DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
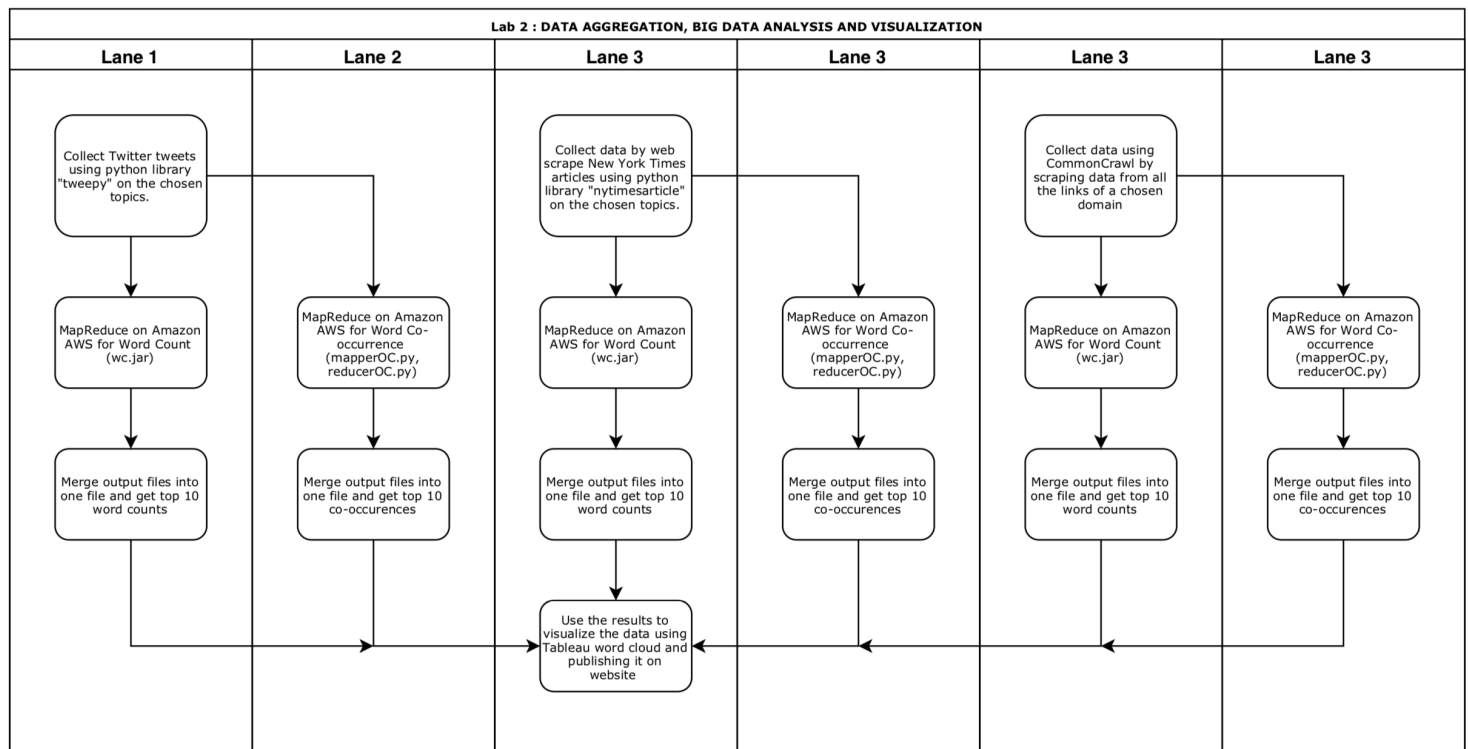# UNIVERSITY AT BUFFALO

DATA INTENSIVE COMPUTING

Lab 2 : DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

Akhil Amit Senroy
UBIT : akhilami
Person # : 50288588

LAB 2 : FLOW DIAGRAM

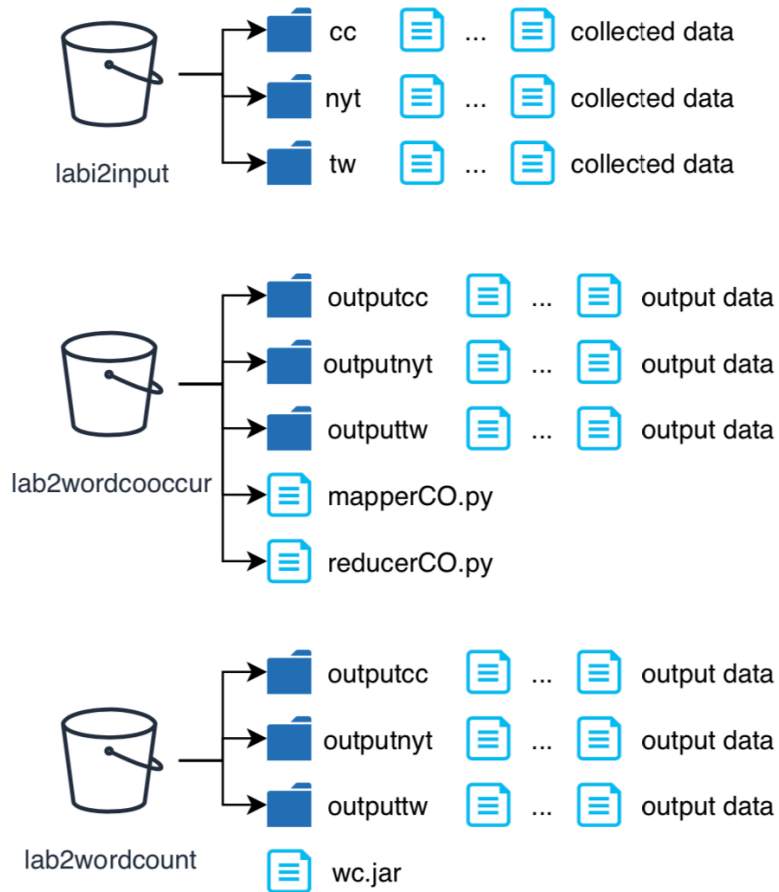| Lab 2 : DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION					
Lane 1	**Lane 2**	**Lane 3**	**Lane 3**	**Lane 3**	**Lane 3**
Collect Twitter tweets using python library "tweepy" on the chosen topics.		Collect data by web scrape New York Times articles using python library "nytimesarticle" on the chosen topics.		Collect data using CommonCrawl by scraping data from all the links of a chosen domain	
MapReduce on Amazon AWS for Word Count (wc.jar)	MapReduce on Amazon AWS for Word Co-occurrence (mapperOC.py, reducerOC.py)	MapReduce on Amazon AWS for Word Count (wc.jar)	MapReduce on Amazon AWS for Word Co-occurrence (mapperOC.py, reducerOC.py)	MapReduce on Amazon AWS for Word Count (wc.jar)	MapReduce on Amazon AWS for Word Co-occurrence (mapperOC.py, reducerOC.py)
Merge output files into one file and get top 10 word counts	Merge output files into one file and get top 10 co-occurences	Merge output files into one file and get top 10 word counts	Merge output files into one file and get top 10 co-occurences	Merge output files into one file and get top 10 word counts	Merge output files into one file and get top 10 co-occurences
		Use the results to visualize the data using Tableau word cloud and publishing it on website			

Chosen Topics : "music", "song", "guitar", "violin", "piano", "sing", "orchestra"

Python Libraries Used :

- import tweepy
- from nytimesarticle import articleAPI
- import requests
- import pandas as pd
- import csv
- import time
- import os
- import glob
- from os.path import join, getsize
- from nltk.tokenize import word_tokenize
- from nltk.corpus import stopwords
- from nltk.stem.porter import PorterStemmer

Amazon AWS : S3 Directory Structure

DATA VISUALIZATION

Twitter - Word Co-occurrence

song|song

sing|song music|gt play|violin

play|guitar music|video

music|best music|one

guitar|solo song|http

F1
- guitar|solo
- music|best
- music|gt
- music|one
- music|video
- play|guitar
- play|violin
- sing|song
- song|http
- song|song

New York Times - Word Count

play

also time album

work music perform

year but song

F1
- album
- also
- but
- music
- perform
- play
- song
- time
- work
- year

New York Times - Word Co-occurrence

F1
- jazz|happen
- music|director
- music|festiv
- music|listen
- music|video
- piano|concerto
- play|guitar
- play|music
- song|like
- song|video

song|video
play|music
music|listen   music|festiv
piano|concerto
play|guitar
jazz|happen   music|director
music|video

CommonCrawl - Word Count

F1
- also
- artist
- field
- get
- like
- music
- piano
- playlist
- time
- way

field also   artist like
piano   music   playlist
get time   way

CommonCrawl - Word Co-occurrence

music|data

music|industri

music|releas

music|project

music|product

music|content

music|video

music|conceiv

music|diy

music|manag