Social Media Analytics – Engagement Prediction and recommendation

Business Scenario :

This project is aimed at increasing engagement on National Geographic's Instagram page (Instagram: natgeo). The final outcome is to advise Natgeo regarding the type of content they should post more and less of – i.e., what types of images increase engagement? What types decrease engagement?

Methodology :

1. Scrape 500-1000 images from the natgeo Instagram page using scrapers available online. Along with the images, scrape captions, the number of likes and the number of comments for each post.
2. Image labels for 1500 pictures were collected using Google Cloud vision API.
3. 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).
4. Build a model to predict engagement with
    a. Image labels (text) as predictors.
    b. using captions to predict the same?
    c. Using both image labels and Captions as predictors
5. Perform topic modeling (LDA) on the original image labels and check engagement score per topic
6. Recommendations to NatGeo

Task A : Scrape Instagram

The package 'instaloader' was used to scrape captions, number of likes and number of comments from Instagram. The code to perform this operation is provided in the code snippet attached in this repository.

Task B : Use the Google cloud vision to obtain image tags for each post.

Image labels for 1500 pictures were collected using Google Cloud vision API. The code to perform this operation is provided in the code snippet attached in this repository.

Task C : Metric creation

#_likes and #_comments were obtained by scraping Instagram page of NatGeo.

Metrics #_likes and #_comments were normalized by dividing by max(#_likes) and max(#_comments) respectively.

The metric for **engagement** by using the normalized #_likes and #_comments was created

engagement score = .4*#_likes (normalized) + .6*#_comments (normalized).

The rationale for weights is that comments are valued more than likes and is also an indication of better engagement than likes.

Task D : Model to predict engagement with image labels (text) as predictors

Accuracy of the model :
0.7479838709677419

Confusion Matrix :

	Predicted 0	Predicted 1
Actual 0	189	50
Actual 1	75	182

Task E : Model to predict engagement using captions as predictors

Accuracy of the model :
0.7419354838709677

Confusion Matrix :

	Predicted 0	Predicted 1
Actual 0	195	55
Actual 1	73	173

Hence,
Model using image labels gives a better accuracy than the model using captions.

Task F : Model to predict engagement using both image labels and captions as predictors

Accuracy of the model :
0.7661290322580645

Confusion Matrix :

	Predicted 0	Predicted 1
Actual 0	187	66
Actual 1	50	193

It is observed that a model that uses both image labels and captions better predicts engagement than a model with either captions alone or just image labels alone.

Task E : Perform topic modeling (LDA) on the original image labels.

The topics identified using topic modelling were :

Topic 0 : Wildlife

Words with the highest weights in topic 0 :

wildlife

animal

mammal

terrestrial

Vertebrate

Topic 1 : Aerial landscape

Words with the highest weights in topic 1 :

sky

landscape

natural

phenomenon

mountain

Tree

Topic 2 : Water

Words with the highest weights in topic 2:

water

sky

sea

architecture

Turtle

Topic 3 : People

Words with the highest weights in topic 3:

photography

monochrome

black

people

adaptation

Task F : Which topic has the highest engagement

Engagement score was bucketized based on quantiles as follows :

Quartile 1 : <25 percentile

Quartile 2 : Between 25 percentile to 75 percentile

Quartile 3 : >75 percentile

The below picture indicates the average topic weights across each of the quantiles.

quartile	engagement_score	topic_0	topic_1	topic_2	topic_3
1	0.0410753	0.0439492	0.165754	0.189811	0.600486
2	0.0827797	0.193854	0.333625	0.223958	0.248563
3	0.193955	0.487746	0.25986	0.168091	0.0843033

The above diagram indicates that Topic_0 (Wildlife) has the highest topic weight in Quartile 3, which indicates the top quantile of engagement score.

Similarly, Topic_3 (People) has the largest topic weight in Quartile 1, which indicates the lowest quantile of engagement score.

Task G : Recommendations to NatGeo

It is observed that Topic_0 (Wildlife) has the highest topic weight in Quartile 3, which indicates the top quantile of engagement score. This is an indication that “Wildlife” related pictures drive more engagement on Instagram for NatGeo.

Similarly, it is also observed that Topic_3 (People) has the largest topic weight in Quartile 1, which indicates the lowest quantile of engagement score. This is an indication that “People” related pictures drive the least engagement on Instagram for NatGeo.

Hence, NatGeo must ensure that they post tons of “Wildlife” related pictures that has many animals in them. NatGeo must also make efforts to avoid “People” related pictures that are monochrome in nature. These strategies will help NatGeo improve engagement on their platform.