

ANALYTICS SYSTEMS ENGINEERING

ASSIGNMENT 1

shreenidhi.bharadwaj@northwestern.edu | ChristopherFiore2015@u.northwestern.edu

Objective

- Learn about AWS platform, source code control, data ingestion techniques, data storage

Submissions

- All project artifacts including README.txt & code.

Question 1: {50 points}

- Create a python script that scrapes the data from a webpage or using an API. You can use any data sources such as Twitter, YouTube, Yelp. etc

References:

- <https://towardsdatascience.com/how-to-web-scrape-with-python-in-4-minutes-bc49186a8460>
- <https://www.dataquest.io/blog/web-scraping-beautifulsoup/>
- <https://github.com/REMitchell/python-scraping>
- <https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf>
- <https://hackernoon.com/web-scraping-tutorial-with-python-tips-and-tricks-db070e70e071>

Question 2: {10 points}

- Using Python, prepare the raw data to be stored in S3 within AWS.

Question 3: {20 points}

- Create the landing zone in S3.

Question 4: {20 points}

- Spin up PostgreSQL server, create a data model and load the curated data.
- Write any 4 SQL queries that provides insights relating to the dataset.