Capstone Project - Data Wrangling

I think an interesting question to answer as a data scientist would be: what skill-set should possess a person which aim to work at Facebook, Microsoft or Google ? What programming languages should she/he know and what soft skills and experience should she have ?

For giving an answer to this question, I will be studying different datasets harvested on career sites on the one hand, and Linkedin and other professional networks on the other. For this “Data Wrangling” submission I have been using the dataset in this link https://www.kaggle.com/niyamatalmass/google-job-skills which shows open positions at Google career site and the skills and/or experience needed to fill them.

I will model the problem as a matching problem. In addition, since my datasets contain mainly unstructured text data, I will use NLP techniques for the analysis and for finding possible ways of automatic matching between the open position features which are advertised at the company’s web site and the candidate’s skills.

More in particular here are the data wrangling steps that I undertook to clean my text data set.

As my dataset is mainly text so as an EDA (Explorative Data Analysis) I did a world count and visualized it as a line graphics.

For this analysis I used the following libraries : nltk, matplotlib and seaborn as well as the tutorial in
https://www.datacamp.com/community/tutorials/web-scraping-python-nlp

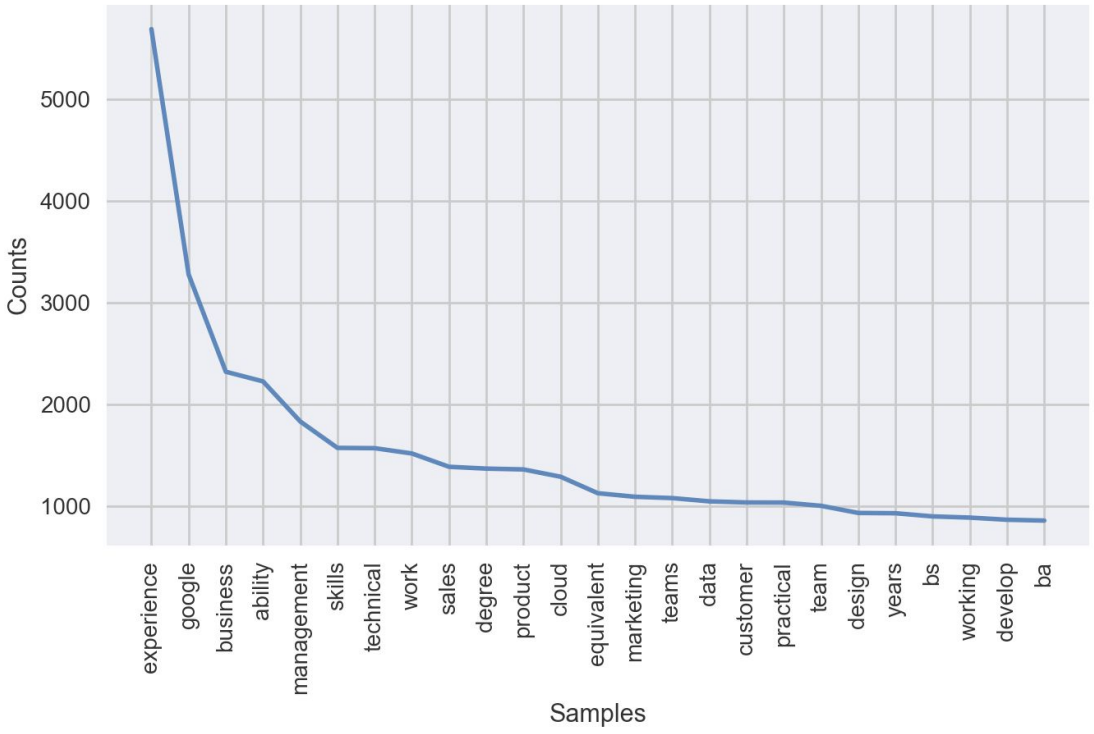
So after loading the data in my PyCharm IDE I :

1 - tokenized the text in single words. For this step the RegexpTokenizer from the nltk.tokenize library is used.

2- put all the words in lower case. This step will help eventually the counting process. The point is that for the counting algorithm, the world “Google” is different from “google” but when all worlds are lower case they can be counted as the same world.

3- removed stop words such as 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'you', "you're". I had no missing values to deal with but the process of removing stop words finally lets me visualize the result as in the graphics below.

The graphics shows the top 20 most frequent words in the dataset and for each world is indicated its frequency number in the dataset.



A first insight is that the world “experience” is the most frequent one in the dataset suggesting that at Google (as in any other company) the open positions require job experience.