**YOUR NAME**  
**YOUR GIT USERNAME**  
**YOUR E-MAIL**

**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:

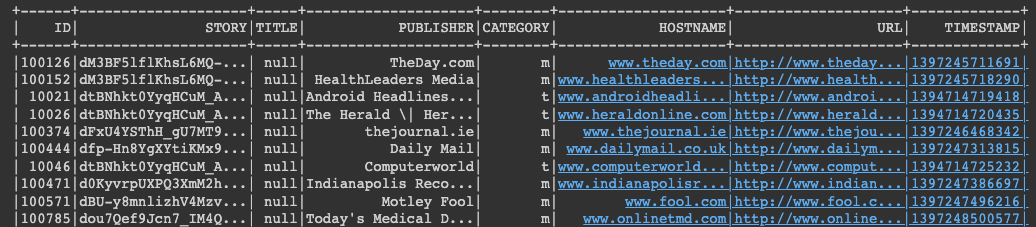
* data/id\_titles.csv
* data/id\_publishers.csv

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 publishers\_no\_titles.txt. This table should look something like the below (ignore the values):  
  

. .  
**Question 1 b.)**

* 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):  
  

# Frequent Itemsets

## What are the frequent itemset rules?

**Question 2** What are the frequent itemset rules

# Parameter Estimation

In the data provided by j

# K-Means

The [normalized automobile distributor timing speed and ignition coil gaps](https://course.ccs.neu.edu/cs6220/homework-3/data/) for production F-150 trucks over the years of 1996, 1999, 2006, 2015, and 2022. We have stripped out the labels for the five years of data.  
**Question 3**  
Implement a simple -means algorithm in Python on Colab with the following initialization:

In order to main consistency between submissions, use a random seed of 27. You can do this with

>> numpy.random.seed(seed=27)

Scatter the results with the different clusters as different colors. What do you notice?  
**Question 4**  
A common distance metric is the *Mahalanobis Distance* with a specialized covariance. Implement -means with a distance metric as defined as:

where **x** and **y** are two points, and is the distance between them.

Implement a specialized -means with the above Mahalanobis Distance. Scatter the results with the different clusters as different colors. What do you notice? You may want to pre-compute so that you aren’t calculating an inverse every single loop of the the -Means algorithm.

# Submission Instructions

* Clearly mark your questions. Submit your **Colab URL**, output files (**publishers\_no\_titles.txt** and **titles\_no\_publishers.txt**) and PDF write-up via from the invited [Github link](https://classroom.github.com/a/Kog9MCRN). Provide repository URL on [Gradescope](https://www.gradescope.com/courses/494275) before 5pm Wednesday, February 1, 2023.
* [this Google Forms](https://docs.google.com/forms/d/e/1FAIpQLSfUfjMu0JPzuYqisZux2cvh1EfMH-D9MbkFFYjxp-yuRtLLvg/viewform) before 5pm Wednesday, February 1, 2023. You will need to follow the instructions about permissions.
* Colab has an extensive Markdown capability, so make sure you document your code while writing it. Code’s legibility is part of our grading criterion, so please make sure it’s readable.
* Include a diagram of your pipeline description in your writeup.
* Include in your writeup the recommendations for the users with following user IDs: 924, 8941, 8942, 9019, 9020, 9021, 9022, 9990, 9992, 9993.