axis by frequency
df_sort = df_term_frequencies.sort_values('frequencies', ascending = False)

# plotly interactive visualization
term = df_sort.term
plotly_data = [go.Bar(x=df_sort.term, y=df_sort.frequencies)]
layout = go.Layout(title='Interactive Visualization Figure by plotly', font={'size':3
fig = go.Figure(data=plotly_data, layout=layout)
plotly.offline.init_notebook_mode()
plotly.offline.iplot(fig,filename='basic-scatter')
```



Since we already have those term frequencies, we can also transform the values in that vector into the log distribution. All we need is to import the `math` library provided by python and apply it to the array of values of the term frequency vector. This is a typical example of attribute transformation. Let's go for it. The log distribution is a technique to visualize the term frequency into a scale that makes you easily visualize the distribution in a more readable format. In other words, the variations between the term frequencies are now easy to observe. Let us try it out!

In [258…
```
import math
term_frequencies_log = [math.log(i) for i in term_frequencies]
```

In [259…
```
plt.subplots(figsize=(100, 10))
g = sns.barplot(x=count_vect.get_feature_names_out()[:300],
                y=term_frequencies_log[:300])
g.set_xticklabels(count_vect.get_feature_names_out()[:300], rotation = 90);
```



Besides observing a complete transformation on the disrtibution, notice the scale on the y-axis. The log distribution in our unsorted example has no meaning, but try to properly sort the terms by their frequency, and you will see an interesting effect. Go for it!

5.6 Discretization and Binarization

In this section we are going to discuss a very important pre-preprocessing technique used to transform the data, specifically categorical values, into a format that satisfies certain criteria required by particular algorithms. Given our current original dataset, we would like to transform one of the attributes, `category_name`, into four binary attributes. In other words, we are taking the category name and replacing it with a `n` asymmetric binary attributes. The logic behind this transformation is discussed in detail in the recommended Data Mining text book (please refer to it on page 58). People from the machine learning community also refer to this transformation as one-hot encoding, but as you may become aware later in the course, these concepts are all the same, we just have different prefrence on how we refer to the concepts. Let us take a look at what we want to achieve in code.

In [260…
```
from sklearn import preprocessing, metrics, decomposition, pipeline, dummy
```

In [261…
```
mlb = preprocessing.LabelBinarizer()
```

In [262…
```
mlb.fit(X.category)
```

Out[262]:
```
▾ LabelBinarizer
LabelBinarizer()
```

In [263…
```
X['bin_category'] = mlb.transform(X['category']).tolist()
```

In [264…
```
X[0:9]
```

Out[264]:

	text	category	category_name	unigrams	bin_cate
0	From: sd345@city.ac.uk (Michael Collier) Subje...	1	comp.graphics	[From, :, sd345, @, city.ac.uk, (, Michael, Co...	[0, 1,
1	From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...	1	comp.graphics	[From, :, ani, @, ms.uky.edu, (, Aniruddha, B....	[0, 1,
2	From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...	3	soc.religion.christian	[From, :, djohnson, @, cs.ucsd.edu, (, Darin, ...	[0, 0,
3	From: s0612596@let.rug.nl (M.M. Zwart) Subject...	3	soc.religion.christian	[From, :, s0612596, @, let.rug.nl, (, M.M, ., ...	[0, 0,
4	From: stanly@grok11.columbiasc.ncr.com (stanly...	3	soc.religion.christian	[From, :, stanly, @, grok11.columbiasc.ncr.com...	[0, 0,
5	From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...	3	soc.religion.christian	[From, :, vbv, @, lor.eeap.cwru.edu, (, Virgil...	[0, 0,
6	From: jodfishe@silver.ucs.indiana.edu (joseph ...	3	soc.religion.christian	[From, :, jodfishe, @, silver.ucs.indiana.edu,...	[0, 0,
7	From: aldridge@netcom.com (Jacquelin Aldridge)...	2	sci.med	[From, :, aldridge, @, netcom.com, (, Jacqueli...	[0, 0,
8	From: geb@cs.pitt.edu (Gordon Banks) Subject: ...	2	sci.med	[From, :, geb, @, cs.pitt.edu, (, Gordon, Bank...	[0, 0,

Take a look at the new attribute we have added to the `X` table. You can see that the new attribute, which is called `bin_category`, contains an array of 0's and 1's. The `1` is basically to indicate the position of the label or category we binarized. If you look at the first two records, the one is places in slot 2 in the array; this helps to indicate to any of the algorithms which we are feeding this data to, that the record belong to that specific category.

Attributes with **continuous values** also have strategies to tranform the data; this is usually called **Discretization** (please refer to the text book for more inforamation).

>>> Exercise 16 (take home):

Try to generate the binarization using the `category_name` column instead. Does it work?

In [265…

```
# Answer here

"""
Yes, it works!
However, category is the category_name who converted to the dummy code,
and bin_category and bin_categoryNameare are the same meaning and the same value.

The results can be justified by the following code.
"""

mlb.fit(X.category_name)
mlb.classes_
X['bin_categoryName'] = mlb.transform(X['category_name']).tolist()

print(X[0:9])
```

```
                                                text  category  \
0  From: sd345@city.ac.uk (Michael Collier) Subje...         1
1  From: ani@ms.uky.edu (Aniruddha B. Deglurkar) ...         1
2  From: djohnson@cs.ucsd.edu (Darin Johnson) Sub...         3
3  From: s0612596@let.rug.nl (M.M. Zwart) Subject...         3
4  From: stanly@grok11.columbiasc.ncr.com (stanly...         3
5  From: vbv@lor.eeap.cwru.edu (Virgilio (Dean) B...         3
6  From: jodfishe@silver.ucs.indiana.edu (joseph ...         3
7  From: aldridge@netcom.com (Jacquelin Aldridge)...         2
8  From: geb@cs.pitt.edu (Gordon Banks) Subject: ...         2

            category_name                                           unigrams  \
0          comp.graphics  [From, :, sd345, @, city.ac.uk, (, Michael, Co...
1          comp.graphics  [From, :, ani, @, ms.uky.edu, (, Aniruddha, B....
2  soc.religion.christian  [From, :, djohnson, @, cs.ucsd.edu, (, Darin, ...
3  soc.religion.christian  [From, :, s0612596, @, let.rug.nl, (, M.M, ., ...
4  soc.religion.christian  [From, :, stanly, @, grok11.columbiasc.ncr.com...
5  soc.religion.christian  [From, :, vbv, @, lor.eeap.cwru.edu, (, Virgil...
6  soc.religion.christian  [From, :, jodfishe, @, silver.ucs.indiana.edu,...
7                sci.med  [From, :, aldridge, @, netcom.com, (, Jacqueli...
8                sci.med  [From, :, geb, @, cs.pitt.edu, (, Gordon, Bank...

   bin_category bin_categoryName
0  [0, 1, 0, 0]     [0, 1, 0, 0]
1  [0, 1, 0, 0]     [0, 1, 0, 0]
2  [0, 0, 0, 1]     [0, 0, 0, 1]
3  [0, 0, 0, 1]     [0, 0, 0, 1]
4  [0, 0, 0, 1]     [0, 0, 0, 1]
5  [0, 0, 0, 1]     [0, 0, 0, 1]
6  [0, 0, 0, 1]     [0, 0, 0, 1]
7  [0, 0, 1, 0]     [0, 0, 1, 0]
8  [0, 0, 1, 0]     [0, 0, 1, 0]
```

---

6. Data Exploration

Sometimes you need to take a peek at your data to understand the relationships in your dataset. Here, we will focus in a similarity example. Let's take 3 documents and compare them.

In [266…

```
# We retrieve 3 sentences for a random record
document_to_transform_1 = []
random_record_1 = X.iloc[50]
random_record_1 = random_record_1['text']
document_to_transform_1.append(random_record_1)

document_to_transform_2 = []
random_record_2 = X.iloc[100]
random_record_2 = random_record_2['text']
document_to_transform_2.append(random_record_2)

document_to_transform_3 = []
random_record_3 = X.iloc[150]
random_record_3 = random_record_3['text']
document_to_transform_3.append(random_record_3)
```

Let's look at our emails.

In [267…

```
print(document_to_transform_1)
print(document_to_transform_2)
print(document_to_transform_3)
```

['From: ab@nova.cc.purdue.edu (Allen B) Subject: Re: TIFF: philosophical significanc
e of 42 Organization: Purdue University Lines: 39 In article <prestonm.735400848@c
s.man.ac.uk> prestonm@cs.man.ac.uk (Martin Preston) writes: > Why not use the PD C
library for reading/writing TIFF files? It took me a > good 20 minutes to start usin
g them in your own app. I certainly do use it whenever I have to do TIFF, and it us
ually works very well. That\'s not my point. I\'m >philosophically< opposed to it
because of its complexity. This complexity has led to some programs\' poor TIFF wri
ters making some very bizarre files, other programs\' inability to load TIFF images
(though they\'ll save them, of course), and a general inability to interchange image
s between different environments despite the fact they all think they understand TIF
F. As the saying goes, "It\'s not me I\'m worried about- it\'s all the >other< ass
holes out there!" I\'ve had big trouble with misuse and abuse of TIFF over the year
s, and I chalk it all up to the immense (and unnecessary) complexity of the format.
In the words of the TIFF 5.0 spec, Appendix G, page G-1 (capitalized emphasis mine):
"The only problem with this sort of success is that TIFF was designed to be powerful
and flexible, at the expense of simplicity. It takes a fair amount of effort to han
dle all the options currently defined in this specification (PROBABLY NO APPLICATION
DOES A COMPLETE JOB), and that is currently the only way you can be >sure< that you
will be able to import any TIFF image, since there are so many image-generating appl
ications out there now." If a program (or worse all applications) can\'t read >eve
ry< TIFF image, that means there are some it won\'t- some that I might have to deal
with. Why would I want my images to be trapped in that format? I don\'t and neithe
r should anyone who agrees with my reasoning- not that anyone does, of course! :-)
ab ']
['From: mathew <mathew@mantis.co.uk> Subject: Re: university violating separation of
church/state? Organization: Mantis Consultants, Cambridge. UK. X-Newsreader: rusnews
v1.01 Lines: 29 dmn@kepler.unh.edu (...until kings become philosophers or philosoph
ers become kings) writes: > Recently, RAs have been ordered (and none have resi
sted or cared about > it apparently) to post a religious flyer entitled _The Soul Sc
roll: Thoughts > on religion, spirituality, and matters of the soul_ on the inside o
f bathroom > stall doors. (at my school, the University of New Hampshire) It is some
sort > of newsletter assembled by a Hall Director somewhere on campus. It poses a >
question about \'spirituality\' each issue, and solicits responses to be > included
in the next \'issue.\' It\'s all pretty vague. I assume it\'s put out > by a Christi
an, but they\'re very careful not to mention Jesus or the bible. > I\'ve heard someo
ne defend it, saying "Well it doesn\'t support any one religion. > " So what??? This
is a STATE university, and as a strong supporter of the > separation of church and s
tate, I was enraged. > > What can I do about this? It sounds to me like it\'s
just SCREAMING OUT for parody. Give a copy to your friendly neighbourhood SubGenius
preacher; with luck, he\'ll run it through the mental mincer and hand you back an ou
trageously offensive and gut-bustingly funny parody you can paste over the original
s. I can see it now: The Stool Scroll Thou
ghts on Religion, Spirituality, and Matters of the Colon (Yo
u can use this text to wipe) mathew ']
['From: lfoard@hopper.virginia.edu (Lawrence C. Foard) Subject: Re: Assurance of Hel
l Organization: ITC/UVA Community Access UNIX/Internet Project Lines: 43 In article
<Apr.20.03.01.19.1993.3755@geneva.rutgers.edu> REXLEX@fnal.fnal.gov writes: > >I dre
amed that the great judgment morning had dawned, > and the trumpet had blown. >I
dreamed that the sinners had gathered for judgment > before the white throne. >O
h what weeping and wailing as the lost were told of their fate. >They cried for the
rock and the mountains. >They prayed, but their prayers were too late. >The soul tha
t had put off salvation, >"Not tonight I\'ll get saved by and by. > No time now to
think of ..... religion," >Alas, he had found time to die. >And I saw a Great Whi
te Throne. If I believed in the God of the bible I would be very fearful of making
this statement. Doesn\'t it say those who judge will be judged by the same measure?
>Now, some have protest by saying that the fear of hell is not good for >motivation,
yet Jesus thought it was. Paul thought it was. Paul said, >"Knowing therefore, th
e terror of the Lord, we persuade men." A God who must motivate through fear is not
a God worthy of worship. If the God Jesus spoke of did indeed exist he would not nee
d hell to convince people to worship him. >Today, too much of our evangelism is not
hing but soft soap and some of >it is nothing but evangelical salesmanship. We don
\'t tell people anymore, that >there\'s such a thing as sin or that there\'s such a
place as hell. It was the myth of hell that made me finally realize that the whol
e thing was untrue. If it hadn\'t been for hell I would still be a believer today. T
he myth of hell made me realize that if there was a God that he was not the all know
ing and all good God he claimed to be. Why should I take such a being at his word, e
ven if there was evidence for his existance? -- ------ Join the Pythagore
an Reform Church! . \\ / Repent of your evil irrational numb
ers . . \\ / and bean eating ways. Accept 10 into your heart!
. . . \\/ Call the Pythagorean Reform Church BBS at 508-793-9568 . . . .
']

In [268…
```
from sklearn.preprocessing import binarize

# Transform sentence with Vectorizers
document_vector_count_1 = count_vect.transform(document_to_transform_1)
document_vector_count_2 = count_vect.transform(document_to_transform_2)
document_vector_count_3 = count_vect.transform(document_to_transform_3)

# Binarize vectors to simplify: 0 for abscence, 1 for prescence
document_vector_count_1_bin = binarize(document_vector_count_1)
document_vector_count_2_bin = binarize(document_vector_count_2)
document_vector_count_3_bin = binarize(document_vector_count_3)

# print vectors
print("Let's take a look at the count vectors:")
print(document_vector_count_1.todense())
print(document_vector_count_2.todense())
print(document_vector_count_3.todense())
```

```
Let's take a look at the count vectors:
[[0 0 0 ... 0 0 0]]
[[0 0 0 ... 0 0 0]]
[[0 0 0 ... 0 0 0]]
```

In [269…
```
from sklearn.metrics.pairwise import cosine_similarity

# Calculate Cosine Similarity
cos_sim_count_1_2 = cosine_similarity(document_vector_count_1, document_vector_count_
cos_sim_count_1_3 = cosine_similarity(document_vector_count_1, document_vector_count_
cos_sim_count_2_3 = cosine_similarity(document_vector_count_2, document_vector_count_

cos_sim_count_1_1 = cosine_similarity(document_vector_count_1, document_vector_count_
cos_sim_count_2_2 = cosine_similarity(document_vector_count_2, document_vector_count_
cos_sim_count_3_3 = cosine_similarity(document_vector_count_3, document_vector_count_

# Print
print("Cosine Similarity using count bw 1 and 2: %(x)f" %{"x":cos_sim_count_1_2})
print("Cosine Similarity using count bw 1 and 3: %(x)f" %{"x":cos_sim_count_1_3})
print("Cosine Similarity using count bw 2 and 3: %(x)f" %{"x":cos_sim_count_2_3})

print("Cosine Similarity using count bw 1 and 1: %(x)f" %{"x":cos_sim_count_1_1})
print("Cosine Similarity using count bw 2 and 2: %(x)f" %{"x":cos_sim_count_2_2})
print("Cosine Similarity using count bw 3 and 3: %(x)f" %{"x":cos_sim_count_3_3})
```

```
Cosine Similarity using count bw 1 and 2: 0.608862
Cosine Similarity using count bw 1 and 3: 0.622050
Cosine Similarity using count bw 2 and 3: 0.565566
Cosine Similarity using count bw 1 and 1: 1.000000
Cosine Similarity using count bw 2 and 2: 1.000000
Cosine Similarity using count bw 3 and 3: 1.000000
```

As expected, cosine similarity between a sentence and itself is 1. Between 2 entirely different sentences, it will be 0.

We can assume that we have the more common features in the documents 1 and 3 than in documents 1 and 2. This reflects indeed in a higher similarity than that of sentences 1 and 3.

---

7. Concluding Remarks

Wow! We have come a long way! We can now call ourselves experts of Data Preprocessing. You should feel excited and proud because the process of Data Mining usually involves 70% preprocessing and 30% training learning models. You will learn this as you progress in the Data Mining course. I really feel that if you go through the exercises and challenge yourself, you are on your way to becoming a super Data Scientist.

From here the possibilities for you are endless. You now know how to use almost every common technique for preprocessing with state-of-the-art tools, such as as Pandas and Scikit-learn. You are now with the trend!

After completing this notebook you can do a lot with the results we have generated. You can train algorithms and models that are able to classify articles into certain categories and much more. You can also try to experiment with different datasets, or venture further into text analytics by using new deep learning techniques such as word2vec. All of this will be presented in the next lab session. Until then, go teach machines how to be intelligent to make the world a better place.

. References

- Pandas cook book (Recommended for starters)
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison Wesley

---

*** Section 2 ***

Data Mining Lab 1

In this lab session we will focus on the use of scientific computing libraries to efficiently process, transform, and manage data. Furthermore, we will provide best practices and introduce visualization tools for effectively conducting big data analysis and visualization.

In [270…

```
# TEST necessary for when working with external scripts
%load_ext autoreload
%autoreload 2
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

1. The Data

The data is downloaded from
https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences# .

The dataset consists of three parts:

- amazon.com
- yelp.com
- imdb.com

---

2. Data Preparation

In the following we will use the built-in dataset loader for 20 newsgroups from scikit-learn. Alternatively, it is possible to download the dataset manually from the website and use the sklearn.datasets.load_files function by pointing it to the 20news-bydate-train sub-folder of the uncompressed archive folder.

In order to get faster execution times for this first example we will work on a partial dataset with only 4 categories out of the 20 available in the dataset:

In [271…
```
# import packages
import os
import pandas as pd
import helpers.data_mining_helpers as dmh
```

In [272…
```
# get file
root = './sentiment labelled sentences/'

files = os.listdir(root)

if '.DS_Store' and 'readme.txt' in files:
    files.remove('.DS_Store')
    files.remove('readme.txt')

# get company name
companies = []
for file in files:
    companies.append(file.split("_")[0])

# read file
df_amazon = pd.read_csv(root+files[0], names=['text', 'score'], header=None,delimiter
df_yelp = pd.read_csv(root+files[1], names=['text', 'score'], header=None,delimiter='
df_imdb = pd.read_csv(root+files[2], names=['text', 'score'], header=None,delimiter='
print(df_amazon)
print(df_yelp)
```

```
                                                text score
0    So there is no way for me to plug it in here i...     0
1                          Good case, Excellent value.     1
2                               Great for the jawbone.     1
3    Tied to charger for conversations lasting more...     0
4                                    The mic is great.     1
..                                                 ...   ...
995  The screen does get smudged easily because it ...     0
996  What a piece of junk.. I lose more calls on th...     0
997                       Item Does Not Match Picture.     0
998  The only thing that disappoint me is the infra...     0
999  You can not answer calls with the unit, never ...     0

[1000 rows x 2 columns]
                                                text score
0                             Wow... Loved this place.     1
1                                   Crust is not good.     0
2            Not tasty and the texture was just nasty.     0
3    Stopped by during the late May bank holiday of...     1
4    The selection on the menu was great and so wer...     1
..                                                 ...   ...
995  I think food should have flavor and texture an...     0
996                           Appetite instantly gone.     0
997  Overall I was not impressed and would not go b...     0
998  The whole experience was underwhelming, and I ...     0
999  Then, as if I hadn't wasted enough of my life ...     0

[1000 rows x 2 columns]
```

In [273…
```
for t in df_amazon.text[:10]:
    print(t)
```

```
So there is no way for me to plug it in here in the US unless I go by a converter.
Good case, Excellent value.
Great for the jawbone.
Tied to charger for conversations lasting more than 45 minutes.MAJOR PROBLEMS!!
The mic is great.
I have to jiggle the plug to get it to line up right to get decent volume.
If you have several dozen or several hundred contacts, then imagine the fun of sendi
ng each of them one by one.
If you are Razr owner...you must have this!
Needless to say, I wasted my money.
What a waste of money and time!.
```

>>> Exercise 1 (5 min):

In this exercise, please print out the *text* data for the first three samples in the dataset. (See the above code for help)

In [274…
```
# Answer here
for t in df_amazon.text[:3]:
    print(t)
```

```
So there is no way for me to plug it in here in the US unless I go by a converter.
Good case, Excellent value.
Great for the jawbone.
```

---

3. Data Transformation

We want to explore and understand our data a little bit better. Before we do that we definitely need to apply some transformations just so we can have our dataset in a nice format to be able to explore it freely and more efficient.

3.1 Converting Dictionary into Pandas Dataframe

Here we will show you how to convert dictionary objects into a pandas dataframe. And by the way, a pandas dataframe is nothing more than a table magically stored for efficient information retrieval.

In [275…
```
# add company name in the dataframe
df_amazon['company'] = companies[0]
df_yelp['company'] = companies[1]
df_imdb['company'] = companies[2]
```

In [276…
```
# merge 3 lables into the dataframe
df = pd.concat([df_amazon,df_yelp,df_imdb])
df
```

Out[276]:

	text	score	company
0	So there is no way for me to plug it in here i...	0	amazon
1	Good case, Excellent value.	1	amazon
2	Great for the jawbone.	1	amazon
3	Tied to charger for conversations lasting more...	0	amazon
4	The mic is great.	1	amazon
...	...	...	...
743	I just got bored watching Jessice Lange take h...	0	imdb
744	Unfortunately, any virtue in this film's produ...	0	imdb
745	In a word, it is embarrassing.	0	imdb
746	Exceptionally bad!	0	imdb
747	All in all its an insult to one's intelligence...	0	imdb

2748 rows × 3 columns

Now we can print and see what our table looks like.

In [277…

```
# Make company Name into dummy code
df["companyLabel"] = pd.factorize(df["company"])[0]

df[0:10]
```

Out[277]:

	text	score	company	companyLabel
0	So there is no way for me to plug it in here i...	0	amazon	0
1	Good case, Excellent value.	1	amazon	0
2	Great for the jawbone.	1	amazon	0
3	Tied to charger for conversations lasting more...	0	amazon	0
4	The mic is great.	1	amazon	0
5	I have to jiggle the plug to get it to line up...	0	amazon	0
6	If you have several dozen or several hundred c...	0	amazon	0
7	If you are Razr owner...you must have this!	1	amazon	0
8	Needless to say, I wasted my money.	0	amazon	0
9	What a waste of money and time!.	0	amazon	0

Nice! Isn't it? With this format we can conduct many operations easily and efficiently since Pandas dataframes provide us with a wide range of built-in features/functionalities. These features are operations which can directly and quickly be applied to the dataset. These operations may include standard operations like **removing records with missing values** and **aggregating new fields** to the current table (hereinafter referred to as a dataframe), which is desirable in almost every data mining project. Go Pandas!

3.2 Familiarizing yourself with the Data

To begin to show you the awesomeness of Pandas dataframes, let us look at how to run a simple query on our dataset. We want to query for the first 10 rows (documents), and we only want to keep the `text` and `category_name` attributes or fields.

In [278…
```
# a simple query
df[:10][["text","company"]]
```

Out[278]:

	text	company
0	So there is no way for me to plug it in here i...	amazon
1	Good case, Excellent value.	amazon
2	Great for the jawbone.	amazon
3	Tied to charger for conversations lasting more...	amazon
4	The mic is great.	amazon
5	I have to jiggle the plug to get it to line up...	amazon
6	If you have several dozen or several hundred c...	amazon
7	If you are Razr owner...you must have this!	amazon
8	Needless to say, I wasted my money.	amazon
9	What a waste of money and time!.	amazon

Let us look at a few more interesting queries to familiarize ourselves with the efficiency and conveniency of Pandas dataframes.

Let's query the last 10 records

In [279…
```
df[-10:]
```

Out[279]:

	text	score	company	companyLabel
738	The opening sequence of this gem is a classic,...	1	imdb	2
739	Fans of the genre will be in heaven.	1	imdb	2
740	Lange had become a great actress.	1	imdb	2
741	It looked like a wonderful story.	1	imdb	2
742	I never walked out of a movie faster.	0	imdb	2
743	I just got bored watching Jessice Lange take h...	0	imdb	2
744	Unfortunately, any virtue in this film's produ...	0	imdb	2
745	In a word, it is embarrassing.	0	imdb	2
746	Exceptionally bad!	0	imdb	2
747	All in all its an insult to one's intelligence...	0	imdb	2

>>> Exercise 2 (take home):

Experiment with other querying techniques using pandas dataframes. Refer to their documentation for more information.

In [280…
```
#Answer here
# string to int
df['score'] = pd.to_numeric(df['score'])

df.query('score > 0')
```

Out[280]:

	text	score	company	companyLabel
1	Good case, Excellent value.	1	amazon	0
2	Great for the jawbone.	1	amazon	0
4	The mic is great.	1	amazon	0
7	If you are Razr owner...you must have this!	1	amazon	0
10	And the sound quality is great.	1	amazon	0
...	...	...	...	...
737	:) Anyway, the plot flowed smoothly and the ma...	1	imdb	2
738	The opening sequence of this gem is a classic,...	1	imdb	2
739	Fans of the genre will be in heaven.	1	imdb	2
740	Lange had become a great actress.	1	imdb	2
741	It looked like a wonderful story.	1	imdb	2

1386 rows × 4 columns

>>> Exercise 3 (5 min):

Try to fetch records belonging to the `comp.graphics` category, and query every 10th record. Only show the first 5 records.

In [281…

```
# Answer here
df.loc[lambda f: f.company == 'yelp'].iloc[::10, :][0:5]
```

Out[281]:

	text	score	company	companyLabel
0	Wow... Loved this place.	1	yelp	1
10	Service was very prompt.	1	yelp	1
20	The Burrittos Blah!	0	yelp	1
30	Also there are combos like a burger, fries, an...	1	yelp	1
40	The shrimp tender and moist.	1	yelp	1

---

4. Data Mining using Pandas

Let's do some serious work now. Let's learn to program some of the ideas and concepts learned so far in the data mining course. This is the only way we can be convince ourselves of the true power of Pandas dataframes.

Program some of the ideas and concepts with Pandas dataframes.

>>> Exercise 4 (5 min):

Let's try something different. Instead of calculating missing values by column let's try to calculate the missing values in every record instead of every column.
$Hint:$ `axis` parameter. Check the documentation for more information.

```
# Answer here
```

In [282…
```
df.isnull().apply(lambda x: dmh.check_missing_values(x), axis=1)
```

Out[282]:
```
0      (The amoung of missing records is: , 0)
1      (The amoung of missing records is: , 0)
2      (The amoung of missing records is: , 0)
3      (The amoung of missing records is: , 0)
4      (The amoung of missing records is: , 0)
                       ...
743    (The amoung of missing records is: , 0)
744    (The amoung of missing records is: , 0)
745    (The amoung of missing records is: , 0)
746    (The amoung of missing records is: , 0)
747    (The amoung of missing records is: , 0)
Length: 2748, dtype: object
```

>>> Exercise 5 (take home)

There is an old saying that goes, "The devil is in the details." When we are working with extremely large data, it's difficult to check records one by one (as we have been doing so far). And also, we don't even know what kind of missing values we are facing. Thus, "debugging" skills get sharper as we spend more time solving bugs. Let's focus on a different method to check for missing values and the kinds of missing values you may encounter. It's not easy to check for missing values as you will find out in a minute.

Please check the data and the process below, describe what you observe and why it happened.
$Hint$: why `.isnull()` didn't work?

In [283…
```
import numpy as np

NA_dict = [{ 'id': 'A', 'missing_example': np.nan },
           { 'id': 'B'                    },
           { 'id': 'C', 'missing_example': 'NaN'  },
           { 'id': 'D', 'missing_example': 'None' },
           { 'id': 'E', 'missing_example':  None  },
           { 'id': 'F', 'missing_example': ''     }]

NA_df = pd.DataFrame(NA_dict, columns = ['id','missing_example'])
NA_df
```

Out[283]:

	id	missing_example
0	A	NaN
1	B	NaN
2	C	NaN
3	D	None
4	E	None
5	F	

In [284…
```
NA_df['missing_example'].isnull()
```

Out[284]:
```
0     True
1     True
2    False
3    False
4     True
5    False
Name: missing_example, dtype: bool
```

In [285…
```
# Answer here
"""
For function of isnull(), there are only np.nan, None, or completely blank considered
In the other words, the column where id is "C" and id is "D", their data of 'NaN' and
```

```
Also, the column where id is "F,'' its data of '' will be regarded as a value with a
Therefore, they won't be seen as missing data and isnull() function won't work.
"""
```

Out[285]:
```
'\nFor function of isnull(), there are only np.nan, None, or completely blank consid
ered missing data.\nIn the other words, the column where id is "C" and id is "D", th
eir data of \'NaN\' and \'None\' will be regarded as string by the computer. \nAlso,
the column where id is "F,\'\' its data of \'\' will be regarded as a value with a b
lank key. \nTherefore, they won\'t be seen as missing data and isnull() function won
\'t work.\n'
```

check duplicate on a specific column "like"

In [286…
```
# check the length of the data
len(df)
```

Out[286]:
```
2748
```

We can also check the sum of duplicate records by simply doing:

In [287…
```
# find duplcate data
sum(df.duplicated('text'))
```

Out[287]:
```
17
```

In [288…
```
# Remove the dublicates from the dataframe
# Inplace applies changes directly on our dataframe
df.drop_duplicates(keep=False, inplace=True)
```

In [289…
```
# Reset the index
df = df.reset_index(drop=True)
df
```

Out[289]:

	text	score	company	companyLabel
0	So there is no way for me to plug it in here i...	0	amazon	0
1	Good case, Excellent value.	1	amazon	0
2	Great for the jawbone.	1	amazon	0
3	Tied to charger for conversations lasting more...	0	amazon	0
4	The mic is great.	1	amazon	0
...	...	...	...	...
2709	I just got bored watching Jessice Lange take h...	0	imdb	2
2710	Unfortunately, any virtue in this film's produ...	0	imdb	2
2711	In a word, it is embarrassing.	0	imdb	2
2712	Exceptionally bad!	0	imdb	2
2713	All in all its an insult to one's intelligence...	0	imdb	2

2714 rows × 4 columns

In [290…
```
# The length of data after removing the dublicates from the dataframe
# 2748 17*2 = 2714
len(df)
```

Out[290]:
```
2714
```

In [291…
```
# Reset the index
df = df.reset_index(drop=True)
df
```

Out[291]:

	text	score	company	companyLabel
0	So there is no way for me to plug it in here i...	0	amazon	0
1	Good case, Excellent value.	1	amazon	0
2	Great for the jawbone.	1	amazon	0
3	Tied to charger for conversations lasting more...	0	amazon	0
4	The mic is great.	1	amazon	0
...	...	...	...	...
2709	I just got bored watching Jessice Lange take h...	0	imdb	2
2710	Unfortunately, any virtue in this film's produ...	0	imdb	2
2711	In a word, it is embarrassing.	0	imdb	2
2712	Exceptionally bad!	0	imdb	2
2713	All in all its an insult to one's intelligence...	0	imdb	2

2714 rows × 4 columns

---

5. Data Preprocessing

In the Data Mining course we learned about the many ways of performing data preprocessing. In reality, the list is quiet general as the specifics of what data preprocessing involves is too much to cover in one course. This is especially true when you are dealing with unstructured data, as we are dealing with in this particular notebook. But let us look at some examples for each data preprocessing technique that we learned in the class. We will cover each item one by one, and provide example code for each category. You will learn how to perform each of the operations, using Pandas, that cover the essentials to Preprocessing in Data Mining. We are not going to follow any strict order, but the items we will cover in the preprocessing section of this notebook are as follows:

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature Subset Selection
- Feature Creation
- Discretization and Binarization
- Attribute Transformation

5.1 Sampling

In [292…

```
# Answer here
# Use for answer Exercise 6
# Copy initial df

df_initial = df
```

In [293…

```
# Random Select 1000 data
df_sample = df.sample(n=1000)
len(df_sample)
```

Out[293]:

```
1000
```

In [294…
```
df_sample[0:5]
```

Out[294]:

	text	score	company	companyLabel
72	Nice docking station for home or work.	1	amazon	0
838	Poor Reliability.	0	amazon	0
198	It makes very strange ticking noises before it...	0	amazon	0
314	I ordered this product first and was unhappy w...	0	amazon	0
634	When I placed my treo into the case, not only ...	0	amazon	0

>>> Exercise 6 (take home):

Notice any changes to the `X` dataframe? What are they? Report every change you noticed as compared to the previous state of `X`. Feel free to query and look more closely at the dataframe for these changes.

In [295…
```
# Answer here

"""
From cell above, we can notice there are not any differences between initial X and X
Then, X_sample is randomly chosen from X, so index is not sorting.
And X_sample will not change X value.

This result can be proved by the following code.
"""
```

Out[295]:
```
'\nFrom cell above, we can notice there are not any differences between initial X an
d X after sample.\nThen, X_sample is randomly chosen from X, so index is not sortin
g.\nAnd X_sample will not change X value.\n\nThis result can be proved by the follow
ing code.\n'
```

In [296…
```
df
```

Out[296]:

	text	score	company	companyLabel
0	So there is no way for me to plug it in here i...	0	amazon	0
1	Good case, Excellent value.	1	amazon	0
2	Great for the jawbone.	1	amazon	0
3	Tied to charger for conversations lasting more...	0	amazon	0
4	The mic is great.	1	amazon	0
...	...	...	...	...
2709	I just got bored watching Jessice Lange take h...	0	imdb	2
2710	Unfortunately, any virtue in this film's produ...	0	imdb	2
2711	In a word, it is embarrassing.	0	imdb	2
2712	Exceptionally bad!	0	imdb	2
2713	All in all its an insult to one's intelligence...	0	imdb	2

2714 rows × 4 columns

In [297…
```
# Answer here
# Compare X_initial and X after .sample
diff = False
```

```
for i in range(len(df)):
    if(df_initial["text"][i] != df["text"][i]):
        print("text", i)
        diff = True
    if(df_initial["score"][i] != df["score"][i]):
        print("label", i)
        diff = True
    if(df_initial["company"][i] != df["company"][i]):
        print("company", i)
        diff = True
    if(df_initial["companyLabel"][i] != df["companyLabel"][i]):
        print("companyLabel", i)
        diff = True

if (diff == True):
    print("There are some different")
else:
    print("There isn't any different")
```

```
There isn't any different
```

In [298…
```
import matplotlib.pyplot as plt
%matplotlib inline
```

In [299…
```
categories
```

Out[299]:
```
['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']
```

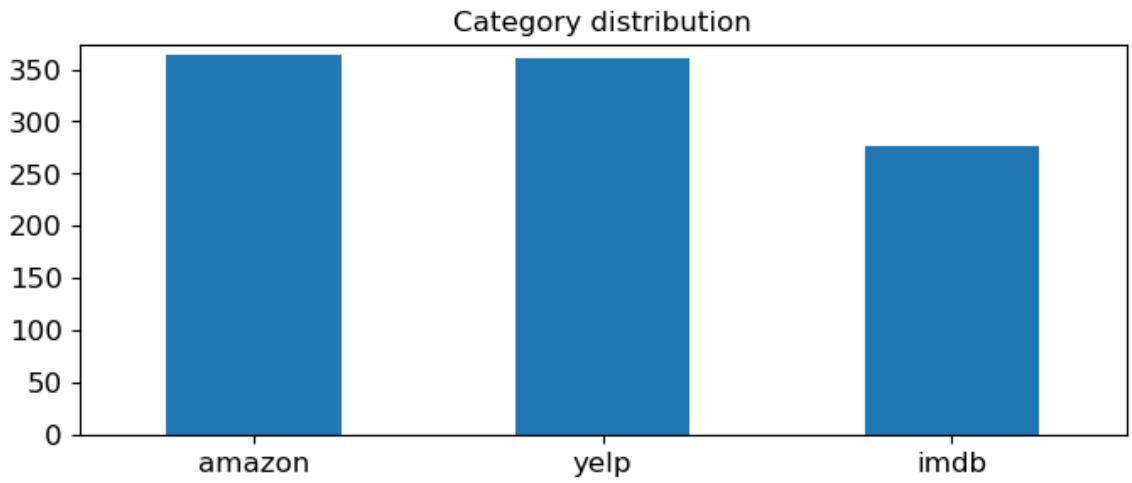
>>> Exercise 7 (5 min):

Notice that for the `ylim` parameters we hardcoded the maximum value for y. Is it possible to automate this instead of hard-coding it? How would you go about doing that? (Hint: look at code above for clues)

In [302…
```
# Answer here
upper_bound = max(df_sample.company.value_counts()) + 10

print(df_sample.company.value_counts())
# plot barchart for X_sample
df_sample.company.value_counts().plot(kind = 'bar',
                                      title = 'Category distribution',
                                      ylim = [0, upper_bound],
                                      rot = 0, fontsize = 12, figsize = (8,3))
```

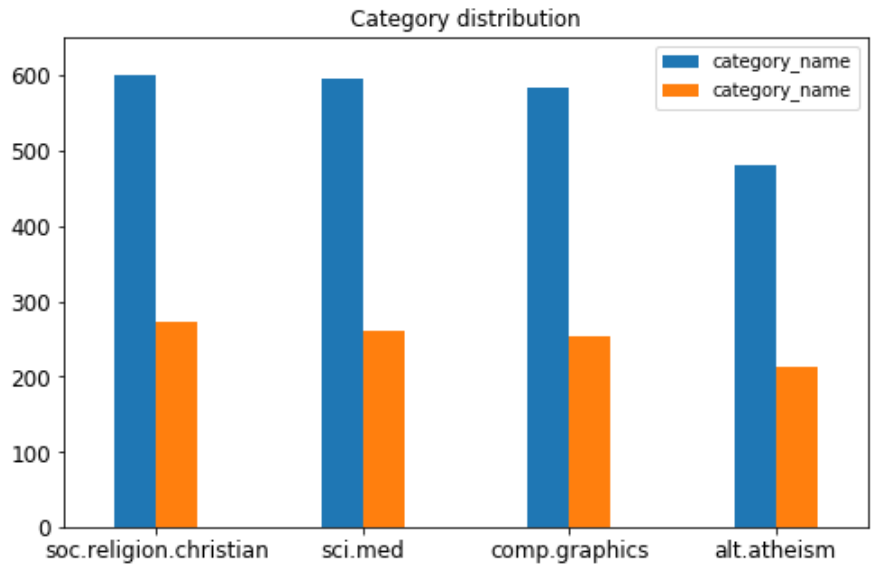
```
amazon    363
yelp      361
imdb      276
Name: company, dtype: int64
```

Out[302]:
```
<AxesSubplot:title={'center':'Category distribution'}>
```



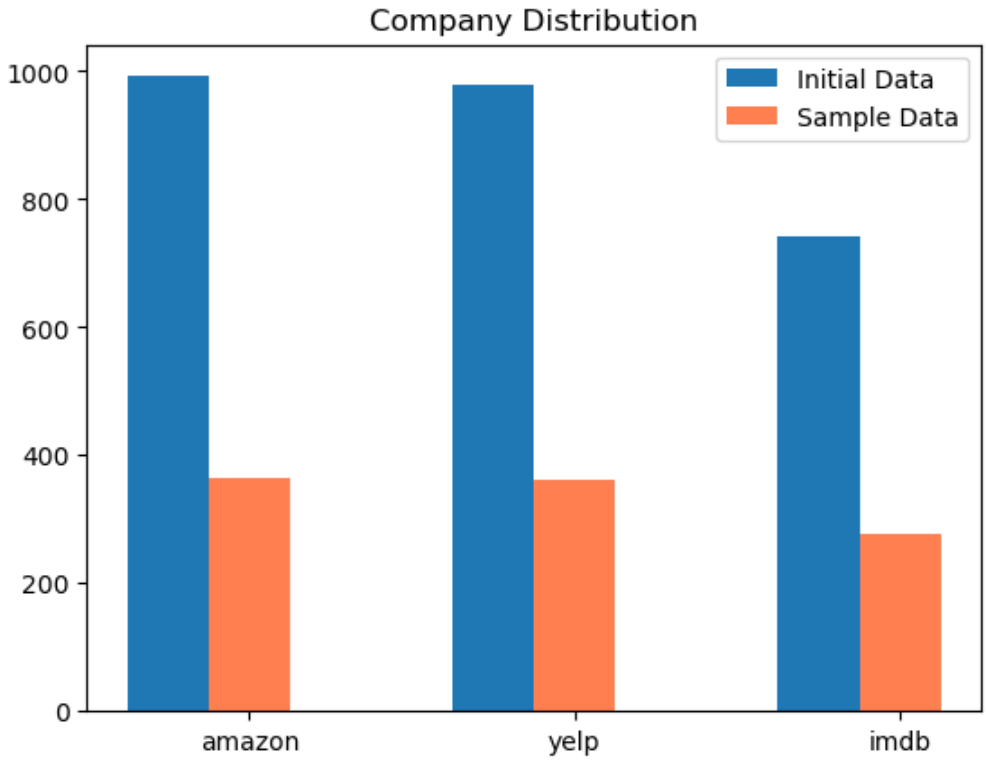
>>> Exercise 8 (take home):

We can also do a side-by-side comparison of the distribution between the two datasets, but maybe you can try that as an excerise. Below we show you an snapshot of the type of chart we are looking for.



In [304…

```
# Answer here
bar1 = df.company.value_counts()
bar2 = df_sample.company.value_counts()

index = np.arange(3)
plt.bar(index,bar1.tolist(), label='Initial Data', width=0.25)
plt.bar(index+0.25, bar2.tolist(), label='Sample Data', width=0.25, color= 'coral')
plt.xticks(index+0.25, companies)
plt.title("Company Distribution")
plt.legend()
plt.show()
```



In [305…
```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/antinghsieh/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[305]:
```
True
```

>>> Exercise 9 (5 min):

Let's analyze the first record of our X dataframe with the new analyzer we have just built. Go ahead try it!

In [308…
```
# Answer here
df_count_vect = CountVectorizer()
df_counts = df_count_vect.fit_transform(df.text)
analyze = df_count_vect.build_analyzer()

analyze(" ".join(list(df[:2].text)))
```

Out[308]:
```
['so',
 'there',
 'is',
 'no',
 'way',
 'for',
 'me',
 'to',
 'plug',
 'it',
 'in',
 'here',
 'in',
 'the',
 'us',
 'unless',
 'go',
 'by',
 'converter',
 'good',
 'case',
 'excellent',
 'value']
```

Now let us look at the term-document matrix we built above.

In [309…
```
df_counts
```

Out[309]:
```
<2714x5153 sparse matrix of type '<class 'numpy.int64'>'
        with 30149 stored elements in Compressed Sparse Row format>
```

In [310…
```
# We can check the shape of this matrix by:
df_counts.shape
```

Out[310]:
```
(2714, 5153)
```

In [311…
```
# We can obtain the feature names of the vectorizer, i.e., the terms
# usually on the horizontal axis
count_vect.get_feature_names_out()[0:10]
```

Out[311]:
```
array(['00', '000', '0000', '0000001200', '000005102000', '0001',
       '000100255pixel', '00014', '000406', '0007'], dtype=object)
```

In [312…
```
df_counts.shape
```

Out[312]:
```
(2714, 5153)
```

```
In [313…  # we convert from sparse array to normal array
          df_counts[0:5, 0:100].toarray()
```

```
Out[313]: array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
                [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
                [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
                [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 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, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
                [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])
```

```
In [315…  count_vect.get_feature_names_out()[0:1]
```

```
Out[315]: array(['00'], dtype=object)
```

>>> Exercise 10 (take home):

We said that the `1` at the beginning of the fifth record represents the `00` term. Notice that there is another 1 in the same record. Can you provide code that can verify what word this 1 represents from the vocabulary. Try to do this as efficient as possible.

```
In [319…  # Answer here

          # Check the second one
          order = 0
          # choose the first record
          firRec = df_counts[0:5,0:100].toarray()

          for i in range(5):
              for j in range(100):
                  if(firRec[i][j] == 1):
                      print("The word of first 1 is", df_count_vect.get_feature_names()[i])
                      break
```

```
The word of first 1 is 11
```

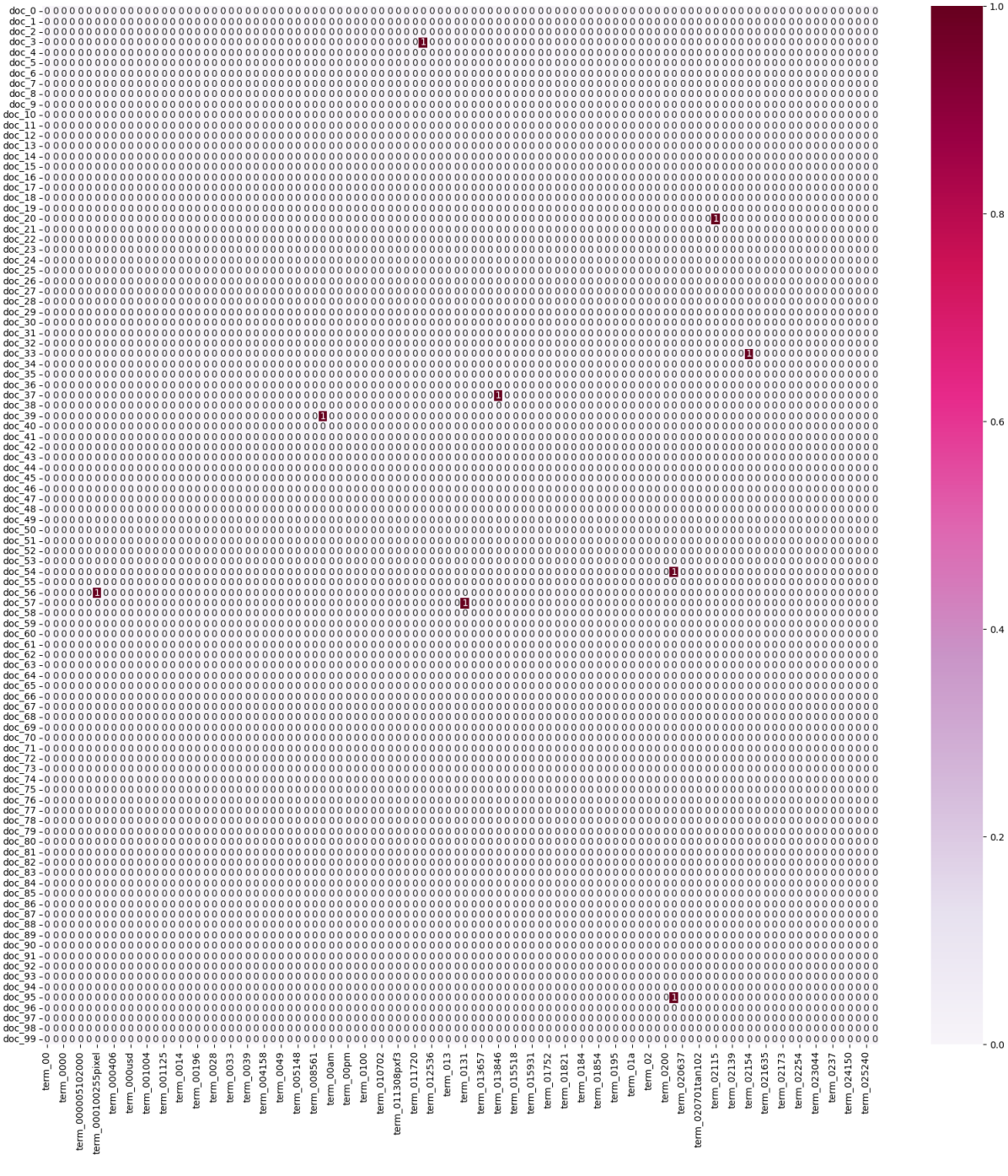
>>> Exercise 11 (take home):

From the chart above, we can see how sparse the term-document matrix is; i.e., there is only one terms with frequency of `1` in the subselection of the matrix. By the way, you may have noticed that we only selected 20 articles and 20 terms to plot the histrogram. As an excersise you can try to modify the code above to plot the entire term-document matrix or just a sample of it. How would you do this efficiently? Remember there is a lot of words in the vocab. Report below what methods you would use to get a nice and useful visualization

In [320…

```
# Answer here
print("GO!")
# x,y,z
plot_x_all = ["term_" + str(i) for i in count_vect.get_feature_names()[0:100]]
plot_y_all = ["doc_" + str(i) for i in list(X.index)[0:100]]
plot_z_all = df_counts[0:100, 0:100].toarray()

# plot
df_all_todraw = pd.DataFrame(plot_z_all, columns = plot_x_all, index = plot_y_all)
plt.subplots(figsize=(20, 20))
ax = sns.heatmap(df_all_todraw,
                 cmap="PuRd",
                 vmin=0, vmax=1, annot=True)

# plt.show()
print('END')
```

```
GO!
END
```



In [321…

```
from sklearn.decomposition import PCA
```

In [327…
```
df_3dim = PCA(n_components = 3).fit_transform(df_counts.toarray())
```

In [328…
```
df_3dim.shape
```

Out[328]:
```
(2714, 3)
```

In [329…
```
categories
```

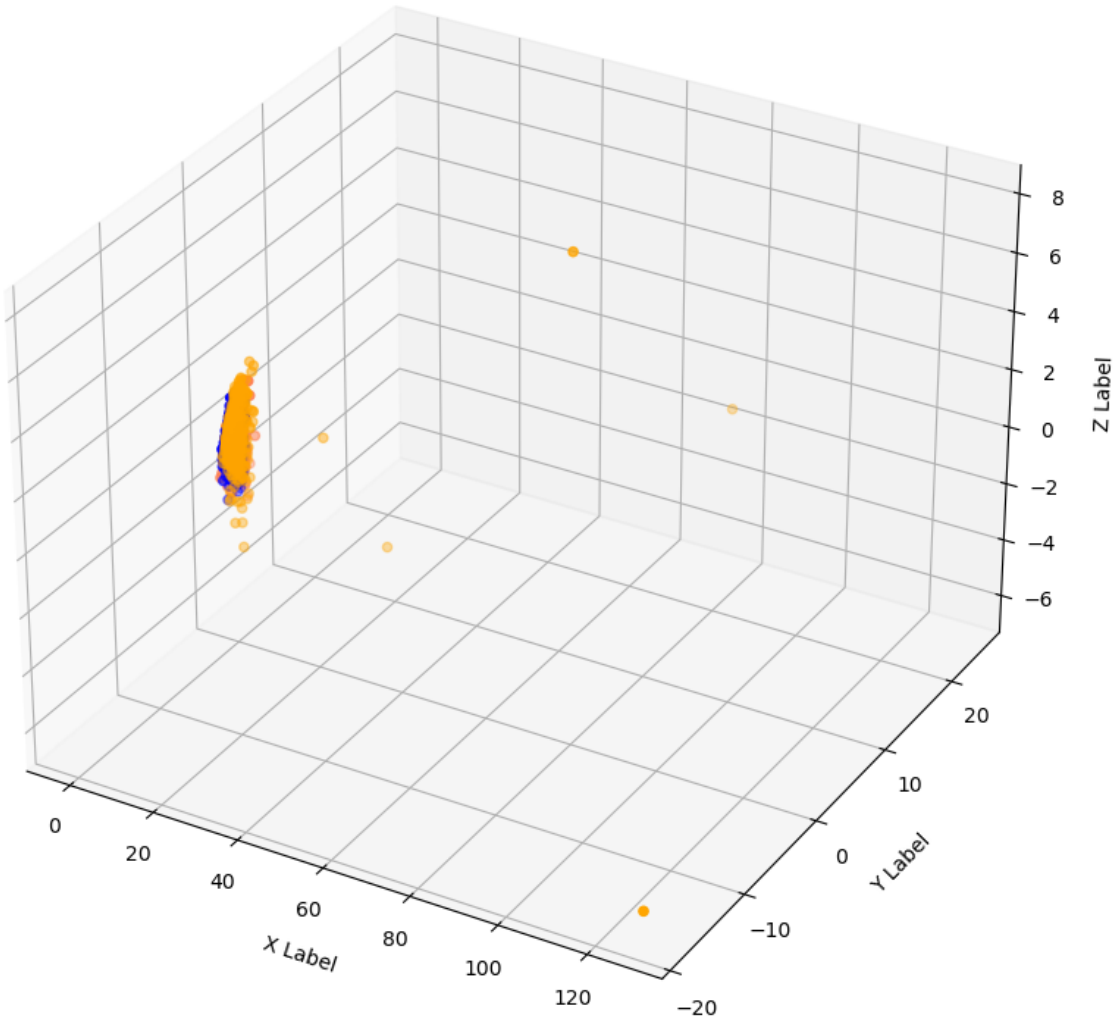
Out[329]:
```
['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med']
```

>>> Exercise 12 (take home):

Please try to reduce the dimension to 3, and plot the result use 3-D plot. Use at least 3 different angle (camera position) to check your result and describe what you found.

Hint: you can refer to Axes3D in the documentation.

In [334…
```
col = ['coral', 'blue', 'orange']

# plot
fig = plt.figure(figsize = (25,10))
ax = fig.add_subplot(111, projection='3d')

for c, company in zip(col, companies):
    xs = df_3dim[df['company'] == company].T[0]
    ys = df_3dim[df['company'] == company].T[1]
    zs = df_3dim[df['company'] == company].T[2]

    ax.scatter(xs, ys, zs, c = c, marker='o')

ax.grid(color='gray', linestyle=':', linewidth=2, alpha=0.2)
ax.set_xlabel('\nX Label')
ax.set_ylabel('\nY Label')
ax.set_zlabel('\nZ Label')

plt.show()
```



5.5 Attribute Transformation / Aggregation

Calculate Frquency Term

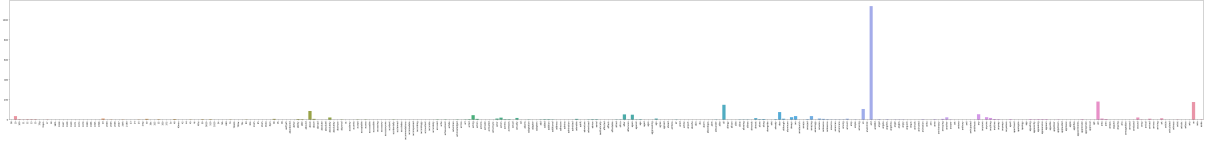
In [335…
```
# note this takes time to compute. You may want to reduce the amount of terms you wan
term_frequencies = []
for j in range(0,df_counts.shape[1]):
    term_frequencies.append(sum(df_counts[:,j].toarray()))
```

In [336…
```
term_frequencies = np.asarray(df_counts.sum(axis=0))[0]
```

In [337…
```
term_frequencies[0] #sum of first term
```

Out[337]: 1

In [338…
```
plt.subplots(figsize=(100, 10))
g = sns.barplot(x=df_count_vect.get_feature_names_out()[:300],
            y=term_frequencies[:300])
g.set_xticklabels(df_count_vect.get_feature_names_out()[:300], rotation = 90);
```



>>> Exercise 13 (take home):

If you want a nicer interactive visualization here, I would encourage you try to install and use plotly to achieve this.

In [208…
```
# Answer here

!pip install plotly
!pip install plotly --upgrade
!pip install chart_studio
```

```
Requirement already satisfied: plotly in /Users/antinghsieh/opt/anaconda3/envs/dataM
ining/lib/python3.9/site-packages (5.9.0)
Requirement already satisfied: tenacity>=6.2.0 in /Users/antinghsieh/opt/anaconda3/e
nvs/dataMining/lib/python3.9/site-packages (from plotly) (8.0.1)
Requirement already satisfied: plotly in /Users/antinghsieh/opt/anaconda3/envs/dataM
ining/lib/python3.9/site-packages (5.9.0)
Collecting plotly
  Downloading plotly-5.10.0-py2.py3-none-any.whl (15.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 15.2/15.2 MB
6.9 MB/s eta 0:00:00m eta 0:00:010:01:01
Requirement already satisfied: tenacity>=6.2.0 in /Users/antinghsieh/opt/anaconda3/e
nvs/dataMining/lib/python3.9/site-packages (from plotly) (8.0.1)
Installing collected packages: plotly
  Attempting uninstall: plotly
    Found existing installation: plotly 5.9.0
    Uninstalling plotly-5.9.0:
      Successfully uninstalled plotly-5.9.0
Successfully installed plotly-5.10.0
Collecting chart_studio
  Downloading chart_studio-1.1.0-py3-none-any.whl (64 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 64.4/64.4 kB
536.6 kB/s eta 0:00:00 kB/s eta 0:00:01
Requirement already satisfied: six in /Users/antinghsieh/opt/anaconda3/envs/dataMini
ng/lib/python3.9/site-packages (from chart_studio) (1.16.0)
Requirement already satisfied: plotly in /Users/antinghsieh/opt/anaconda3/envs/dataM
ining/lib/python3.9/site-packages (from chart_studio) (5.10.0)
Collecting retrying>=1.3.3
  Downloading retrying-1.3.3.tar.gz (10 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: requests in /Users/antinghsieh/opt/anaconda3/envs/dat
aMining/lib/python3.9/site-packages (from chart_studio) (2.28.1)
Requirement already satisfied: tenacity>=6.2.0 in /Users/antinghsieh/opt/anaconda3/e
nvs/dataMining/lib/python3.9/site-packages (from plotly->chart_studio) (8.0.1)
Requirement already satisfied: charset-normalizer<3,>=2 in /Users/antinghsieh/opt/an
aconda3/envs/dataMining/lib/python3.9/site-packages (from requests->chart_studio)
 (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /Users/antinghsieh/opt/anaconda3/env
s/dataMining/lib/python3.9/site-packages (from requests->chart_studio) (3.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /Users/antinghsieh/opt/anaco
nda3/envs/dataMining/lib/python3.9/site-packages (from requests->chart_studio) (1.2
6.11)
Requirement already satisfied: certifi>=2017.4.17 in /Users/antinghsieh/opt/anaconda
3/envs/dataMining/lib/python3.9/site-packages (from requests->chart_studio) (2022.9.
14)
Building wheels for collected packages: retrying
  Building wheel for retrying (setup.py) ... done
  Created wheel for retrying: filename=retrying-1.3.3-py3-none-any.whl size=11431 sh
a256=896a067c776ed8e260e9113078e52e848478ee6e135370e08da44c062824074a
  Stored in directory: /Users/antinghsieh/Library/Caches/pip/wheels/ce/18/7f/e9527e3
e66db1456194ac7f61eb3211068c409edceecff2d31
Successfully built retrying
Installing collected packages: retrying, chart_studio
Successfully installed chart_studio-1.1.0 retrying-1.3.3
```

In [339…
```
# Answer here
```

```
import plotly
import plotly as py
import plotly.graph_objs as go
import chart_studio
import plotly.offline as pyo
from plotly.offline import iplot
import plotly.figure_factory as ff
```

>>> Exercise 14 (take home):

The chart above contains all the vocabulary, and it's computationally intensive to both compute and visualize. Can you efficiently reduce the number of terms you want to visualize as an exercise.

In [342…

```
# Construct the df with term and frequencies
df_term_frequencies = pd.DataFrame(columns = ["term", "frequencies"])

for i in range(300):
    df_term_frequencies.loc[i, "term"] = str(df_count_vect.get_feature_names()[i])
    df_term_frequencies.loc[i, "frequencies"] = int(term_frequencies[i])
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
```

```
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:
```

```
Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
```

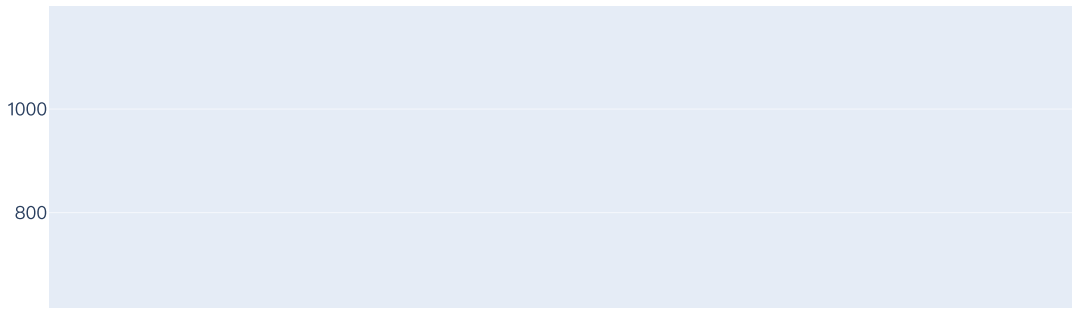
```
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

```
/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.

/Users/antinghsieh/opt/anaconda3/envs/dataMining/lib/python3.9/site-packages/sklear
n/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and
will be removed in 1.2. Please use get_feature_names_out instead.
```

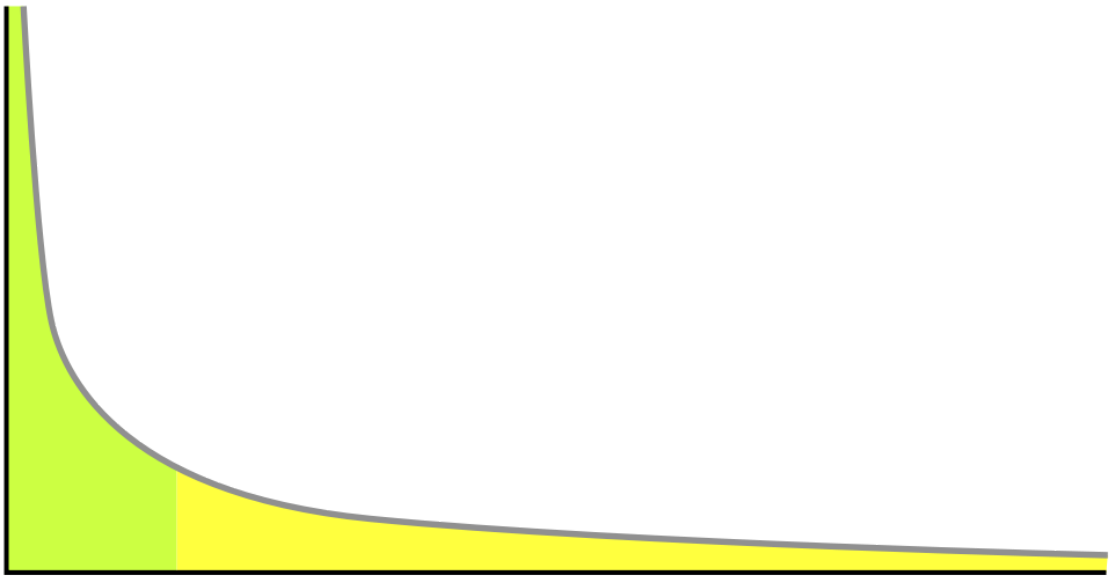
In [343…

```
# Answer here

# Screening by the frequency of the term > 15
# Time Complexity of Screening by this function only O(1)
df_reduce = df_term_frequencies[df_term_frequencies["frequencies"]>15]

# plotly interactive visualization
reduce_data = [go.Bar(x=df_reduce.term, y=df_reduce.frequencies)]
reduce_layout = go.Layout(font={'size':10,'family':'sansserif'})
reduce_fig = go.Figure(data=reduce_data, layout=reduce_layout)
plotly.offline.init_notebook_mode()
plotly.offline.iplot(reduce_fig,filename='basicscatter')
```



>>> Exercise 15 (take home):

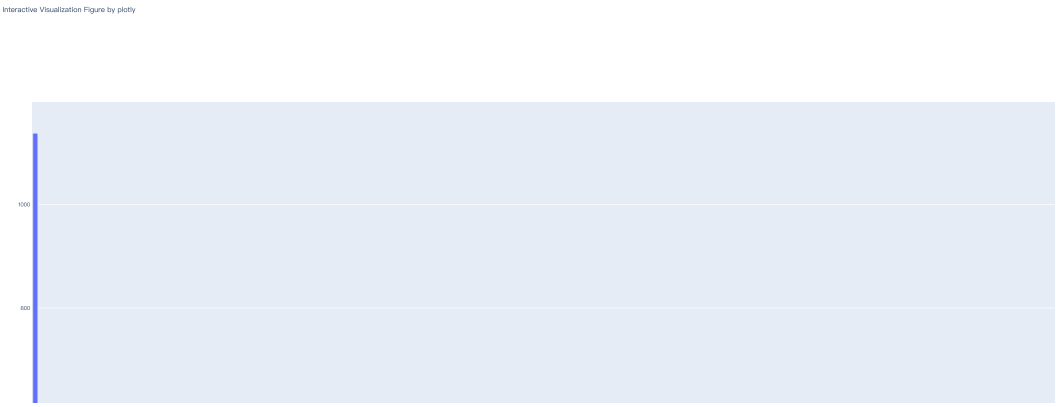
Additionally, you can attempt to sort the terms on the `x-axis` by frequency instead of in alphabetical order. This way the visualization is more meaninfgul and you will be able to observe the so called long tail (get familiar with this term since it will appear a lot in data mining and other statistics courses). see picture below



In [344…

```
# Answer here

# Sort the terms on the x-axis by frequency
df_sort = df_term_frequencies.sort_values('frequencies', ascending = False)
```

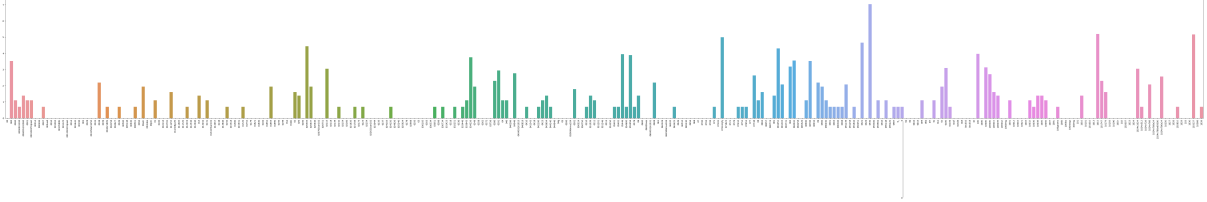
```
# plotly interactive visualization
term = df_sort.term
plotly_data = [go.Bar(x=df_sort.term, y=df_sort.frequencies)]
layout = go.Layout(title='Interactive Visualization Figure by plotly', font={'size':3
fig = go.Figure(data=plotly_data, layout=layout)
plotly.offline.init_notebook_mode()
plotly.offline.iplot(fig,filename='basic-scatter')
```



Since we already have those term frequencies, we can also transform the values in that vector into the log distribution. All we need is to import the `math` library provided by python and apply it to the array of values of the term frequency vector. This is a typical example of attribute transformation. Let's go for it. The log distribution is a technique to visualize the term frequency into a scale that makes you easily visualize the distribution in a more readable format. In other words, the variations between the term frequencies are now easy to observe. Let us try it out!

In [345…
```
import math
term_frequencies_log = [math.log(i) for i in term_frequencies]
```

In [346…
```
plt.subplots(figsize=(100, 10))
g = sns.barplot(x=df_count_vect.get_feature_names_out()[:300],
                y=term_frequencies_log[:300])
g.set_xticklabels(count_vect.get_feature_names_out()[:300], rotation = 90);
```



5.6 Discretization and Binarization

- Compare the binary of company Name and company Lable by using mlb package.

>>> Exercise 16 (take home):

Try to generate the binarization using the `category_name` column instead. Does it work?

In [365…
```
# Answer here

"""
Yes, it works!
However, category is the category_name who converted to the dummy code,
and bin_category and bin_categoryNameare are the same meaning and the same value.

The results can be justified by the following code.
"""
```

Out[365]:
```
'\nYes, it works!\nHowever, category is the category_name who converted to the dummy code, \nand bin_category and bin_categoryNameare are the same meaning and the same value.\n\nThe results can be justified by the following code.\n'
```

In [353…
```
from sklearn import preprocessing, metrics, decomposition, pipeline, dummy
```

In [354…
```
# Make df['companyLabel'] into binary code
mlb = preprocessing.LabelBinarizer()
```

In [355…
```
mlb.fit(df.companyLabel)
```

Out[355]:
```
▾ LabelBinarizer
LabelBinarizer()
```

In [362…
```
df['bin_companyName'] = mlb.transform(df['companyLabel']).tolist()
```

In [363…
```
df[0:10]
```

Out[363]:

	text	score	company	companyLabel	bin_companyLable	bin_companyName
0	So there is no way for me to plug it in here i...	0	amazon	0	[1, 0, 0]	[1, 0, 0]
1	Good case, Excellent value.	1	amazon	0	[1, 0, 0]	[1, 0, 0]
2	Great for the jawbone.	1	amazon	0	[1, 0, 0]	[1, 0, 0]
3	Tied to charger for conversations lasting more...	0	amazon	0	[1, 0, 0]	[1, 0, 0]
4	The mic is great.	1	amazon	0	[1, 0, 0]	[1, 0, 0]
5	I have to jiggle the plug to get it to line up...	0	amazon	0	[1, 0, 0]	[1, 0, 0]
6	If you have several dozen or several hundred c...	0	amazon	0	[1, 0, 0]	[1, 0, 0]
7	If you are Razr owner...you must have this!	1	amazon	0	[1, 0, 0]	[1, 0, 0]
8	Needless to say, I wasted my money.	0	amazon	0	[1, 0, 0]	[1, 0, 0]
9	What a waste of money and time!.	0	amazon	0	[1, 0, 0]	[1, 0, 0]

In [364…

```
df[["bin_companyName", "bin_companyLable"]]
```

Out[364]:

	bin_companyName	bin_companyLable
0	[1, 0, 0]	[1, 0, 0]
1	[1, 0, 0]	[1, 0, 0]
2	[1, 0, 0]	[1, 0, 0]
3	[1, 0, 0]	[1, 0, 0]
4	[1, 0, 0]	[1, 0, 0]
...	...	...
2709	[0, 0, 1]	[0, 0, 1]
2710	[0, 0, 1]	[0, 0, 1]
2711	[0, 0, 1]	[0, 0, 1]
2712	[0, 0, 1]	[0, 0, 1]
2713	[0, 0, 1]	[0, 0, 1]

2714 rows × 2 columns

---

6. Data Exploration

Sometimes you need to take a peek at your data to understand the relationships in your dataset. Here, we will focus in a similarity example. Let's take 3 documents and compare them.

In [366…

```
# We retrieve 3 sentences for a random record
document_to_transform_1 = []
random_record_1 = df.iloc[50]
random_record_1 = random_record_1['text']
document_to_transform_1.append(random_record_1)
```

```
document_to_transform_2 = []
random_record_2 = df.iloc[100]
random_record_2 = random_record_2['text']
document_to_transform_2.append(random_record_2)

document_to_transform_3 = []
random_record_3 = df.iloc[150]
random_record_3 = random_record_3['text']
document_to_transform_3.append(random_record_3)
```

In [367…
```
print(document_to_transform_1)
print(document_to_transform_2)
print(document_to_transform_3)
```

```
['good protection and does not make phone too bulky.']
['Buyer Beware, you could flush money right down the toilet.']
['Audio Quality is poor, very poor.']
```

In [368…
```
from sklearn.preprocessing import binarize

# Transform sentence with Vectorizers
document_vector_count_1 = count_vect.transform(document_to_transform_1)
document_vector_count_2 = count_vect.transform(document_to_transform_2)
document_vector_count_3 = count_vect.transform(document_to_transform_3)

# Binarize vectors to simplify: 0 for abscence, 1 for prescence
document_vector_count_1_bin = binarize(document_vector_count_1)
document_vector_count_2_bin = binarize(document_vector_count_2)
document_vector_count_3_bin = binarize(document_vector_count_3)

# print vectors
print("Let's take a look at the count vectors:")
print(document_vector_count_1.todense())
print(document_vector_count_2.todense())
print(document_vector_count_3.todense())
```

```
Let's take a look at the count vectors:
[[0 0 0 ... 0 0 0]]
[[0 0 0 ... 0 0 0]]
[[0 0 0 ... 0 0 0]]
```

In [377…
```
from sklearn.metrics.pairwise import cosine_similarity

# Calculate Cosine Similarity
cos_sim_count_1_2 = cosine_similarity(document_vector_count_1, document_vector_count_
cos_sim_count_1_3 = cosine_similarity(document_vector_count_1, document_vector_count_
cos_sim_count_2_3 = cosine_similarity(document_vector_count_2, document_vector_count_

cos_sim_count_1_1 = cosine_similarity(document_vector_count_1, document_vector_count_
cos_sim_count_2_2 = cosine_similarity(document_vector_count_2, document_vector_count_
cos_sim_count_3_3 = cosine_similarity(document_vector_count_3, document_vector_count_

# Print
print("Cosine Similarity using count bw 1 and 2: %(x)f" %{"x":cos_sim_count_1_2})
print("Cosine Similarity using count bw 1 and 3: %(x)f" %{"x":cos_sim_count_1_3})
print("Cosine Similarity using count bw 2 and 3: %(x)f" %{"x":cos_sim_count_2_3})

print("Cosine Similarity using count bw 1 and 1: %(x)f" %{"x":cos_sim_count_1_1})
print("Cosine Similarity using count bw 2 and 2: %(x)f" %{"x":cos_sim_count_2_2})
print("Cosine Similarity using count bw 3 and 3: %(x)f" %{"x":cos_sim_count_3_3})
```

```
Cosine Similarity using count bw 1 and 2: 0.000000
Cosine Similarity using count bw 1 and 3: 0.000000
Cosine Similarity using count bw 2 and 3: 0.000000
Cosine Similarity using count bw 1 and 1: 1.000000
Cosine Similarity using count bw 2 and 2: 1.000000
Cosine Similarity using count bw 3 and 3: 1.000000
```

In [ ]: