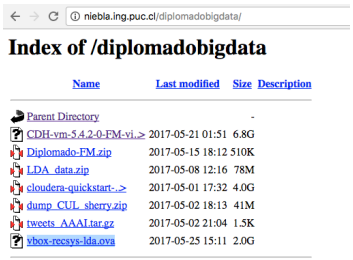
Actividad: LDA (versión máquina virtual)

Objetivo: Aprender a usar el método Latent Dirichlet Allocation (LDA) tanto para hacer clustering como para hacer recomendaciones basadas en contenido.

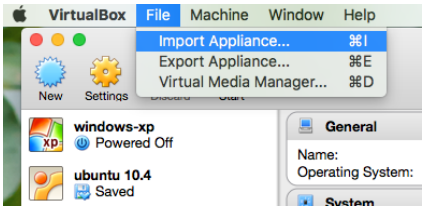
Requisitos:
Para llevar a cabo esta actividad debe descargar e instalar una máquina virtual. Para eso, ejecute los pasos siguientes:

1. Descargar máquina virtual desde:

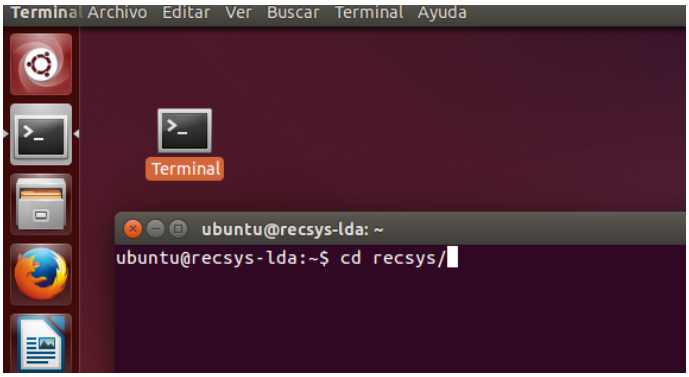
```
http://niebla.ing.puc.cl/diplomadobigdata/vbox-recsys-lda.ova

o desde

http://dparra.sitios.ing.uc.cl/classes/diplomadoBigData/vbox-recsys-lda.ova
  (estará disponible alrededor de las 13:30horas del sábado 27 de mayo)
```



2. Importar máquina en virtualBox. Aparecerá una máquina llamada **vbox-recsys-lda**

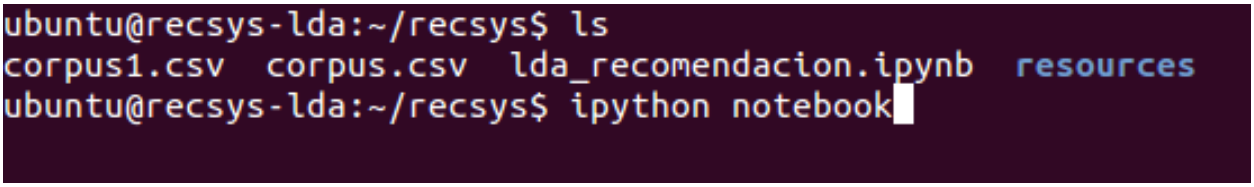


3. Luego de importar la máquina, iniciarla y hacer doble click en icono **Terminal**. En el terminal ingrese a la carpeta recsys tipeando `cd recsys`

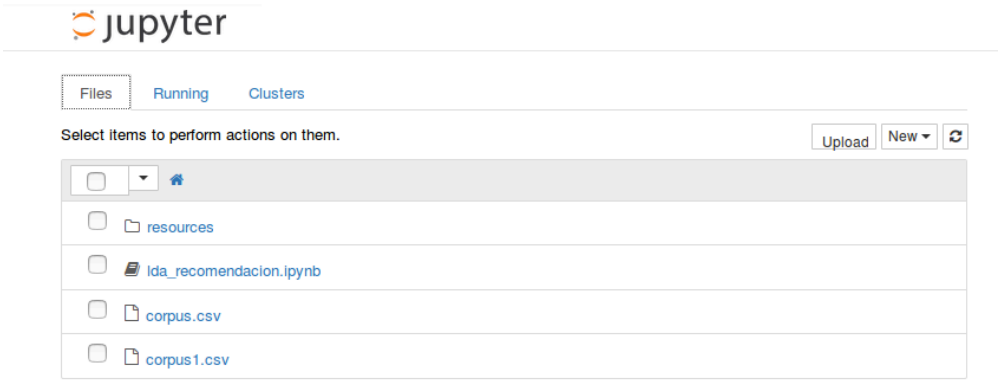


4. Estando ya en la carpeta recsys, inicia el ipython notebook tipeando

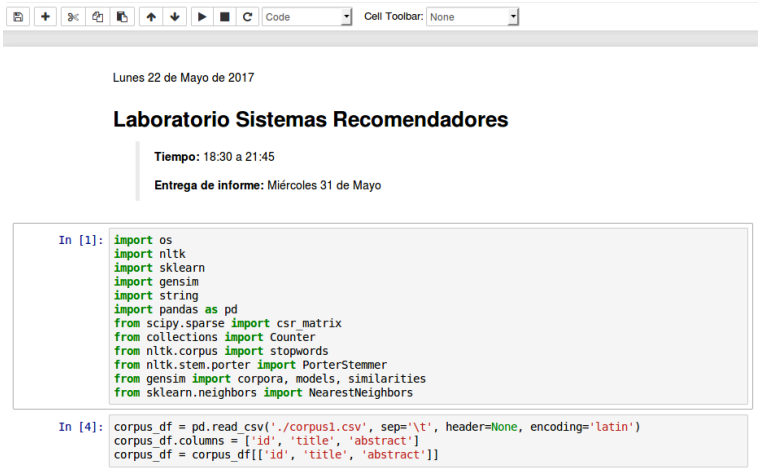
```
ipython notebook
```



a continuación se abre automáticamente Firefox en la dirección `http://localhost:8888/tree`



haga click en **lda_recomendacion.ipynb**, debería ver esto



A continuación, ejecute cada celda, asegurándose que termine de ejecutar antes de pasar a la siguiente. Una celda está en ejecución aún cuando muestra el símbolo **[*]**. Cuando llegue a la sección titulada **Actividad**, siga cada uno de los pasos indicados.

Actividad (Recomendación con TF-DF y LDA)

Ejecute el código siguiente haciendo los siguientes cambios de parametros:

nearest_neighbors : 5, 10, 20 ¿qué efecto tiene el modelo en las recomendaciones observadas?

Eligiendo un valor fijo para nearest neighbors, ejecute **metric = 'cosine'** ¿qué efecto tiene la métrica de distancia en las recomendaciones observadas?

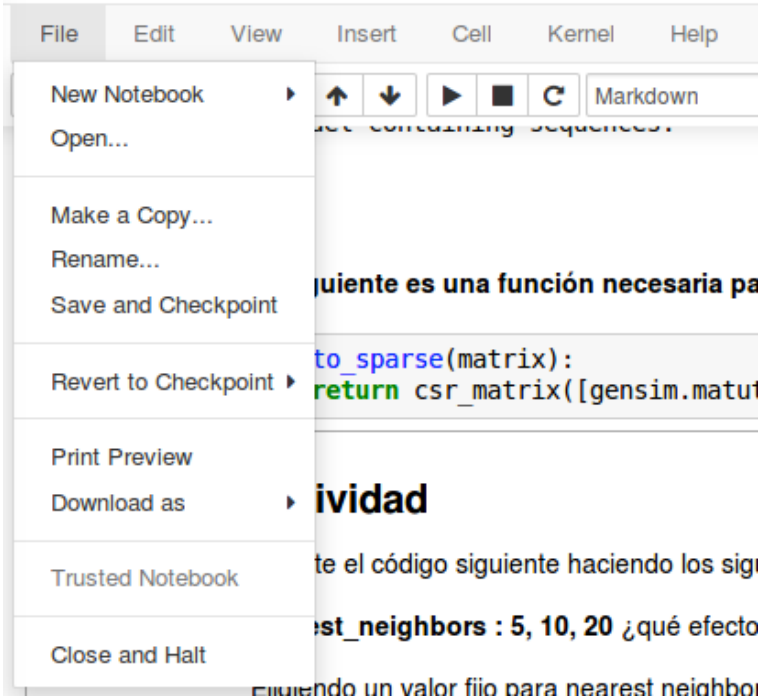
Eligiendo un valor fijo de nearest_neighbors y metric **model : 'lda'** ¿qué efecto tiene el usar LDA versus TF-IDF en las recomendaciones observadas?

Pruebe nuevamente con LDA usando sólo 5 tópicos, rehacer modelo más arriba en **(2)** ¿qué efecto tiene el número de tópicos en las recomendaciones observadas?

Cree todas las celdas adicionales que sea necesario para entregar la tarea probando todas las opciones mencionadas.

Entregable:

Entregue los ipython notebooks ejecutados con sus salidas respectivas, agregando una discusión respecto de los parámetros que modificó. En File > Download as > ipynb


File
Edit
View
Insert
Cell
Kernel
Help
New Notebook
Open...
Make a Copy...
Rename...
Save and Checkpoint
Revert to Checkpoint
Print Preview
Download as
Trusted Notebook
Close and Halt
Markdown
guiente es una función necesaria pa
to_sparse(matrix):
return csr_matrix([gensim.matut
ividad
te el código siguiente haciendo los sigu
est_neighbors : 5, 10, 20 ¿qué efecto