



College of Computer Science and Engineering
Department of Computer Science and Artificial Intelligence

CCAI-413: Natural Language Processing
Lab#1 Introduction to Kaggle and NLTK

Objectives

- Introduction to Kaggle and NLTK

Lab Tool(s)

<https://www.kaggle.com/>

What is Kaggle?

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to **find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.**

Kaggle got its **start in 2010 by offering machine learning competitions** and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Goldbloom and Jeremy Howard. Nicholas Gruen was founding chair succeeded by Max Levchin. Equity was raised in 2011 valuing the company at \$25 million. On 8 March 2017, Google announced that they were acquiring Kaggle.

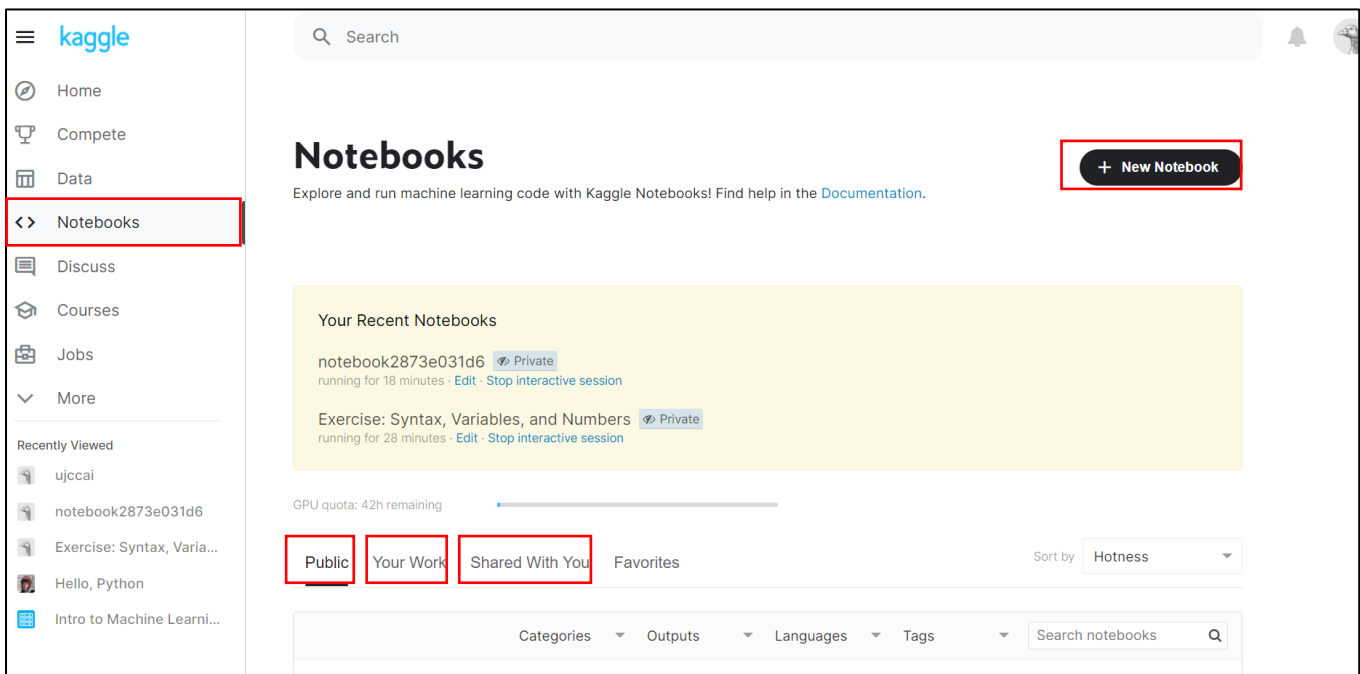
Getting Started With Kaggle

1. Open Kaggle and create an account
2. Watch: [A Quick tour on Kaggle](#)

Notebooks

Explore and run machine learning code with Kaggle Notebooks, a cloud computational environment that enables reproducible and collaborative analysis.

Creating a Notebook



Supported Languages

Python and R

Types of Notebooks

There are two different types of Notebooks on Kaggle.

1. Scripts

The first type is a script. Scripts are files that execute everything as code sequentially. To start a script, click on “Create Notebook” and select “Script”. This will open the Scripts editing interface.

2. Notebooks

The second type is Jupyter notebooks (usually just “notebooks”). Jupyter notebooks consist of a sequence of cells, where each cell is formatted in either Markdown (for writing text) or in a programming language of your choice (for writing code). To start a notebook, click on “Create Notebook”, and select “Notebook”. This will open the Notebooks editing interface.

Renaming, Saving and Sharing Notebooks



Using the Notebook Editor



The Notebook editor allows you to write and execute both traditional Scripts (for code-only files ideal for batch execution or Rmarkdown scripts) and Notebooks (for interactive code and markdown editor ideal for narrative analyses, visualizations, and sharing work).

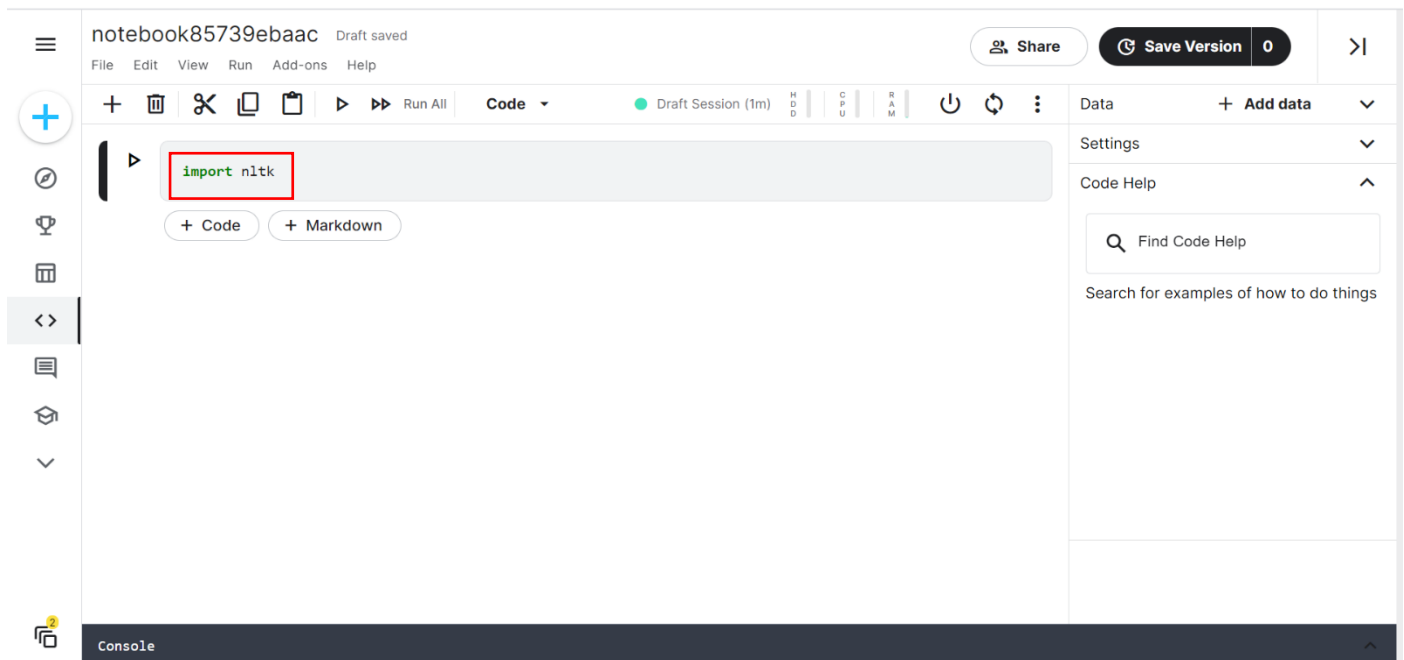
Getting Started with NLTK

What is NLTK?

“Natural Language Toolkit (NLTK) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.”

Import NLTK in Python

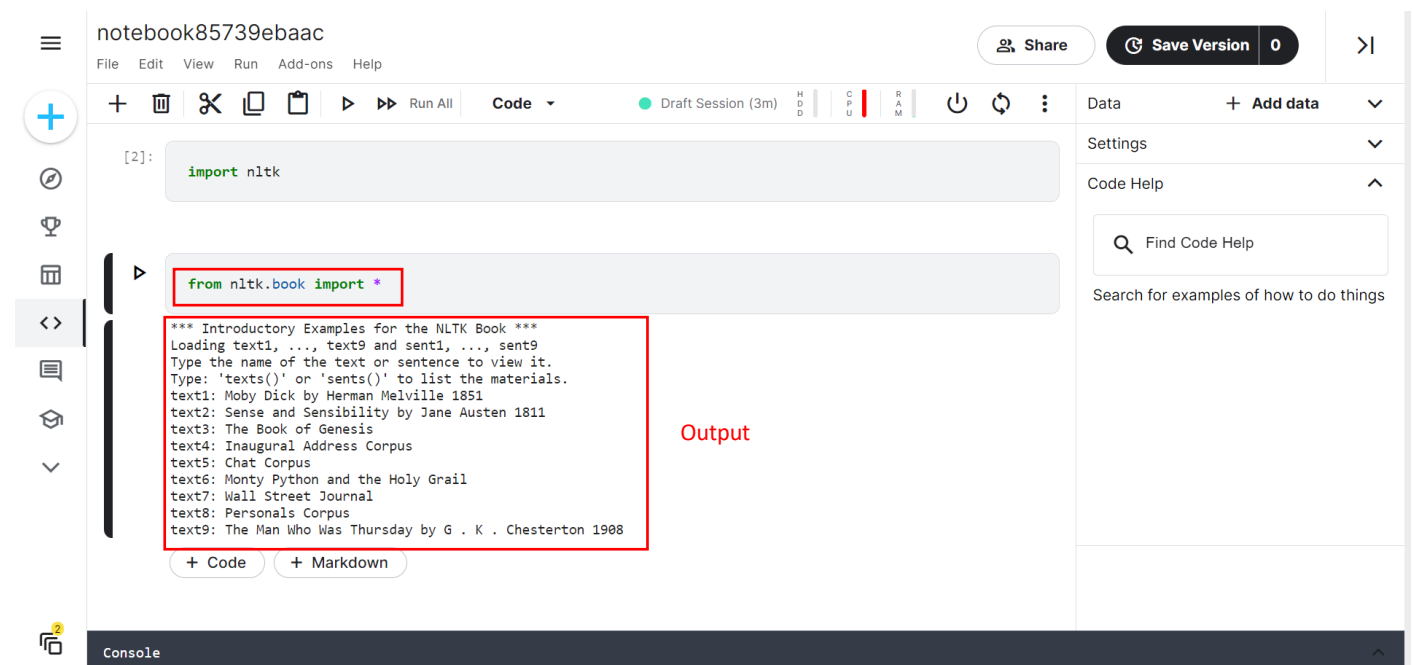
Type the following command to import NLTK at the Python prompt:



The screenshot shows a Jupyter Notebook interface. The top bar indicates the notebook is named 'notebook85739ebaac' and is a 'Draft saved'. The menu bar includes 'File', 'Edit', 'View', 'Run', 'Add-ons', and 'Help'. The toolbar shows various icons for file operations and execution. The code cell contains the command `import nltk`, which is highlighted with a red box. Below the code cell are buttons for '+ Code' and '+ Markdown'. The right sidebar contains sections for 'Data', 'Settings', and 'Code Help', with a search bar for 'Find Code Help'. The bottom status bar shows 'Console'.

Download NLTK book Data

NLTK offers data of book collections that required for the examples and exercises in this lab. Type the following command to import the data, where (*) means to import all the data.



The screenshot shows a Jupyter Notebook interface. The top bar indicates the notebook is named 'notebook85739ebaac' and is a 'Draft saved'. The menu bar includes 'File', 'Edit', 'View', 'Run', 'Add-ons', and 'Help'. The toolbar shows various icons for file operations and execution. The code cell contains the command `from nltk.book import *`, which is highlighted with a red box. Below the code cell are buttons for '+ Code' and '+ Markdown'. The output of the command is displayed in a separate cell, showing a list of text and sentence corpora. The output is highlighted with a red box. The right sidebar contains sections for 'Data', 'Settings', and 'Code Help', with a search bar for 'Find Code Help'. The bottom status bar shows 'Console'.

Output

```
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

Type the text name to find out more about it:

The screenshot shows a Jupyter Notebook titled 'notebook85739ebaac'. The interface includes a top bar with 'File', 'Edit', 'View', 'Run', 'Add-ons', and 'Help' menus. Below the menu bar is a toolbar with icons for adding, deleting, copying, pasting, and running code. The main area displays a code cell with the following text:

```
[3]: from nltk.book import *  
  
*** Introductory Examples for the NLTK Book ***  
Loading text1, ..., text9 and sent1, ..., sent9  
Type the name of the text or sentence to view it.  
Type: 'texts()' or 'sents()' to list the materials.  
text1: Moby Dick by Herman Melville 1851  
text2: Sense and Sensibility by Jane Austen 1811  
text3: The Book of Genesis  
text4: Inaugural Address Corpus  
text5: Chat Corpus  
text6: Monty Python and the Holy Grail  
text7: Wall Street Journal  
text8: Personalads Corpus  
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

Below the code cell, a variable lookup box shows 'text1' selected, and the output is displayed as '<Text: Moby Dick by Herman Melville 1851>'. The output is labeled 'Output' in red text. At the bottom of the notebook, there are buttons for '+ Code' and '+ Markdown'. The right sidebar contains a 'Data' panel with '+ Add data', 'Settings', and 'Code Help' options, along with a search bar for 'Find Code Help'.

References:

<https://www.kaggle.com/>