***Follow these steps to run program on Python 2 on a Linux OS***

Note: A Python3 version of the code exists and it yields good results on a Windows machine, however, this readme file is not intended to help setup a windows machine. If facing installation issues email [jarlaufer@gmail.com](mailto:jarlaufer@gmail.com)

1. Download and extract .zip

2. Create a new virtual environment

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**

7. 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

**import nltk**

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

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

- run script

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**

f) Install scikit-learn dependencies

**pip install scikit-learn**

8. Run the TCmix.py file

**python TCmix.py**

9. Run the doc2vec.py file

**python doc2vec.py**