Assignment 1

Please submit the code and report (in pdf format) on Blackboard system before 23:59 Oct. 15. Report can be written either in English or Chinese. Name your report as `studentID_Name.pdf`.

Note:

- If you use ChatGPT etc., list all the prompts you use. Note that the LLM may have hallucination problem and you are responsible for the correctness of the answers.
- You can use your laptop or Google Colab computing platform, etc.

Q1

Papers with Code is a famous platform that collects published papers and their code. Select **three** topics you are interested from the NLP area, e.g., text classification, machine translation, etc.

For each task,

- [3 points] give a brief introduction, including the model input and output, evaluation metric, etc. Introduce with details instead of just mentioning their names.
- [3 points] Select one benchmark you like, list top-2 models/methods and their performance.

Q2

FastText is a famous word embedding toolkit. Write code to train a word embedding model with FastText. The dataset is attached.

- [2 points] Data preprocessing. Train a sentencepiece model (vocabulary size = 3000) and tokenize the dataset with sentencepiece toolkit. You may refer to the python tutorial here. You can also use the command line.
- [2 point] Train word embedding based on the tokenized data using skip-gram method and FastText toolkit. You can refer to the tutorial here.