**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](https://www.youtube.com/watch?v=O1P4r0Iy55U)

# Notebooks

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

## Creating a Notebook

Graphical user interface, text, application

Description automatically generated

### 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.

1. 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

Graphical user interface, text, application

Description automatically generated

### Using the Notebook Editor

# Graphical user interface, text, email Description automatically generated

3. Settings

2. Console

1. Editing Window

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:

Graphical user interface, text, application, email

Description automatically generated

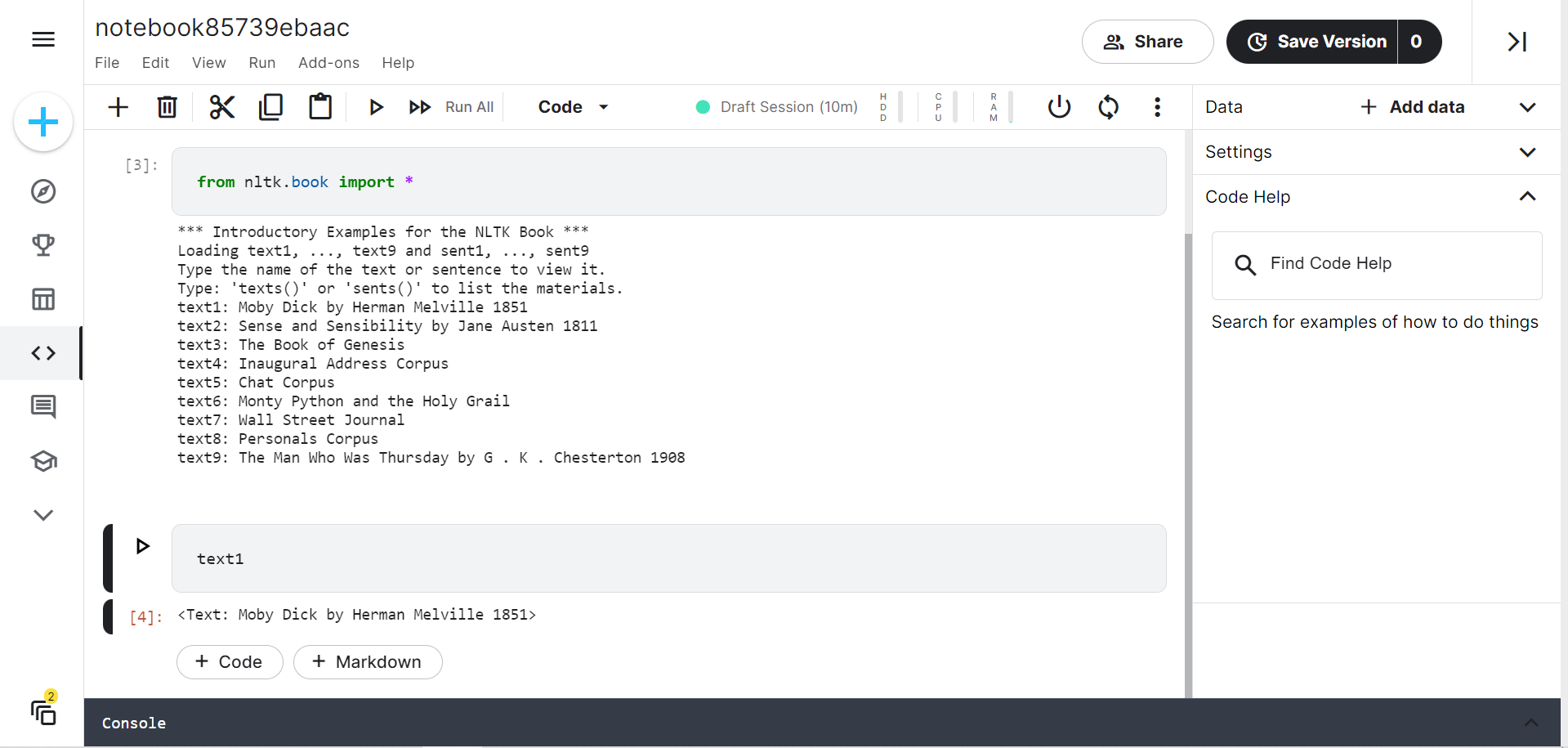
## Download NLTK book Data

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.

# 

Output

Type the text name to find out more about it:



Output

# References:

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