- Aditya Agrawal : adityaai@bu.edu
- Melissa Zhen : mzhen001@bu.edu
- Liangzhuo Zhang : zhanglz@bu.edu
- Cindy Lu : cindy19@bu.edu

Project Scope

The goal of this project is to identify inventors by race and to create an algorithm that predicts the gender of the inventors based on Chinese and Korean character recognition. Using that algorithm, then perform gender analysis of inventors (for example, to see the percentage of male & female inventors). We would like to compare the performance when using English translated names and using the original Chinese/Korean names.

Progress

We have read through the previous team's datasets and deliverables.

They contain information of Patents filed in USPTO (The United States Patent and Trademark Office). They also contain information about the country from which they originated.

We have then found a way to extract the inventor's name using the patent numbers (https://pypi.org/project/google-patent-scraper/). All the patent numbers in the dataset are from the US, so we have created a program to scrape the US patent webpages (on Google Patents) and extract the patent numbers from other countries. We then filter those for only China (CN) and Korea (KR) patent numbers and use them to return the names in the original characters.

The code for this is in the
"reading_spring22_team_2_data_extracting_kr_cn_names_examples.ipynb" file.

Below is an example, where the first line is the US patent number, and the list contains the worldwide patent numbers. Using those numbers, we can get the original names.

```
https://patents.google.com/patent/US6347144
12
['KR0152788B1/en', 'CN1822668B/en', 'CN1123226C/en', 'CN1567993B/en', 'CN1145576C/en', 'CN1822166A/en', 'CN1801919A/en', 'CN1268127C/en', 'CN1822168A/en', 'CN1822667B/en', 'CN1268126C/en', 'CN1822167A/en']
https://patents.google.com/patent/KR0152788B1/en
Patent inventor : 박태준


https://patents.google.com/patent/CN1822668B/en
Patent inventor : 朴兑浚


https://patents.google.com/patent/CN1123226C/en
Patent inventor : 朴兑浚


https://patents.google.com/patent/CN1567993B/en
Patent inventor : 朴兑浚


https://patents.google.com/patent/CN1145576C/en
Patent inventor : ���鱣
Patent inventor : 姜祥保
```

Data

We have a small batch of data on the Chinese or Korean names of inventors with their corresponding English names.

We used Namsor's Name Gender Guesser API (https://namsor.app) to predict the gender probabilities of each name.

Chinese & Korean Characters

	Original Name	Gender	Probability
1	程滋颐 (original - CN)	Female	50.2%
2	김민겸 (Original - KR)	Male	94.2%
3	安丽华 (Original - CN)	Female	90.88%
4	황광조 (Original - KR)	Male	99.72 %
5	송기환 (Original - KR)	Male	99.75%

US translated names

	US translated name	Gender	Probability
1	Ziyi Cheng	Female	64.83%
2	Min-kyum Kim	Male	52.44%
3	Lihua An	Female	92.59%
4	Kwang-Jo Hwang	Male	69.59 %
5	Ki-whan Song	Male	86.27 %

Analysis

We were able to answer one of the key questions from our initial exploration of the data: What is the difference in gender analysis when analyzing characters instead of US translated names?

According to the API we used, while each name was categorized into the same genders, there were different levels of certainty in that classification. For example, 程滋颐 had 50.2% likely to be female, whereas the English pinyin of her name Ziyi Cheng, had a higher likelihood at 64.83% female.

Future Work

In order to analyze gender ratios within both the English letters dataset and converted Chinese and Korean characters dataset, we will leverage the use of Namsor (https://namsor.app/api-documentation/#genderize-chinese-name-batch). Namsor is a library that can classify names by gender, including Chinese and Korean character names.

Because Namsor doesn't offer unlimited free use, we plan to code our own gender sorter in the future. We plan to pattern match traditionally masculine and feminine characters with characters within names of the dataset. While we might miss some more unique or gender neutral characters, it should catch most cases. We also noticed that Namsor sees slightly less accuracy with Chinese character names, so we plan to use cultural knowledge to create a more accurate gender sorter.

From this we can run our modified dataset (with Chinese and Korean characters as names) and an English name dataset through the gender sorters. We will then compare the gender ratios of the two datasets against each other to see any differences.

Limitations

In our analysis so far we have identified a few limiting factors to our work. Technological limits we have are processing power and speed. Currently, scraping data is taking extensive computation time for the size of our datasets. A potential workaround to this is processing in batches.

Cultural factors that pose a challenge to us are the existence of gender neutral names. With these there would be a great degree of difficulty in sorting a gender, especially if the characters used are widespread within both males and females.

References

Characters in Chinese Names - https://blog.tutorabcchinese.com/mandarin-chinese-learning-tips/difference-chinese-male-and-female-names-gender

Characters in Korean Names - https://www.topikguide.com/find-korean-name-gender-male-or-female/