Team 2: Edona Mujaj edona@bu.edu, Andrew Choi dooby@bu.edu , Colette Li ruojiali@bu.edu, Sangjoon Lee, sj0726@bu.edu

- Collect and pre-process a preliminary batch of data
- Perform a preliminary analysis of the data
    - Based on the data collected, there are 828,409 Korean patent inventors and 436,917 Chinese patent inventors totaling 1,265,326 inventors.
    - Out of the total Chinese/Korean inventors, there are 429,959 entries whose gender was set to ‘andy’ (short for androgynous), and 502,562 entries whose gender was unknown.
    - We examined two datasets from the World Gender Name Dictionary 2.0: one with {name, gender, language code} and another with {name, gender, country code, weight}.
    - From the {name, gender, language code} dataset, we filtered rows based on the language code for China and Korea (zh and ko). Then we filtered rows out that had names that were in English characters (romanized names), so that we were left with only rows that had names in Chinese and Korean characters.
    - For the {name, gender, country code, weight} dataset, we first had to figure out a list of country codes that use Chinese and Korean language codes because some countries (like Taiwan) also use Chinese characters.
    - These are the dataset we will
- Answer one key question
    - What is the difference in gender analysis when analyzing characters instead of US translated names?
        - There are masculine and feminine sounding characters in Korean and Chinese romanized characters. There are also a lot of the same characters, and the order of the characters can help determine the gender. US names often have more variations.
- Refine project scope and list of limitations with data and potential risks of achieving project goal
    - Project scope: CN and KR data. We will parse a list of names, and then compare it with the World Gender Name Dictionary, if the name isn’t contained within the dictionary, it is labeled as unknown.
    - Limitations with data
        - Chinese romanization will not be taken into consideration since the romanization of different characters can be the same spelling.
        - Possibility that the names written in Korean and Chinese characters do not exist in the World Gender Name Dictionary.
        - There is a significant amount of names with unknown genders in the dataset with the inventors
    - Since there is a significant amount of data to parse, we might run into issues with runtime as well as storage. Solving these problems may take us a while and impede on our ability to meet project deadlines, however these risks can be mitigated with proper time management and efficient code.

Team 2: Edona Mujaj edona@bu.edu, Andrew Choi dooby@bu.edu , Colette Li ruojiali@bu.edu, Sangjoon Lee, sj0726@bu.edu

Submit a PR with the above report and modifications to original proposalters.