Project 1

Final Report

What are the most common words on the internet?

**Noble prize**:

The Nobel Prizes are a series of international awards given each year to recognize achievements in physics, chemistry, physiology or medicine, literature, and peace. In 1895, Alfred Nobel, a Swedish innovator, and businessman established the Nobel Prize. The Nobel Prizes are widely regarded as the highest honors bestowed upon individuals in their respective fields of study. What is a few of the words that are commonly searched on the internet? Noble Prizes are widely spread in different fields like physics, chemistry, physiology or medicine, literature, and peace. I’m determined to do project 1 on ‘The most common words on the internet on Noble Prize’.

**Analysis Steps**:

* Choose a topic
* Search for data sources (3)
* Extract the most common words
* Compare the results and see if there is any similarity present in the results

In this fast-moving world where every tiny detail is being recorded, a lot of data is available on a lot of topics. A lot of various and interesting ideas were popping up. For example, Games, Movie actors, Technologies, Facebook marketplace, social media, and many more. After giving a thorough thought, Nobel Prize suited best for this project.

There were various factors involved in choosing this topic. For example, whether the data available contained the information required to do the fundamental analysis, whether are there any multiple data sources available to confirm the results, do I have the basic knowledge about the topic at hand to continue working, etc. I gave ample amount of time to these fundamental questions and decided to go ahead with Noble Prize.

Now coming to the data sources, I wanted to go with a Kaggle dataset, a Wikipedia page, and an article or a blog. Finding the Kaggle dataset and Wikipedia page was not difficult. But an article was difficult to find. There were many articles on Noble prizes, but it was specific to that current year. There was no hope of finding an article that discussed generally on noble prizes. I was left with no other option is drop the idea of doing an analysis on a blog/article. I went with a webpage that contained information very much similar to the first two data sources. Once after obtaining all three data sources, a thorough look at the data available was given.

**Data Sources**:

What is a few of the words commonly searched on the internet? Noble Prizes are widely spread in fields like physics, chemistry, physiology or medicine, literature, and peace. I’m determined to do project 1 on ‘The most common words on the internet on Noble Prize’.

**Data Source 1-Kaggle dataset:**

The first dataset is a **Kaggle dataset** for noble “Nobel Prize Winners (1900-2020)”:

Source Link: <https://www.kaggle.com/datasets/rishidamarla/nobel-prize-winners-19002020>

The data present in Kaggle is in CSV format. It consists of various columns which provide information on the noble prize winners’ names, born country, winning year, and born/death day. The Kaggle dataset needs to be uploaded to Google Colab to perform various operations to check the frequency of the most searched words. As Google collab is user-friendly and open source, there are various options to explore the data.

This Kaggle dataset has information about the noble prize winners. For example, their born date, death date (if any), and field in which they won. Based on the data provided, I’m going to perform an analysis using Google Colab. A Few tasks are as follows:

* On which country has the most prize holders (Country frequency),
* Common words used in their titles, etc..,.
* Plotting a histogram using the age at which they received

**Screen Shots**:

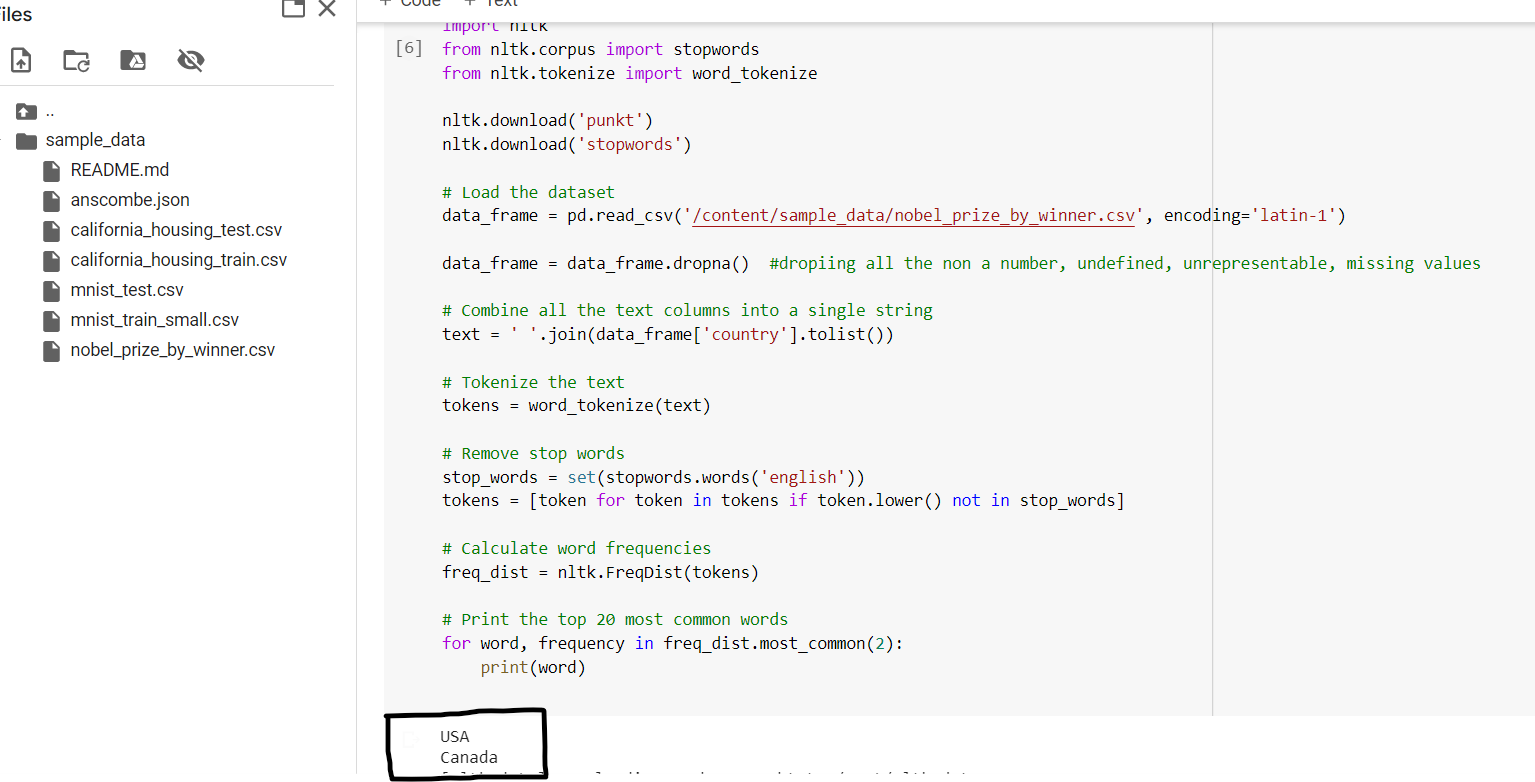


Figure.1.1

Here, above code snippet gives the top two countries which hold the highest number of noble prizes. For this, I'm using the column Country to extract the information. As the USA holds the highest number of noble prizes from the given data followed by Canada, the output displays the same as highlighted in the snapshot.

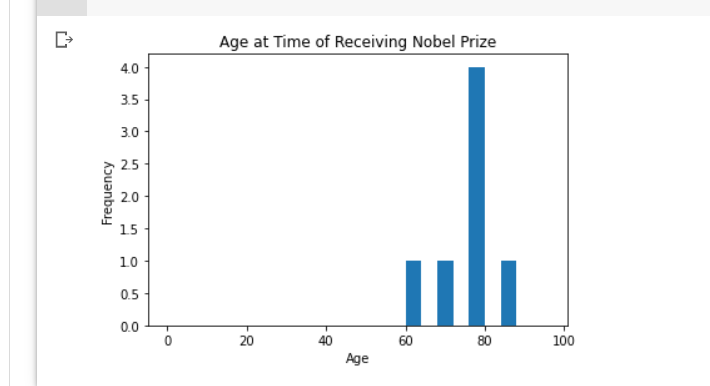


Figure.1.3

As shown in the figure 1.3, a histogram is plotted to extract the data at what age they received the award.

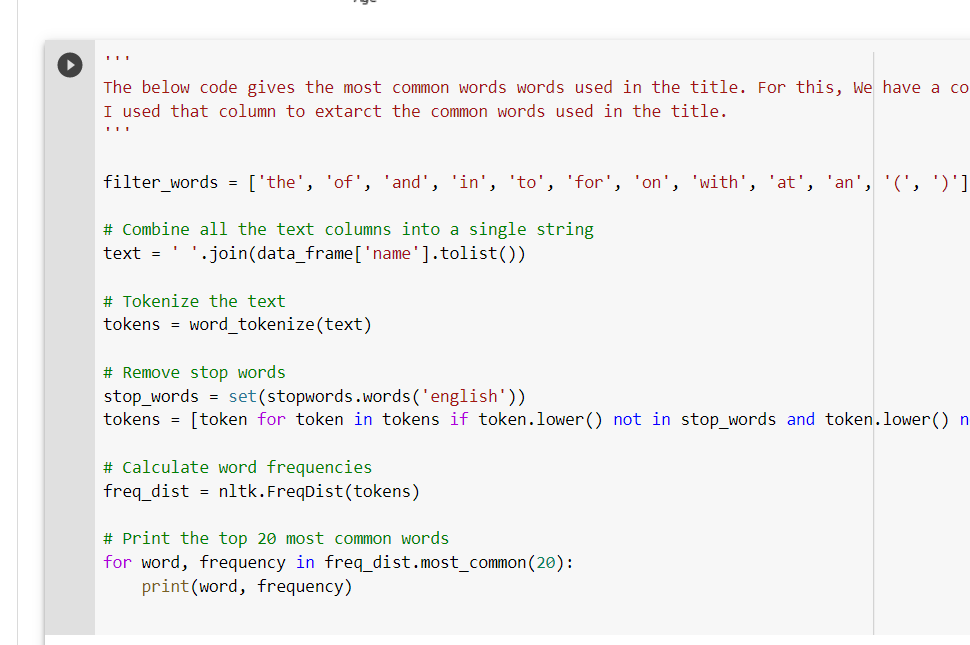
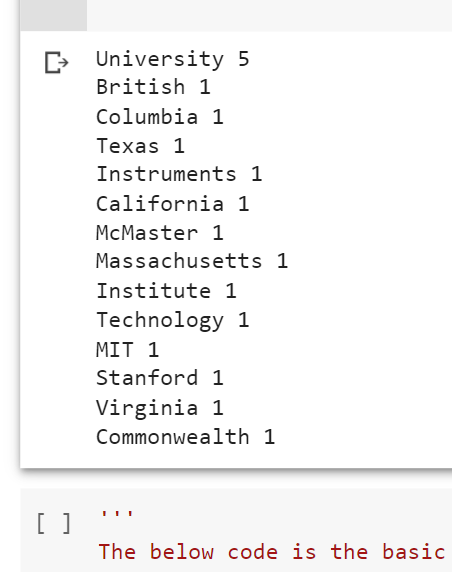
 

Figure.1.4 Figure.1.5

Figures 1.4 and 1.5 give the data about the most common words used in the title. First, all the filtered words like for, and many more needs to be excluded in this extraction. Then use the “Name” column to find the most common words used in the title. As shown in figure 1.5, these are a few words used in the title by many scientists and researchers.

**Data Source 2-Wikipedia Page:**

A **Wikipedia page** on “List of Nobel laureates by country”

Source: <https://en.wikipedia.org/wiki/List_of_Nobel_laureates_by_country>

A list of Nobel laureates by country Wikipedia which has information on which country holds the noble awards to the exact number. Wikipedia holds a lot of information on which country has how many Nobel prize winners. It also holds information about the full names, categories/fields in which they won. It also holds a few references to other websites which have similar information.

Again, this page holds information about noble prize winners country-wise. For this data source, I’m going to be using web scraping. Performing web scraping is going to be a bit of a challenge as I’ve never done it before. I will be using beautifulsoup4, bs4, and requests libraries. Using these libraries, I will extract the basic data from the website by writing the code. For example, getting the country details which holds the highest winners in general, and field-wise.

**Screen Shots:**

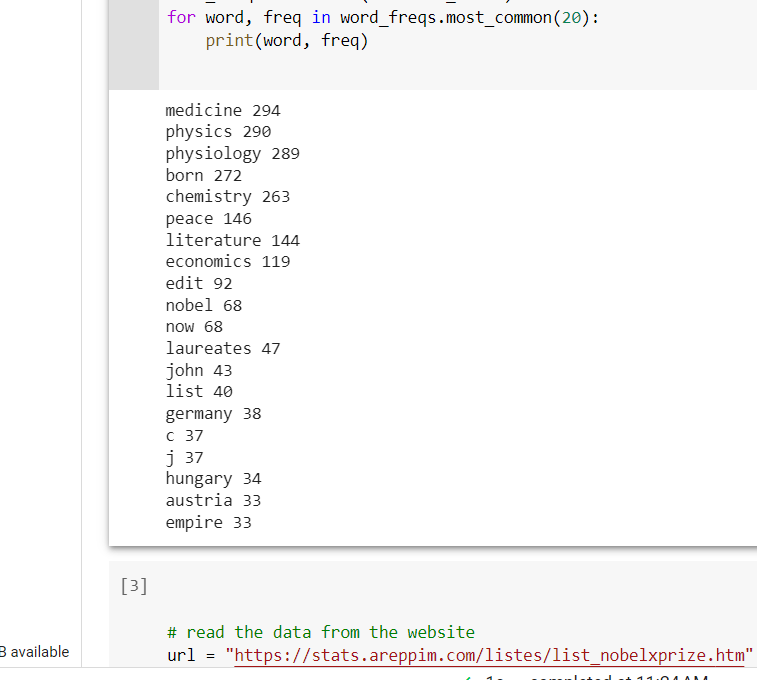


Figure.2.1

From experiment 2, we can confirm the results from experiment 1 by extracting the common words used the Wikipedia page about Noble Laurates. From experiment 1, where extraction was done on the Kaggle dataset to get the top two countries holding the Nobel prizes, the outputs were compared, as it confirmed that the USA in fact has the record of holding the highest number of Nobel laureates. A large number of awards have been given to the area of physics as a result of the numerous notable discoveries and scientific advances that have been made by its scholars. Hence proving the result again that the field of Physics has the highest number of nobel prizes.

**Data Source 3-Website:**

Web scraping the **website** containing information list on noble prizes by country.

Source: [https://stats.areppim.com/listes/list\_nobelxprize.htm](https://stats.areppim.com/listes/list_nobelxprize.htm%20)

This data source has information about noble laureates. This information is categorized by the field of study, the year in which they won the award, born date, and name

Data source 3 is another website that contains data similar to data source 2. This website has detailed information and additional data. It mentions gender and it is classified according to the field of study along with citizenship. I will perform web scraping on this website to check the following:

* Who holds the highest number of awards in that particular field, Men/Women
* It also has information on Affiliation at the time of the award. Based on this information, I will try to extract the data about which university has the highest affiliation when they received the prize

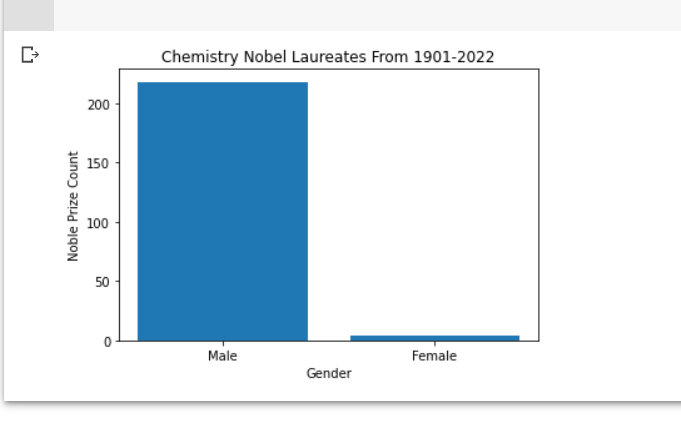


Figure.3.1

For the third experiment, a webpage is used which holds the noble prize winners. The special thing about this website it holds the gender information as well. Here, a bar graph is plotted which compares the male-female count from 1901 to 2022 in the field of Chemistry.

**Big Take Aways:**

Data is everything. Without data, any kind of analysis is impossible. We recognize the work people have done and try to be better in the future only if it is stored somewhere. Otherwise, it would just be a memory.

This project was quite challenging. At every point, there was a new thing to learn. A lot of research was put in for every experiment. I gained valuable hands-on experience using Google Colab, a cloud-based platform for data analysis that I had not previously worked with. I learned how to analyze and visualize data from various sources, and the importance of data cleaning and preparation.

As we have data on the Noble prizes, the field of work, and the amount of effort they put into it, every single detail is considered valuable, and this level of analysis was possible. Data can be stored in any form. For example, books, files, tables, Wikipedia pages, blogs, articles, and many more.

**Conclusion**:

I began by employing web scraping methods to gather information from various sources, such as news stories and Wikipedia pages. In order to better comprehend the data, preprocessed it by removing stop words, punctuation, and special characters. Then performed exploratory data analysis.

To determine the most frequently occurring terms associated with the Nobel Prize, I used a variety of data visualization approaches, such as word clouds and bar charts. According to the analysis, the most often occurring words are "prize" "Nobel”, "winners", "physics", and many more.

I now have a better understanding of web scraping, data pre-processing, and data analysis methodologies overall thanks to this research.