GEITEC – CURSO EJECUTIVO

LORENA RECALDE

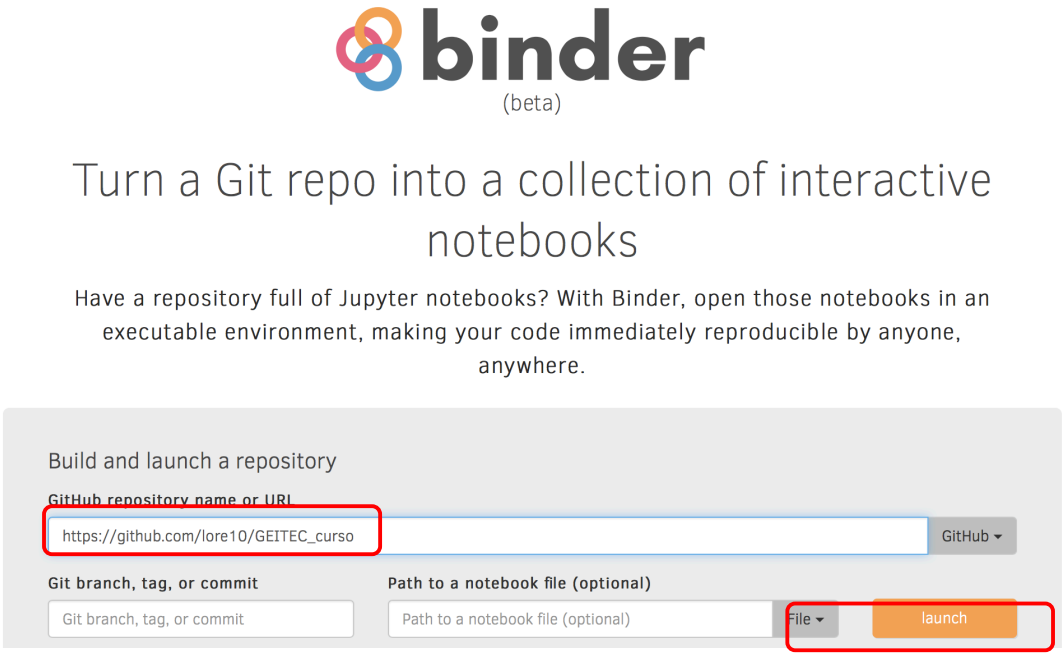
Datos de redes sociales para el Gobierno Electrónico

Caso de Estudio: Digital Citizens and their Degree of Interest in Politics

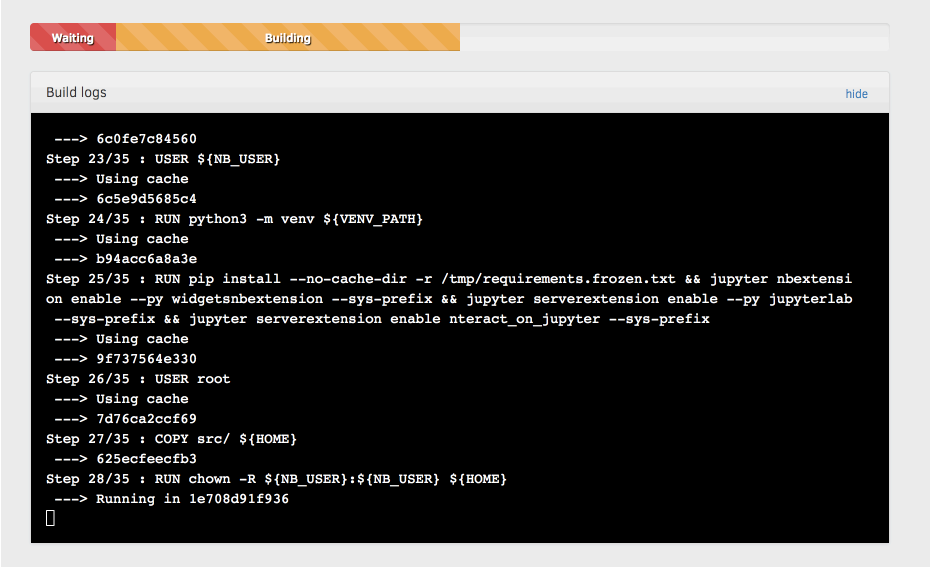
(Si no tiene IPython Notebook, siga los siguientes pasos. De lo contrario, descargue el proyecto desde Github en https://github.com/lore10/GEITEC_curso e instale el módulo llamado word2vec en su entorno de Python. Por favor, vaya al paso 4)

1. Inicie JupyterLab para ejecutar (y hacer) algo de código. Abra su navegador web y vaya al siguiente enlace: https://mybinder.org/

2. Copie el siguiente enlace del repositorio en la casilla <GitHub repo or URL>. Click en el botón <launch>
** ESPERE unos minutos hasta que se cargue el proyecto.
Repository: https://github.com/lore10/ICEDEG_tutorial



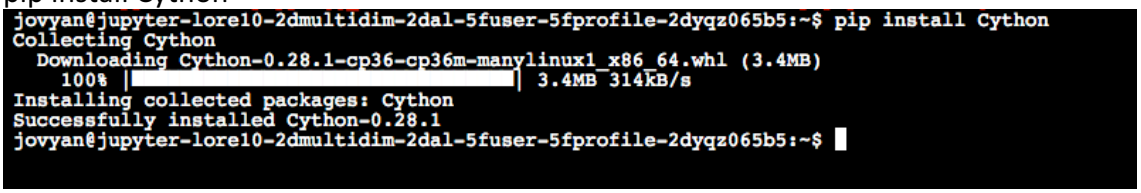
Mientras espera verá:



3. Vaya a New – Terminal. Instale los paquetes que serán necesarios a continuación

Sig alas instrucciones para la instalación:

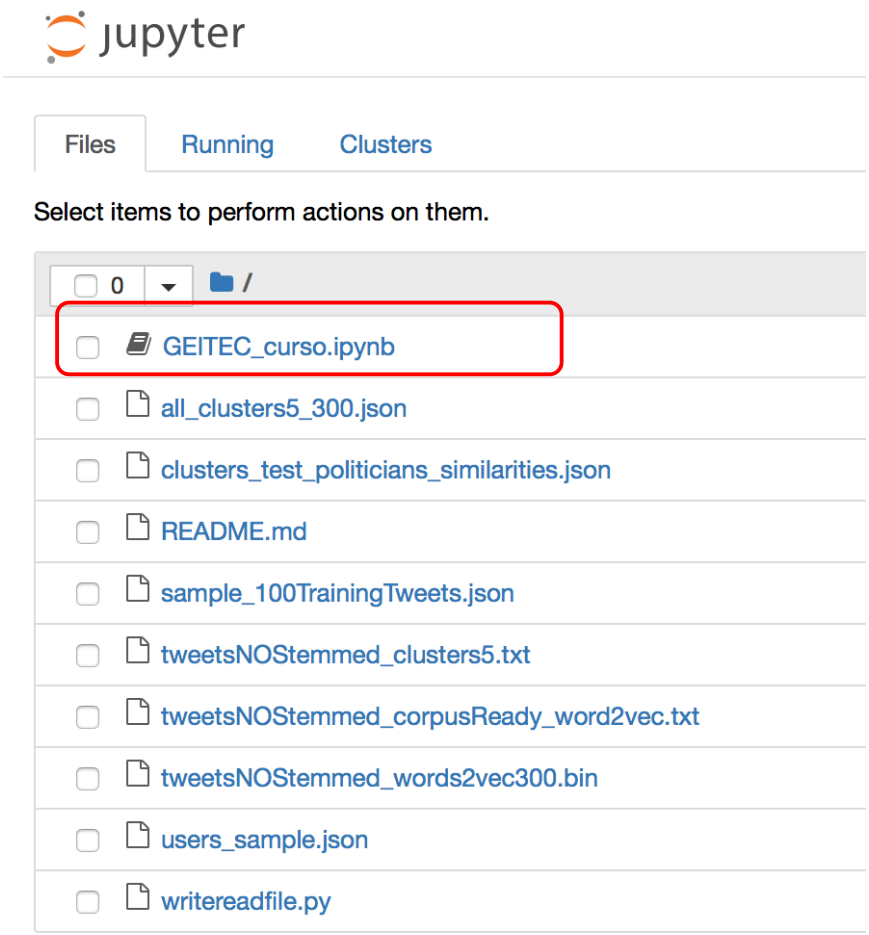
pip install Cython



pip install word2vec

Instale también:
pip install unidecode
pip install nltk
pip install stop_words
pip install sklearn
pip install scipy
pip install matplotlib

4. Abra el archivo GEITEC_curso.ipynb con un click.



PARTE 1

Training the word embeddings model, word2vec

5. Ver las decisiones tomadas sobre la limpieza / preprocesamiento de datos.

6. Cargue los tweets preprocesados que se utilizaron para entrenar el modelo. Esta es solo una muestra de 100 tweets. Aquí tienes el tweet original y el resultado cuando está preprocesado.

6. Para entrenar el modelo debemos proporcionar como entrada un archivo txt con los tweets de entrenamiento (cada tweet por línea). Este archivo ya está preparado para ejecutarse.
>> word2vec.word2vec(input_file.txt, output_file.bin, OTROS PARAMETROS)

PARTE 2
Explorando el trabajo de Word embeddings

** No olvide importar el módulo word2vec y cargar el modelo.

7. Siga el ítem 5) en el JupyterNotebook para ver cómo se encuentran las similitudes. Esto es útil para verificar la calidad de su modelo entrenado.

8. Siga el elemento 6) y 7) en el JupyterNotebook para ver algunos plots en un espacio 2D.

9. Siga el ítem 8) para trabajar con analogías y el ítem 9) para encontrar qué tan parecidas son dos palabras.

PARTE 3
Clustering words in the vocabulary
10. Explore cómo se agruparon las palabras y si los grupos tienen sentido. ¿Qué tan subjetivo puede ser el agrupamiento? Recuerde que en nuestra propuesta trabajamos con 5 clusters.

PARTE 4
¿Qué tan similar es un tweet al centroide relacionado con la política?

11. Pruebe algo de código para clasificar los tweets dados en el clúster correspondiente

PARTE 5
DoIP ya calculado para algunos de los usuarios

12.¿Cómo se clasifican los tweets de los políticos?

Gracias…