**INSY 670**

**Yichen Wang 260761601**

**You will have to scrape 500 images from the company’s Instagram page using the Insta\_download.py script. The script allows you to scrape captions of the number of comments for each post, along with the images.**

The company’s Instagram I chose is the Discovery Channel, which is an American multinational mass media factual television conglomerate based in New York City.

**Task A. Perform topic modeling (LDA) on the image labels from Google Vision. Choose an appropriate number of topics. Describe the process of finding the best number of topics in detail, report the top 25 words for each topic, and decide on suitable names for each topic.**

The process of finding the best number of topics based on two approaches. First is to check whether each topic intersects with each other. If there’s such intersection, we may need to change the number of topics. Another criterion is to see the top labels of each topic. If labels within topics are too repetitive, then we may need to reduce the number of topics. After testing and checking from the diagram, I came up with the most appropriate number of topics to be 4.

Chart

Description automatically generated

For the top 25 words within each topic, it is shown below:

for cluster 0 , the top 25 labels are:

[('Plant', 0.038543392), ('Natural', 0.038082536), ('Terrestrial', 0.03143545), ('Carnivore', 0.024615074), ('animal', 0.024532482), ('landscape', 0.024078459), ('Sky', 0.022479665), ('tiger', 0.021312715), ('Fawn', 0.017994486), ('Felidae', 0.017057633), ('plant', 0.015039114), ('Cloud', 0.015026487), ('Organism', 0.014766953), ('Water', 0.014248578), ('Grass', 0.013465532), ('Nature', 0.013107425), ('Whiskers', 0.012780199), ('Tree', 0.011877354), ('environment', 0.011328196), ('Bengal', 0.010935523), ('Siberian', 0.010854695), ('Tiger', 0.010739982), ('Ecoregion', 0.010375121), ('cats', 0.010147236), ('Vertebrate', 0.00958678)]

-------------------------------------------------------------------

for cluster 1 , the top 25 labels are:

[('Water', 0.03279741), ('Marine', 0.023532387), ('Organism', 0.019021414), ('Underwater', 0.018699924), ('Plant', 0.017474327), ('biology', 0.01701429), ('Terrestrial', 0.016221559), ('Natural', 0.014183614), ('plant', 0.013678569), ('Sky', 0.013509034), ('Fluid', 0.013358776), ('landforms', 0.013257207), ('of', 0.013203066), ('Bird', 0.012003221), ('Liquid', 0.011622954), ('Beak', 0.011303527), ('blue', 0.0111320065), ('Electric', 0.01094785), ('Wing', 0.010931505), ('Fin', 0.010834294), ('Feather', 0.010131557), ('Event', 0.009831716), ('animal', 0.009668235), ('landscape', 0.008608635), ('Vertebrate', 0.008473522)]

-------------------------------------------------------------------

for cluster 2 , the top 25 labels are:

[('Water', 0.031236868), ('Sky', 0.030940412), ('Natural', 0.02264504), ('Terrestrial', 0.015147776), ('landscape', 0.01503523), ('Cloud', 0.014538403), ('animal', 0.012735454), ('Mountain', 0.012503015), ('Bird', 0.009826946), ('Nature', 0.009820403), ('Ecoregion', 0.009671515), ('Dog', 0.0093991775), ('Carnivore', 0.009389588), ('Eye', 0.008646824), ('Blue', 0.008259303), ('Highland', 0.0081264), ('Lake', 0.008086761), ('Atmosphere', 0.007900251), ('Plant', 0.007867926), ('Wood', 0.0075501353), ('Whiskers', 0.0074472516), ('cats', 0.0073379846), ('Head', 0.0071925856), ('Liquid', 0.0069393357), ('Light', 0.006792024)]

-------------------------------------------------------------------

for cluster 3 , the top 25 labels are:

[('Water', 0.04111455), ('Sky', 0.025487406), ('Natural', 0.02493806), ('landscape', 0.019753853), ('Plant', 0.01810372), ('Automotive', 0.015906129), ('Art', 0.013101616), ('Liquid', 0.012704737), ('Marine', 0.012603915), ('Cloud', 0.011925615), ('Fluid', 0.010832307), ('Landscape', 0.010566318), ('phenomenon', 0.010080457), ('resources', 0.009568168), ('and', 0.009389439), ('Nature', 0.0092626205), ('Tree', 0.009067205), ('Organism', 0.009051102), ('biology', 0.008371516), ('blue', 0.007959557), ('Insect', 0.0077169742), ('object', 0.007567145), ('Astronomical', 0.0074553327), ('Electric', 0.0073679285), ('Vehicle', 0.007146427)]

**According to the top words within each topic, names are given.**

cluster 0: Natural Environment

cluster 1: Ocean

cluster 2: Mountain

Cluster 3: Human & Earth

**Now sort the data from high to low engagement (# of comments), and take the highest and the lowest quartiles. 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.**

topic weights for high quartile: 0.2172023823200646 0.4119819876248554 0.13477917975288325 0.23221000826964944

topic weights for low quartile: 0.2555057130053512 0.4344980562599296 0.14954822529584932 0.1899228637289797

the result is organized into a table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Natural Environment | Ocean | Mountain | Human & Earth |
| High Quartile | 0.2172023823200 | 0.4119819876248 | 0.13477917975288 | 0.23221000826964 |
| Low Quartile | 0.2555057130053 | 0.4344980562599 | 0.14954822529584 | 0.18992286372897 |

From the result we could see that those posts with high interactions differ from those within low quartile in 2 topics: natural environments and Human & Earth. For posts in high quartile, more posts within Human & Earth are posted, where for posts in low quartile, more in natural environment are posted. But overall, the difference is small, and require more explanatory factors to explain the difference in the number of interactions.

**Task B. What advice would you give the company if it wants to increase engagement on its Instagram page based on your findings?**

According to the table above, there isn’t such obvious differences between topics for posts in high and low quartiles. Therefore, the problem may not be the topic of the post, but the quality of the post instead. However, according to the result, discovery inc. could try to post more related to Human & Earth related posts, such as technology and arts.