Data Science Publications(Phase1)

# Scraping:

## Things done:

1. For 600 data points, 10 publications are scraped for each point which can be easily extended to all the publications for each point if required.
2. For 600 data points, publication titles, affiliations, cited by and interests of each reporter is collected.
3. 641 user urls scraped

**NB:** With minute tinkering of the script, more publications and their date of publishing along with individual citation number and citedby sites can be obtained wherever applicable.

## Things to be done:

1. There are 1967 emails which are missing in the original DS.
2. There are 2018 Twitter Links which needs to be added to the DS
3. Locations of 56 datapoints is missing
4. Gender of 444 reporters needs to be classified
5. There are 1843 LinkedIn urls that needs to be added to the DS
6. For each reporter the Category needs to be verified and added wherever applicable

# Analysing:

## Things done (which can be extended):

1. Ratio of male/female reporters from each country
2. Maximum keyword counter for each country (from the DS of 600 points)
3. The publication keyword counter for each country
4. Reporters on various domains
5. Top 10 countries getting highly reported from
6. The reporters and their respective countries with highest citations

## Things to be done(or, that can be done):

1. How regularly do reporters report from various countries? Are the reports original or copied from other reporters in the Dataset?
2. Which group of reporters are most active on social media and does their domain and content have anything to do with it? If so, which domain or subset of the domain are they posting in social media.
3. Which country has more craze towards a particular field of reporting so there  
   are more reporters working on that field out of health, technology, environment  
   or others?
4. Is there a disproportionate area of coverage on the basis of – race, institution, country?
5. Is reporting correlated with science funding? Are there any biases?
6. Are there Certain places / institutions/ groups where more researchers are coming up ?  
   Based on - Locations, Nepotism, Ethnicity, etc
7. Are Science Journalists covering from press releases or are they actually involved in going through research papers and talking to individuals?
8. Ratio of reporters for different languages and the pattern they follow in terms of location.

**NB: Please add any relevant points needed for this study to be more concrete and if possible, an approach**

**For Reference:**

|  |  |  |
| --- | --- | --- |
| #Prob | **Problem** | **Possible Approach** |
| 1 | Twitter Links | Using Maltego or similar software |
| 2 | Gender Classification | Manual work as many have unisex names |
| 3 | Category Classification | Manual / NLP if there is a versatile corpus |
| 4 | How regularly do reporters report from various countries? | Get individual publications of each reporter with timestamp and measure the frequency on a scale |
| 5 | Are the reports original or copied from other reporters in the Dataset? | A possible approach is clustering and matching the clusters for exact duplicates |
| 6 | Which group of reporters are most active on social media and does their domain and content have anything to do with it? If so, which domain or subset of the domain are they posting in social media. | Manual Work |
| 10 | Are there Certain places / institutions/ groups where more researchers are coming up ? | Plot a graph of Affiliations vs Researcher count |
| 11 | Are Science Journalists covering from press releases or are they actually involved in going through research papers and talking to individuals? | We can refer citations of reporters and possibly answer this question |
|  |  |  |
|  |  |  |

Suggested Workflow:

