CS 613: NLP

Assignment 1: Data Preparation and De-duplication

Total marks: **100 Pts.**	**Submission deadline: 23:59:59 Hrs, September 22, 2024 (Sunday)**

Assignment Instructions

1. Regarding the late submission, we will be following the penalty as per the table:

Late Submission	**Penalty (Out of 100)**
Till 1-hour past the deadline	5 points
1 to 12 hours past the deadline	10 points
12 to 24 hours past the deadline	20 points
24 to 36 hours past the deadline	40 points
36+ hours past the deadline	100 points

2. We will follow the zero plagiarism policy, and any act of plagiarism will result in a zero score for the assignment.
3. Please cite and mention others' work and give credit wherever possible.
4. If you seek help and discuss it with the stakeholders or individuals, please ask their permission to mention it in the report/submission.
5. Compute requirement: Use Colab and write the answers in the colab itself.

Problem Statement

You are supposed to curate, process, and deduplicate the dataset in your assigned languages.

Tasks (100 Pts.)

1. Download and curate the datasets for the assigned languages. Just make sure that you follow the following steps **[25 pts]**:
    a. Only crawl data that is publicly accessible and not copyrighted.
    b. You can download data from existing corpora like ROOTS, CC, OSCAR, C4, etc. However, the marks will depend on the data you curate by not downloading from these existing Corpuses.

2. You need to prepare the list of the dataset source names, volume in GBs, and total of articles in each source as a table **[5 pts]**.
3. One article/page should correspond to a text file (in .txt) **[5 pts]**.
4. Prepare a list of bad-word dictionaries for the respective language **[10 pts]**.
    a. Feel free to create your own and use existing dictionaries.
5. You need to clean the dataset by removing articles containing bad words, pornographic content, hate, abuse, etc. **[30 pts]**
    a. Feel free to use any tool to remove the bad words.
6. Prepare the table with the statistics (total articles, GBs) and datasets before and after ***cleaning***. **[5 pts]**
7. Deduplicate the dataset by removing the duplicated articles **[10 pts]**.
    a. The TAs will share a codebase for the same.
    b. Bonus [5 pts] for each robust technique to deduplicate. Maximum 4 techniques can be shown.
8. Prepare the table with the statistics (total articles, GBs) and datasets before and after ***deduplication***. **[10 pts]**
9. ~~Train the Tokenizer and measure the fertility scores of the trained tokenizer **[20 pts]**.~~

Points Split: 25+5+5+10+30+5+10+10 = 100

Submission

1. Submit your code (GitHub) or colab notebook with proper comments to this link.
    a. Ensure the individual contribution is appropriately added (OTHERWISE PENALTY OF 10 MARKS).

Expectations from the team:

1. Properly divide the team into sub-groups and distribute your tasks equally.
2. Negative marks for documentation and justifications!
3. Write the contributions or tasks completed by each team member. Scores might be different among team members if the tasks are not equally distributed.

References

1. https://github.com/AamodThakur/NLP_Scraping
2. https://github.com/AamodThakur/NLP_Pre_Training/tree/main

TAs to Contact

1. Himanshu Beniwal (himanshubeniwal@iitgn.ac.in)
2. Indrayudh Mandal (24210041@iitgn.ac.in)
3. Jenil Pradipkumar Patel (24210048@iitgn.ac.in)
4. Mithlesh Singla (24210063@iitgn.ac.in)
5. Nirmalkumar Jitendrabhai Patel (24210070@iitgn.ac.in)

FAQs

1. *We will add clarifications to doubts here. Please check periodically, as someone might have already asked about the doubt, which will be appended here.*