When discussing our plan of attack for this project and what models to use, Jaccard Similarity immediately stuck out as a valuable tool that would be of value in matching specific resume skills and attributes to highly sought after attributes of an ideal job candidate or employee.

Jaccard Similarity uses a relatively simple concept to compare similarities and differences between sets of data. In this case, we used a vector consisting of words we determined (both by other models deployed as well as tools including custom Dictionaries, etc.) were frequently listed in data job postings, either as a minimum requirement, a preferred attribute, or even something that would give a candidate a leg up in the job search process. This vector provided us with our “key words”, or the words that would ideally be present in a resume of someone seeking a job in the data field. These “key words” are commonly present in data job postings and if also present in a resume, would in theory indicate the candidate was qualified for the job requirements. However, as mentioned in class, a data model is only as good as its attributes, making this part of the process one of the first determinants of how truly valuable the model would be in real world settings.

We felt the Jaccard method was appropriate and sufficient for our specific business question and question of interest. We were not as concerned about “meaning” in the words within a resume, but simply seeing if the words we felt were “key words” in data job postings (and therefore contained in our vector list ) were present at all in an individual’s resume. To do this, the Jaccard method uses binary scores of 1 and 0 to denote if a word has been matched between data sets. If a word from the tested dataset has a 0, it has not been matched and if the word from the tested set has a 1 it has been matched to a word within the predetermined key words. If Jaccard is not only a relatively simple method to determine but also gave us a straightforward “metric”, defined as a number between a 0 and a 1 by taking a ratio of the similarities and the total cases. A Jaccard similarity “score” that fell closer to 1 meant there were a relatively high amount of matched words or similarity between the two datasets.

For data sets to compare to our vector of key words, we used actual resumes provided as text files. We imported the text from each resume and removed any words in the resume that were not in our vector list. Finally, we used the resulting data (after removing the words that did not appear in our vector) to run the Jaccard method. We were interested in looking at resumes with a variety of education backgrounds, rather than testing only resumes of data science students. We felt that this would help us assess whether our model was truly a good model or if it was only good in very specific cases. If the model did not perform well in real world cases then there would really be no use for it. For this project, the resume collection is an area of concern simply due to lack of data quantity in resume form. In the future, we would advise, as always in data modeling, to ensure not only having a large sample size N of resumes, but also a varied random sample set within the group collected. This is an area we felt could be improved in the future.

We expected the relationship between the amount of words a resume had that were also in our vector list to suggest that the individual, at the very least, had the “bare minimum” qualifications required by most data job postings to pass a “first screen” test. Usually the first screening in the interview process is done within HR and essentially is simply a straightforward process of checking off a list of qualifications determined by their company in comparison with the candidates resume information. Therefore, we concluded that while there are too many outside variables ( i.e. poor interview performance, not fitting with the company culture, etc.) in whether the individual would ultimately be determined as “good for the job” and as a result be more likely to receive a job offer. However, we felt that we could at the very least determine if they would likely pass the first “screen” of the interview process by having the essential attributes desired by the interviewing company. Getting your foot in the door is sometimes half the battle In the job interview process, and without at the very least the key qualifications, a candidate is likely to be overlooked or dismissed within a few minutes. Our hope is to maximize a candidate’s likelihood of at least passing the screen round of an interview, which in this case would mean having similarity between their resume and the qualifications frequently listed for job they want.

As mentioned earlier, a benefit of using the Jaccard similarity method is its simplicity. It looks for specific words within a data set that match a predetermined list of “target words” in another dataset. The end result the Jaccard method gives is an overall percentage (with a score between 1 and 0) that tells you how similar the two data sets are based on the amount of matches or similarities found between the two. While for our case this was sufficient, it did not address issues such as the relationship between words, the meaning of words, etc. It simply scores matched cases. It also depends highly on the formatting, language, misspellings, etc. of the two datasets and may miss matches due to one of these or other factors.

Some of the resumes run through our model contained a high amount (usually equal to a 0.6 or above Jaccard score), while others did not. The more similarities or matches cases found, the more likely we felt an individual would be to passing the screen test of the interview process. Interestingly, we found that some of our classmates and group members resumes would have little to no matches found between their resume and common data job posting qualifications previous to being a part of Bellarmine’s MSA program. However, after participating in the MSA program, these highly preferred attributes were more likely to be found in the same person’s resume which is an encouraging sign that this program is preparing students well for a future in the data field.