Data Science Assignment 1

Proposal A -

Q1.

a.

This proposal deals with the collection of the dataset of UFO sightings around the globe for a period of 2010 through 2013. Details such as place of sighting, time of sighting, perceived shape of the UFO, did disparate sources reported the sighting etc. would be collected over the course of pursuit of this proposal. The mode of collection would be Observation as the data collected would be collected by crawling through news feeds to get the data. Requirement for doing this data collection is driven by personal interest to analyze the pattern between sightings and place,time of their occurrence.

b.

Management of data would comprise of archiving the data in useful format for its future and long term use. Also data would be licensed to be used publicly in order for other researchers to deploy the collected data to get more inferences on the same.

The initial curation plan for 'Proposal A' would be driven by following markers :

(I). **Creation of Logical Collection :-**

Collected data would essentially be organized in a tabular form for easy access. To keep the data and its meta-data coupled together, putting them in a file format(preferably JSON format) for long term access seems to be the best plan to manage them.

(II). **Physical Data Handling :-**

Data would be collected by crawling on news feeds and would have to arranged in file format as mentioned in the previous point. Data would essentially be in a tabular format.

(III). **Interoperability :-**

Data collected and arranged in a tabular format would later on be converted into a JSON format. This would increase the interoperability quotient of the collected dataset with JSON format having its wide reach across all development and research domains.

(IV). **Security :-**

Every amount of data that is collected is fetched from public sources and hence the collected data has to be pushed in public domains so that it could act as a source for other research ideas/implementations. Security is no concern in this collection of data.

(V). **Data Ownership :-**

Data has a public ownership and licensed to be pushed in open domain.

(VI). **Metadata Collection Management and access :-**

Other than the basic data, additional data such mode of evidence which would truthify the occurrence of the event is of prime importance. Collection of this meta-data would come inherent with collecting the data But the format of this meta-data would be unorganized and would need re-organization to stabilize it with the converted organized data. Accessing the meta-data would be as easy as accessing the original data because this also would pushed in a JSON file format.

(VII). **Persistence :-**

One of the important reasons of proposing pushing the collected data in JSON format is to optimize its persistence capability. JSON data so collected would not lose any part of the data while transferred from one processing engine to other. Also, the conversion capability of JSON data in contemporary technical world is among the highest.

(VIII). **Discovery :-**

A website visually correlating the identified events based on the collected data would be made public. Everybody would have the access to that link. This would help generating or confirming relations between the identified events visually.

(VIII). **Dissemination :-**

The visual implementation of the collected data would be pushed public on my personal domain. A node application would redirect any access for the implementation to the respective page. Advertisement would be provided over UFO-linked domains. Paying some amount to Google to increase the page rank of the implemented page would help in advertising the implementation.

Q2. Collected data would be in a tabular format essentially. This tabular format could be later transformed into a JSON format. But the basic details collected in this data won't have any unique id attached to it. For the data set to be used efficiently across the domains, a sequential unique id should be pushed in with every unit of data. If the latitude and longitude details of the sightings could be deduced out of the news feeds or tweets, this JSON data could be easily converted to GSON data which could be easily exploited visually using the Google API.

Q3. The data format for meta data is same as that of the data. It would be assembled in a tabular format which if need would be clubbed with the original data which is supposedly going to be pushed as in a JSON format. If needed the meta data could be easily edited corresponding to a unique id against each unit in the tabular format and JSON file could be regenerated with the updated meta data.

Q4. Requirement for doing this data collection is driven by personal interest to analyze the pattern between sightings and place,time of their occurrence. This data collected would be documented in detail sufficient so that it could be reproduced from scratch by anyone. Also , it could be used and extended for further researches.

Proposal B -

Q1.

a. This proposal deals with the collection of data about National Expenditure on Research and Development split sector wise in India. The details such as amount invested per year, sector in which money is invested etc. would prove to be useful to execute this proposal through. The mode of collection for this proposal would be 'Observation' as the data is collected from public government data provided by the Government of India and is available over the Internet. Requirement for this data collection is driven by personal interest of setting up a comparison of R&D investment in India compared to those of other countries for the past 5 years.

b. Management of the collected data is not a concern as the data is available all the time with over the Internet as public data which is published by the Government of India.

The initial curation plan of the Proposal 'B' would be driven by following markers :

(I). **Creation of Logical Collection :-**

Collected data would essentially be organized in a tabular form for easy access. To keep the data and its meta-data coupled together, putting them in a file format(preferably JSON format) for long term access seems to be the best plan to manage them.

(II). **Physical Data Handling :-**

Data would be collected by pulling up from “data.gov.in/” which is India's open public data published

by the Government of India. This data exists in XML, CSV, JSON, JSONP, and ODS format.

(III). **Interoperability :-**

Data collected is arranged in a tabular format would later majorly be used as a JSON format. This would increase the interoperability quotient of the collected dataset with JSON format having its wide reach across all development and research domains. Also with the other possible formats that the data exists in, it could be used and compared with other data that exists in other formats as well.

(IV). **Security :-**

Every amount of data that is collected is fetched from public sources and hence the collected data has to be pushed in public domains so that it could act as a source for other research ideas/implementations. Security is no concern in this collection of data.

(V). **Data Ownership :-**

Indian government has the ownership of the data. All the updates to data would be driven by Indian Government. Ideally it is legally valid to display the data in different rendered formats as it holds a public license. Ownership would not migrate to any other holder.

(VI). **Metadata Collection Management and access :-**

There exists a very blur line between data and metadata in this proposal because the data that is to be collected would be self-explanatory and would give a blur set of meta-data for the data. For more meta-data if needed would be added externally by relating the data with other existing data. The only concern in this context is that the data to be related to be should be published by Government of India. Ideally data would be used in JSON format , which would encapsulate and preserve the meta-data for the collected data.

(VII). **Persistence :-**

One of the important reasons of proposing pushing the collected data in JSON format is to optimize its persistence capability. JSON data so collected would not lose any part of the data while transferred from one processing engine to other. Also, the conversion capability of JSON data in contemporary technical world is among the highest.

(VIII). **Discovery :-**

A visual time-line displaying the R&D investments sector wise in India for the last 5 years. This gives clear picture of what amount of investment needs to be provided per sector to meet the international average investments in R&D so that India can survive in this contemporary world where in every country is driven by its Research quotient.

Q2. The data to be collected already exists in the JSON format. Almost all the required data and meta-data are encapsulated in the collected data. It is already in suitable format. The collected data would consist of the sector name, year of investment and amount of investment done in research & development by India for the last five years.

Q3. Some of the meta data is the integral part of the collected data thus would go along with the data. If needed, more data for the meta data would be collected. The challenge that lies here is to convert the newly collected meta data to the existing meta data format which is coupled with the data. For this result, the newly collected meta data would be pushed in a rough tabular form initially so that it could be easily transformed to a JSON format.

Q4. Requirement for this data collection is driven by personal interest of setting up a comparison of R&D investment in India compared to those of other countries for the past 5 years. This data collected would be documented in detail sufficient so that it could be reproduced from scratch by anyone. Also , it could be used and extended for further researches. End result of the analysis of the collected data would give a clear picture of what amount of investment needs to be provided per sector to meet the

international average investments in R&D so that India can survive through this contemporary competitive world where in every country is driven by its Research quotient.