Architectural Decisions Document

1. Dataset

https://www.kaggle.com/new-york-city/new-york-city-current-job-postings

2. Use Case

To make a job recommender system using the dataset.

3. Architectural Choices

3.1 I will be using Apache Spark accompanied by Pandas.
Why? I originally planned to use spark data frame all the way through. Unfortunately, I was not able to complete everything I wanted to achieve. As a result, I had to switch to pandas data frame instead. The good news is that I achieved what I wanted to do. In the near future, I will try to do it with spark.

3.2 Using matplotlib for visualizations.
Why? I simply wanted to review what I learned from the previous specialization courses, including bar chart and horizontal bar chart. I also used word cloud for the most mentioned skills and qualifications.

4. Data Exploration

4.1 Found missing data. But I cannot drop those data as rows or columns since it reduces the dataset drastically.

4.2 Found wrong data types of certain values. For example, salary should be in integer type instead of string.

4.3 Data visualizations using plots. I can find the highest paid jobs for example, as well as most required skills or qualifications.

4.4 Correlation matrix allows us to see how relevant is a certain measurement.

5. Data Cleaning and Feature Engineering

5.1 Drop the column with unformatted and irrelevant data such as location.
Why? Because all the jobs are posted within NYC and the locations are not in uniform format.

5.2 Impute some missing value such as full-time/part-time indicator.
Why? Because some jobs are labeled as pay per hour and it is only natural to assume those are part-time jobs.

5.3 Fix data types such as salary and number of positions.
Why? Because salary and number of positions should be in integer type rather than strings.

5.4 Combine multiple columns into one and Extract keywords.
Why? Because it promotes one-hot encoding.

6. Model Training

6.1 Machine learning model – TfidVectorizer and Cosine Similarity Matrix.
Why? Because I need to weigh a keyword in the combined column and assign the importance to that keyword based on the number of times it appears in the column.

6.2 Deep Learning model – Keras Sequential.
Why? Because it is the simplest model I know and it allows us to create models layer-by-layer for most problems.

6.3 Evaluation metric used – Accuracy.
Why? Because this recommender is content based. The best way to evaluate the model performance is using the idea of relevance of the information. Namely, how relevant/similar between two jobs in this example.

6.4 Model Improvement – removed punctuations and digits.
Why? Because it is difficult extract keywords when having different types of punctuations. In addition, digits in keyword provides no information for our model. In fact, we lost information simply by extract keywords.