Report.md

Summer Internship Report - 2017

```
                          Resume Filtering Using Machine Learning
```

Submitted By:

- Mradul Dubey
- Shubham Kumar Singh Rajput
- Rahul Singh

Project Introduction

```
The project aims to automate the filtering process of the influx of CVs/resumes for
an IT company.
```

This is, in its current form, achieved by assigning a score to each CV by intelligently comparing them against the corresponding Job Description. This reduces the window to a fraction of an original size of applicants. Resumes in the final window can be manually checked for further analysis. The project uses recent advances in Data Science and Machine Learning to automate the procedure. The project was challenging not only because filtering CV is a subjective matter but because of the variety of resume writing and the unavailability of processed data.

Overview

- Not unlike any other Data Science project, we started off at Data Collection. We mainly required three datasets.
- Then we trained the Word2Vec Model using the StackOverflow data dump.
- We extracted sections from the CVs like Education, Experience etc.
- Finally, the CVs were awarded scores against each Job Descriptions available.

Data Collection

Mainly we required three datasets:

StackExchange Network Posts

- This dataset was required to trains the word2vec model. Fortunately, StackExchange network dumps it's data in xml format under Creative Commons License. One can find a download link for the dataset(44 GB) on Internet Archive.

- This is an anonymized dump of all user-contributed content on the Stack Exchange network. Each site is formatted as a separate archive consisting of XML files zipped via 7-zip using bzip2 compression. The following sites were included:

```
3dprinting.stackexchange.com.7z                   13-Jun-2017 13:49     2.8M
Sites.xml                                         13-Jun-2017 15:53     327.4K
academia.stackexchange.com.7z                     13-Jun-2017 13:50     71.1M
ai.stackexchange.com.7z                           13-Jun-2017 13:50     2.4M
android.stackexchange.com.7z                      13-Jun-2017 13:50     75.5M
anime.stackexchange.com.7z                        13-Jun-2017 13:51     20.6M
apple.stackexchange.com.7z                        13-Jun-2017 13:51     149.4M
arabic.stackexchange.com.7z                       13-Jun-2017 13:51     326.8K
arduino.stackexchange.com.7z                      13-Jun-2017 13:52     30.1M
askubuntu.com.7z                                  13-Jun-2017 13:56     546.7M
astronomy.stackexchange.com.7z                    13-Jun-2017 13:56     13.2M
aviation.stackexchange.com.7z                     13-Jun-2017 13:56     34.7M
avp.stackexchange.com.7z                          13-Jun-2017 13:56     9.6M
beer.stackexchange.com.7z                         13-Jun-2017 13:56     2.2M
bicycles.stackexchange.com.7z                     13-Jun-2017 13:56     31.5M
biology.stackexchange.com.7z                      13-Jun-2017 13:57     40.7M
bitcoin.stackexchange.com.7z                      13-Jun-2017 13:57     26.0M
blender.stackexchange.com.7z                      13-Jun-2017 13:57     48.3M
boardgames.stackexchange.com.7z                   13-Jun-2017 13:57     21.9M
bricks.stackexchange.com.7z                       13-Jun-2017 13:58     4.6M
buddhism.stackexchange.com.7z                     13-Jun-2017 13:58     18.4M
chemistry.stackexchange.com.7z                    13-Jun-2017 13:58     47.5M
chess.stackexchange.com.7z                        13-Jun-2017 13:58     10.8M
chinese.stackexchange.com.7z                      13-Jun-2017 13:58     12.0M
christianity.stackexchange.com.7z                 13-Jun-2017 13:59     52.7M
civicrm.stackexchange.com.7z                      13-Jun-2017 13:59     8.8M
codegolf.stackexchange.com.7z                     13-Jun-2017 14:01     142.4M
codereview.stackexchange.com.7z                   13-Jun-2017 14:04     274.7M
coffee.stackexchange.com.7z                       13-Jun-2017 14:04     2.2M
cogsci.stackexchange.com.7z                       13-Jun-2017 14:04     14.1M
computergraphics.stackexchange.com.7z             13-Jun-2017 14:04     3.8M
cooking.stackexchange.com.7z                      13-Jun-2017 14:04     45.3M
craftcms.stackexchange.com.7z                     13-Jun-2017 14:05     11.1M
crafts.stackexchange.com.7z                       13-Jun-2017 14:05     1.8M
crypto.stackexchange.com.7z                       13-Jun-2017 14:05     38.8M
cs.stackexchange.com.7z                           13-Jun-2017 14:06     51.1M
cstheory.stackexchange.com.7z                     13-Jun-2017 14:06     27.6M
datascience.stackexchange.com.7z                  13-Jun-2017 14:06     12.7M
dba.stackexchange.com.7z                          13-Jun-2017 14:08     151.0M
diy.stackexchange.com.7z                          13-Jun-2017 14:08     63.9M
drupal.stackexchange.com.7z                       13-Jun-2017 14:10     113.7M
dsp.stackexchange.com.7z                          13-Jun-2017 14:10     30.4M
earthscience.stackexchange.com.7z                 13-Jun-2017 14:10     7.9M
ebooks.stackexchange.com.7z                       13-Jun-2017 14:10     2.5M
economics.stackexchange.com.7z                    13-Jun-2017 14:10     11.2M
electronics.stackexchange.com.7z                  13-Jun-2017 14:13     224.7M
elementaryos.stackexchange.com.7z                 13-Jun-2017 14:13     6.3M
ell.stackexchange.com.7z                          13-Jun-2017 14:14     79.5M
emacs.stackexchange.com.7z                        13-Jun-2017 14:14     20.0M
engineering.stackexchange.com.7z                  13-Jun-2017 14:14     11.3M
english.stackexchange.com.7z                      13-Jun-2017 14:16     232.3M
es.stackoverflow.com.7z                           13-Jun-2017 14:17     73.7M
esperanto.stackexchange.com.7z                    13-Jun-2017 14:17     2.0M
ethereum.stackexchange.com.7z                     13-Jun-2017 14:17     12.6M
expatriates.stackexchange.com.7z                  13-Jun-2017 14:17     6.4M
expressionengine.stackexchange.com.7z             13-Jun-2017 14:17     16.6M
fitness.stackexchange.com.7z                      13-Jun-2017 14:17     19.8M
freelancing.stackexchange.com.7z                  13-Jun-2017 14:17     5.3M
french.stackexchange.com.7z                       13-Jun-2017 14:17     17.5M
gamedev.stackexchange.com.7z                      13-Jun-2017 14:19     105.5M
gaming.stackexchange.com.7z                       13-Jun-2017 14:21     151.8M
gardening.stackexchange.com.7z                    13-Jun-2017 14:21     18.5M
genealogy.stackexchange.com.7z                    13-Jun-2017 14:21     8.9M
german.stackexchange.com.7z                       13-Jun-2017 14:21     27.2M
gis.stackexchange.com.7z                          13-Jun-2017 14:24     171.3M
graphicdesign.stackexchange.com.7z                13-Jun-2017 14:25     48.5M
ham.stackexchange.com.7z                          13-Jun-2017 14:25     5.2M
hardwarerecs.stackexchange.com.7z                 13-Jun-2017 14:25     4.0M
health.stackexchange.com.7z                       13-Jun-2017 14:25     7.6M
hermeneutics.stackexchange.com.7z                 13-Jun-2017 14:25     29.1M
hinduism.stackexchange.com.7z                     13-Jun-2017 14:25     21.4M
history.stackexchange.com.7z                      13-Jun-2017 14:26     30.8M
homebrew.stackexchange.com.7z                     13-Jun-2017 14:26     10.2M
hsm.stackexchange.com.7z                          13-Jun-2017 14:26     5.4M
iot.stackexchange.com.7z                          13-Jun-2017 14:26     1.8M
islam.stackexchange.com.7z                        13-Jun-2017 14:26     24.2M
italian.stackexchange.com.7z                      13-Jun-2017 14:27     4.3M
```

```
ja.stackoverflow.com.7z                           13-Jun-2017 14:27     25.4M
japanese.stackexchange.com.7z                     13-Jun-2017 14:27     28.5M
joomla.stackexchange.com.7z                       13-Jun-2017 14:27     8.7M
judaism.stackexchange.com.7z                      13-Jun-2017 14:28     61.9M
korean.stackexchange.com.7z                       13-Jun-2017 14:28     1.4M
languagelearning.stackexchange.com.7z             13-Jun-2017 14:28     1.7M
latin.stackexchange.com.7z                        13-Jun-2017 14:28     4.0M
law.stackexchange.com.7z                          13-Jun-2017 14:29     15.3M
license.txt                                       23-Jan-2014 17:05     1.9K
lifehacks.stackexchange.com.7z                    13-Jun-2017 14:29     6.7M
linguistics.stackexchange.com.7z                  13-Jun-2017 14:29     15.8M
literature.stackexchange.com.7z                   13-Jun-2017 14:29     3.2M
magento.stackexchange.com.7z                      13-Jun-2017 14:31     95.4M
martialarts.stackexchange.com.7z                  13-Jun-2017 14:31     6.5M
math.stackexchange.com.7z                         13-Jun-2017 14:49     1.4G
matheducators.stackexchange.com.7z                13-Jun-2017 14:51     9.6M
mathematica.stackexchange.com.7z                  13-Jun-2017 14:53     138.0M
mathoverflow.net.7z                               13-Jun-2017 14:56     236.4M
mechanics.stackexchange.com.7z                    13-Jun-2017 14:56     28.1M
meta.3dprinting.stackexchange.com.7z              13-Jun-2017 14:56     207.5K
meta.academia.stackexchange.com.7z                13-Jun-2017 14:56     2.9M
meta.ai.stackexchange.com.7z                      13-Jun-2017 14:56     293.6K
meta.android.stackexchange.com.7z                 13-Jun-2017 14:56     2.2M
meta.anime.stackexchange.com.7z                   13-Jun-2017 14:56     3.2M
meta.apple.stackexchange.com.7z                   13-Jun-2017 14:56     3.0M
meta.arabic.stackexchange.com.7z                  13-Jun-2017 14:56     73.5K
meta.arduino.stackexchange.com.7z                 13-Jun-2017 14:56     559.8K
meta.askubuntu.com.7z                             13-Jun-2017 14:57     13.4M
meta.astronomy.stackexchange.com.7z               13-Jun-2017 14:57     441.3K
meta.aviation.stackexchange.com.7z                13-Jun-2017 14:57     1.3M
meta.avp.stackexchange.com.7z                     13-Jun-2017 14:57     420.5K
meta.beer.stackexchange.com.7z                    13-Jun-2017 14:57     193.1K
meta.bicycles.stackexchange.com.7z                13-Jun-2017 14:57     1.2M
meta.biology.stackexchange.com.7z                 13-Jun-2017 14:57     2.1M
meta.bitcoin.stackexchange.com.7z                 13-Jun-2017 14:57     749.0K
meta.blender.stackexchange.com.7z                 13-Jun-2017 14:57     1.6M
meta.boardgames.stackexchange.com.7z              13-Jun-2017 14:57     1.7M
meta.bricks.stackexchange.com.7z                  13-Jun-2017 14:57     346.0K
meta.buddhism.stackexchange.com.7z                13-Jun-2017 14:57     1.4M
meta.chemistry.stackexchange.com.7z               13-Jun-2017 14:57     2.8M
meta.chess.stackexchange.com.7z                   13-Jun-2017 14:57     425.9K
meta.chinese.stackexchange.com.7z                 13-Jun-2017 14:58     501.5K
meta.christianity.stackexchange.com.7z            13-Jun-2017 14:58     5.4M
meta.civicrm.stackexchange.com.7z                 13-Jun-2017 14:58     107.1K
meta.codegolf.stackexchange.com.7z                13-Jun-2017 14:58     12.1M
meta.codereview.stackexchange.com.7z              13-Jun-2017 14:58     7.1M
meta.coffee.stackexchange.com.7z                  13-Jun-2017 14:58     173.4K
meta.cogsci.stackexchange.com.7z                  13-Jun-2017 14:58     1.5M
meta.computergraphics.stackexchange.com.7z        13-Jun-2017 14:58     276.4K
meta.cooking.stackexchange.com.7z                 13-Jun-2017 14:58     2.7M
meta.craftcms.stackexchange.com.7z                13-Jun-2017 14:58     138.9K
meta.crafts.stackexchange.com.7z                  13-Jun-2017 14:58     329.3K
meta.crypto.stackexchange.com.7z                  13-Jun-2017 14:58     1.2M
meta.cs.stackexchange.com.7z                      13-Jun-2017 14:59     2.1M
meta.cstheory.stackexchange.com.7z                13-Jun-2017 14:59     2.0M
meta.datascience.stackexchange.com.7z             13-Jun-2017 14:59     289.6K
meta.dba.stackexchange.com.7z                     13-Jun-2017 14:59     2.1M
meta.diy.stackexchange.com.7z                     13-Jun-2017 14:59     1.1M
meta.drupal.stackexchange.com.7z                  13-Jun-2017 14:59     2.3M
meta.dsp.stackexchange.com.7z                     13-Jun-2017 14:59     539.1K
meta.earthscience.stackexchange.com.7z            13-Jun-2017 14:59     619.2K
meta.ebooks.stackexchange.com.7z                  13-Jun-2017 14:59     209.1K
meta.economics.stackexchange.com.7z               13-Jun-2017 14:59     681.6K
meta.electronics.stackexchange.com.7z             13-Jun-2017 14:59     4.6M
meta.elementaryos.stackexchange.com.7z            13-Jun-2017 14:59     176.6K
meta.ell.stackexchange.com.7z                     13-Jun-2017 15:00     3.6M
meta.emacs.stackexchange.com.7z                   13-Jun-2017 15:00     627.5K
meta.engineering.stackexchange.com.7z             13-Jun-2017 15:00     806.2K
meta.english.stackexchange.com.7z                 13-Jun-2017 15:00     11.7M
meta.es.stackoverflow.com.7z                      13-Jun-2017 15:00     3.0M
meta.esperanto.stackexchange.com.7z               13-Jun-2017 15:00     120.6K
meta.ethereum.stackexchange.com.7z                13-Jun-2017 15:00     449.9K
meta.expatriates.stackexchange.com.7z             13-Jun-2017 15:00     278.5K
meta.expressionengine.stackexchange.com.7z        13-Jun-2017 15:00     318.0K
meta.fitness.stackexchange.com.7z                 13-Jun-2017 15:00     696.8K
meta.freelancing.stackexchange.com.7z             13-Jun-2017 15:00     295.9K
meta.french.stackexchange.com.7z                  13-Jun-2017 15:00     931.8K
meta.gamedev.stackexchange.com.7z                 13-Jun-2017 15:00     2.6M
meta.gaming.stackexchange.com.7z                  13-Jun-2017 15:00     11.6M
meta.gardening.stackexchange.com.7z               13-Jun-2017 15:00     785.8K
meta.genealogy.stackexchange.com.7z               13-Jun-2017 15:01     1.5M
meta.german.stackexchange.com.7z                  13-Jun-2017 15:01     1.6M
meta.gis.stackexchange.com.7z                     13-Jun-2017 15:01     3.1M
meta.graphicdesign.stackexchange.com.7z           13-Jun-2017 15:01     2.5M
meta.ham.stackexchange.com.7z                     13-Jun-2017 15:01     313.6K
meta.hardwarerecs.stackexchange.com.7z            13-Jun-2017 15:01     774.4K
meta.health.stackexchange.com.7z                  13-Jun-2017 15:14     970.2K
meta.hermeneutics.stackexchange.com.7z            13-Jun-2017 15:14     2.3M
meta.hinduism.stackexchange.com.7z                13-Jun-2017 15:14     1.2M
meta.history.stackexchange.com.7z                 13-Jun-2017 15:14     1.6M
meta.homebrew.stackexchange.com.7z                13-Jun-2017 15:14     291.0K
meta.hsm.stackexchange.com.7z                     13-Jun-2017 15:14     337.4K
meta.iot.stackexchange.com.7z                     13-Jun-2017 15:14     316.7K
meta.islam.stackexchange.com.7z                   13-Jun-2017 15:14     2.8M
meta.italian.stackexchange.com.7z                 13-Jun-2017 15:14     307.0K
meta.ja.stackoverflow.com.7z                      13-Jun-2017 15:14     1.5M
meta.japanese.stackexchange.com.7z                13-Jun-2017 15:14     2.0M
meta.joomla.stackexchange.com.7z                  13-Jun-2017 15:15     241.2K
meta.judaism.stackexchange.com.7z                 13-Jun-2017 15:15     4.0M
meta.korean.stackexchange.com.7z                  13-Jun-2017 15:15     157.5K
meta.languagelearning.stackexchange.com.7z        13-Jun-2017 15:15     359.9K
meta.latin.stackexchange.com.7z                   13-Jun-2017 15:15     359.1K
meta.law.stackexchange.com.7z                     13-Jun-2017 15:15     715.2K
meta.lifehacks.stackexchange.com.7z               13-Jun-2017 15:15     933.8K
meta.linguistics.stackexchange.com.7z             13-Jun-2017 15:15     780.3K
meta.literature.stackexchange.com.7z              13-Jun-2017 15:15     1.0M
meta.magento.stackexchange.com.7z                 13-Jun-2017 15:15     1.0M
meta.martialarts.stackexchange.com.7z             13-Jun-2017 15:15     589.7K
meta.math.stackexchange.com.7z                    13-Jun-2017 15:15     26.1M
meta.matheducators.stackexchange.com.7z           13-Jun-2017 15:15     626.6K
meta.mathematica.stackexchange.com.7z             13-Jun-2017 15:15     2.7M
meta.mathoverflow.net.7z                          13-Jun-2017 15:15     4.1M
meta.mechanics.stackexchange.com.7z               13-Jun-2017 15:16     949.0K
meta.moderators.stackexchange.com.7z              13-Jun-2017 15:16     426.9K
meta.monero.stackexchange.com.7z                  13-Jun-2017 15:16     135.3K
meta.money.stackexchange.com.7z                   13-Jun-2017 15:16     1.5M
meta.movies.stackexchange.com.7z                  13-Jun-2017 15:16     3.5M
meta.music.stackexchange.com.7z                   13-Jun-2017 15:16     1.8M
meta.musicfans.stackexchange.com.7z               13-Jun-2017 15:16     320.6K
meta.mythology.stackexchange.com.7z               13-Jun-2017 15:16     463.5K
meta.networkengineering.stackexchange.com.7z      13-Jun-2017 15:16     854.4K
meta.opendata.stackexchange.com.7z                13-Jun-2017 15:16     342.6K
meta.opensource.stackexchange.com.7z              13-Jun-2017 15:16     726.9K
meta.outdoors.stackexchange.com.7z                13-Jun-2017 15:16     896.9K
meta.parenting.stackexchange.com.7z               13-Jun-2017 15:16     1.5M
meta.patents.stackexchange.com.7z                 13-Jun-2017 15:16     364.5K
meta.pets.stackexchange.com.7z                    13-Jun-2017 15:16     1,012.2K
meta.philosophy.stackexchange.com.7z              13-Jun-2017 15:17     1.7M
meta.photo.stackexchange.com.7z                   13-Jun-2017 15:17     3.2M
meta.physics.stackexchange.com.7z                 13-Jun-2017 15:17     9.6M
meta.pm.stackexchange.com.7z                      13-Jun-2017 15:17     948.5K
meta.poker.stackexchange.com.7z                   13-Jun-2017 15:17     247.6K
meta.politics.stackexchange.com.7z                13-Jun-2017 15:17     1.2M
meta.portuguese.stackexchange.com.7z              13-Jun-2017 15:17     446.0K
meta.productivity.stackexchange.com.7z            13-Jun-2017 15:17     416.4K
meta.programmers.stackexchange.com.7z             15-Dec-2016 18:04     9.7M
meta.pt.stackoverflow.com.7z                      13-Jun-2017 15:17     7.5M
meta.puzzling.stackexchange.com.7z                13-Jun-2017 15:17     3.9M
meta.quant.stackexchange.com.7z                   13-Jun-2017 15:17     507.5K
meta.raspberrypi.stackexchange.com.7z             13-Jun-2017 15:17     1.2M
meta.retrocomputing.stackexchange.com.7z          13-Jun-2017 15:17     318.4K
meta.reverseengineering.stackexchange.com.7z      13-Jun-2017 15:17     382.3K
meta.robotics.stackexchange.com.7z                13-Jun-2017 15:18     415.3K
meta.rpg.stackexchange.com.7z                     13-Jun-2017 15:18     7.2M
meta.ru.stackoverflow.com.7z                      13-Jun-2017 15:18     7.2M
meta.rus.stackexchange.com.7z                     13-Jun-2017 15:18     245.5K
meta.russian.stackexchange.com.7z                 13-Jun-2017 15:18     389.1K
meta.salesforce.stackexchange.com.7z              13-Jun-2017 15:18     1.4M
meta.scicomp.stackexchange.com.7z                 13-Jun-2017 15:18     499.6K
meta.scifi.stackexchange.com.7z                   13-Jun-2017 15:18     10.2M
meta.security.stackexchange.com.7z                13-Jun-2017 15:18     3.3M
meta.serverfault.com.7z                           13-Jun-2017 15:18     8.3M
meta.sharepoint.stackexchange.com.7z              13-Jun-2017 15:18     1.1M
meta.sitecore.stackexchange.com.7z                13-Jun-2017 15:18     265.3K
meta.skeptics.stackexchange.com.7z                13-Jun-2017 15:19     4.7M
meta.softwareengineering.stackexchange.com.7z     13-Jun-2017 15:19     12.2M
meta.softwarerecs.stackexchange.com.7z            13-Jun-2017 15:19     2.1M
meta.sound.stackexchange.com.7z                   13-Jun-2017 15:19     364.7K
meta.space.stackexchange.com.7z                   13-Jun-2017 15:19     1.2M
meta.spanish.stackexchange.com.7z                 13-Jun-2017 15:19     811.8K
meta.sports.stackexchange.com.7z                  13-Jun-2017 15:19     794.7K
meta.sqa.stackexchange.com.7z                     13-Jun-2017 15:19     394.5K
meta.stackexchange.com.7z                         13-Jun-2017 15:22     239.8M
meta.stackoverflow.com.7z                         13-Jun-2017 15:23     142.7M
meta.startups.stackexchange.com.7z                13-Jun-2017 15:23     397.5K
meta.stats.stackexchange.com.7z                   13-Jun-2017 15:23     5.3M
meta.superuser.com.7z                             13-Jun-2017 15:24     13.7M
meta.sustainability.stackexchange.com.7z          13-Jun-2017 15:24     365.5K
meta.tex.stackexchange.com.7z                     13-Jun-2017 15:24     7.8M
meta.tor.stackexchange.com.7z                     13-Jun-2017 15:24     236.3K
meta.travel.stackexchange.com.7z                  13-Jun-2017 15:24     3.9M
meta.tridion.stackexchange.com.7z                 13-Jun-2017 15:24     386.6K
meta.unix.stackexchange.com.7z                    13-Jun-2017 15:24     4.3M
meta.ux.stackexchange.com.7z                      13-Jun-2017 15:24     2.3M
meta.vi.stackexchange.com.7z                      13-Jun-2017 15:24     507.5K
meta.webapps.stackexchange.com.7z                 13-Jun-2017 15:24     1.7M
meta.webmasters.stackexchange.com.7z              13-Jun-2017 15:24     1.4M
meta.windowsphone.stackexchange.com.7z            13-Jun-2017 15:24     281.2K
meta.woodworking.stackexchange.com.7z             13-Jun-2017 15:24     291.9K
meta.wordpress.stackexchange.com.7z               13-Jun-2017 15:24     2.6M
meta.workplace.stackexchange.com.7z               13-Jun-2017 15:24     5.1M
meta.worldbuilding.stackexchange.com.7z           13-Jun-2017 15:24     5.1M
meta.writers.stackexchange.com.7z                 13-Jun-2017 15:24     1.4M
moderators.stackexchange.com.7z                   13-Jun-2017 15:24     2.0M
monero.stackexchange.com.7z                       13-Jun-2017 15:25     2.7M
money.stackexchange.com.7z                        13-Jun-2017 15:25     49.3M
movies.stackexchange.com.7z                       13-Jun-2017 15:25     44.1M
music.stackexchange.com.7z                        13-Jun-2017 15:26     36.3M
musicfans.stackexchange.com.7z                    13-Jun-2017 15:26     2.9M
mythology.stackexchange.com.7z                    13-Jun-2017 15:26     3.1M
networkengineering.stackexchange.com.7z           13-Jun-2017 15:26     21.7M
opendata.stackexchange.com.7z                     13-Jun-2017 15:26     6.7M
opensource.stackexchange.com.7z                   13-Jun-2017 15:26     5.0M
outdoors.stackexchange.com.7z                     13-Jun-2017 15:26     14.8M
parenting.stackexchange.com.7z                    13-Jun-2017 15:27     25.1M
patents.stackexchange.com.7z                      13-Jun-2017 15:27     6.9M
pets.stackexchange.com.7z                         13-Jun-2017 15:28     10.2M
philosophy.stackexchange.com.7z                   13-Jun-2017 15:28     39.1M
photo.stackexchange.com.7z                        13-Jun-2017 15:29     63.7M
physics.stackexchange.com.7z                      13-Jun-2017 15:33     270.8M
pm.stackexchange.com.7z                           13-Jun-2017 15:36     14.7M
poker.stackexchange.com.7z                        13-Jun-2017 15:36     3.8M
politics.stackexchange.com.7z                     13-Jun-2017 15:36     19.4M
portuguese.stackexchange.com.7z                   13-Jun-2017 15:36     4.6M
productivity.stackexchange.com.7z                 13-Jun-2017 15:37     9.7M
programmers.stackexchange.com.7z                  15-Dec-2016 18:10     198.8M
pt.stackoverflow.com.7z                           13-Jun-2017 15:40     192.3M
puzzling.stackexchange.com.7z                     13-Jun-2017 15:41     44.7M
quant.stackexchange.com.7z                        13-Jun-2017 15:42     19.7M
raspberrypi.stackexchange.com.7z                  13-Jun-2017 15:42     40.9M
readme.txt                                        23-Jan-2014 00:46     4.6K
retrocomputing.stackexchange.com.7z               13-Jun-2017 15:43     3.4M
reverseengineering.stackexchange.com.7z           13-Jun-2017 15:43     11.4M
robotics.stackexchange.com.7z                     13-Jun-2017 15:44     9.3M
rpg.stackexchange.com.7z                          13-Jun-2017 15:44     99.8M
ru.stackoverflow.com.7z                           13-Jun-2017 15:46     293.7M
rus.stackexchange.com.7z                          13-Jun-2017 15:46     21.2M
russian.stackexchange.com.7z                      13-Jun-2017 15:46     7.5M
salesforce.stackexchange.com.7z                   13-Jun-2017 15:47     110.7M
scicomp.stackexchange.com.7z                      13-Jun-2017 15:47     17.9M
scifi.stackexchange.com.7z                        13-Jun-2017 15:49     154.7M
security.stackexchange.com.7z                     13-Jun-2017 15:49     130.7M
serverfault.com.7z                                13-Jun-2017 15:53     520.5M
sharepoint.stackexchange.com.7z                   13-Jun-2017 15:53     108.0M
sitecore.stackexchange.com.7z                     13-Jun-2017 15:53     5.1M
skeptics.stackexchange.com.7z                     13-Jun-2017 15:54     40.7M
softwareengineering.stackexchange.com.7z          13-Jun-2017 15:55     215.1M
softwarerecs.stackexchange.com.7z                 13-Jun-2017 15:55     24.8M
sound.stackexchange.com.7z                        13-Jun-2017 15:55     19.0M
space.stackexchange.com.7z                        13-Jun-2017 15:57     21.2M
spanish.stackexchange.com.7z                      13-Jun-2017 15:57     12.8M
sports.stackexchange.com.7z                       13-Jun-2017 15:57     8.9M
sqa.stackexchange.com.7z                          13-Jun-2017 15:57     16.6M
stackapps.com.7z                                  13-Jun-2017 15:57     8.6M
```

```
stackexchange_files.xml                           14-Jul-2017 15:57     96.3K
stackexchange_meta.sqlite                         13-Jun-2017 21:56     431.0K
stackexchange_meta.xml                            28-Jun-2017 04:11     3.3K
stackexchange_reviews.xml                         14-Jul-2017 15:57     11.3K
stackoverflow.com-Badges.7z                       13-Jun-2017 15:59     166.2M
stackoverflow.com-Comments.7z                     13-Jun-2017 16:24     3.2G
stackoverflow.com-PostHistory.7z                  13-Jun-2017 19:32     18.3G
stackoverflow.com-PostLinks.7z                    13-Jun-2017 19:33     56.9M
stackoverflow.com-Posts.7z                        13-Jun-2017 21:14     10.5G
stackoverflow.com-Tags.7z                         13-Jun-2017 21:14     685.0K
stackoverflow.com-Users.7z                        13-Jun-2017 21:16     284.3M
stackoverflow.com-Votes.7z                        13-Jun-2017 21:22     757.9M
startups.stackexchange.com.7z                     13-Jun-2017 21:22     7.5M
stats.stackexchange.com.7z                        13-Jun-2017 21:23     243.5M
superuser.com.7z                                  13-Jun-2017 21:28     688.6M
sustainability.stackexchange.com.7z               13-Jun-2017 21:28     4.3M
tex.stackexchange.com.7z                          13-Jun-2017 21:31     373.8M
tor.stackexchange.com.7z                          13-Jun-2017 21:31     6.8M
travel.stackexchange.com.7z                       13-Jun-2017 21:31     69.0M
tridion.stackexchange.com.7z                      13-Jun-2017 21:31     10.9M
unix.stackexchange.com.7z                         13-Jun-2017 21:33     291.0M
ux.stackexchange.com.7z                           13-Jun-2017 21:33     76.4M
vi.stackexchange.com.7z                           13-Jun-2017 21:34     9.9M
webapps.stackexchange.com.7z                      13-Jun-2017 21:34     38.3M
webmasters.stackexchange.com.7z                   13-Jun-2017 21:34     56.1M
windowsphone.stackexchange.com.7z                 13-Jun-2017 21:34     4.7M
woodworking.stackexchange.com.7z                  13-Jun-2017 21:34     5.6M
wordpress.stackexchange.com.7z                    13-Jun-2017 21:35     144.9M
workplace.stackexchange.com.7z                    13-Jun-2017 21:35     72.0M
worldbuilding.stackexchange.com.7z                13-Jun-2017 21:36     104.4M
writers.stackexchange.com.7z                      13-Jun-2017 21:36     20.4M
```

- All the zip files were extracted using a `bash script`.
- This colection of sites is referenced as `stackechange/` folder from hereafter.
- Each site archive includes Posts, Users, Votes, Comments, PostHistory and PostLinks (all in .xml files). The `README.md` file of the dataset is given below:

README.md

- Format: 7zipped
- Files:
  - **badges**.xml
    - UserId, e.g.: "420"
    - Name, e.g.: "Teacher"
    - Date, e.g.: "2008-09-15T08:55:03.923"
  - **comments**.xml
    - Id
    - PostId
    - Score
    - Text, e.g.: "@Stu Thompson: Seems possible to me - why not try it?"
    - CreationDate, e.g.:"2008-09-06T08:07:10.730"
    - UserId
  - **posts**.xml
    - Id
    - PostTypeId
      - 1: Question
      - 2: Answer
    - ParentID (only present if PostTypeId is 2)
    - AcceptedAnswerId (only present if PostTypeId is 1)
    - CreationDate
    - Score
    - ViewCount
    - Body

    - OwnerUserId
    - LastEditorUserId
    - LastEditorDisplayName="Jeff Atwood"
    - LastEditDate="2009-03-05T22:28:34.823"
    - LastActivityDate="2009-03-11T12:51:01.480"
    - CommunityOwnedDate="2009-03-11T12:51:01.480"
    - ClosedDate="2009-03-11T12:51:01.480"
    - Title=
    - Tags=
    - AnswerCount
    - CommentCount
    - FavoriteCount
  - **posthistory**.xml
    - Id
    - PostHistoryTypeId
    - 1: Initial Title - The first title a question is asked with.
    - 2: Initial Body - The first raw body text a post is submitted with.
    - 3: Initial Tags - The first tags a question is asked with.
    - 4: Edit Title - A question's title has been changed.
    - 5: Edit Body - A post's body has been changed, the raw text is stored here as markdown.
    - 6: Edit Tags - A question's tags have been changed.
    - 7: Rollback Title - A question's title has reverted to a previous version.
    - 8: Rollback Body - A post's body has reverted to a previous version - the raw text is stored here.
    - 9: Rollback Tags - A question's tags have reverted to a previous version.
    - 10: Post Closed - A post was voted to be closed.
    - 11: Post Reopened - A post was voted to be reopened.
    - 12: Post Deleted - A post was voted to be removed.
    - 13: Post Undeleted - A post was voted to be restored.
    - 14: Post Locked - A post was locked by a moderator.
    - 15: Post Unlocked - A post was unlocked by a moderator.
    - 16: Community Owned - A post has become community owned.
    - 17: Post Migrated - A post was migrated.
    - 18: Question Merged - A question has had another, deleted question merged into itself.
    - 19: Question Protected - A question was protected by a moderator
    - 20: Question Unprotected - A question was unprotected by a moderator
    - 21: Post Disassociated - An admin removes the OwnerUserId from a post.
    - 22: Question Unmerged - A previously merged question has had its answers and votes restored.
  - PostId
  - RevisionGUID: At times more than one type of history record can be recorded by a single action. All of these will be grouped using the same RevisionGUID
  - CreationDate: "2009-03-05T22:28:34.823"
  - UserId
  - UserDisplayName: populated if a user has been removed and no longer referenced by user Id
  - Comment: This field will contain the comment made by the user who edited a post
  - Text: A raw version of the new value for a given revision
    - If PostHistoryTypeId = 10, 11, 12, 13, 14, or 15 this column will contain a JSON encoded string with all users who have voted for the PostHistoryTypeId
    - If PostHistoryTypeId = 17 this column will contain migration details of either "from " or "to "
  - CloseReasonId
    - 1: Exact Duplicate - This question covers exactly the same ground as earlier questions on this topic; its answers may be merged with another identical question.
    - 2: off-topic
    - 3: subjective
    - 4: not a real question

  - 7: too localized
- **postlinks**.xml
  - Id
  - CreationDate
  - PostId
  - RelatedPostId
  - PostLinkTypeId
    - 1: Linked
    - 3: Duplicate
- **users**.xml
  - Id
  - Reputation
  - CreationDate
  - DisplayName
  - EmailHash
  - LastAccessDate
  - WebsiteUrl
  - Location
  - Age
  - AboutMe
  - Views
  - UpVotes
  - DownVotes
- **votes**.xml
  - Id
  - PostId
  - VoteTypeId
    - `1` : AcceptedByOriginator
    - `2` : UpMod
    - `3` : DownMod
    - `4` : Offensive
    - `5` : Favorite - if VoteTypeId = 5 UserId will be populated
    - `6` : Close
    - `7` : Reopen
    - `8` : BountyStart
    - `9` : BountyClose
    - `10` : Deletion
    - `11` : Undeletion
    - `12` : Spam
    - `13` : InformModerator
  - CreationDate
  - UserId (only for VoteTypeId 5)
  - BountyAmount (only for VoteTypeId 9)

Job Descriptions

- A Kaggle dataset containing Job Descriptions for several job openings was used.
- We used NLP to filter out the Job Descriptions related to IT industry.
- Finally, 5000+ JDs including JDs for positions like 'Web devloper', 'C++ software developer', 'Software developer', 'Enbedded Software Engineer' were filtered out and saved as ***jd.csv***.

Resumes

- No open source/dataset for Resumes was found.
- We needed resumes in text format. Since extracting proper text from PDF files is a complex problem on it's own.

- Indeed.com was the only site which displayed the resumes openly.
- So, a Python Script(collectCV.py) was used to collect around 300 resumes of applicants for positions like 'Software Developer' , 'Data Scientist', 'Web Developer' etc.

Training Word2Vec Model

Word2Vec models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space

Requirement of training our own models

- There are pre-trained models available both in `gensim` and `spaCy` packages in Python. These models are trained over Google News Data. This implies that they are not suitable for the technically aware context distinction required for this project. For e.g. HTML and Ruby may have higher similarity value in these models than the model we trained.

- Therefore, we required a dataset which was both technically aware and also has sufficient amount of unique words present for the non-technical functioning of the model.

```
The dataset used to train Word2Vec model becomes more crucial considering the fact
that Word2Vec models can be retrained over and over, however,
new Vocabulary cannot be added to the model.
```

- Therefore, we decide to use the `stackexchange/` network data.
- We used the `gensim` implementation (in Python) of Word2Vec to train our model.

Cleaning and Extracting data

- From the `stackexchange/` dataset the `Posts.xml` for each site was used to extract each Post irrespective of whether it's a Question or an Answer. These Posts were extracted as HTML para tags and saved as `paras.txt` in the corresponding subfolder of the site.

- At this stage, each subdirectory of `StackExchange/` which corresponds to the site under StackExchange network, has a new file called `paras.txt`.

- For training the Word2Vec model, we required a sequence of sentences to be streamed from the disk. Each sentence is represented as a list i.e. each element of this list is the word of the sentence.

- So, the `paras.txt` files were used to extract sentences using BeautifulSoup(a Python Library), and saved into `sentences.txt` (for each site), such that the final result is free of formatting and mathematics, code etc.

- These sentences were streamed into the Word2Vec train method for training the model.

Generated Word2Vec Word Embeddings

- Each word is represented as a `300-sized` numpy array (vector).

- Collected **1237328 unique words** from a **corpus of 565919447 raw words** and **32701720 sentences.**

- Running time for the training was around 3hrs.

Extracting sections

- We have some collection of words that are usually the heading in the resumes. For example 'education', 'academic', 'school', 'study', etc will mark the start of the education section

- We iterate over all lines of all resumes, one by one.

- For each line, first, we remove all the blank lines or the lines containing just symbols. Some resumes have a line denoted to just asterisks or dashes.
- Next, we categorize each line into one of the four sections. This is done by calculating its similarity to the existing words. If the similarity is higher than the threshold, we update the section and mark that point, on the other hand, if the similarity if below the threshold, we continue with the previous section.
- This enables us to separate the sections with good enough accuracy.
- Finally, we write each section of a resume in a .csv file after removing the stop words and doing lemmatization.

Assigning scores

- For a given Job Description, we remove all the stop words and do lemmatization, to get a selected few keywords.
- For each keyword found, we find `5` similar words and their corresponding similarity.
- Now, we find `tf-idf` for each word, that we got in step 2.
- The score of the CV is the sum of `tf-idf * similarity` for all words we got in step 2.

Suggestions for Subsequent Work

This section describes our suggestions for the next iteration of the development:

Identifying Sections in Resumes

- After the first iteration, we have enough resumes to create a training dataset. That is, from the resumes, we can extract sentences and assign them labels according to the section they are in. For eg. 'MS from Cambridge' will be labeled as 'Education'.
- This is possible because most of the resumes have similar structure since they share the source.
- This training set can be used to train a sentence classification algorithm (SVM is recommended).
- This algorithm can be used to classify sentences of resumes into different sections.

Word Embeddings

- Word2Vec has two popular implementations:
    - The C Google implementation
    - The Python Gensim implementation The vectors can not be retrained in C implementation. The vectors in Python implementation can be retrained but the Vocabulary can't be added to the model.
- So, the gensim implementation of Doc2Vec should be used instead. It is similar but more flexible. The model can be retrained and Vocabulary can be added to the model as well. Further, vectors for Phrases can be generated more easily since the averaging algorithm is inbuilt.
- Better Tokenization while training model. For eg. Identification of common phrases and generating a single token for it instead of individual words. Like 'New York' is better tokenized as 'new_york' than 'new' and 'york'. This can be achieved by using gensim implementation of Phrases or spaCy.

Scoring Algorithm

- The division of section can be improved by using the currently sorted sections, that is, we can use them for classifying the lines.
- Instead of considering each word individually, we can take phrases together. Like 'software developer' should be treated as a single entity instead of two.
- We give a higher value to words 'python', 'java', etc. over words like 'knowledge', 'experience' etc in keywords of the job description. This can be done by extracting the tags from the `stackoverflow/` data.

- The Algorithm can use any meta-data (if available) about any preferences for the candidates.

Conclusion

However, there is definitely room for improvements, the result is satisfactory enough for the first iteration of the project. Further, most of the pivotal improvements have been mentioned in the previous section. We have learned a lot during the project and hopefully, the project will serve it's purpose in SkyBits as well. The filtering up of CVs has always been subjective process, although, the use of Machine Learning can certainly reduce the unnecessary amount of human effort.

Resources

- spaCy Documentation: https://spacy.io/
- spaCy GitHub Issue Page: https://github.com/explosion/spaCy/issues
- Gensim Word2Vec Documentation: http://radimrehurek.com/gensim/models/word2vec.html
- Gensim Word2Vec GitHub repository: link
- Google Word2Vec: https://code.google.com/archive/p/word2vec/
- GitHub Repository for Doc2Vec Illustration: https://github.com/linanqiu/word2vec-sentiments