**CPSC 6780-01\_24FA**

**Big Data Proc & Analytics**

**Fall 2024**

**PROJECT**

**“Analysis of Nobel Prize Winners”**

**Student Name: Kodanda Ramu Yanamalachintala**

Advanced Data-Driven Analysis of Nobel Prize Winners’ Demographic and Institutional Trends

**Problem**

This project aims to conduct a **comprehensive analysis of the Nobel Prize winners’ dataset** with a focus on identifying and quantifying factors associated with Nobel success across multiple dimensions, such as country affiliation, age demographics, funding sources, and institutional connections. Unlike general demographic studies, this project will systematically explore **"common pathways to Nobel success"** by integrating **Natural Language Processing (NLP)**, **time-series analysis**, and **clustering**. This includes investigating if specific awards, institutions, or grants are frequently associated with Nobel laureates, and if such associations have changed over time.

**Key Research Questions:**

1. **What countries have the highest Nobel laureates per capita, and how do these rates change over time?**
2. **What are the age trends among Nobel laureates, and do specific categories show distinct age patterns?**
3. **Which awards, funding sources, or institutions are frequently associated with future Nobel laureates, and do these associations correlate with higher chances of winning the prize?**

**Tasks to Finish**

1. **Data Collection and Integration**:
   * **Primary Dataset**: Download the Nobel Prize winners dataset from Kaggle.
   * **Supplementary Data**: Integrate additional data sources, such as historical population data (e.g., from the World Bank or UN) for per capita calculations, and databases or lists of prominent awards and research grants.
2. **Data Cleaning and Preprocessing**:
   * **Standardization**: Clean and format data, addressing missing values, normalizing country names, and standardizing date formats.
   * **Text Preprocessing for NLP**:
     + Extract and preprocess biographical descriptions to prepare for NLP tasks.
     + Apply tokenization, stopword removal, and lemmatization using **SpaCy** or **NLTK** to ensure text is ready for entity recognition.

**Exploratory Data Analysis (EDA)**:

* + **Country Analysis**: Visualize country-wise distribution of laureates and identify top-performing countries.
  + **Age Trends**: Calculate the age at award receipt, analyze category-wise distributions, and observe changes in the age profile over decades.
  + **Category Patterns**: Conduct an initial analysis of prize distribution across categories, noting any significant changes in category prominence.

**Feature Engineering**:

* + **Nobel Laureates per Capita**: Calculate per capita laureate counts for each country by merging historical population data.
  + **Age at Award**: Generate an age field from birthdate and award year.
  + **Funding and Affiliation Mentions**: Use Named Entity Recognition (NER) to extract mentions of awards, institutions, and funding sources in laureates' biographies.
  + **Institutional Affiliation Score**: Develop a scoring metric to assess “institutional prominence” by counting repeated mentions of institutions that have produced multiple laureates.

1. **Advanced Analysis and Visualization**:
   * **Time-Series Analysis**: Use time-series plots to examine how laureate demographics, such as age and country representation, change over time. This can reveal any generational shifts in Nobel recognition.
   * **Named Entity Recognition (NER) for Awards and Institutions**:
     + Use **SpaCy's Named Entity Recognition (NER)** to extract entities like “MacArthur Fellowship,” “Harvard University,” “Howard Hughes Medical Institute,” and “ERC Grant” from biographical texts.
     + Analyze the frequency of each award, institution, or funding source to identify common pathways.
   * **Clustering Analysis with Scikit-Learn**:
     + Apply **K-means clustering** or **DBSCAN** to group countries based on laureate density, diversity trends, and field prominence.
     + Develop cluster profiles to identify patterns among countries or institutions that frequently produce Nobel laureates.
2. **Evaluation and Conclusion**:
   * **Visual Summaries**: Present findings through bar charts, heatmaps, and time-series plots to illustrate country-wise achievements, age trends, and category patterns.
   * **Descriptive Statistics**: Provide statistical summaries (e.g., average age by category, laureates per capita) for clear insight into demographic and category-based findings.
   * **Comparative Analysis**: Benchmark findings against peer-reviewed studies on Nobel demographics, emphasizing new insights or deviations.

**Tech/Tools/Systems**

* **Programming Language**: Python
* **Data Handling**: **Pandas** and **NumPy** for efficient data processing and manipulation.
* **Data Visualization**:
  + **Matplotlib** and **Seaborn** for high-quality static visualizations.
  + **Plotly** for interactive, drill-down visualizations, if required.
* **Natural Language Processing (NLP)**:
  + **SpaCy** or **NLTK** for entity recognition and text processing.
  + **NER Models** for identifying mentions of awards, institutions, and grants in the biographies.
* **Clustering and Predictive Modeling**: **Scikit-Learn** for clustering analysis and potential predictive modeling.
* **Platform**: Jupyter Notebook for interactive coding, analysis, and documentation.
* **Data Sources**:
  + Kaggle’s Nobel Prize dataset.
  + Historical population data (e.g., World Bank).
  + Supplementary lists or databases of awards and research grants, if available.

**Implementation / Evaluation Plan (Justification)**

This project will proceed in a structured, modular format:

1. **Data Collection and Cleaning**:
   * A clean and standardized dataset ensures high-quality analysis. Integrating additional data sources like population data and award information enriches the dataset, providing more context for per capita calculations and institutional analysis.
2. **Feature Engineering for Institutional and Award Analysis**:
   * Generating features such as “Nobel Laureates per Capita” and “Institutional Affiliation Score” will add depth to the analysis, offering insight into common institutional pathways.
   * Using **NER for Award and Institution Recognition** allows us to pinpoint common affiliations and funding sources, setting up potential pathways to Nobel success.
3. **Trend Analysis and Clustering**:
   * Trend analysis will help in detecting time-based patterns, while clustering enables identification of groups (e.g., high-output countries or prominent institutions) with similar characteristics, providing a deeper understanding of global patterns.
4. **Conclusion and Comparative Benchmarking**:
   * Findings will be documented with a focus on comparison to existing literature. This includes statistical validation and a visual presentation of unique insights, strengthening the contribution of this project to Nobel demographic studies.

**Timeline and Expected Outcome**

|  |  |
| --- | --- |
| **Week** | **Tasks** |
| **Week 1** | **Data Collection and Cleaning** - Download datasets from Kaggle and other sources. - Perform data cleaning, handling missing values, and standardizing formats. - Merge with supplementary data (e.g., population data). |
| **Week 2** | **Exploratory Data Analysis (EDA)** - Conduct initial analysis and visualizations. - Focus on examining country distributions, age trends, and category patterns to gain preliminary insights. |
| **Week 3** | **Feature Engineering and Advanced Analysis** - Develop advanced features such as per capita laureate counts, age at award, and institutional affiliations. - Use NLP and Named Entity Recognition (NER) to extract notable awards, grants, or institutional mentions. |
| **Week 4** | **Final Analysis, Conclusion, and Documentation** - Perform clustering analysis and finalize visualizations. - Compile the final report with data-backed conclusions. - Benchmark findings against peer-reviewed studies to validate insights. |

1. Bottom of Form

**Expected Outcomes**

1. **Detailed Report and Visualizations**:
   * A comprehensive report that includes visualizations on demographics, such as age trends, country-based analysis, and prominent institutional affiliations.
   * Time-series analysis showing shifts in age and category trends, including any identified patterns linked to historical events.
2. **Significant Findings on Pathways to Nobel Success**:
   * Identification of common “pathways to Nobel success,” such as affiliations with certain institutions or prior receipt of specific awards (e.g., **MacArthur Fellowship**, **Howard Hughes Medical Institute**).
   * A ranked list or network graph of the most common awards, grants, or institutions found in the dataset.
3. **Contributions to Knowledge**:
   * This project will make a unique contribution by identifying specific affiliations or funding sources linked to a higher probability of Nobel recognition, a perspective that is largely unexplored in current literature.
   * Insights on global and demographic shifts over time, offering valuable information on how Nobel recognition patterns have evolved across different fields and demographics.