UGA CSCI4795/6795: Cloud Computing (Spring 2019)

Programming Assignment 5 (PA#5) (due: May 04 2019 – 11:59 p.m.)

Goals

Celebrate our senior students' graduation and gain more hands-on experience with MapReduce (Hadoop)

Introduction

Spring 2019 graduation commencement will be held on May 10th. CSCI 4795/6795 cloud computing class has a number of UGA seniors (4th year undergrads). Congratulations to UGA class of 2019 on your great achievement!!! (You will face real challenges in your life though 😆) Anyways... Social network services (e.g., Facebook, Instagram, Twitter) are everywhere. It is very common to post your story with specific hashtags to the services and share the post with your friends. In this assignment, you are required to analyze social network data (json files from Instagram) using Hadoop and find interesting posts/statistics related to **UGA 2019 graduation commencement**. Again, **you must use Hadoop for this assignment (using any add-ons like HBase, NoSQL like MongDB, or Spark are not allowed).** Dataset includes 43,645 json files from four hashtags related to the University of Georgia. The hashtags used for data collection include `#godawgs`, `#uga`, `#ugabulldogs`, `#universityofgeorgia`, and others.

Questions

- Find top 5 most frequently used hashtags related to "UGA's Spring 2019 graduation commencement". Note that your top 5 hashtags shouldn't include the following hashtags: `#godawgs, #uga, #ugabulldogs, #universityofgeorgia`
  - The list of hashtags and their statistics (e.g., ranking, frequency)
  - Time series graphs, representing the frequency of top 5 hashtags
- Find top 3 most influencers on Instagram regarding "UGA's Spring 2019 graduation commencement"
  - The list of top 3 influencers and their ranking
  - Time series graphs, representing the frequency of top 3 most influencers
- Find any interesting statistics related to "UGA's Spring 2019 graduation commencement"
  - List of your finding and statistics
  - Time series graphs, visualizing your findings

Similar to PA #4, you should use python to write Map/Reduce programs using Hadoop 2.7.6 (pseudo-distribution mode) on GCP instance to compute and analyze the following json dataset: http://cobweb.cs.uga.edu/~kim/classes/S19-CSCI4795-6795/PA5/uga-hashtags-jsons.tar.gz Again, **you cannot use any "add-on" to Hadoop (such as Hive), NoSQL (such as MongoDB) or Spark. The great majority of the data processing must be performed within Hadoop; relatively minor "post-processing" (e.g., sort) is allowed outside of Hadoop.** You are required to submit two items via eLC:

1. A PDF of "what-to-submit.doc"
2. A zip file containing all of your Map/Reduce programs.

Additional Notes

- Json file information
  - Fine name format: *timestamp_UTC.json* e.g., 2019-03-05_20-11-32_UTC.json
  - Total # of json files: 43,645 files
  - Total size of json files: 295MBytes (uncompressed), 49MBytes (compressed tar.gz)
  - Json example is as below:


```
{'instaloader': {'node_type': 'Post', 'version': '4.2.3'},
 'node': {'__typename': 'GraphImage',
          'accessibility_caption': 'Image may contain: text',
          'comments_disabled': False,
          'dimensions': {'height': 1080, 'width': 1080},
          'display_url': 'https://scontent-atl3-…',
          'edge_liked_by': {'count': 15},
          'edge_media_preview_like': {'count': 15},
          'edge_media_to_caption': {'edges': [{'node': {'text': 'Are you
graduating this semester? Be sure to register for the Lamar Dodd School of Art
graduation ceremony! 🎓 \nLearn more through the link in our bio\n#doddlife
#thedodd #artatuga #uga'}}]},
          'edge_media_to_comment': {'count': 0},
          'id': '1997248124665835303',
          'is_video': False,
          'owner': {'id': '3041334883'},
          'shortcode': 'Bu3ppZnBa8n',
          'taken_at_timestamp': 1552310551,
          'thumbnail_resources': [{'config_height': 150,
                                   'config_width': 150,
                                   'src': 'https://scontent-atl3-…'},
                                  {'config_height': 240,
                                   'config_width': 240,
                                   'src': 'https://scontent-atl3-…'},
                                  {'config_height': 320,
                                   'config_width': 320,
                                   'src': 'https://scontent-atl3-…'},
                                  {'config_height': 480,
                                   'config_width': 480,
                                   'src': 'https://scontent-atl3-…'},
                                  {'config_height': 640,
                                   'config_width': 640,
                                   'src': 'https://scontent-atl3-…'}],
          'thumbnail_src': 'https://scontent-atl3-…'}}
```

- Interesting attributes
    - 'edge_liked_by': # of like
    - 'edge_media_to_caption' → 'edges' → 'node' → 'text': Text of the post that often includes hashtags
    - 'owner' → 'id': Instagram owner ID
    - 'taken_at_timestamp': Unix timestamp of the post. Let's assume that every post was from UTC time zone

- Minimum pre-processing (aka merging) is allowed.
    - Do not update or remove content of json files
    - Consider block size of Hadoop
- **No late or email submission allowed**
- **You are required to do this assignment alone**