Age dataset: A structured general-purpose dataset on life, work, and death of 1.22 million distinguished people

Issa Annamoradnejad,[1] **Rahimberdi Annamoradnejad** [2]

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[2] Associate Professor, Urban Planning, University of Mazandaran, Babolsar, Iran
imoradnejad@ce.sharif.edu, r.moradnejad@umz.ac.ir

Abstract

Several fields of study can benefit from a large, structured, and accurate dataset of historical figures. Due to a lack of such a dataset, in this paper, we aim to use machine learning and text mining approaches to collect, predict, and clean online data with a focus on age and gender. We developed a five-step method to use the community-submitted data from all language versions (languages) of the Wikipedia project to infer the birth and death years, binary gender, and occupation of more than 1 million historical figures. The dataset is the largest one on notable deceased people and includes individuals from a variety of social groups, including but not limited to 107k females, 90k researchers, and 124 non-binary individuals, spread across more than 300 contemporary or historical regions. The technical method demonstrates the usability of the latest text mining approaches to accurately clean historical data and reduce the missing values. The final product provides new insights into the demographics of mortality in relation to gender and profession in a historical context.

Introduction

Several fields of study in the social sciences could benefit from a large and accurate dataset of historical figures. This outcome is not specific to datasets about people, and any large general dataset, just like a general predictive model, can speed up scientific discoveries in several research areas. This positive outcome is multiplied for minority and underrepresented groups, such as women and non-binary people.

Previously, there have been a few attempts on aggregating information on historically significant people, many of which focused on a specific society, application, or aspect of life, and included information on a few thousand handpicked individuals. With respect to ease of access, some previous researchers preferred not to publish the collected datasets and the existing open data are provided as RDF triples, a data format that is not suitable for manual preview, analysis, and processing in the current data science projects. Furthermore, previous works collected their data from specific versions of Wikipedia (mostly English Wikipedia), which results in an Anglophone bias towards the dataset.

In this paper, we present a novel dataset that contains information on 1,223,009 historical figures from all walks of life and all over the world. To achieve this goal, we filtered out all human subjects from the open data RDF triples hosted by the Wikimedia Foundation and collected submitted properties. Since the original RDF source contains all entries submitted via all Wikipedia websites, there is less bias in submission compared to previous works. We devised text mining and machine learning models (including BERT language model and XGBoost classifier) to unify values and minimize the number of missing data for age of death, birth year, and death year, gender and occupation properties.

Numerous applications of this dataset for research and entertainment activities are plausible. We prepared the dataset to foster research in several areas of study, including cultural and cross-cultural studies, health-related studies, popularity ranking tasks, and world event analysis.

Method

In this section, we briefly explore methods used in the process. It covers the entire process of dataset creation– data collection, data filtering/selection, data conversion, and post-production. The creation procedure is composed of five major steps, as described in the following sub-sections.

Retrieving base data

The novel dataset is based on values submitted by the community, available as open data[1] and hosted by the Wikimedia Foundation (Wikimedia 2021). We used Wikidata RDF data dump, that is available as a single 96GB+ file in JSON format[2]. Working with such files, due to their size and format, is impossible using common data processing tools, and cumbersome using programming languages.

The Wikidata project, as the original source of this work, contains all entities, human or otherwise, submitted through all Wikipedia projects. Each entity is assigned a unique identifier, such as Q12340927, which resolves problems that could arise due to duplicate full names or possible name changes. RDF triples provide information in subject-predicate-object relationship, meaning that a subject is described through multiple relationships (or predicates) with other objects. We filtered the data to only include relationships on all human subjects. For individuals that hold more

[1] Under Creative Commons (CC) BY-SA licence.
[2] https://dumps.wikimedia.org/wikidatawiki/entities/

than one value for a property (e.g. country), we kept all values, concatenated with a delimiter.

Determining age of death

Birth and death date was available for 86% and 75% of individuals in the original source, respectively. For the remaining people, we noted that short descriptions of around 300k subjects concluded by stating the exact lifespan of the deceased person (e.g. "1452-1509" or "(30 BC-23 AD)"). Based on this idea, we were able to extract birth year and death year from the textual values by using a few regular expressions. Based on the extracted two values, we calculated the age of death for every deceased person that had a valid birth and death year.

Determining gender

In the original source, gender was available for 50% of the people. To predict this property for the rest of the dataset, we trained a binary classification model to detect the gender of given subjects based on their given name, country, and century. While the given name is the most important factor, we included country and era to generate more accurate results (similar to previous efforts on predicting gender based on given names, such as (Wais 2016; Blevins and Mullen 2015)). For this goal, we chose XGBoost classifier, a state-of-the-art tree-based boosting algorithm. The inputs of the model are century (number) and encoded categorical values of given name and country.

For training and evaluation, we utilized labeled part of the dataset, by splitting it into two sections: 67% for training and 33% for evaluation. Based on our evaluation, the trained model achieves 97.51% accuracy and 98.89% F1-score in predicting binary gender. It is important to note that the classifier achieves a high value of F1-score, a metric proper for skewed train datasets. Based on this model, we were able to identify gender for more than 512k additional individuals.

Determining occupation

While occupation property was available for close to 66% of the people in the original data source, values were inconsistent with each other in two regards: first, they included several synonyms (like businessman/businessperson; professor/university professor) and second, different occupation levels were used (like military personnel vs. soldier or athlete vs. footballer). To address the inconsistency problem, we performed a unification step and converted all values to coarse-grained classes using a manually generated mapping function. This resulted in unified values for occupation.

To determine occupation for the rest of the dataset, we created a multiclass text classification model that predicts occupation based on the short description of an individual. This step is possible as the main field of work is usually included in the short description of individuals (e.g., "USA president", "Photographer", and "Ski player" would be classified to "Politician", "Artist" and "Athlete", respectively). Our method uses BERT language model to generate numerical embedding for short descriptions and trains an SVM classifier using these generated vectors. Since this is a multiclass classification task, the model generates a probability for each class (between 0 and 1) which indicates the confidence of the model for that specific class/occupation. We used a threshold for confidence and considered the high-scoring class with more than 0.6 probability as occupation. Our model was able to predict for 84% of the dataset with 93.4% accuracy.

Generating final output

As the final processing step, we converted the dataset into a single tabular file which is provided in CSV (comma-separated values), and JSON formats. The dataset is publicly available at https://dx.doi.org/10.17632/2sfz4tt88g.

Content

This section explores the key aspects of the final product. The dataset contains nine important properties for 1,223,009 notable deceased individuals. The properties that describe the individuals are full name, birth year, country, gender, a short description, occupation of the individual, manner of death (a general method such as suicide, natural causes or capital punishment), death year, and age of death.

There are 1,130,871 unique full names in the dataset (92k repetitions, such as "John Smith"), available for everyone included in the dataset. "John" with 26k, "William" with 17k, and "Robert" with 12k are the most common given names, and "Li" with 7.5k, "Liu" with 4.3k, and "Chen" with 4.1k are the most recurring family names in the dataset. Percentage of women included in the dataset increases steadily from 10% for people born in the eighteenth century to close to 30% in the twentieth century.

The dataset provides the relevant country for 72% of the rows of the dataset, where USA, Germany, UK, France and Russia are the most repeated countries with 153k, 107k, 93k, 78k, and 36k people, respectively. It should be noted that a historical country may have existed in the past and the name of the territory at the time of the individual is used in the dataset (e.g. "Roman Empire"). With regards to occupation, the dataset includes more than 281k artists, 195k politicians, 110k athletes, 90k researchers, 52k military personnel, 37k religious figures, 19k businesspersons, and 215k other less frequent occupations (such as journalists and judges)

Manner of death is available for 53,284 individuals. Natural causes (such as myocardial infarction, cancer, and pneumonia) as the most common reason for death is reported for 62.9% of the available data. Suicide with 10.9%, accidents with 9.4%, and homicide with 8.8% are in the next places.

As seen in Figure 1, number of women included in the dataset are much lower than men (0.11x). Ages between 70 and 90 are the top recurring death ages for all individuals, and the average age for females is relatively higher than males (71.3 y compared to 69.1 y). With respect to occupation, businesspersons (76.0 y), religious figures (75.9 y) and researchers (75.8 y) have the highest average ages, while military personnel (63.8 y), athletes (69.0 y) and journalists (70.6 y) have the lowest averages. The given statistics are for people born after 1870, to produce results merely based on occupation and neutralize the impact of higher life expectancy in recent centuries on relatively new occupations

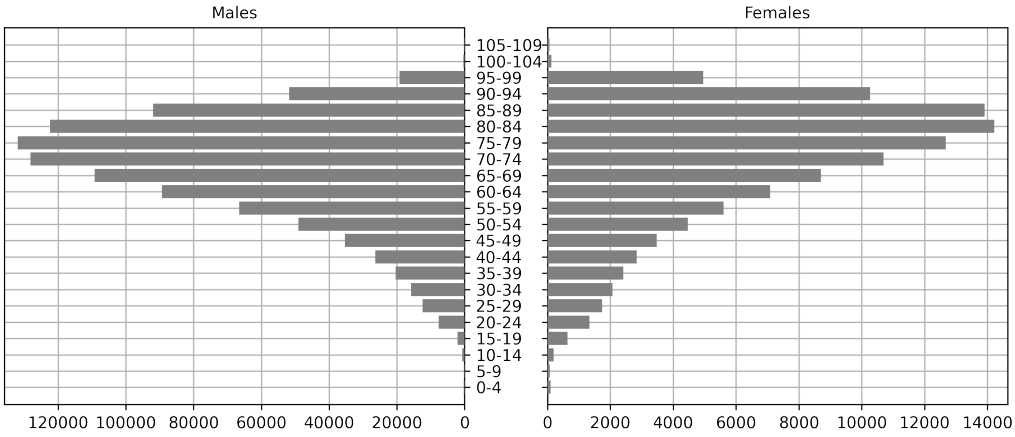

Figure 1: Number of people in the dataset based on gender and age group

(e.g. journalists). Finally, the dataset includes people from all eras with a high peak for people born in 1940s and 1950s.

Discussion

The dataset can be used as the primary or secondary source in performing cultural and cross-cultural studies. There are countless ways to infer statistics for a specific goal by filtering time, location, gender, ... Gender studies can be performed with a focus on the relationships of gender with other properties such as occupation (similar to (Cortes and Pan 2018)) and given name ((To et al. 2020)), based on the time or place.

Several health-related contributions can be attained with relation to the determinants of age and death. More specifically, determining the impact of factors such as gender, occupation, geographical location, weather, and education on the age and reason of death ((Whalley and Deary 2001)). It is also possible to study the impact of properties such as occupation and gender on specific reasons of death, such as suicide probability, alongside longitudinal studies (similar to (Bando et al. 2012; Bridge et al. 2020)).

Figures 2 and 3 are some examples of this usage. Figure 2 displays the average age of males and females in the past millennium. Based on our data, there was a period of sharp increase in the average age of both genders for people born in the fourteenth century. Figure 3 shows a comparison between the average age of males and females for several countries in our dataset. The X-axis and Y-axis represent the average age for males and females, respectively. Canada is shown to have on of the highest averages, both for males and females.

The dataset is subject to a few limitations, such as incorrect data submissions and missing values, common in all works based on the crowd-sourced content. Despite our effort to reduce the number of missing values through predictive models, the dataset contains some blank cells for certain individuals (such as the cause of death). Lost in history, not being reported by the community and disputes over accuracy are among the main reasons for not being present. Regarding bias and distributions, such datasets on famous people represent the distinguished people and not the demographics of general population. Meaning that if the goal of a particular research is to achieve insights on world populations (and not on famous population), the differences between the two (famous vs. all) should be acknowledged or calculated using additional reported statistics. Such research goals should be pursued based on the goal and context of the study.

Conclusions

In this paper, we presented a unified large dataset on the known properties of life and death of more than 1.22 million historically significant people. This work shows the usability of the latest text mining models to accurately clean historical data and reduce the missing values, especially to retrieve accurate and large information on the relation of well-being with other physical and social properties. We unified occupation values in two aspects and reduced the number of missing values to less than 18%. We also predicted gender for 82% of the dataset using a model on first names.

While the focus of the dataset is on the age of people (100% availability), subsets of the dataset can be used as the primary or a secondary source in several areas of study, including cultural and cross-cultural studies, health-related studies, popularity ranking tasks, and world event analysis. Since the data is much larger than the previous data sources used in related studies, the results based on the novel dataset can be more accurate and defendable. In general, we hope and believe that the emergence of this dataset can shed a light on several research paths that have been ignored by researchers for the lack of proper data.

References

Bando, D. H.; Brunoni, A. R.; Fernandes, T. G.; Benseñor, I. M.; and Lotufo, P. A. 2012. Suicide rates and trends in São Paulo, Brazil, according to gender, age and demographic aspects: a joinpoint regression analysis. *Brazilian Journal of Psychiatry* 34(3): 286–293.

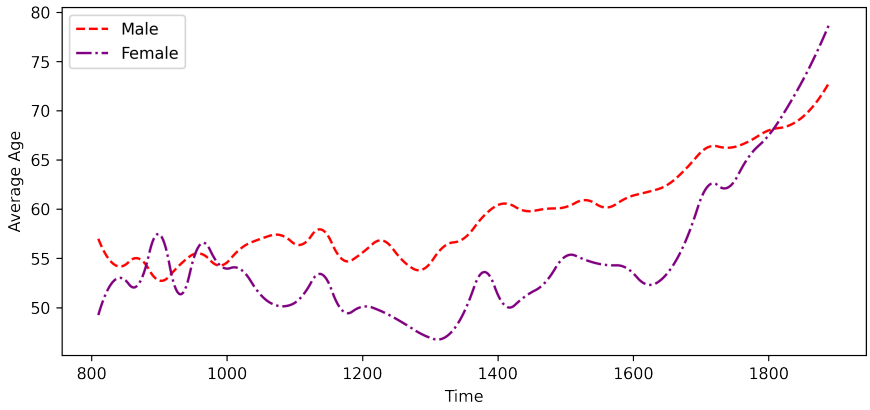

Figure 2: Average age of males and females, since 800AD

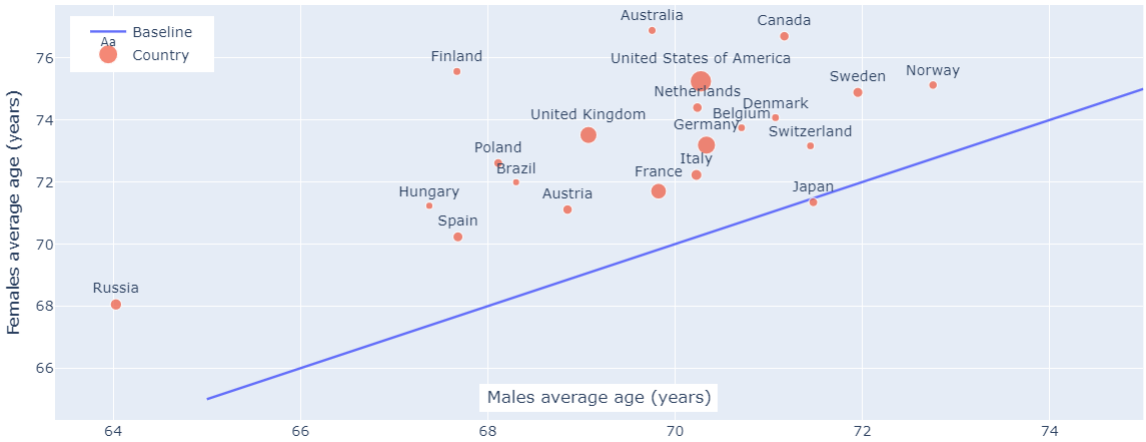

Figure 3: Gap between average age of males and females for most populous countries of the dataset. The size of dots is proportional to the number of famous people from that country.

Blevins, C.; and Mullen, L. 2015. Jane, John... Leslie? A Historical Method for Algorithmic Gender Prediction. *DHQ: Digital Humanities Quarterly* 9(3).

Bridge, J. A.; Greenhouse, J. B.; Ruch, D.; Stevens, J.; Ackerman, J.; Sheftall, A. H.; Horowitz, L. M.; Kelleher, K. J.; and Campo, J. V. 2020. Association between the release of Netflix's 13 Reasons Why and suicide rates in the United States: An interrupted time series analysis. *Journal of the American Academy of Child & Adolescent Psychiatry* 59(2): 236–243.

Cortes, P.; and Pan, J. 2018. Occupation and gender. *The Oxford handbook of women and the economy* 425–452.

To, H. Q.; Van Nguyen, K.; Nguyen, N. L.-T.; and Nguyen, A. G.-T. 2020. Gender Prediction Based on Vietnamese Names with Machine Learning Techniques. *arXiv preprint arXiv:2010.10852* .

Wais, K. 2016. Gender Prediction Methods Based on First Names with genderizeR. *R J.* 8(1): 17.

Whalley, L. J.; and Deary, I. J. 2001. Longitudinal cohort study of childhood IQ and survival up to age 76. *Bmj* 322(7290): 819.

Wikimedia. 2021. Wikimedia Foundation. URL https://wikimediafoundation.org/.