

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](https://course.ccs.neu.edu/cs6220/homework-3/). 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.)

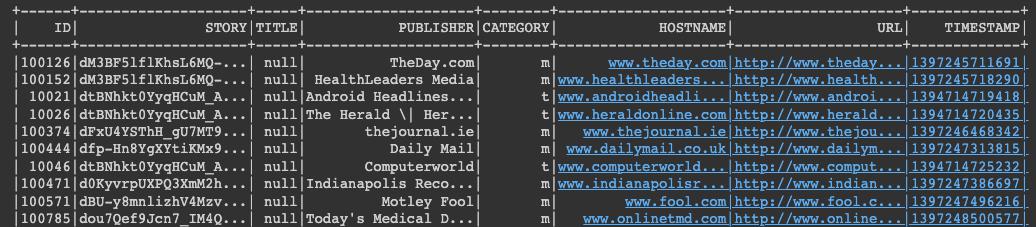
* Write out a file that has all the publishers for which there are no titles, called publish-ers no titles.txt. This table should look something like the below (ignore the values):

1



Question 1 b.)

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



.

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.

Question 2 [15 pts total]

[5 pts] Question 2a.) List all candidate 4-itemsets obtained by a candidate generation pro-cedure using the Fk−1 × F1 merging strategy.

[5 pts] Question 2b.) List all candidate 4-itemsets obtained by the candidate generation procedure in A Priori, using Fk−1 × Fk−1.

[5 pts] Question 2c.) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

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. An architect has

2

outfitted a building with 32,000 of the same lightbulb.

Question 3 [15 pts total]

Assuming the Poisson distribution has the form:

|  |  |  |  |
| --- | --- | --- | --- |
| p(X|λ) = | exp−λ λxi | (0.1) |  |
| xi! |  |

derive the maximum likelihood estimate of the parameter λ in terms of xi.

Submission Instructions

When you have finished, follow the instructions on the [homework main page](https://course.ccs.neu.edu/cs6220/homework-3/) commit your code, outputs, and PDF writeup to your repository and provide the repository link to [Gradescope](https://www.gradescope.com).

3