***Follow these steps to run the program on a Linux machine using Python 2.7***

Note: A Python 3 version of the code exists which was tested on Windows, however, best results are not guaranteed. This readme file is not intended to be followed step for step if not running on Linux OS. If facing installation issues email [jarlaufer@gmail.com](mailto:jarlaufer@gmail.com)

1. Download and extract .zip

2. Create a new Python 2.7 virtual environment (i.e. for anaconda run **conda create -n yourenvname python=2.7 anaconda**)

3. In the virtual environment navigate to location of master folder

4. Check Java install version

**java -version**

5. If Java is not installed on your system, or the version of Java is earlier than 1.7, install Java, try

**apt install openjdk-8-jre-headless**

6. Install the following packages using pip for the virtual environment by following the steps below

a) Install NLTK and the necessary items:

**pip install nltk**

- open Python and run

**import nltk**

**nltk.download('punkt')**

**nltk.download('stopwords')**

b) Install tabula-py

**pip install tabula-py**

c) Install numpy

**pip install numpy==1.16.1**

d) Install pandas

**pip install pandas**

e) Install gensim

**pip install gensim**

f) Install tqdm

**pip install tqdm**

g) Install scikit-learn dependencies

**pip install scikit-learn==0.19.1**

7. Run the TCmix.py file

**python TCmix.py**

8. Run the doc2vec.py file

**python doc2vec.py**

NOTE: if running on Python 3 run **python doc2vec\_python3.py**