ICEDEG TUTORIAL

SOCIAL MEDIA DATA FOR E-GOVERNMENT
Digital Citizens and their Degree of Interest in Politics

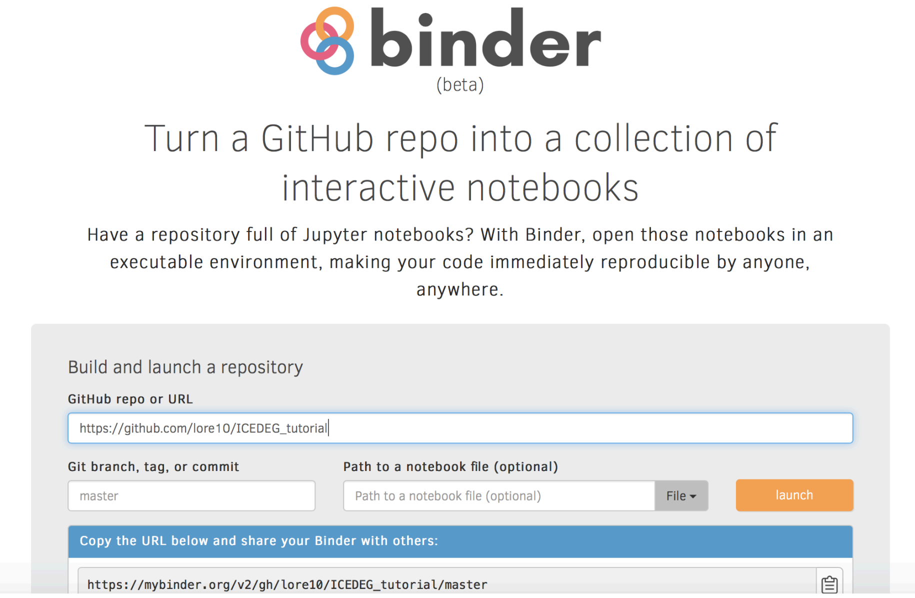
LORENA RECALDE

(If you don’t have IPython Notebook, please follow next steps. Otherwise, download the project from Github in ……… and install the module named word2vec in your python environment. Please go to step 4)

1. Lounch JupyterLab to run (and do) some code. Open your web browser and go to next link. https://mybinder.org/

2. Copy next repository link in box <GitHub repo or URL>. Click on button <launch>
** WAIT few minutes until the project is uploaded.

Repository: https://github.com/lore10/ICEDEG_tutorial



Turn a GitHub repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Build and launch a repository

GitHub repo or URL: https://github.com/lore10/ICEDEG_tutorial

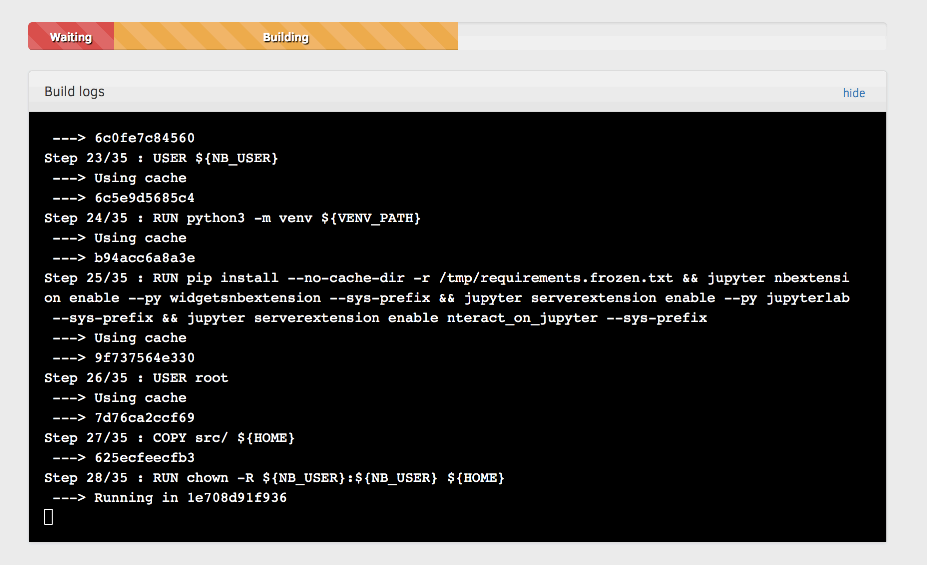
Git branch, tag, or commit: master

Path to a notebook file (optional): Path to a notebook file (optional)

Copy the URL below and share your Binder with others:

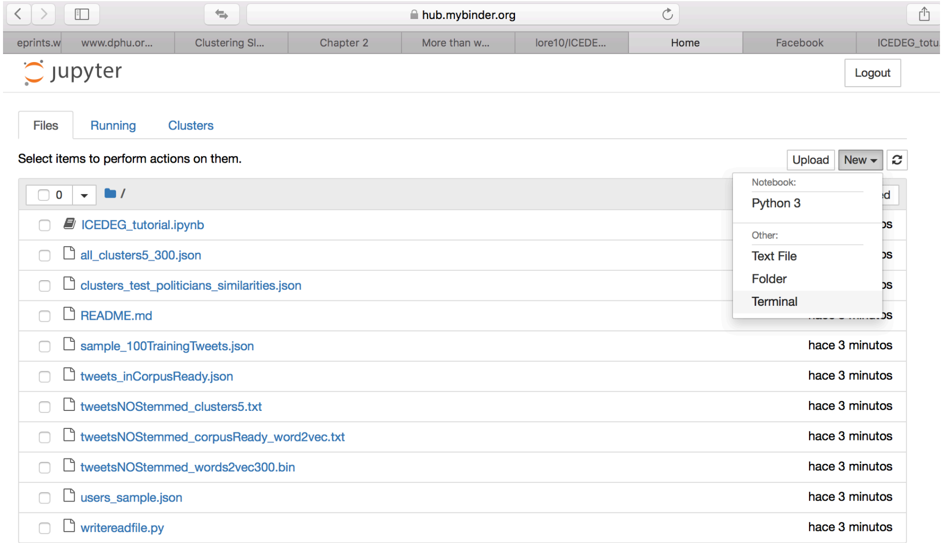
https://mybinder.org/v2/gh/lore10/ICEDEG_tutorial/master

While waiting you will see:



```
 ---> 6c0fe7c84560
Step 23/35 : USER ${NB_USER}
 ---> Using cache
 ---> 6c5e9d5685c4
Step 24/35 : RUN python3 -m venv ${VENV_PATH}
 ---> Using cache
 ---> b94acc6a8a3e
Step 25/35 : RUN pip install --no-cache-dir -r /tmp/requirements.frozen.txt && jupyter nbextensi
on enable --py widgetsnbextension --sys-prefix && jupyter serverextension enable --py jupyterlab
 --sys-prefix && jupyter serverextension enable nteract_on_jupyter --sys-prefix
 ---> Using cache
 ---> 9f737564e330
Step 26/35 : USER root
 ---> Using cache
 ---> 7d76ca2ccf69
Step 27/35 : COPY src/ ${HOME}
 ---> 625ecfeecfb3
Step 28/35 : RUN chown -R ${NB_USER}:${NB_USER} ${HOME}
 ---> Running in 1e708d91f936
```

3. Go to New – Terminal. Install required packages

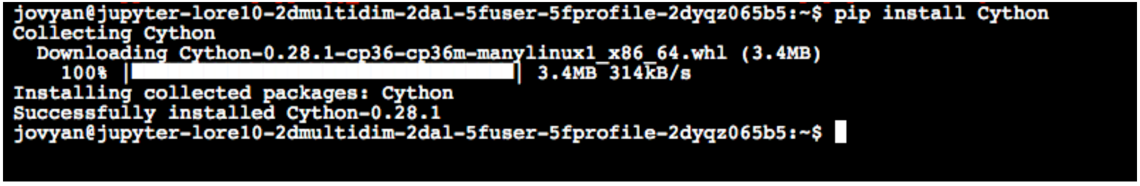


Select items to perform actions on them.

ICEDEG_tutorial.ipynb
all_clusters5_300.json
clusters_test_politicians_similarities.json
README.md
sample_100TrainingTweets.json hace 3 minutos
tweets_inCorpusReady.json hace 3 minutos
tweetsNOStemmed_clusters5.txt hace 3 minutos
tweetsNOStemmed_corpusReady_word2vec.txt hace 3 minutos
tweetsNOStemmed_words2vec300.bin hace 3 minutos
users_sample.json hace 3 minutos
writereadfile.py hace 3 minutos

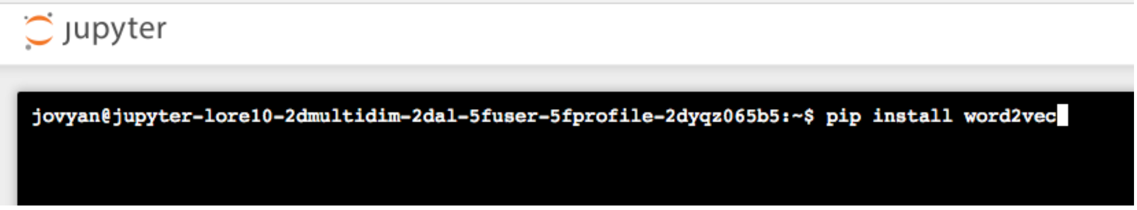
Follow next instructions:

pip install Cython



```
jovyan@jupyter-lore10-2dmultidim-2dal-5fuser-5fprofile-2dyqz065b5:~$ pip install Cython
Collecting Cython
  Downloading Cython-0.28.1-cp36-cp36m-manylinux1_x86_64.whl (3.4MB)
    100% |████████████████████████████████| 3.4MB 314kB/s
Installing collected packages: Cython
Successfully installed Cython-0.28.1
jovyan@jupyter-lore10-2dmultidim-2dal-5fuser-5fprofile-2dyqz065b5:~$
```

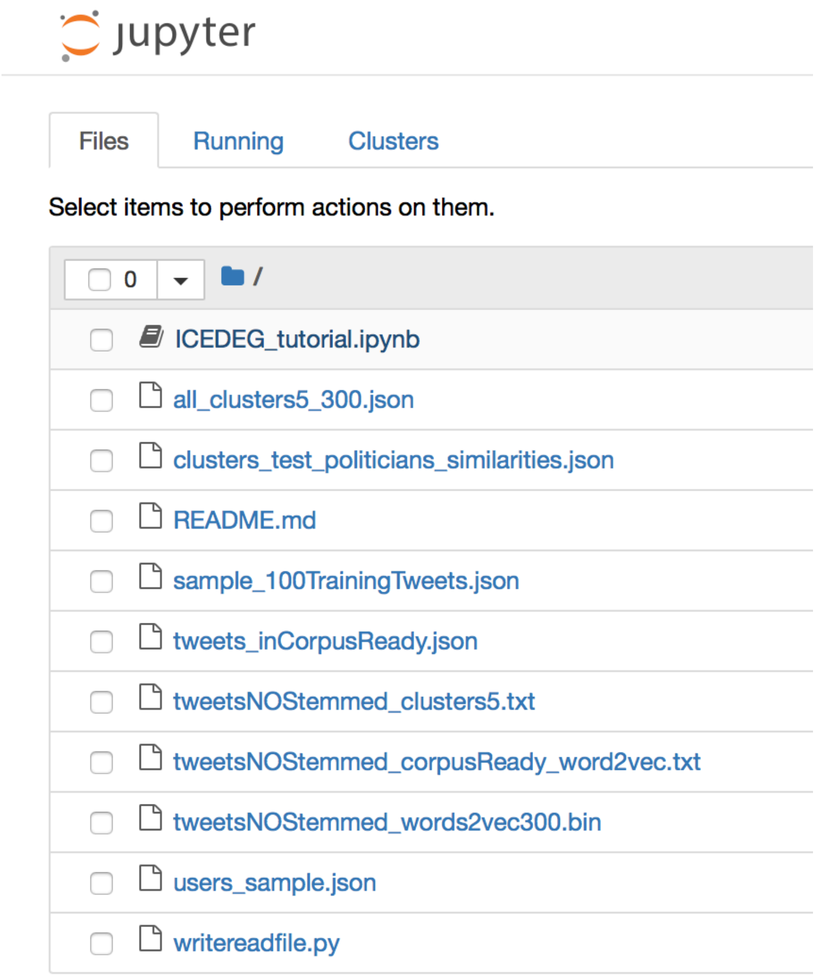
pip install word2vec



```
jovyan@jupyter-lore10-2dmultidim-2dal-5fuser-5fprofile-2dyqz065b5:~$ pip install word2vec
```

Install also:
pip install unidecode
pip install nltk
pip install stop_words
pip install sklearn
pip install scipy
pip install matplotlib

4. Open the file ICEDEG_tutorial.ipynb with one click on it.



Select items to perform actions on them.

- ICEDEG_tutorial.ipynb
- all_clusters5_300.json
- clusters_test_politicians_similarities.json
- README.md
- sample_100TrainingTweets.json
- tweets_inCorpusReady.json
- tweetsNOStemmed_clusters5.txt
- tweetsNOStemmed_corpusReady_word2vec.txt
- tweetsNOStemmed_words2vec300.bin
- users_sample.json
- writereadfile.py

PART 1

Training the word embeddings model, word2vec

5. See the decisions made about data cleaning/preprocessing.

6. Load the preprocessed tweets that were used to train the model. This is only a sample of 100 tweets. Here you have the original tweet and the result when it is preprocessed.

6. To train the model we need to provide as input a txt file with the training tweets (every tweet per line). This file is already prepared to run

```
>> word2vec.word2vec(input_file.txt, output_file.bin, OTHER PARAMETERS)
```

PART 2
Exploring the work of Word embeddings

- Do not forget to import the module word2vec and load the model

7. Follow item 5) in the notebook to see how similarities are found. This is useful to verify the quality of your trained model.

8. Follow item 6) and 7) in the notebook to see some plots in a 2D space.

9. Follow item 8) to work with analogies and item 9) to find how similar two words are.

PART 3
Clustering words in the vocabulary

10. Explore how the words were grouped and if the clusters make sense. How subjective may clustering be? Remember that in our proposal we worked with 5 clusters.

PART 4
How similar is a tweet to the Politics-related centroid?

11. Try some code to classify the given tweets into the corresponding cluster

PART 5
DoIP already calculated for some users

12. How the tweets of politicians are classified?

THANKS…