Steps to execute :

1. Download ElasticSearch (version 7.11.2, to be safe) from https://www.elastic.co/downloads/elasticsearch
2. Start ./bin/elasticsearch from a terminal and keep it running.
3. Find the host url and port name from the elastic search command line and paste the values in coomon\_files/constants.json
4. Git clone https://github.com/rovatsos/opportunity\_match (or pull if you already have it)
5. Make sure python3 is installed (or create a virtual environment with Anaconda
   1. conda create -n my\_python3 python=3.7 anaconda
   2. conda activate my\_python3)
6. Go to opportunity\_match/document-embeddings/production, and run
   * 1. pip install -r requirements.txt
     2. python -m spacy download en\_core\_web\_lg

The actual data files for Step 7 can be obtained from:

1. <https://www.dropbox.com/s/4sp8c0rxvotzd9b/expertise.json?dl=0>
2. https://www.dropbox.com/s/9f8z471yr80p0hj/experts.json?dl=0
3. Run “python create\_load\_indexes.py”   —> this will create two Elastic Search indices (experts and expertise) and load the json files (experts.json and expertise.json) in Elastic Search. (Ignore warnings.)
4. Copy these two files into common\_files folder.
5. Run curl 'localhost:9200/\_cat/indices?v' -> this checks that the indices and data have been created correctly in Elastic Search.
6. Run “python train\_model\_es.py”.  —> Trains a doc2vec model using the data previously loaded to Elastic Search — note this may take up to 2 hours to run.
7. Run “python test\_load\_search\_model\_es.py” —> Tests the doc2vec model and retrieves the most similar documents (of a given text) from Elastic Search.
8. Change directory to production folder.
9. Always first run “python exp\_bias.py” in exp\_bias folder, because this creates exp\_dict.json in common\_files directory which is used by the other \*bias.py files.
10. To fetch results for bias evaluation of gender, ethnicity, experience and race, go to the respective folder and run the \*\_bias\_exec.py file.
11. This creates \*\_queries.json (short and long queries for all schools) and \*\_bias\_results.json file which has the following structure for the results.
12. School\_distr : it’s the distribution of various categories attributes of gender, ethnicity, experience or race.
13. For gender: We get School Male to Female ratio
14. For experience: we School distribution of { beginner, moderate, experienced and very experienced}
15. For Ethnicity: we get School distribution of {British, Non British}
16. For Race: we get School distribution of { API (Asia Pacific islander), Black, Hispanic, White}
17. Average\_publishing\_ratio/count: This gives the average publishing ratio or count of experts across a given school.
18. Tfidf: This has bias results for each school for tf-idf based short queries.
19. Doc2vec: This has bias results for each school for doc2vec based long queries.