110-2
Natural Language Processing

HW3

TA: Kuei-Chun Kao

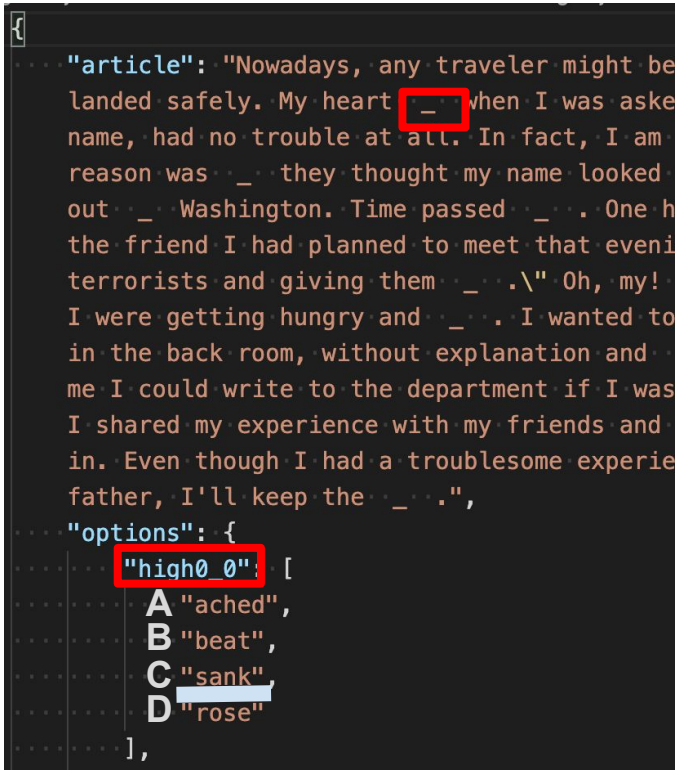
Task introduction (Cloze test)

- What is cloze test?
  - Under a typical setting, a cloze test requires examinees to fill in missing words (or sentences) to best fit the surrounding context.
- Learning objectives: train your own N-gram language model
- Not using any NN model~

Dataset

- Dataset format: json format (need to parse by yourself)
- article: A string. There are several blanks (denoted as "_") within each passage, where each blank represents a cloze question.
- options: A list of options for each question. There are four options for each blank.
- answers: A list, representing the golden labels of the questions. The answer can be A, B, C or D.
- id: an unique id of the passage.

Example (high0.json)



Outputs

- Each line consists of two fields separated by horizontal whitespace (a single tab or space character). The first field is the ID of a article from the Json file. The second field is the option of the blank.
- Kaggle Link: https://www.kaggle.com/t/584e537cf7634e57911d52e3827729ac
- Displayed name: <student_ID>
- Submission format: .csv file (You can also see from sample_submission.csv)
- Evaluation metric: Accuracy

submission.csv 0.54862

Kaggle submission

- You may submit up to 5 results each day (UTC).
- Up to 2 submissions will be considered for the private leaderboard

remember to select **2** results for your final scores before the competition ends!

Reference (You can follow some ideas here)

- Paper: A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
- You can split your train folder into training set and validation set
- You can use external corpus if you want
- Use the packages/tools I allow:
  - Python 3.8 / 3.9 and Python Standard Library
  - SpaCy, sentencepiece, nltk, stanfordcorenlp
  - datasets, json, tqdm, itertools
  - Dependencies of above packages/tools.

Requirements

- Python only
- No plagiarism!
- At the top of your Source code

```
#Author: Kuei-Chun Kao

#Student ID: 1234567

#HW ID: Hw3

#Due Date: 01/30/2020
```

Submission

- Deadline
  - Submit Zip to E3 before 5/11 11:59 PM
  - No Late Submission, thanks!
- Format
  - Source code: Hw3_<StudentID>.py (py only, your main file), you can split anther utils files, if you needed. Be careful for your import path.
  - Report file: Hw3_<StudentID>.pdf (pdf only)
  - Make sure the .py file contains the correct execution results and formats.
  - If can’t compile correctly, no score for you
  - Zip file: Hw3_<StudentID>.zip (zip only)
- Any question can ask me on E3, answer your question ASAP

Grading policy

- Ranking score in Kaggle Leaderboard (40%)
- Report (50%)
- Code comment, file format, display name (10%)
- I can only see your last submission.
- Do not submit your model or dataset.
- If your code is not reasonable, your final grade will be multiplied by 0.8!
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.

Ranking score in Kaggle Leaderboard (40%)

- Public leaderboard (20%): Your public leaderboard score > baseline, you can get 20% of this part; Otherwise, you can only get 10% of this part.
- Private leaderboard (20%): Your private leaderboard score * 20%
- This part score = public leaderboard + private leaderboard

Report (50%)

1. Your model design and concept (8%)
2. Error Analysis and Discussion (7%)
3. Compare and implement unsupervised method and supervised method (7%)
4. 請描述嘗試過的方法，並且討論曾經遇到的問題以及解決的方法(7%)
5. 請討論使用不同訓練資料量訓練n-gram language model對於預測克漏字的效能影響(7%)
6. 請討論使用不同domain的訓練資料訓練n-gram language model對於預測克漏字的效能影響(7%)
7. 請用n-gram language model實作next word prediction，分析使用不同數量、不同domain的資料訓練model後，生成的句子有什麼差異。請以“This is”、“He said”、“She said”為prompt進行討論。(7%)

Q1: Data processing (4%)

1. Tokenizer and build window (2%):
    a. Describe in detail about the tokenization algorithm you use. You need to explain what it does in your own ways.

2. Answer Option (2%):
    a. How did you convert the answer on characters to options on tokens after LM tokenization?
    b. After your model predicts the probability of answer, what rules did you apply to determine the final option?

Q2: Modeling with LMs and their variants (4%)

Describe (4%)

a. your model (configuration of the model) and details
b. performance of your model.

Bonus

- If your ranking is top 3 in class, you can get 3 points bonus in this hw final score!