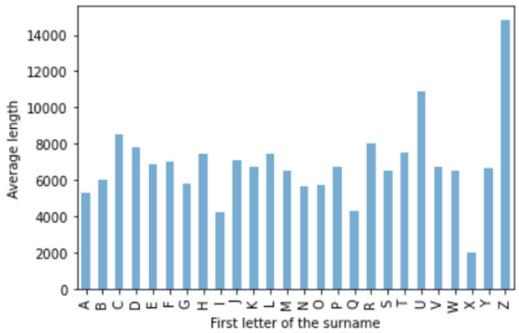
SCHWEIZER JUGEND FORSCHT
SCIENCE ET JEUNESSE
SCIENZA E GIOVENTÙ

PROJECT REPORT
STUDY WEEK “FASCINATING INFORMATICS”

The Unfair Advantage of Alphabetically Early People Based on Wikipedia Article Length



M. Apel, M. Mol

Deutsche Schule Genf, Geneva

Supervised by: Tiziano Piccardi

Date: 9 September 2021

Abstract

A hypothesis claims that people with a surname that is earlier in the alphabet are more likely to have more success in life. This success is in this study based on the length of their Wikipedia articles.

In this study we did an analysis and compared the Wikipedia to prove or dismiss the hypothesis. In order to do so we had access to 3 different pickle files, one with a Dataframe containing information about the text length of the articles, one with information about the people featured in those articles (date of birth/death, gender, birthplace, gender, name, occupation and the articles wikiID) and a last one with the people’s location (countries and cities). Those 3 Dataframes then had to be merged in order to have the people assigned to their own article length and location. After doing so the first letter of the surname had to be put in a separate column to then plot the Dataframe into a graph that showed the first letter with and the corresponding average article length. The results will be mentioned in the rest of this report.

But to really be sure about the result we concluded that a deeper analysis of the correlation of the first letter of a person's surname and their success had to be done.

We started by adding columns that would simplify the plotting later on such as a person's age and if their surname was in the first or last half of the alphabet, this then made it possible for us to compare the average length of articles more concisely by now being able to compare article length based on not only the letters but for example people that live in the same country, worked the same jobs and have lived for a similar amount of time. Plotting this into a graph gave the whole study way more depth. The results will also later be mentioned in the report.

1. Introduction/Question

The topic of our study was if alphabetically early people had an unfair advantage, we based the measurement of people's success on Wikipedia article length. We had 3 different pickle files for us to use which had information about people's location, detailed information about people that were featured in Wikipedia articles and a last one with the length of the articles. In order to create and plot the Dataframes etc. we used Jupyter Notebook where we imported Pandas to make use of the Dataframes, NumPy to use variables and Name Parser to filter out the surnames of each person.

2. Materials & Methods

We started by loading the Dataframe with the text length and the one with the people's details. After that we added an extra column to the people Dataframe which only had the people's surnames using the Name Parser API. We then merged the people Dataframe and the length Dataframe on the "title" column having every name associated with the right article length. To then narrow down the number of articles and get rid of false Data we removed everyone who did not have a surname as well as all the people that did not have a surname that started with a Latin letter. You can then group the article length by the first surname letter and print the average of the first and last 13 letters of the alphabet the average text length should then be: 6836.675973412959 for the first 13 and 6709.310631544491 for the last 13 letters.

This would maybe already be enough to say that yes, the hypothesis is right, although that result would be based on a very superficial analysis which is why we decided to go into more detail to really get a deeper analysis of the Data we had access to.

So, to get a more detailed analysis more columns were required we started by removing everyone that did not have a date of birth and then chose the first 5 numbers of the date of birth line which looks like this: +1925-06-02T00:00:00Z, in order to create an extra column that only had the persons year of birth in it.

We then did the same thing with the date of death column and then subtracted the 2 in order to get the age of the individuals into another column.

We then added another Dataframe that contained the country id and the location id the new Dataframe was then merged with the old one on the birthplace column giving us detailed information about a person's country and city.

To make the Dataframe clearer we added a column named" beginning of alphabet" that printed "True" when the first letter was in the first 13 and "False" if it was in the last 13 letters.

Now to the deeper analysis, we wrote a script that would not only take everyone with the same first surname letter and calculate their average article length but calculate the average article length of the people that are the same age, gender and nationality and of course have the same first surname letter surname.

After doing so a way more complex comparison was possible and then calculating the average article length showed completely different results then the more superficial analysis. The new results should then be
7095.401974612129 average article length for the first 13 letters and
7093.418899858956 for the other 13.

3. Results

The first superficial analysis showed that the first 13 letters of the alphabet are more likely to succeed according to the average Wikipedia article length which was 127.365341868 words longer than the average length of the last 13 letters. This would indeed confirm the hypothesis.
The second results we got from the deeper analysis showed something different, with only a difference of only 1.98307475317 words which can be neglected. This goes to show that the hypothesis is, considering the new data, wrong.

4. Discussion

Although a simple analysis of the data proved the hypothesis right, a deeper analysis is always recommended as it might, just like in this case, prove an initial statement wrong. Our initial statement proved the hypothesis right which is, as we later found out, wrong. It is a mistake that probably occurs quite often when trying to prove a hypothesis wrong/right, the first and often easiest result to get is chosen

5. Acknowledgements

Thank you Tiziano for introducing us to python and further on showing us how to work with pandas and guiding us through this project by giving us help and ideas.
We would also like to thank the Swiss Youth in Science for giving us the opportunity to get a look into university life and expanding our knowledge on informatics.

References


location_country.pkl


text_len.pkl


people_features.pkl