Your name:

B565 HW1 (Spring 2022)

Sumit Lakhawat

Questions

1. Tan Chapter 2, Problem 8. Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features. [10 pts]

Solution –

A discrete attribute has finite set of values or countably infinite set of values. A document-term matrix is an example of that. Only the presence of a term in the document is regarded as important.
As the occurrence of all potential phrases in a document is extremely rare, there are many zero entries in a document matrix that are useless. A document-term matrix is an example of a dataset with asymmetric discrete characteristics, as indicated by this. As the values indicating if the term is present in the matrix or not is represented by an integer (0/1), it is discrete in nature.

Now, since the document-term matrix is not an efficient way to represent big collection of documents, if we apply TF-IDF normalization to each of the terms then the matrix obtained will be a document term matrix with continuous values. The features present will still be asymmetric, as we give a value of 1 if the term is present in the document and 0 if it is not. But even after normalization, the zero entries will remain zero. This will make the document-term matrix an example of a dataset that has asymmetric continuous features.

This way a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

2. Tan Chapter 2, Problem 12. Distinguish between noise and outliers. Be sure to consider the following questions. [10 pts]

Answer -

- Is noise ever interesting or desirable? Outliers?

  So, Noise is not desirable or interesting as it is a disturbance that gets added to the data sometimes. We eliminate noise before we process our data further. On the other hand, Outliers are something from which our model can learn. These are anomalies that we have to watch out for. For example, in credit card fraud detection, we can watch outlier data points in order to predict a suspicious credit card usage. So, outliers are our values of interest, thus they are desirable and interesting.
- Can noise objects be outliers?

When noise is present in the dataset, it is possible that it modifies some data points which will seem rather unusual. So, it is possible that some of the data points appear outliers because of the noise present in the dataset.

- Are noise objects always outliers?

  No. Noise can also cause data points to act like normal data points. So, noise objects are not always outliers.

- Are outliers always noise objects?

  No, outliers are legitimate data points that are different from the actual dataset.

- Can noise make a typical value into an unusual one, or vice versa?

  Yes, if there is noise in the dataset, we will not be able to accurately judge if a data point is an outlier or typical. Noise can make an outlier a typical data point and a typical data point an outlier.

3. Implement a notebook on Kaggle to explore this dataset. This dataset lists the number of antibiotic resistance genes (AMR), and the presence or absence of the CRISPR-Cas systems in the genomes included in the file. Report what you have learned by including the html output from your notebook in the PDF file you are going to submit. What to check? The distribution of the two variables (AMR, CRISPR-Cas), and if there is any correlation between the two variables. [25 pts]

Answer –

Your name:

In [40]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [41]:
```
dataset = pd.read_csv("../input/efaeciumamrc/Efaecium_AMRC.csv")
dataset.head()
```

Out[41]:

	genome_ID	CRISPR_Cas	AMR
0	GCA_010120755.1_ASM1012075v1	0	8
1	GCA_001720945.1_ASM172094v1	0	21
2	GCA_009697285.1_ASM969728v1	0	13
3	GCA_900639535.1_E8202_hybrid_assembly	0	11
4	GCA_002007625.1_ASM200762v1	0	18

In [42]:
```
dataset.shape
```

Out[42]:
```
(2223, 3)
```

In [43]:
```
dataset.dtypes
```

Out[43]:
```
genome_ID     object
CRISPR_Cas     int64
AMR            int64
dtype: object
```

Your name:

In [44]:
```
dataset.describe()
```

Out[44]:

	CRISPR_Cas	AMR
count	2223.000000	2223.000000
mean	0.024741	10.330184
std	0.155371	6.661470
min	0.000000	0.000000
25%	0.000000	3.000000
50%	0.000000	12.000000
75%	0.000000	16.000000
max	1.000000	31.000000

In [45]:
```
dataset.nunique()
```

Out[45]:
```
genome_ID      2223
CRISPR_Cas        2
AMR              27
dtype: int64
```

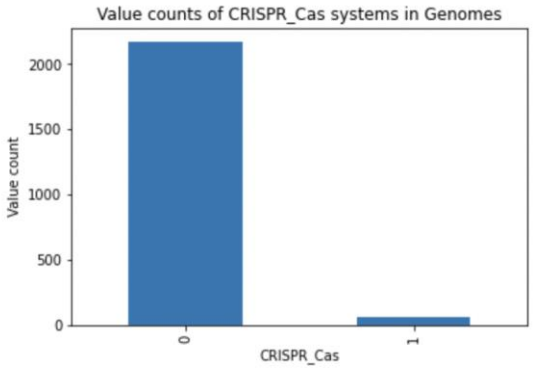
In [46]:
```
unique_CRISPR_Cas = dataset.CRISPR_Cas.unique()
print("Unique values of CRISPR_Cas: ",unique_CRISPR_Cas)
```

```
Unique values of CRISPR_Cas:  [0 1]
```

In [47]:
```
CRISPR_Cas_present = dataset[dataset.CRISPR_Cas == 1]
CRISPR_Cas_absent = dataset[dataset.CRISPR_Cas == 0]
```

In [48]:
```
dataset.CRISPR_Cas.value_counts().plot.bar(title = 'Value counts of CRISPR_Cas systems in Genomes')
plt.xlabel('CRISPR_Cas')
plt.ylabel('Value count')
```

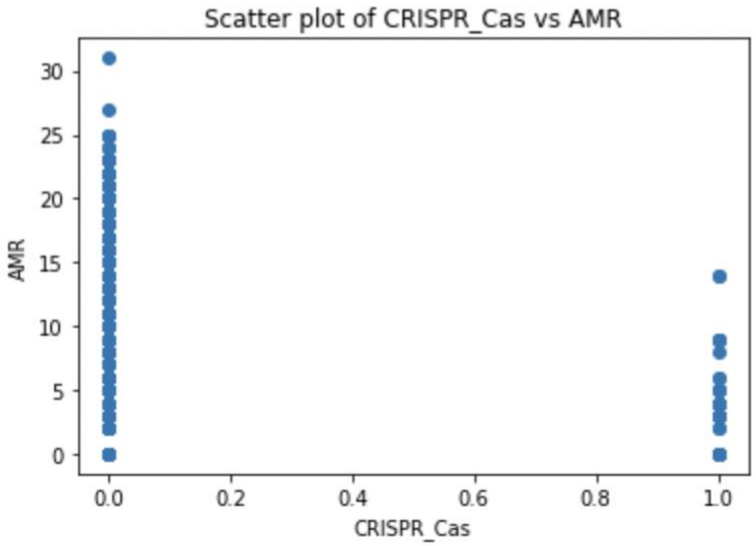
Out[48]:
```
Text(0, 0.5, 'Value count')
```



Your name:

In [49]:
```
plt.plot(dataset.CRISPR_Cas,dataset.AMR,'o')
plt.xlabel("CRISPR_Cas")
plt.ylabel("AMR")
plt.title("Scatter plot of CRISPR_Cas vs AMR")
```

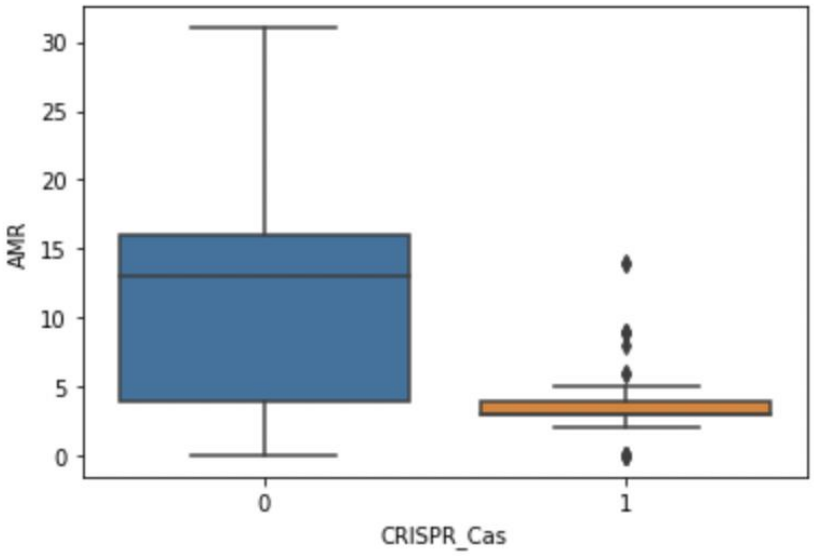
Out[49]: Text(0.5, 1.0, 'Scatter plot of CRISPR_Cas vs AMR')



Your name:

```
In [50]: sns.boxplot(x = dataset.CRISPR_Cas,y = dataset.AMR)

Out[50]: <AxesSubplot:xlabel='CRISPR_Cas', ylabel='AMR'>
```



From the above statistics we observe that:

1. The genomes which do not have the CRISPR-Cas systems have a higher count of antibiotic resistance genes on an average, relative to the genomes which have the CRISPR-Cas systems.

2. There are 253 or 11.6% genomes which do not have the CRISPR-Cas systems and do not have the antibiotic resistance genes. 16.36% genomes which have the CRISPR-Cas systems and do not have the antibiotic resistance genes.

Conclusion The genomes which do not have the CRISPR-Cas systems have lesser number of antibiotic resistant genes on an average, compared to genomes which do not have the CRISPR-Cas systems.

4. Learn about Omicron using Google Trend. Write a brief summary including four highlights about what you have learned. [25 pts]

Answer –

The trend over searching the word omicron has somewhat obvious graph. It starts to emerge around late 2021 as omicron came into light. As we move forward, it has a peak followed by a valley. As we witness around the world, the cases have increased exponentially in lot of places in United States, the trend begins to emerge and reaches a new peak. It's currently descending once more.

We can explore more from this trend as we can learn where these searches originated in the United States by looking at interest by subregion. By doing so, we observe that it shows light and dark colors at the borders of the country to reflect that those areas were more involved in leading these searches. These searches were most from the following three states - Columbia, Washington, and New Jersey.

The related queries information enables us to see what users are looking for in addition to omicron. It's unsurprising that it's followed by “omicron symptoms”, implying that people are using the Internet to some part to know more about their symptoms and maybe trying to figure out if they have the same symptoms as the omicron variant one. This also explains the third most searched word “omicron variant” as people are trying to learn more about this novel variant of COVID-19.

The queries data set and related subjects are inextricably linked. The rising feature present there allows us to look at terms that have rapidly increased in popularity in a short period of time. This data contains values that are substantially similar to those found in the queries section.

The downward trend now shows us that people have searched a lot about it in the previous few months and are either more knowledgeable about omicron variant or they are tired to search more about it.

This allows us to see lot of different features about a particular word and gives lot of insights into smaller more niche things.

Your name:

5. Write a summary for this paper: COVID-19 or Flu? Discriminative Knowledge Discovery of COVID-19 Symptoms from Google Trends Data. [30 pts]

   - The length of your summary is about one page.
   - State main ideas: problem that the paper tries to address, what data was used,and what was the method that was applied/developed to solve the problem.
   - Add your personal opinion. Do you like the paper or not? Why? How do youthink about the paper?
   - Write the summary in your own words; don't copy and paste.

Answer –

Overview –

This paper uses Google Trend data to find symptoms that are exclusive to COVID-19, By comparing data from two time periods - years when COVID-19 and flu are both prevalent (2020) and years when just flu is prevalent (2018, 2019).

The dataset has data points from all 51 states over 3 years, 2018,2019,2020 and split into Background dataset and Target dataset, where Background dataset has the data points when there was only flu and Target dataset has data points when there was both flu and COVID-19.

Method –

The goal is to maximize the variance between Target and background dataset. The paper uses two methods, discriminative principal component analysis (dPCA) and contrustive principal component analysis (cPCA). Here, dPCA tries to maximize the ratio of target data variance to background data variance. And, the goal of cPCA is to maximize the difference between target data variance and background data variance. The difficulty faced with dPCA to find the discriminative symptoms of Covid 19 and flu was Sign ambiguity. Finally, to address the problem, a novel non-negative discriminative analysis (DNA) approach was presented, which performs non-negative matrix factorization (NNMF) on Cy-1 Cx, where Cx, Cy are the covariance matrices of the target and background datasets, respectively.

Evaluation –

According to the data obtained from the experimental test cases, DNA was able to successfully identify three unique symptoms that are specific to Covid 19, which are ageusia, shortness of breath, and anosmia. And, six symptoms common to both flu and COVID-19, which are vomiting, diarrhea, cough, fever, fatigue and headache. DNA method outperformed the existing methods tested like dPCA, cPCA and NNMF.

Your name:

Personal opinion –

I like the paper as it first analyzes the existing methods and solve the issue by proposing a novel method. I find it informative that it shows why and how the existing methods like dPCA, cPCA and NNMF have failed to give the desired result. This was a unique study to distinguish symptoms unique to Covid 19 from the flu and the proposed algorithm, DNA was successfully able to distinguish symptoms unique to Covid 19.