

# Lab 5 Kaggle Introduction

COMP4901K and MATH 4824B

Fall 2018

## Prerequisites

- You need to sign up for <https://www.kaggle.com> before this lab.

## 1 Kaggle Brief Introduction

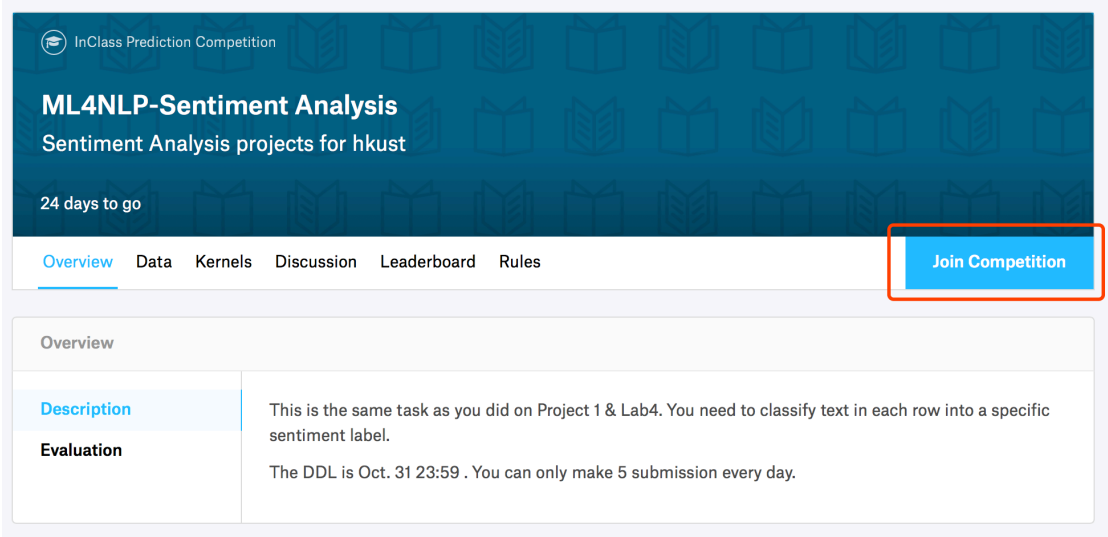
### 1.1 What is Kaggle

**Kaggle** is the world's largest community of data scientists and machine learners. Kaggle got its start by offering machine learning competitions now also offers a public data platform, a cloud-based workbench for data science and short form AI education. On 8 March 2017, Google announced that they were acquiring Kaggle.

### 1.2 Join in our In-class Competition

Now join in our in-class Kaggle competition! Click this url

<https://www.kaggle.com/t/6d1e19c4d1064b189f76fa83a8b4ce4b>



The screenshot shows the Kaggle competition interface for 'ML4NLP-Sentiment Analysis'. The header includes the competition title, a subtitle 'Sentiment Analysis projects for hkust', and a countdown '24 days to go'. A navigation bar contains links for Overview, Data, Kernels, Discussion, Leaderboard, and Rules. A prominent blue 'Join Competition' button is highlighted with a red rectangle. Below the navigation bar, the 'Overview' section is visible, containing a 'Description' and an 'Evaluation' section.

| Overview    |  |
|-------------|--|
| Description | This is the same task as you did on Project 1 & Lab4. You need to classify text in each row into a specific sentiment label. |
| Evaluation  | The DDL is Oct. 31 23:59 . You can only make 5 submission every day.   |

Then click “Join Competition”.

## 1.3 Explore the Competition

### 1.3.1 Overview

The task of this competition is the same as Lab4 / Project 1, just to use machine learning algorithm to do sentiment analysis. You can use any program language you like and any machine learning algorithms to boost the performance of your classifier.

### 1.3.2 Data

You need to first download the overall dataset from “Data” section of the competition.

Overview Data Kernels Discussion Leaderboard Rules [Submit Predictions](#)

#### Data Description

##### File descriptions

- **train.csv** - the training set and labels (16,000 data). Each row contains three fields: "id", "text", "label"
- **test.csv** - the test set and labels (4,491 data). Each row contains two fields: "id", "text", "label". "label" is set to -1 by default. It should be predicted by your model
- **sample\_submission.csv** - a sample submission file in the correct format

##### Data fields

- **id** - Id of this sample
- **text** - Text you need to analyze
- **label** - the sentiment of the text

Data (5 MB) [API](#) `kaggle competitions download -c m14nlp-sentiment...` [Download All](#)

| Data Sources   | About this file    | Columns   |
|--|--------------------|---|
| <ul style="list-style-type: none"><li>sample_submission.csv 4491 x 2</li><li>test.csv 4491 x 3</li><li>train.csv 16.0k x 3</li></ul> | No description yet | <ul style="list-style-type: none"><li>id</li><li>pred</li></ul> |

### 1.3.3 Kernel

Overview Data Kernels Discussion Leaderboard Rules Team Host [New Kernel](#)

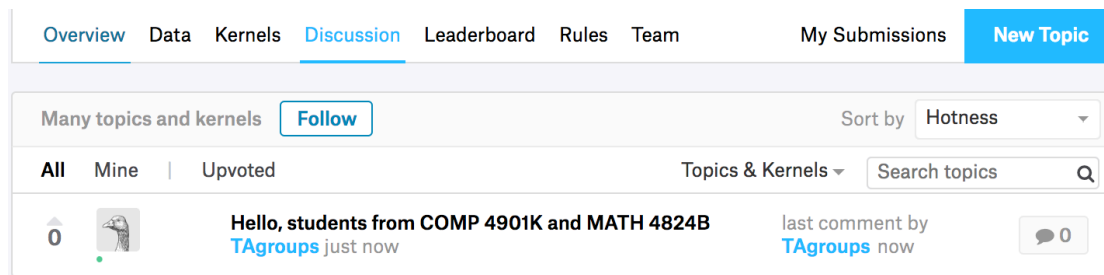
Public Your Work Favorites Sort by Hotness

Outputs Languages Types Tags Search kernels

1 **A simple EDA**  
~10s ago Py 0

You can use Kaggle kernel to run your scripts or notebook. Also, you can share your kernels in public.

### 1.3.4 Discussion



You can share your ideas and questions in the discussion field.

### 1.3.5 Leaderboard

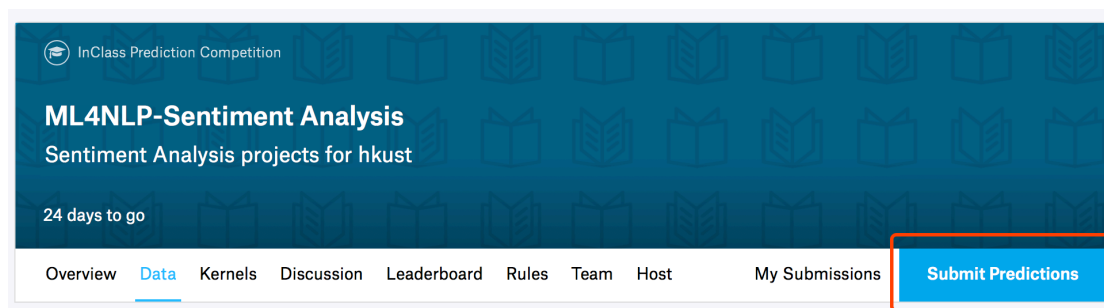
Your Score & Rank are shown in the Leaderboard. Every time you make a submission, the leaderboard will be updated.

## 1.4 Make a valid submission

Open the `sample_submission.csv`, your job is to use your model to generate a submission of this format.

We first Run the Lab 4 code to generate a submission `lab4_submission.csv`

Then click “Submit Predictions”.

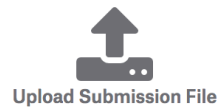


And then upload `lab4_submission.csv`,

You have 5 submissions remaining today. This resets 17 hours from now (00: 00 UTC).

### Step 1

Upload submission file



lab4\_submission.csv (29.62 KB)

Complete 100% 29.62 KB

#### File Format

Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

#### Number of Predictions

We expect the solution file to have 4491 prediction rows. This file should have a header row. Please see sample submission file on the [data page](#).

### Step 2

Describe submission

**B** *I* **H** Styling with Markdown supported

this is my first submission!

Make Submission


Then go back to the leaderboard to check our result

Public LeaderboardPrivate Leaderboard

This leaderboard is calculated with all of the test data.

Raw Data

Refresh

| # | Δ1w | Team Name             | Kernel | Team Members  | Score ? | Entries | Last |
|---|-----|-----------------------|--------|---|---------|---------|------|
| 📍 |     | baseline - bonus.csv  |        |   | 0.63259 |         |      |
| 📍 |     | baseline - 100pts.csv |        |   | 0.63126 |         |      |
| 📍 |     | baseline - 90pts.csv  |        |   | 0.61990 |         |      |
| 📍 |     | baseline - 80pts.csv  |        |   | 0.59652 |         |      |
| 1 | new | TAgroups              |        |  | 0.58427 | 1       | 14m  |
| 📍 |     | baseline - 60pts.csv  |        |   | 0.58383 |         |      |

We already got 60 pts! Cheers !

## 1.5 Rules

- (1) You're supposed to finish this project on your own. Plagiarism and Team work is not allowed.
- (2) You can only make 5 submissions every day.
- (3) Grades are given according to your scores on the leaderboard.

## 2 Intel Cloud Sign up

### 2.1 What is Intel Cloud

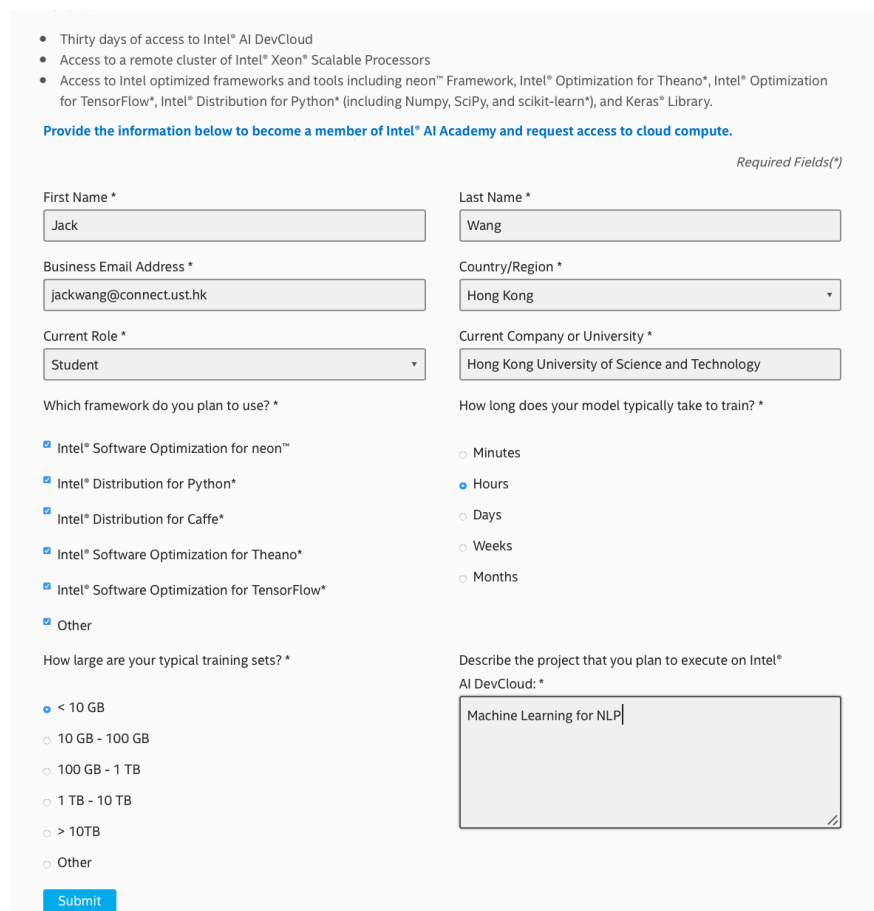
Intel Cloud provides you with remote computing resources to finish your project. Details will be introduced in Lab 6.

### 2.2 Sign up for intel cloud

Before using Intel Cloud, you need to sign up with the following link

[https://plan.seek.intel.com//ww\\_en\\_software\\_registration-form-intelaidevcloudsignup.html?registration\\_source=salesforce&activityID=OPTY-0015197](https://plan.seek.intel.com//ww_en_software_registration-form-intelaidevcloudsignup.html?registration_source=salesforce&activityID=OPTY-0015197)

Sample form is as follows:



The image shows a sample registration form for Intel AI DevCloud. At the top, there are three bullet points listing benefits: 30 days of access to Intel AI DevCloud, access to a remote cluster of Intel Xeon Scalable Processors, and access to Intel optimized frameworks and tools. Below this is a blue link to provide information to become a member of Intel AI Academy and request access to cloud compute. The form itself is titled 'Required Fields(\*)' and contains several sections. The first section has two columns: 'First Name \*' with the value 'Jack' and 'Last Name \*' with the value 'Wang'. The second section has 'Business Email Address \*' with the value 'jackwang@connect.ust.hk' and 'Country/Region \*' with a dropdown menu showing 'Hong Kong'. The third section has 'Current Role \*' with a dropdown menu showing 'Student' and 'Current Company or University \*' with the value 'Hong Kong University of Science and Technology'. The fourth section has 'Which framework do you plan to use? \*' with several checkboxes, all of which are checked: 'Intel Software Optimization for neon', 'Intel Distribution for Python', 'Intel Distribution for Caffe', 'Intel Software Optimization for Theano', 'Intel Software Optimization for TensorFlow', and 'Other'. The fifth section has 'How long does your model typically take to train? \*' with radio buttons for 'Minutes', 'Hours' (selected), 'Days', 'Weeks', and 'Months'. The sixth section has 'How large are your typical training sets? \*' with radio buttons for '< 10 GB' (selected), '10 GB - 100 GB', '100 GB - 1 TB', '1 TB - 10 TB', '> 10TB', and 'Other'. The seventh section has 'Describe the project that you plan to execute on Intel AI DevCloud: \*' with a text area containing the text 'Machine Learning for NLP'. At the bottom left of the form is a blue 'Submit' button.

- Thirty days of access to Intel® AI DevCloud
- Access to a remote cluster of Intel® Xeon® Scalable Processors
- Access to Intel optimized frameworks and tools including neon™ Framework, Intel® Optimization for Theano®, Intel® Optimization for TensorFlow®, Intel® Distribution for Python® (including Numpy, SciPy, and scikit-learn®), and Keras® Library.

[Provide the information below to become a member of Intel® AI Academy and request access to cloud compute.](#)

Required Fields(\*)

First Name \*  
Jack

Last Name \*  
Wang

Business Email Address \*  
jackwang@connect.ust.hk

Country/Region \*  
Hong Kong

Current Role \*  
Student

Current Company or University \*  
Hong Kong University of Science and Technology

Which framework do you plan to use? \*

- ☒ Intel® Software Optimization for neon™
- ☒ Intel® Distribution for Python\*
- ☒ Intel® Distribution for Caffe\*
- ☒ Intel® Software Optimization for Theano\*
- ☒ Intel® Software Optimization for TensorFlow\*
- ☒ Other

How long does your model typically take to train? \*

- ☐ Minutes
- ☒ Hours
- ☐ Days
- ☐ Weeks
- ☐ Months

How large are your typical training sets? \*

- ☒ < 10 GB
- ☐ 10 GB - 100 GB
- ☐ 100 GB - 1 TB
- ☐ 1 TB - 10 TB
- ☐ > 10TB
- ☐ Other

Describe the project that you plan to execute on Intel® AI DevCloud: \*

Machine Learning for NLP

Submit