**UGCA F2018 Assignment 4 (Group work)**

**Date Handed out: 5th November 2018**

**Date Due: 19th November 2018 by 11:59 p.m.**

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

On Instagram, choose the National Geographic (natgeo) page (do not use hashtags). Write a scraper or use the Web Scraper to extract (i) image URLs (do not extract video URLs, it may end up costing you a lot of money to run analytics on video), (ii) post caption (the text description of a post), (iii) # likes and (iv) # comments. You don’t need actual comments for this assignment. Scrape 250-500 image posts.

Using the image URLs, obtain **image labels** from Google Vision cloud (you will have to create an account with Google to get your credentials as a json file, though the first $300 are free, which should be more than plenty for this assignment). You will need to write a script to access the Google Vision API. You can also use IBM Watson Vision Analytics (the basic account is free) as an alternative (if you are not a Google fan). Do not use Microsoft image analytics, which I have demonstrated in class to be inferior to Google Vision. Read about Google Vision here: For Google Vision API, look here: <https://cloud.google.com/vision/docs/quickstart>

Task A. Create a metric for **engagement** by using a weighted sum of # likes and # comments. However, first normalize # likes and # comments such that they both have values between 0 and 1. You can scale the # likes by dividing by the maximum # likes (for a post) in your data and do the same for # comments, so that # likes and comments will be in the range [0,1]. Now create an engagement score = .4\*# likes (normalized) + .6\*# comments (normalized). Define High (1) and Low (0) engagement based on whether the engagement score is above or below the median value.

Task B. Run a logistic regression with Engagement (binary) as the dependent variable, and the image labels as independent variables. What is the accuracy (show the confusion matrix)?

What accuracy do you get by using the post caption words as the independent variables instead of image labels? Finally, what accuracy do you get by combining the image labels and post captions and using them as independent variables? What can you conclude from your analysis?

Note: Doing a word frequency analysis and word replacement on the image labels as well as captions will increase the accuracy of prediction. Needless to say, TF-IDF scores should be used.

Task C. Perform topic modeling (LDA) on the image labels. Choose an appropriate number of topics. You may want to start with 5, but adjust the number up or down depending on the word distributions you get. LDA should produce two outputs: (i) A file showing which words load on which topics, and (ii) a file showing topic weights for each image.

Now take the quartiles with highest and lowest engagement scores. What are the differences in the average topic weights of pictures across the two quartiles (e.g., greater proportion of some topics in highest engagement quartile)? Show the main results in a table.

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

**Deliverables:** Create a Python notebook with all code. Also submit all data files. Write the names of all team members **inside** the notebook.