Problem 1

- Chinese word segmentation: 20 points
  - This task provides PKU data as training set and test set (e.g., you can use 80% data for model training and other 20% for testing), and you are free to use data learned or model trained from any resources.
  - Evaluation Metrics:
    - Precision = (Number of words correctly segmented)/(Number of words segmented) * 100%
    - Recall = (Number of words correctly segmented)/(Number of words in the reference) * 100%
    - F measure = 2*P*R / (P+R)

Problem 2

- N-gram Language Models: 20 points
  - In this assignment you will explore a simple, typical N-gram language model.
  - This model can be trained and tested on sentence-segmented data of a Chinese text corpus. “Word Perplexity” is the most widely-used evaluation metric for language models.
  - Additional points: if you can test how does the different "Word Perplexity" of the different “N” grams, you will get additional 10 points
  - Additional points: if you can test how does the different "Word Perplexity" of the different smoothing methods, you will get additional 10 points

Problem 2’

- N-gram Language Models: 20 points
  - In this assignment you will explore a simple neural network language model.
  - This model can be trained and tested on sentence-segmented data of a Chinese text corpus. “Word Perplexity” is the most widely-used evaluation metric for language models.
  - Additional points: if you can test how does the different "Word Perplexity" of the different network architectures, you will get additional 10 points

Problem 3

- Part-of-speech tagging: 30 points
  - This data set contains one month of Chinese daily which are segmented and POS tagged under Peking Univ. standard.
  - Project ideas:
    - Design a sequence learning method to predicate a POS tags for each word in sentences.
    - Use 80% data for model training and other 20% for testing (or 5-fold cross validation to test learner's performance. So it could be interesting to separate dataset.)