Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 3

Due: February 15, 2023(100 points)

YOUR NAME
YOUR GIT USERNAME
YOUR E-MAIL

Multisource Joins

News articles are commonly aggregated from multiple sites and companies. The landscape of news has been evolving ever since social media has amplified its effects. In politics, Congress has explored the topic of bias with the diversity of news sources. That is, news articles may cover news stories with differing perspectives and language.

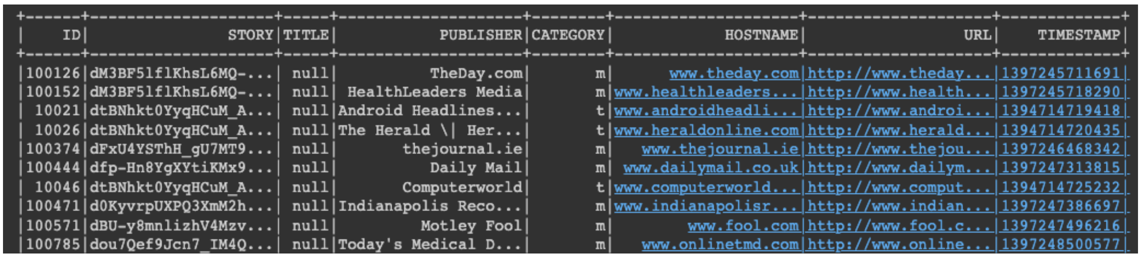
The data that we will be using today comes from Kaggle, and it is available here. There are two CSV files that we wish to join in this week's homework:

- `data/id_titles.csv`
- `data/id_publishers.csv`

As there name suggests, there is publishing data associated with articles and there is title and description information associated with the same articles. Each table has many instances, and each instance for both tables have an associated ID, where it is possible to join the two data sources.

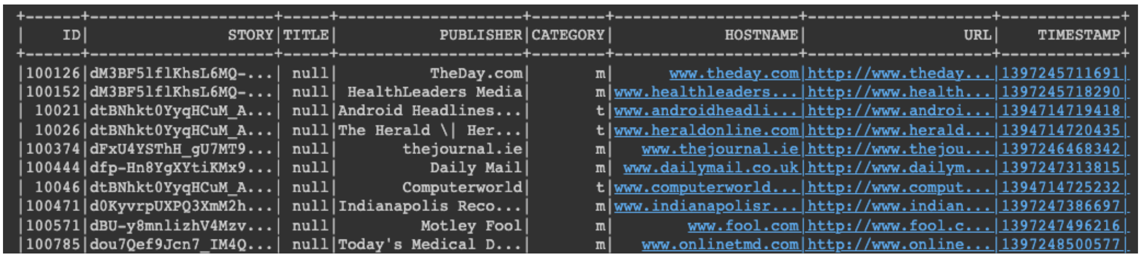
In this particular case, there is some missing information in the join. Your task is as follows.

1. Write out a file that has all the data columns, but where the rows are only those articles where there are titles but no publisher information. Call it titles_no_publishers.txt. In your PDF writeup, include the first 10 rows ordered by ID. This table should look something like the below (ignore the values):



```
+------+--------------------+-----+--------------------+--------+--------------------+--------------------+-------------+
|    ID|               STORY|TITLE|           PUBLISHER|CATEGORY|            HOSTNAME|                 URL|    TIMESTAMP|
+------+--------------------+-----+--------------------+--------+--------------------+--------------------+-------------+
|100126|dM3BF5lflKhsL6MQ-...| null|          TheDay.com|       m|      www.theday.com|http://www.theday...|1397245711691|
|100152|dM3BF5lflKhsL6MQ-...| null| HealthLeaders Media|       m|www.healthleaders...|http://www.health...|1397245718290|
| 10021|dtBNhkt0YyqHCuM_A...| null|Android Headlines...|       t|www.androidheadli...|http://www.androi...|1394714719418|
| 10026|dtBNhkt0YyqHCuM_A...| null|The Herald \| Her...|       t|www.heraldonline.com|http://www.herald...|1394714720435|
|100374|dFxU4YSThH_gU7MT9...| null|       thejournal.ie|       m|   www.thejournal.ie|http://www.thejou...|1397246468342|
|100444|dfp-Hn8YgXYtiKMx9...| null|          Daily Mail|       m| www.dailymail.co.uk|http://www.dailym...|1397247313815|
| 10046|dtBNhkt0YyqHCuM_A...| null|       Computerworld|       t|www.computerworld...|http://www.comput...|1394714725232|
|100471|d0KyvrpUXPQ3XmM2h...| null|Indianapolis Reco...|       m|www.indianapolisr...|http://www.indian...|1397247386697|
|100571|dBU-y8mnlizhV4Mzv...| null|         Motley Fool|       m|        www.fool.com|http://www.fool.c...|1397247496216|
|100785|dou7Qef9Jcn7_IM4Q...| null|Today's Medical D...|       m|   www.onlinetmd.com|http://www.online...|1397248500577|
```

2. Write out a file that has all the data columns, but where the rows are only those articles where there are publishers but not title information. Call it titles_no_publishers.txt. In your PDF writeup, include the first 10 rows ordered by ID. That table should look something like the below (ignore the values):



```
+------+--------------------+-----+--------------------+--------+--------------------+--------------------+-------------+
|    ID|               STORY|TITLE|           PUBLISHER|CATEGORY|            HOSTNAME|                 URL|    TIMESTAMP|
+------+--------------------+-----+--------------------+--------+--------------------+--------------------+-------------+
|100126|dM3BF5lflKhsL6MQ-...| null|          TheDay.com|       m|      www.theday.com|http://www.theday...|1397245711691|
|100152|dM3BF5lflKhsL6MQ-...| null| HealthLeaders Media|       m|www.healthleaders...|http://www.health...|1397245718290|
| 10021|dtBNhkt0YyqHCuM_A...| null|Android Headlines...|       t|www.androidheadli...|http://www.androi...|1394714719418|
| 10026|dtBNhkt0YyqHCuM_A...| null|The Herald \| Her...|       t|www.heraldonline.com|http://www.herald...|1394714720435|
|100374|dFxU4YSThH_gU7MT9...| null|       thejournal.ie|       m|   www.thejournal.ie|http://www.thejou...|1397246468342|
|100444|dfp-Hn8YgXYtiKMx9...| null|          Daily Mail|       m| www.dailymail.co.uk|http://www.dailym...|1397247313815|
| 10046|dtBNhkt0YyqHCuM_A...| null|       Computerworld|       t|www.computerworld...|http://www.comput...|1394714725232|
|100471|d0KyvrpUXPQ3XmM2h...| null|Indianapolis Reco...|       m|www.indianapolisr...|http://www.indian...|1397247386697|
|100571|dBU-y8mnlizhV4Mzv...| null|         Motley Fool|       m|        www.fool.com|http://www.fool.c...|1397247496216|
|100785|dou7Qef9Jcn7_IM4Q...| null|Today's Medical D...|       m|   www.onlinetmd.com|http://www.online...|1397248500577|
```

3. Explore the data further, and identify potential problems that could arise if we were to further analyze the data (e.g., apply a machine learning algorithm). Is there still missing data? That is to say, do all the columns have the correct data? What could have gone wrong in the data creation step? (You needn't code anything, but conceptually describe any issues you see and how you would remedy it.)

.

Frequent Itemsets

Consider the following set of frequent 3-itemsets:

```
{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4},
{2, 3, 4}, {2, 3, 5}, {3, 4, 5}.
```

Assume that there are only five items in the data set. This question was taken from Tan et al., which may help in reviewing Candidate Generation.

4. List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

5. List all candidate 4-itemsets obtained by the candidate generation procedure in A Priori, using $F_{k-1} \times F_{k-1}$.

6. List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

Principle Components Analysis

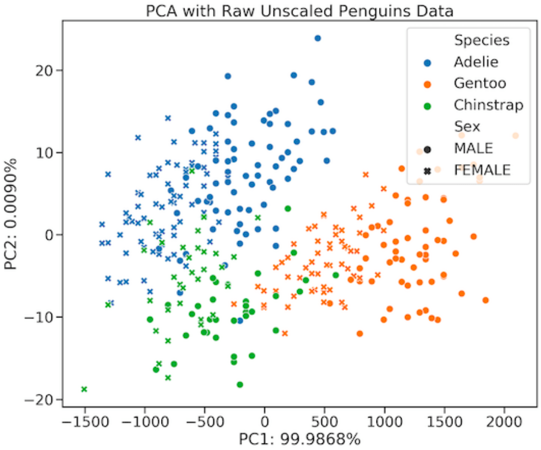
Italy is home to over 2000 grape varieties. Even within a single region, wines exhibit distinct attributes from different cultivators that can be measured with objective and numerical features. Notably, in the dataset we are exploring today, there are thirteen different measurements taken for different constituents found in the three types of wine. We would like to visualize how well-separated the data is for the different wineries in a 2D scatter plot.

We will be using the UCI Wine's dataset. Please review `sklearn`'s description of wine data, and load it in with the following code:

```
from sklearn.datasets import load_wine
wine = load_wine()
```

6. Preprocess the the data with **z-score normalization** and scatter the data that's been projected onto the first two principle components with different colors for each target/-class of wine. Include your code (linked or inline).

   The below scatter plot is *an example* of displaying multiple classes with different colors on a single plot. (Ignore the contents of the plot.)


PCA with Raw Unscaled Penguins Data
Species
Adelie
Gentoo
Chinstrap
Sex
MALE
FEMALE
PC2: 0.0090%
PC1: 99.9868%