Kaggle

Reference: Kaggle presentation

About Kaggle

- Biggest platform for competitive data science in the world
- 18.1 Million Kaggle user accounts
- Great platform to learn about the latest techniques and avoid under/overfitting
- Great platform to share and meet up with other data nerds
- Kaggle ≠ All Data Scientist Job Tasks
  - Try not to confuse Kaggle objectives versus all aspects of Data Scientist job
  - Certain aspects overlap but not everything

https://www.kaggle.com/code/carlmcbrideellis/kaggle-in-numbers

What is Kaggle used for?

Kaggle is primarily used for data science competitions, where participants can compete with each other to create the best models for solving specific problems. Organizations from around the world sponsor these competitions, and they cover a wide range of topics, such as image classification, natural language processing, and predictive modeling.

Kaggle is also used for:

- **Learning**: Kaggle provides resources such as public data sets, machine learning tutorials, and code notebooks that allow users to learn and practice data science skills.
- **Collaboration**: Kaggle allows users to form teams and collaborate on submissions, share code and data sets, and provide feedback to each other.
- **Community building**: Kaggle has a large community of data scientists, machine learning engineers, and data enthusiasts, providing a platform for users to connect, share ideas, and collaborate on projects.
- **Research**: Kaggle's data sets and competitions are impactful for research purposes, making it a platform for testing and improving machine learning algorithms.

Overall, Kaggle is a versatile platform that offers a range of opportunities for data scientists and machine learning engineers, from learning and collaboration to research.

General Strategy for Competitions

- Get a good score as fast as possible
  - Failing fast and failing often / Agile sprint / Iteration
- Using versatile libraries, ensembling, experiment tracking, etc.
- Create ML / AI models with optimized pipelines that are:

Data agnostic	(Sparse, dense, missing values, larger than memory [VRAM and/or RAM])
Problem agnostic	(Classification, regression, clustering)
Solution agnostic	(Production-ready, PoC, latency)
Automated	(Turn on and go to bed)
Memory-friendly	(Don’t want to pay lots of $)
Robust	(Good for generalization, concept drift, consistent)

The Learning Agency Lab - PII Data Detection

Kaggle Competition
Jan. 17, 2024 – April 23, 2024

References were taken from Kaggle

Overview

The goal of this competition is to develop a model that detects personally identifiable information (PII) in student writing. Your efforts to automate the detection and removal of PII from educational data will lower the cost of releasing educational datasets. This will support learning science research and the development of educational tools.

Reliable automated techniques could allow researchers and industry to tap into the potential that large public educational datasets offer to support the development of effective tools and interventions for supporting teachers and students.

For this competition, Vanderbilt has partnered with The Learning Agency Lab, an Arizona-based independent nonprofit focused on developing the science of learning-based tools and programs for the social good.

Evaluation: F5-Score

Submissions are evaluated on micro F_β, which is a classification metric that assigns value to recall and precision. The value of β is set to 5, which means that recall is weighted 5 times more heavily than precision.

F_β score [edit]

A more general F score, F_β, that uses a positive real factor β, where β is chosen such that recall is considered β times as important as precision, is:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}.$$

Dataset Description - PII Types

The competition asks competitors to assign labels to the following seven types of PII:

- `NAME_STUDENT` - The full or partial name of a student that is not necessarily the author of the essay. This excludes instructors, authors, and other person names.
- `EMAIL` - A student's email address.
- `USERNAME` - A student's username on any platform.
- `ID_NUM` - A number or sequence of characters that could be used to identify a student, such as a student ID or a social security number.
- `PHONE_NUM` - A phone number associated with a student.
- `URL_PERSONAL` - A URL that might be used to identify a student.
- `STREET_ADDRESS` - A full or partial street address that is associated with the student, such as their home address.

Dataset Description - File and Field Information

annotations. The documents were tokenized using the SpaCy English tokenizer.

Token labels are presented in BIO (Beginning, Inner, Outer) format. The PII type is prefixed with "B-" when it is the beginning of an entity. If the token is a continuation of an entity, it is prefixed with "I-". Tokens that are not PII are labeled "O".

	document	full_text	tokens	trailing_whitespace	labels
0	7	Design Thinking for innovation reflexion-Avril...	[Design, Thinking, for, innovation, reflexion,...	[True, True, True, True, False, False, True, F...	[O, O, O, O, O, O, O, O, O, B-NAME_STUDENT, I-...
1	10	Diego Estrada\n\nDesign Thinking Assignment\n\...	[Diego, Estrada, \n\n, Design, Thinking, Assig...	[True, False, False, True, True, False, False,...	[B-NAME_STUDENT, I-NAME_STUDENT, O, O, O, O, O...
2	16	Reporting process\n\nby Gilberto Gamboa\n\nCha...	[Reporting, process, \n\n, by, Gilberto, Gambo...	[True, False, False, True, True, False, False,...	[O, O, O, O, B-NAME_STUDENT, I-NAME_STUDENT, O...
3	20	Design Thinking for Innovation\n\nSindy Samaca...	[Design, Thinking, for, Innovation, \n\n, Sind...	[True, True, True, False, False, True, False, ...	[O, O, O, O, O, B-NAME_STUDENT, I-NAME_STUDENT...
4	56	Assignment: Visualization Reflection Submitt...	[Assignment, :, , Visualization, , Reflecti...	[False, False, False, False, False, False, Fal...	[O, O, O, O, O, O, O, O, O, O, O, B-NAME_ST...

Example Data Showing the BIO Formatting

Waseem [B-NAME_STUDENT] Mabunda [I-NAME_STUDENT] 591 [B-STREET_ADDRESS] Smith [I-STREET_ADDRESS] Centers [I-STREET_ADDRESS] Apt [I-STREET_ADDRESS] . [I-STREET_ADDRESS] 656 [I-STREET_ADDRESS]
[I-STREET_ADDRESS] Joshuamouth [I-STREET_ADDRESS] , [I-STREET_ADDRESS] RI [I-STREET_ADDRESS] 95963 [I-STREET_ADDRESS] (The Netherlands) 410.526.1667 [B-PHONE_NUM]
vpi@mn.nl

Mind Mapping, Challenge: For several years I have been working for an Asset manager in the Netherlands. During this period I have been involved in many projects. Certainly in the world of asset management, much has changed in recent years in the area of Law and Regulations. What I mainly

Image taken from: https://www.kaggle.com/code/snassimr/pii-data-detection-eda

Dataset Quantity and Leaderboard (LB)

Dataset Quantity

- Total # of Essays: ~22K essays
- Training = 4,768 essays
- Leaderboard
  - Public: ~4,300
  - Private: ~12,900

Leaderboard (LB)

- Public LB scores are shown throughout the competition.
- Private LB scores determines your final placement.
  - You select 3 submission files.
  - Trust your Cross-Validation because its easy to overfit to the Public LB!
  - Think of the Private LB as putting a model into production and seeing how it performs

Label Counts in Training Data

```
Num. Labels: 13
labels
O                      3488563
B-NAME_STUDENT             972
I-NAME_STUDENT             739
B-URL_PERSONAL              72
B-ID_NUM                    56
B-EMAIL                     34
I-STREET_ADDRESS            20
I-PHONE_NUM                 12
B-USERNAME                   5
B-PHONE_NUM                  5
I-URL_PERSONAL               2
B-STREET_ADDRESS             2
I-ID_NUM                     2
```

* Considerable class imbalance
* Few observations for certain classes

Submissions are Made Using Kaggle's Infrastructure

Kaggle's Infrastructure is used to make a submission but you can develop with any compute.

- Jupyter Notebooks (similar feel to Colab)
- Kaggles hardware (similar to going to a Prod. env. with hardware requirements)
  - Limited GPU (16GB - single GPU), CPU (2 cores), Memory (32GB), etc.
- You can upload data files (e.g., models, scripts, data sets, etc.)
- It puts everyone on a level playing field for submissions!

Submissions to this competition must be made through Notebooks. In order for the "Submit" button to be active after a commit, the following conditions must be met:

- CPU Notebook <= 9 hours run-time
- GPU Notebook <= 9 hours run-time
- Internet access disabled
- Freely & publicly available external data is allowed, including pre-trained models
- Submission file must be named `submission.csv`

Thank you!

Myles Dunlap
May 2024
https://github.com/mddunlap924