Natural Language Processing and Text Mining: HW#1

By J. H. Wang

Mar. 15, 2023

Programming Exercise: Sentiment Analysis

- Goal: Sentiment classification on open source datasets
- Input: TSATC: Twitter Sentiment Analysis Training Corpus (to be detailed later)
- Output: Training classifiers to classify the sentiment of tweets (to be detailed later)

Tasks and Data

- Tasks
  - Performing sentiment classification on twitter data (as detailed in the following slides)

- Data: an open dataset from Huggingface

- You have to submit the classification output

Input Data

- Data:
  - [**TSATC: Twitter Sentiment Analysis Training Corpus**] from Hugging Face
    - 1,578,627 tweets, about 15MB in size
  - Available at:
    - https://huggingface.co/datasets/carblacac/twitter-sentiment-analysis
- Format:
  - Two text files consisting of lines of records
  - Each record contains 2 columns:
    feeling, text

Tasks in this Homework

- To train a classifier using the training set in any programming language
- To test the classification result for the test set

Output Format

- Classification results
  - Precision
  - Recall
  - F-measure
  - Accuracy

Homework Submission

- Due: Mar. 29, 2023 (Wed.)

- For programming exercises, please submit it online to iSchool+
  - Under the item [Assignments]\[HW#1]

- Please include program source codes and documents
  - specifying your team members and responsible parts in the homework
  - Indicating configuration and installation steps of necessary packages on the specified platform

Thanks for Your Attention!