
Machine Learning HW5

MLTAs

ntumlta2019@gmail.com

UPDATES

(11/24)

Allowed Spacy model for word segmentation: en_core_web_sm, en_core_web_md, en_core_web_lg

(11/25)

Delete <dict.txt.big file>

(11/26)

Update early baseline deadline.

Update TA hour.

(12/04)

You can use “micro f1 score” in sklearn to evaluate your result.

(12/12)

We will download spacy models in advance, you don't need to upload it.

Outline

- Task Description - Malicious Comments Identification
- Data Format
- Kaggle
- Requirements and Regulation
- Grading Policy
- FAQ

Outline

- Task Description - Malicious Comments Identification
- Data Format
- Kaggle
- Requirements and Regulation
- Grading Policy
- FAQ

Task Description

- This task requires the annotators to give their judgements on whether a tweet is offensive or not.
- Example:
 - Hey @LIRR , you are disgusting. - **Offensive**
 - A true American literary icon. #PhilipRoth will be missed. - **Not offensive**

Method Overview

- Common Workflow
 - Word Segmentation
 - Word Representation
 - Word embedding / Bag of words
 - Training
 - RNN / DNN

Word Segmentation

- “Apple is looking at buying U.K. startup for \$1 billion”

=> [“Apple”, “is”, “looking”, “at”, “buying”, “U.K.”, “startup”, “for”, “\$1”, “billion”]

- Use [SpaCy](#) or implement yourself !
- Other python packages are not allowed !

Word Vectors

- 將每個字/詞轉換為 vector 以利後續 model training 。
- 如何將字/詞轉換為 vector ?
 - One-hot Encoding
 - Word Embedding

One-hot Encoding

- 假設有一個五個字的字典[1,2,3,4,5]

我們可以用不同的one-hot vector來代表這個字

1 -> [1,0,0,0,0]

2 -> [0,1,0,0,0]

3 -> [0,0,1,0,0]

4 -> [0,0,0,1,0]

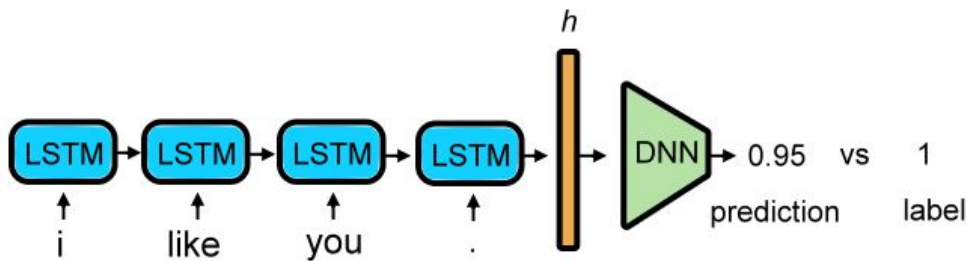
- Issue :

- a. 缺少字與字之間的關聯性(當然你可以相信NN很強大他會自己想辦法)
- b. 很吃記憶體

$200000(\text{data}) * 30(\text{length}) * 20000(\text{vocab size}) * 4(\text{Byte}) = 4.8 * 10^{11} = 480 \text{ GB}$

Word Embedding

- 用一個向量(vector)表示字(詞)的意思
- 用一些方法 pretrain 出 word embedding (ex: skip-gram、CBOW)
可使用 Word2Vec 實做(開放套件:[gensim](#))
 - ※ 如果要實作這個方法, pretrain 的 data 也要是作業提供的!
 - ※ 本次作業 **不開放** 使用現成的 word embedding (GloVe, etc.)
- 或是跟 model 的其他部分一起 train



Bag of Words (BOW)

- BOW的概念就是將**句子**裡的文字變成一個袋子裝著這些詞的方式表現, 這種表現方式不考慮文法以及詞的順序。

例如：

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

在BOW的表示方法下, 會變成:

(1) -> [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]

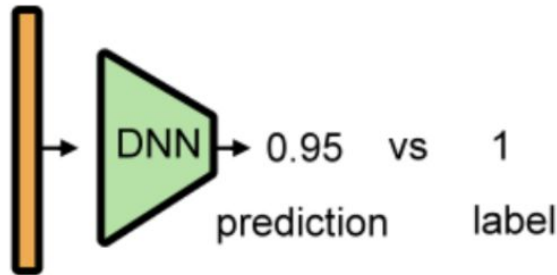
(2) -> [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

此 vector 即代表整個句子, 可以餵進 DNN 訓練。

dictionary

["John", "likes", "to",
"watch", "movies",
"also", "football",
"games", "Mary", "too"]

BOW



Outline

- Task Description - Malicious Comments Identification
- Data Format
- Kaggle
- Requirements and Regulation
- Grading Policy
- FAQ

Data Format ^{1/2}

- 本次作業將提供 train_x.csv, train_y.csv 跟 test_x.csv 等 3 個 csv 檔。
- Label
 - 1: 惡意留言
 - 0: 非惡意留言

Data Format 2/2

```
id,comment
0,Trump... an Ive league billionaire and negotiator extraordinaire!! Jan... #Resister !! 🙄 I'll
1,@user Just ask them how does having some of the toughest gun control laws working out for chid
2,@user Gosh dangit!!! IF we only had reasonable replica gun control this craziness wouldn't
3,@user @user @user Means nothing!! Your always so dramatic and none of your promises ever come
ize you are full of crap.
4,4 dead in mass shooting at Fifth Third Bank building in downtown Cincinnati'. Deep State Fake
einstein/Obama! URL via @
5,@user We thank our God for you every single day. #MAGA to #KAG
```

```
id,label
0,0
1,0
2,0
3,1
4,1
5,0
6,0
7,0
"data/train_y.csv" 13241L, 94819C
```

Outline

- Task Description - Malicious Comments Identification
- Data Format
- Kaggle
- Requirements and Regulation
- Grading Policy
- FAQ

Kaggle - Info ^{1/2}

- Kaggle 連結: [link](#)
 - 個人進行, 不需組隊
 - 隊名:
 - 修課學生: 學號_任意名稱 (ex: b08901666_好想看復仇者聯盟)
 - 旁聽: 旁聽__任意名稱
 - 每天上傳上限5次
 - Leaderboard上所顯示為public score, 在Kaggle Deadline前可以選擇2份submission作為private score的評分依據。
 - 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2筆自行選擇的結果, 測試在private set上的準確率。
- ★ kaggle名稱錯誤者的分數將x0.7。

Kaggle - format 2/2

- 預測 860 筆 testing data 是否為惡意留言，將預測結果上傳至kaggle
 - Upload format : csv file
 - 第一行必須是 id,label
 - 第二行開始，每行分別為id值及預測結果 (binary)，以逗號隔開
 - Evaluation: Mean-F1-Score
- 範例格式如右

```
sample_submission.csv x
1 id,label
2 0,0
3 1,0
4 2,0
5 3,0
6 4,0
7 5,0
8 6,0
9 7,0
10 8,0
11 9,0
12 10,0
13 11,0
14 12,0
15 13,0
16 14,0
17 15,0
18 16,0
19 17,0
20 18,0
21 19,0
22 20,0
```

Evaluation

mean F1-score

Outline

- Task Description - Malicious Comments Identification
- Data Format
- Kaggle
- Requirements and Regulation
- Grading Policy
- FAQ

Requirements

- hw6_test.sh 必須實作 **RNN** 或者 **BOW+DNN** 其中一種(以最後kaggle private baseline高者為準)。
- 本次作業**不開放**使用作業以外的 dataset，也**不開放**使用其他 data pretrain 好的 word embedding。
- 以**任何方式**接觸到testing data的label皆視為作弊，可能會處0分及校規處分。

Regulation 1/3

- Only Python 3.6 available !!!!
- 開放使用套件([conda file](#))
 - python standard library
 - numpy == 1.16.5
 - pandas == 0.25.1
 - pytorch == 1.2.0, keras == 2.2.4, tensorflow == 1.14.0
 - spacy=2.2.2
 - gensim == 3.8.0 (只可使用 word2vec api !!)
 - emoji == 0.5.3
- 其他作圖類套件請不要寫在我們會執行的code 裡面。
- 請注意 gensim **只可使用 word2vec api** (gensim.models.Word2Vec), 使用其他 api 將視為違規使用套件。
- 若需使用其他套件, 請儘早寄信至助教信箱詢問, 並請闡明原因。

Regulation - GitHub 2/3

- 請注意 github commit 為 local 端之時間，務必注意本機的電腦時間設定，助教群將在 deadline 一到就 clone 所有程式以及報告，並且**不再重新 clone 任何檔案**
- 你的 github 上 ML2019FALL/hw5/ 中請包含：
 - report.pdf
 - hw5_test.sh
 - hw5_train.sh
 - your python files
 - models (包括 embedding file)
- **請勿上傳 train.csv, test.csv 等 dataset !!!**
- 批改時將只執行 testing, 請自行跑完 training 部分並且儲存相關模型參數並上傳至 github。
- 若 model 超過 github 容量限制，請傳到其他地方(ex. Dropbox) 並在 script 中寫好下載的 command (請參考 <http://slides.com/sunprinces/deck-16#/2>)

Regulation - Script Usage ^{3/3}

- 助教在批改程式部分時，會執行以下指令：
 - `bash hw5_test.sh <test_x file> <output file>`
 - test file 為助教提供的 test.csv 路徑
 - output file 為助教提供的 output file 路徑
 - E.g. 如果助教執行了 `bash hw5_test.sh ~/data/test.csv ~/ans.csv`，則應該要產生一個檔名為 ans.csv 的檔案
- hw5_test.sh 需要在 10 分鐘內執行完畢，否則該部分將以 0 分計算。
- 切勿於程式內寫死 test.csv 或者是 output file 的路徑，否則該部分將以 0 分計算。
- Script 所使用之模型，如 hdf5 檔、pickle 檔等，可以於程式內寫死路徑，助教會 cd 進 hw5 資料夾執行 reproduce 程序。
- 原則上助教只會跑 testing，不會跑 training，但請還是要上傳 training script/code
 - 助教執行指令：`bash hw5_train.sh <train_x file> <train_y file> <test_x file>`

Outline

- Task Description - Malicious Comments Identification
- Data Format
- Kaggle
- Requirements and Regulation
- Grading Policy
- FAQ

Grading Policy - Deadline ^{1/5}

- Early Simple Deadline: 2019/12/05 23:59:59 (GMT+8)
- Kaggle Deadline: 2019/12/12 23:59:59 (GMT+8)
- Github Deadline: 2019/12/13 23:59:59 (GMT+8)

助教會在deadline一到就clone所有程式, 並且**不再重新clone任何檔案**

Grading Policy - Evaluation (5% + Bonus 1%) ^{2/5}

- (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%) 2019/11/28 23:59:59 (GMT+8)前超過public simple baseline
- (BONUS 1%) private leaderboard 排名前五名且於助教時間上台分享的同學

Grading Policy - Report ^{3/5}

1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法, 回報模型的正確率並繪出訓練曲線*
2. (1%) 請實作 BOW+DNN 模型, 敘述你的模型架構, 回報模型的正確率並繪出訓練曲線*。
3. (1%) 請敘述你如何 improve performance (preprocess, embedding, 架構等), 並解釋為何這些做法可以使模型進步。
4. (1%) 請比較不使用斷詞套件 (e.g., 用空白分開) 與有使用斷詞套件, 兩種方法實作出來的效果差異, 並解釋為何有此差別。
5. (1%) 請比較 RNN 與 BOW 兩種不同 model 對於 "Today is hot, but I am happy" 與 "I am happy, but today is hot" 這兩句話的分數 (model output), 並討論造成差異的原因。

* 訓練曲線 (Training curve): 顯示訓練過程的 loss 或 accuracy 變化。橫軸為 step 或 epoch, 縱軸為 loss 或 accuracy。

Grading Policy - Report 4/5

- 限制
 - 檔名必須為 report.pdf !!!
 - 檔名必須為 report.pdf !!!
 - 檔名必須為 report.pdf !!!
 - 請用中文撰寫 report (非中文母語者可用英文)
 - 頁數建議不超過2頁
 - 保留各題標題
 - 請標明系級、學號、姓名, 並按照report模板回答問題, 切勿隨意更動題號順序
 - 若有和其他修課同學討論, 請務必於題號前標明collaborator (含姓名、學號)
- Report模板連結
 - 連結: [report](#), [數學題](#)
<https://hackmd.io/@2sU0pH8TT7Wh3dZV9OMSDA/H1D2zdG3r>
- 截止日期同 Github Deadline: **2019/12/13 23:59:59 (GMT+8)**

Grading Policy - Other Policy ^{5/5}

- **Lateness**

- Github 遲交一天(不足一天以一天計算) hw6 所得總分將 $\times 0.7$
- **不接受程式 or 報告單獨遲交**
- 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。
- Github 遲交表單: 遲交請先上傳遲交檔案至自己的github 後再填寫遲交表單, 助教群會以表單填寫時間作為繳交時間手動clone 檔案。
- [link](#)

- **Script Error**

- 當 **script 格式錯誤**, 造成助教無法順利執行, 請在公告時間內寄信向助教說明, 修好之後重新執行所得 kaggle 部分分數將 $\times 0.7$ 。
- 可以更改的部分僅限syntax及io的部分, 不得改程式邏輯或是演算法, 至於其他部分由助教認定為主。
- 不接受任何 py 檔的 coding 錯誤更改

TA Hours

12/6(同手把手) 上課後

地點: 博理113(原教室)

FAQ

- 若有其他問題，請寄信至助教信箱，**請勿直接私訊助教**。
- 有問題建議可以在 FB Group 裡面留言發問，可能很多人都有一樣的問題
- 本次作業可能需要花較多時間訓練，請同學儘早開始。
- 助教信箱 ntumlta2019@gmail.com

