**Social Media Analytics S2019 Assignment 2 (Group work)**

**Date Handed out: 11th February 2019**

**Date Due: 25th February 2019 by 11:59 p.m.**

**Is a Picture Worth a Thousand Words?**

The good folks at National Geographic (Instagram: natgeo) have hired you as a social media analytics consultant to help them increase engagement on their Instagram page. Natgeo expects you to advise them on what type of content they should post more and less of – i.e., what **types** of images increase engagement? What types decrease engagement?

You will have to scrape around 500-1000 images from the natgeo Instagram page. There are some Instagram scrapers on github. However, Instagram changed its webpage design late last year, and so some of those scrapers may have to be revised. Web Scraper may also be made to work here. Along with the images, scrape captions, the number of likes and the number of comments for each post (if we had more time, I’d ask you to scrape actual comments as well for additional analysis, but ….).

**Task A.** Use the Google cloud vision (they are better than Azure and AWS in image analytics for sure) to obtain image tags for each post. You can create an account with Google, and the first $300 are free, which should be more than plenty for this assignment.

**Task B.** Create a metric (score) for **engagement** by using a weighted sum of #\_likes and #\_comments. Be sure to normalize #\_likes and #\_comments. Now create an engagement score = .4\*#\_likes (normalized) + .6\*#\_comments (normalized).

Build a model to predict engagement with image labels (text) as predictors. Is this model better than using captions to predict the same? What if you used both image labels and captions to predict engagement?

**Task C.** Perform topic modeling (LDA) on the original image labels. Choose an appropriate number of topics. You may want to start with 5 topics, but adjust the number up or down depending on the word distributions you get. **Decide on a suitable name for each topic.**

Now sort the data from high to low engagement score, and take the highest and the lowest quartiles (by engagement score). What are the main differences in the **average** topic weights of images across the two quartiles (e.g., greater weight of some topics in the highest versus lowest engagement quartiles)? Show the main results in a table.

Task D. What advice would you give Natgeo to increase engagement on its Instagram page based on your findings in Tasks B and C?

**Deliverables:** A word or pdf file with tables and answers. Write the names of all team members **inside** the document.