PreCog Recruitment Tasks

Disclaimer: We know that the tasks can be very time consuming and not straight-forward. All of us come across such brick walls (Randy Pausch describes this in his last lecture). We want you to make the best use of your time and try to do each task the best you can, given the time constraints. All the very best! :)

Applicant has to complete the following three tasks:

1. Paper Summary
2. Social Media Data Analysis
3. Data Parsing and Storage

1. Paper Summary

Choose one of the following papers, and summarize the paper in maximum 500 words, while critically analysing the paper, what is unique, what does not work, anything you would do differently. In the summary focus should be on the methodology and insights from the paper, rather than on being unnecessarily descriptive and verbose. Following are the minimum expectations from the submission:

a) Summary of paper
b) valuable contributions of the paper
c) Critique of the paper
d) future work/improvements

Submissions should be maximum 500 words.

List of papers:

1. Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles
2. How Community feedback shapes users behaviour.
3. Signals Matter: Understanding Popularity and Impact of Users on Stack Overflow

2. Social media Data Collection

1. From Twitter, identify the top trending hashtag in New Delhi or Hyderabad.
2. Pick the top hashtag, and collect unique tweets around that hashtag. The number of tweets should be sufficiently large (>=10,000).
3. Draw insights from the dataset collected. Insights should be conveyed using visualisations. Mandatory tasks:

What can we say about the users who tweeted this hashtag?

Can we draw any insights about their user's followers and friends?

Can we comment on the language aspect of the corpus collected?

Please add the graphs / inferences from above analysis in your submission.

4. You get Brownie points (bonus) for going beyond these tasks and creating interesting and insightful visualisation. For Brownie points (bonus), you can also deploy the interactive visualisation in a web application on Heroku or Digital Ocean Droplet with a user interface to move around different visualisations.

3. Data parsing and storage :

Participant will have to carry out the following subtasks :

a. Table Extraction

i. PDFs can be downloaded here .

Test on multiple tables

ii. Each of these is a 1 pager containing 1 or more tables.

iii. Come up with some python 3.6+ code that, given a path to local pdf file location, captures ONLY the tables into and stores them in a mongo database.

b. XMLs to DB and data visualisations

i. StackExchange periodically releases it's data at https://archive.org/download/stackexchange. For this subtask, we took the stackoverflow dump and removed some rows from it. So that the size is small and easier to analyze. Download the data from here. [Optional: once extracted you can verify the files using MD5 Hashes. You just need to run `openssl md5 filename` and compare it with the ones here]

ii. Data dump (once extracted) consists of XMLs. Convert/Parse the XMLs into a mongo database. (Add the script used for conversion in the submission)

iii. Exploratory data analysis on the mongodb collection:

1. Figure out what criteria did we use for subsampling (i.e. what is common factor amongst the whole dataset)
2. Include one meaningful Word Cloud
3. Draw a barplot of the top-10 occuring tags. The bar plot will show the number of questions on the y axis, and the name of each tag on the x axis.
   Hint: You will only need data from Posts.xml for this task
4. The above points are just to get you started. Please try to explore the data and come up with some visualizations and inferences. Surprise us! :)

Submissions:

1. Create a GitHub repo. Ensure that the repo has a README.md file duly explaining the structure of the submission, code structure, reference to the external resources used. The work should be carried out individually, there will be checks for plagiarism.
2. Submissions for Task 1 : Add a PDF report to the repo.
3. Submissions for Task 2 :
   a. Share the JSON dumps of tweets via repo or a Google Drive link.
   b. Submit a .ipynb file, and also a pdf version of the same .ipynb file. All your visualisations/outputs should be a part of the pdf.

   Place the webapp link (brownie points) and Code in the GitHub repo.
4. Submissions for Task 3:
   a. For task a:
      i. Code

    ii. Mongo db collection

b. For task b:

    i. Code

    ii. Mongo db collection

    iii. A report summarizing the exploratory data analysis on the dump.

    iv. Make sure to include your answers in a point-wise format.

Please make your submission here by sharing the link to the GitHub repo. By submitting this task, you are agreeing that you did all the tasks / coding, etc. yourself and did not plagiarise. If found plagiarised, you will be removed from the selection process.
Do not make any commits after the deadline - it will lead to automatic disqualification.