MIDAS@IIITD

Multimodal Digital Media Analysis Lab

# MIDAS@IIITD Internship/RA Task 2020

## GENERAL INSTRUCTIONS

Failing to follow any of the instructions below will lead to rejection of your submission.

1. This task has 1 problem with 5 parts. All the Parts are mandatory.
2. Submission would Include link to the public github repository, link to the Heroku application and a link to the automated check endpoint.
3. Solution of Part I,II and III should be a Jupyter Notebook. Our requirements from each notebook are mentioned in respective questions. These should be present inside the Github repository.
4. Make sure your code is properly documented. We recommend the following:-
    a. Before each code block have a markdown block/docstring which mentions the following
        i. What the code block is doing.
        ii. What is your intuition behind doing this? Why do you think it is useful?
    b. Keep an experiment log - document everything that worked or failed. This document(preferably a jupyter notebook) should be a snapshot of the process you follow to solve each problem.
5. Github repository should include a requirement and README file which can be used to reproduce your development environment and code.
6. You have **20 days (Midnight, 26th April IST)** to submit the solutions. No extensions will be provided.
7. Google Form submission link - https://forms.gle/s7h68xZf6QNdrbFD6
8. If you have any doubts feel free to email at midas@iiitd.ac.in or hitkuli@iiitd.ac.in

# Problems - Reddit Flare Detection

## Part I - Reddit Data Collection

As a starting step you have to write a script to collect data from r/india. This data would be used in future parts of the problem to build the classifier. You have to yourself decide *how to, how much* and *what all* data to collect.

## Part II - Exploratory Data Analysis (EDA)

Perform EDA on the data collected in Part I. This is helpful for understanding the data you have collected. Explain in detail about the analysis you did, intuition behind doing it, output of the analysis (in terms of graphs or tables), your inference from the output and how it shapes your future system decisions.

## Part III - Building a Flare Detector

Posts in r/india can be corresponding to multiple topics. Each post is tagged for filtering purposes. These tags are called a flares in the reddit world. r/india has flairs like Politics, AskIndia, Science/Technology etc. You have to build a classifier which can predict the flare of a reddit post. Use data collected in Part I as your training and validation data. Report detailed analysis of performance of your classifier. We are looking for answers to questions along the line of *"How well it works?"*, "*What is it good at?"*, "*What is it bad at?"*

## Part IV - Building a Web Application

Build a web application which can be used to predict Flare of a r/india post. Application should have an input field which expects a link to a reddit post from r/india. On submission it should predict the flare of the post.

Web applications should also have an endpoint called `/automated_testing`. This endpoint will be used for testing performance of your classifier. We will send an automated POST request to the end point with a .txt file which contains a link of a r/india post in every line. Response of the request should be a json file in which key is the link to the post and value should be predicted flare.

## Part V - Deployment

Deploy web applications built in Part IV on Heroku.

For regular updates on tasks, results, and other updates from our lab, we recommend you to Like/Follow our social media pages https://www.facebook.com/midasiiitd/, https://twitter.com/midasiiitd, and https://linkedin.com/company/midasiiitd. We also frequently share useful materials/resources and openings worldwide on these pages which could be of great interest to you.