Required Software:

1. JRuby (http://jruby.org/)

   JRuby is a Java implementation of Ruby. At the time of working on this project in May 2015, I used a JRuby API for Tabula. I used JRuby version 1.7.20 along with Java 8, although later versions are now available. A direct Windows installer for JRuby can be downloaded at http://jruby.org/download. Java should be set up prior to installing and working with JRuby.

   Examples of JRuby code for tabula can be seen at https://docs.omniref.com/github/tabulapdf/tabula-extractor/0.7.6

2. Tabula

   Tabula (http://tabula.technology/) is a tool to extract tables from PDF documents into a CSV file format. Tabula can be used through a GUI tool or accessed through JRuby and Java APIs.

   For this project, I have used Tabula for Windows, version 0.9.7. Different versions of Tabula can be downloaded at https://github.com/tabulapdf/tabula/releases.

3. Python

   I used Python 2.7 to convert the raw output CSV files from the JRuby-Tabula program into a readable and well organized final output file. This step is necessary because of the structure of the Election Commission of India documents. Data in different parts of the same page is often spaced and presented differently. For example the top half of a page may have multiple columns of data whereas the bottom half may have only one or two columns with different indentations and spacing than the top half.

   Tabula was not able to pick up most of these finer differences and as a consequence, names and numbers are often split across multiple columns or lines and subsequent processing with the Python code I have written is necessary to clean and organize the election data.