作业总体要求

- 每人至少完成候选题目中的三个
- 使用的方法模型不限、编程语言不限
- 要求提交物：
  - 文件：任务定义、输入输出、方法描述、结果分析、源码运行环境
  - 代码：源码及可执行文件

Problem 1

- Chinese word segmentation: 10 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

- Text classification: 10 points
  - This data set contains 1000 text articles posted to each of 20 online newgroups, for a total of 20,000 articles. For documentation and download, see http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html.
  - The "label" of each article is which of the 20 newsgroups it belongs to. The newsgroups (labels) are hierarchically organized (e.g., "sports", "hockey").

Problem 3

- Part-of-speech tagging: 20 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.)

Problem 4

- Named entity recognition: 20 points
  - Named entities: people names, organizations, locations, numerals, etc
  - Your objective is to build a machine learning named entity recognition system, which when given a new previously unseen text can identify and classify the named entities in the text. This means that your system should annotate each word in the text with one of the four possible classes.
  - You will be given labeled data sets to train and test your model.

Problem 5

- Web Content Identification: 20 points
  - This dataset contains webpages from 4 universities, labeled with whether they are professor, student, project, or other pages. For data and documents, see http://www-2.cs.cmu.edu/~webkb/
  - Project ideas:
    - Learning classifiers to predict the type of webpage from the text
    - Can you improve accuracy by exploiting correlations between pages that point to each other using graphical models?

Problem 6

- Detecting sentiment polarity: 20 points
  - Given text about movie reviews
  - Can we detect sentiment, like whether a comment is
    - Positive?
    - Negative?
  - Can we tell to what extent is a comment positive of negative?
- Data:
  - 5331 positive snippets
  - 5331 negative snippets
- Other resources:
  - The Subjectivity Lexicon

Problem 7

- Word sense disambiguation: 20 points
  - Implement the simplified word sense disambiguation algorithm, and apply it to disambiguate a target ambiguous word in context.
  - For evaluation, use the dataset provided and the sense definitions provided by Wikipedia.
  - Note that you have to apply your own pre-processing to the content of the Wikipedia page (e.g., include the entire page or only certain sections; include the titles of the linked articles or not; etc.).
  - The quality of the pre-processing may affect the quality of your results. Report the accuracy of each word (i.e., number of instances correctly disambiguated).

More to be added…