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.

Question 1 a.) [15 pts]

- Write out a file that has all the publishers for which there are no titles, called publishers_no_titles.txt. This table should look something like the below (ignore the values):

ID	STORY	TITLE	PUBLISHER	CATEGORY	HOSTNAME	URL	TIMESTAMP
100068	dJ_k5DjBr5MzK0MHf...	Networks: Kathlee...	null	null	null	null	null
100176	dM3BF51flKhsL6MQ-...	Medicare data giv...	null	null	null	null	null
100192	dM3BF51flKhsL6MQ-...	Medicare Records ...	null	null	null	null	null
100422	duBSqD7s8phcPsMQK...	Sales get leaner ...	null	null	null	null	null
100442	dfp-Hn8YgXYtiKMx9...	More than 100 pas...	null	null	null	null	null
100570	dBU-y8mnlizhV4Mzv...	Today's Pre-Marke...	null	null	null	null	null
100653	dwnBgdLk-3bzGBMNi...	Aid workers back ...	null	null	null	null	null
100716	dwnBgdLk-3bzGBMNi...	WHO says West Afr...	null	null	null	null	null
100850	dk_vhtrqQFe_dsMiu...	Flu Drugs Tamiflu...	null	null	null	null	null
100939	dk_vhtrqQFe_dsMiu...	Study Questions O...	null	null	null	null	null
100969	dk_vhtrqQFe_dsMiu...	Tamiflu use calle...	null	null	null	null	null
101119	dDtTmiUm0P1qeMMK8...	US close: Sell-of...	null	null	null	null	null
101301	d4p273oepCNzWtMV5...	Can Family Dollar...	null	null	null	null	null
101330	dhpby_46Ae5iB8ME-...	A Turbulent Week ...	null	null	null	null	null
10152	dOQvzWTEFn4NkVM9c...	T. rex's 'pygmy' ...	null	null	null	null	null
101704	dq4CkE5dd_NRkmMCB...	Ron Agostini: Col...	null	null	null	null	null
101839	dSAALz3YglIjh5MZV...	Fitch: JPMorgan 1...	null	null	null	null	null
10191	dA0ddnisozIS59MZa...	Earth has a secre...	null	null	null	null	null
101912	dJVPX-uN99u_nuMNq...	GGG-GAME CHANGER:...	null	null	null	null	null

Question 1 b.) [15 pts]

- Write out a file that has all the publishers for which there are no titles, called titles_no_publishers.txt. That table should look something like the below (ignore the values):

ID	STORY	TITLE	PUBLISHER	CATEGORY	HOSTNAME	URL	TIMESTAMP
100126	dM3BF51flKhsL6MQ-...	null	TheDay.com	m	www.theday.com	http://www.theday...	1397245711691
100152	dM3BF51flKhsL6MQ-...	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

Question 1 c.) [5 pts]

- Explore the data further, and identify any potential problems that could arise if we were to apply any algorithm to it. Is there still missing data? If there are any issues with the data, what might you do to fix it? (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 in Rule Generation.

Question 2a.) [10 pts]

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

Question 2b.) [10 pts]

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

Question 2c.) [10 pts]
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()
```

Question 3 [20 pts]

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.

Parameter Estimation

It is well-known that light bulbs commonly go out according to a Poisson distribution, and are independent regardless of whether or not they're made in the same factory. The Poisson distribution has the form:

$$p(X|\lambda) = \frac{\exp^{-\lambda} \lambda^{x_i}}{x_i!}$$

An architect has outfitted a building with 32,000 of the same lightbulb. The factory has provided him with data on when N of these lightbulbs have gone out over their lifetimes. They've been measured with $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$

Question 4 [20 pts]

Derive the maximum likelihood estimate of the parameter λ in terms of x_i. Please show your work.

Submission Instructions

When you have finished, follow the instructions on the homework main page. Commit your code, outputs, and PDF writeup to your repository and provide the repository link to Gradescope.