Machine Learning HW5

TAs
ntu.mlta@gmail.com

Outline

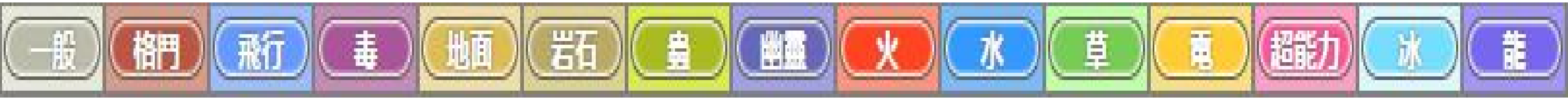
1. Task Introduction
2. Kaggle
3. Deadline and Policy
4. FAQ

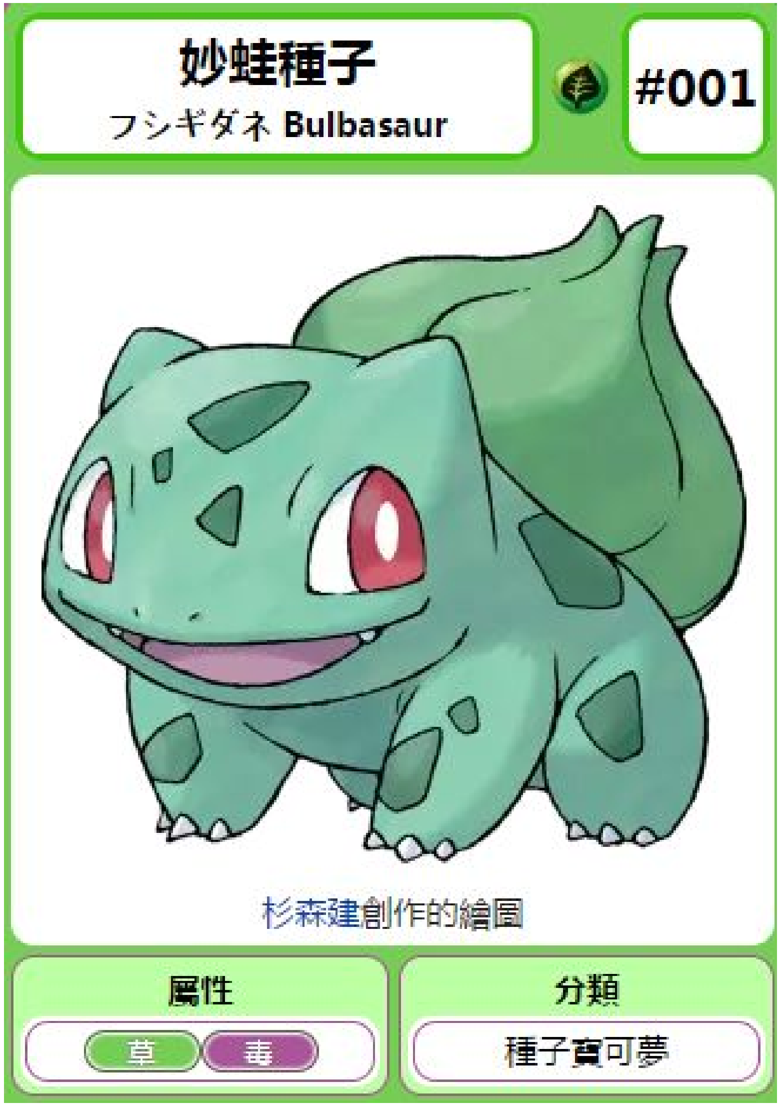
Task Introduction

Multi-class & multi-label article classification

Task Introduction

- Multi-class:
    - 有很多種類
- Multi-label:
    - 一筆資料可能屬於多個種類
- Multi-class & multi-label:
    - 有很多種類且一筆資料可能屬於多個種類





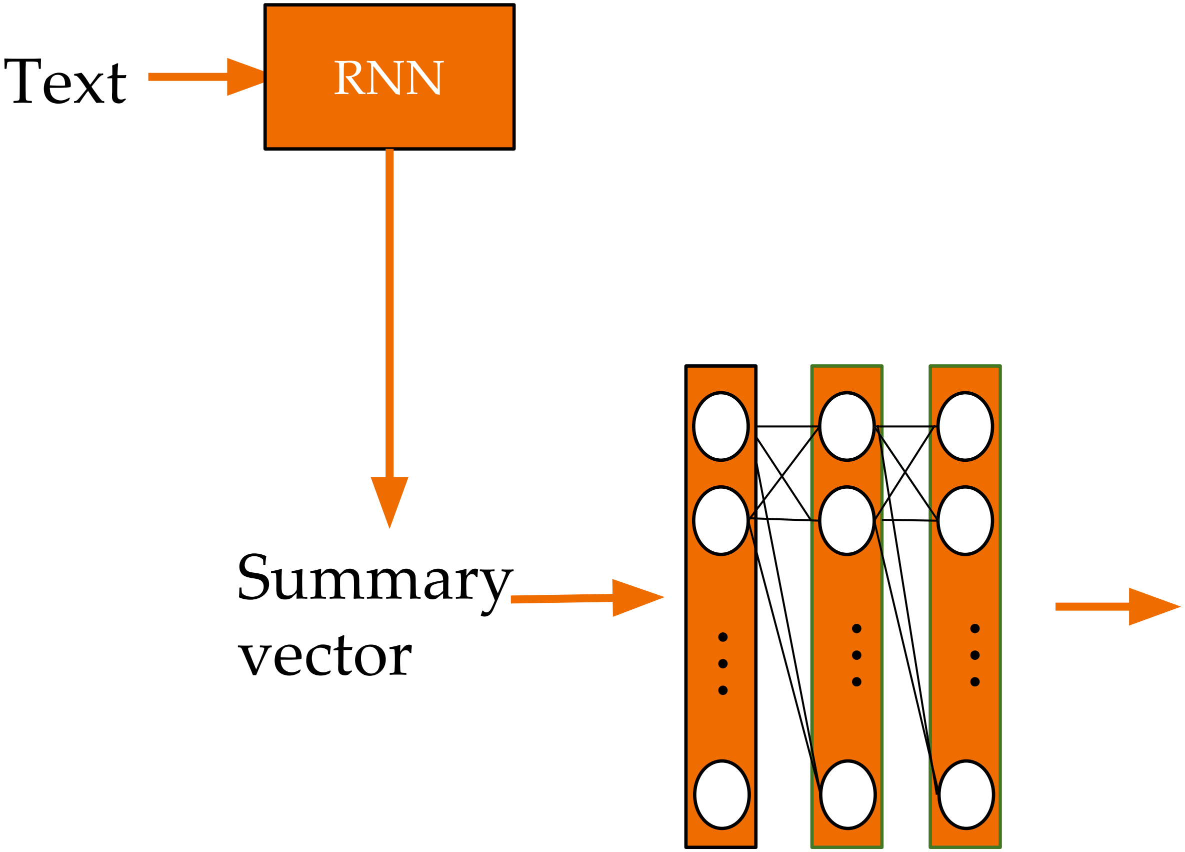
妙蛙種子

フシギダネ Bulbasaur

#001

杉森建創作的繪圖

屬性	分類
草 毒	種子寶可夢



Text

RNN

Summary vector

1 science fiction

0 children's literature

0 fantasy

0 historical fiction

1 speculative fiction

How to implement RNN in keras

1. Preprocessing
   a. Index words in your data
   b. Convert word sequences to word index sequences
   c. Padding sequences to equal length

example:

“I have a pen.” -> [1, 2, 3, 4]

“I have an apple.” -> [1, 2, 5, 6]

Useful function in Keras

1. Tokenizer
   a. fit_on_texts
   b. texts_to_sequences

2. pad_sequences
   a. pad sequence to equal length

How to implement RNN in keras

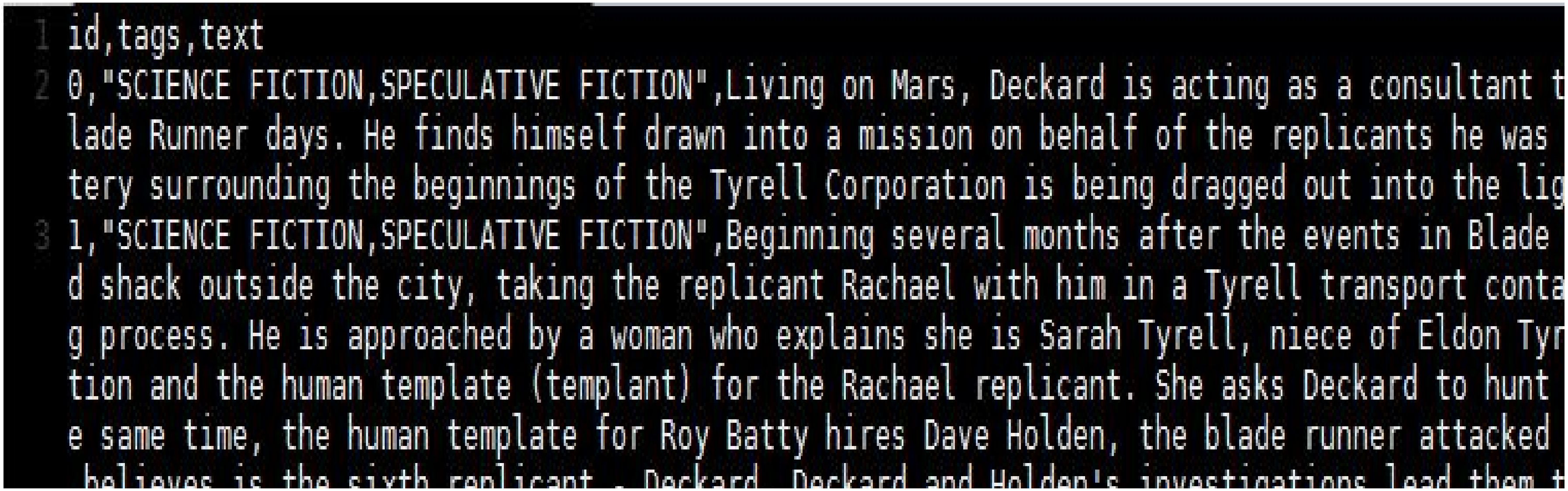
2. Embedding layer
    a. map word index to word vector
3. RNN cell
    a. LSTM
    b. GRU

Word Embedding

1. 1-of-N encoding
   a. 4964(training data)*300(text length)*50000(vocabulary size) *4(Byte) = 2.97*10^11 = 297 GB
2. Train by yourself(use training and testing data only!!)
   a. Training with your model
   b. Trained before training your model
3. Use others’ word embedding
   a. Glove

Reference : http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2017/Lecture/word2vec%20(v2).pdf

Data Format



```
id,tags,text
0,"SCIENCE FICTION,SPECULATIVE FICTION",Living on Mars, Deckard is acting as a consultant t
lade Runner days. He finds himself drawn into a mission on behalf of the replicants he was
tery surrounding the beginnings of the Tyrell Corporation is being dragged out into the lig
1,"SCIENCE FICTION,SPECULATIVE FICTION",Beginning several months after the events in Blade
d shack outside the city, taking the replicant Rachael with him in a Tyrell transport conta
g process. He is approached by a woman who explains she is Sarah Tyrell, niece of Eldon Tyr
tion and the human template (templant) for the Rachael replicant. She asks Deckard to hunt
e same time, the human template for Roy Batty hires Dave Holden, the blade runner attacked
 believes is the sixth replicant - Deckard. Deckard and Holden's investigations lead them t
```

Data Format

1. Text length : between 30 to 300 words
2. Vocabulary size : about 50000 words
3. Train data : 4964 筆
4. Test data : 1234 筆
    a. 一半為private set

Evaluation

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

precision & recall reference : https://en.wikipedia.org/wiki/Precision_and_recall

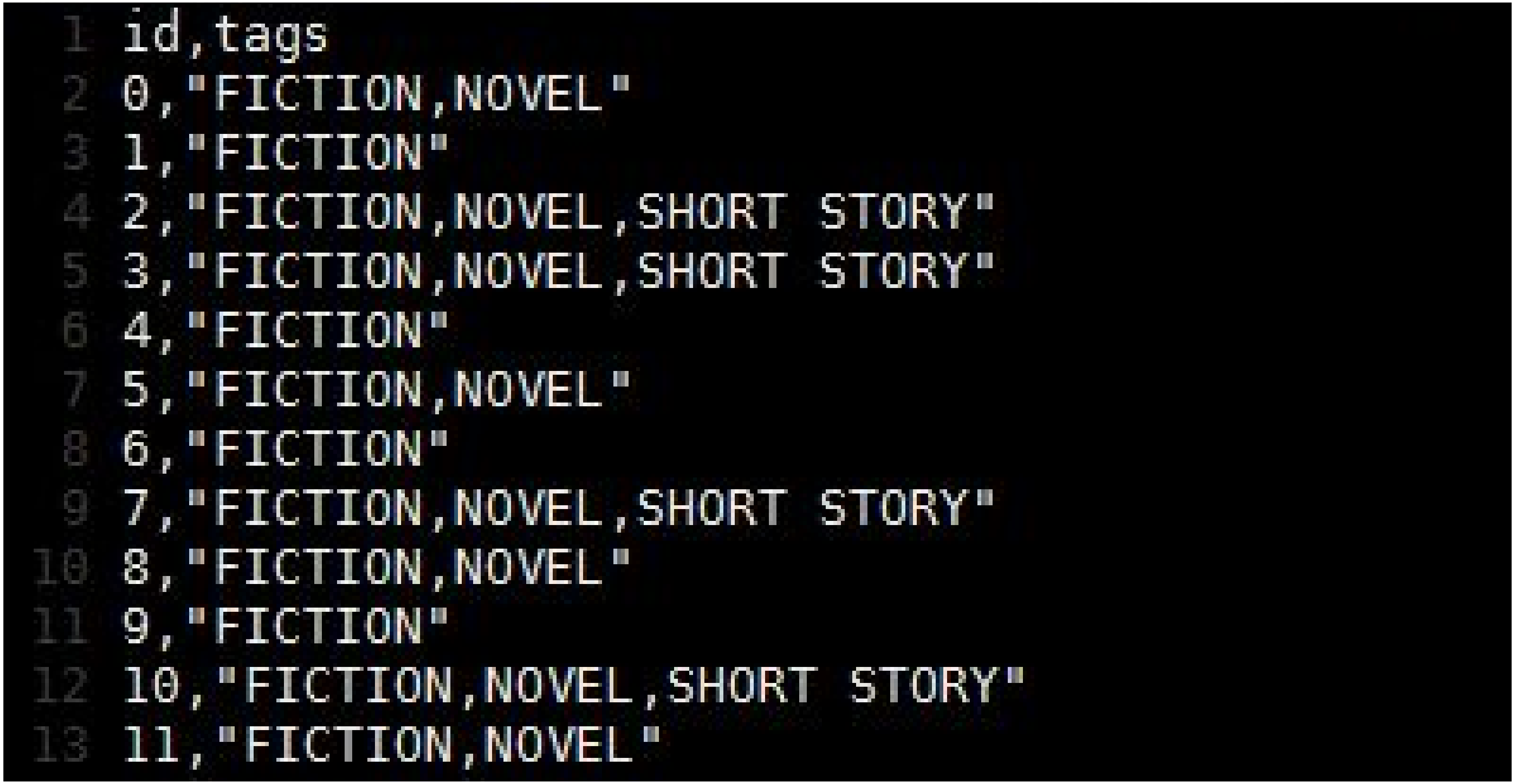
keras metrics : https://keras.io/metrics/

Kaggle

Kaggle

- kaggle_url : https://inclass.kaggle.com/c/ml2017-hw5
- 請至kaggle創帳號登入，需綁定NTU信箱。
- 個人進行，不需組隊。
- 隊名:學號_任意名稱(ex. b02902000_日本一級棒)，旁聽同學請避免學號開頭。
- 每日上傳上限5次。
- test set的資料將被分為兩份，一半為public，另一半為private。
- 最後的計分排名將以1筆自行選擇的結果，測試在private set上的準確率為準。
- kaggle名稱錯誤者將不會得到任何kaggle上分數。

Submission Format



```
id,tags
0,"FICTION,NOVEL"
1,"FICTION"
2,"FICTION,NOVEL,SHORT STORY"
3,"FICTION,NOVEL,SHORT STORY"
4,"FICTION"
5,"FICTION,NOVEL"
6,"FICTION"
7,"FICTION,NOVEL,SHORT STORY"
8,"FICTION,NOVEL"
9,"FICTION"
10,"FICTION,NOVEL,SHORT STORY"
11,"FICTION,NOVEL"
```

Deadline and Policy

Deadline

1. Kaggle: 5/25 23:59 (GMT+8)
2. Report and source code: 5/26 21:00 (GMT+8)

Policy I - Repository

- github上ML2017/hw5/裡面請至少包含:
  - Report.pdf
  - hw5_rnn.sh
  - hw5_best.sh
  - your python files
  - model (can be loaded by your python file)
  - requirements.txt (optional)
- **請不要上傳dataset**
- 如果你的model超過github的最大容量，可以考慮把model放在其他地方(同hw3)。

Policy II – Source Code

- **Python Only**，請使用Python 3.6, Python 2.7, Tensorflow 1.1，Keras 2.0.4
- 可使用現成package (Keras、Tensorflow ...)
- 不能使用額外data來training (包括 pre-training)
- 不能call 其他線上 API (Project Oxford...)
- 請附上訓練好的model (及其參數)，hw5_rnn.sh 和 hw5_best.sh要在10分鐘內跑完
- 請將需要用到的package寫在requirements.txt中(example)，若沒寫在requirements.txt 而出現ImportError會扣點分數。

Policy II – Source Code

- 與之前作業相同，請在script中寫清楚使用python版本
- 以下的路徑，助教在跑的時候會另外指定，請保留可更改的彈性，不要寫死
  - Script usage:

```
bash  hw5_rnn.sh <test data>   <prediction file>
bash  hw5_best.sh  <test data>  <prediction file>
```

    testing data: test.csv的路徑

    prediction file: 結果的csv檔路徑

Policy III - Report

- 請使用中文作答，若使用英文請確定語意清楚☺。
- 請交pdf檔，檔名為Report.pdf
- Report Template： Link

Policy IV - Score

- Kaggle Rank
  - (1%) 超過public leaderboard的simple baseline分數
  - (1%) 超過public leaderboard的strong baseline分數
  - (1%) 超過private leaderboard的simple baseline分數
  - (1%) 超過private leaderboard的strong baseline分數
  - (1%) 5/17 23:59 (GMT+8)前超過public simple baseline
  - (BONUS) kaggle排名前五名(且願意上台跟大家分享的同學)

前五名排名以public以及private平均為準，屆時助教會公布名單

Policy IV - Score

- Report problem
  1. (1%)請問softmax適不適合作為本次作業的output layer?寫出你最後選擇的output layer並說明理由。
  2. (1%)請設計實驗驗證上述推論。
  3. (1%)請試著分析tags的分布情況(數量)。
  4. (1%)本次作業中使用何種方式得到word embedding?請簡單描述做法。
  5. (1%)試比較bag of word和RNN何者在本次作業中效果較好。

Other Policy

1. script 錯誤，直接0分。若是格式錯誤，請在公告時間內找助教修好，修完kaggle分數*0.7。
2. Kaggle超過deadline直接shut down，可以繼續上傳但不計入成績。
3. Github遲交一天(*0.7)，不足一天以一天計算，不得遲交超過兩天，有特殊原因請找助教。
4. Github遲交表單：https://goo.gl/forms/rPgZ73Z8F1xcpiA92(遲交才需填寫)
5. 遲交請「先上傳程式」Github再填表單，助教會根據表單填寫時間當作繳交時間。

FAQ

FAQ

- 作業網址: https://inclass.kaggle.com/c/ml2017-hw5
- 若有其他問題，請po在FB社團裡或寄信至助教信箱，**請勿直接私訊助教**。
- 助教信箱：ntu.mlta@gmail.com