**Readme**

**Software To be Installed (Linux)**

- Python (v2.7.2 or lesser)

- NLTK with corpus package included

**Installation Steps**

- Extract the files in some folder “Project”.

- Install python, using ‘sudo apt-get install python’

- Install NLTK, using ‘sudo apt-get install python-nltk’

- Install NLTK stopwords package:

- start python on command prompt.

> python

- do ‘import nltk’

- type ‘nltk.download()’

- type ‘d stopwords’, this would install stopwords package

- Install matplotlib (Optional to run zipf’s law)

$ curl -O <https://raw.github.com/pypa/virtualenv/master/virtualenv.py>

$ python virtualenv.py my\_new\_env

$ . my\_new\_env/bin/activate

$ (my\_new\_env)$ pip install numpy

$ (my\_new\_env)$ pip install scipy

If the above command fails, go to <http://www.scipy.org/Installing_SciPy/> and read the installation guide or read the steps written below if you are on OSX or Linux.

OSX specific installation for scipy :

$ git clone <https://github.com/scipy/scipy.git>

$ python setup.py build

$ python setup.py install

RPM-based linux OS installation for scipy :

$ yum install python-scipy

Debian-based linux OS installation for scipy :

$ sudo apt-get install python-numpy python-scipy

$ (my\_new\_env) pip install matplotlib

**Running Tasks**

Here are the details to run each program for different tasks (through command line). We assume the python executable is included in the systems path. Please also download and place all the datasets ‘Dataset1’, ‘Dataset2’, ‘Dataset3’, ‘Dataset4’ & ‘EnronDataset’ where the python files have been extracted (inside “Project” folder).

Task 1 (Unigram/Bigram Generation)

> python task\_one.py 2’

Task 2 (Random Sentence Generation)

> python task\_two.py 2

Task 3 (Smoothing/Unknown words/Perplexity)

> python task\_three.py

Task 4 (Perplexity computation)

> sh task\_four.sh

Task 5 (Email Author prediction)

> python email.py

Task 6 (Extension)

> python task\_one.py 4 (for N=4)

> python task\_two.py 4 (For N = 4)

> sh task\_four.sh

> python extension.py (Requires matplotlib library)