DSP HW3

B04902079 資工三 甯芝蓪
environment: *Mac OSX 10.11*

Compilation & Execution

- `make map` can generate `ZhuYin-Big5.map` from `Big5-ZhuYin.map`
- `make` compile the execution file `mydisambig` from `mydisambig.cpp`
- `make run` run through all the test data `testdata/1~10.txt`, and store the result in `result/1~10.txt`
- To test other data, please run:
  ```
  $ ./mydisambig -text $file -map $map -lm $LM -order $order
  ```
  Currently, `mydisambig` only support bigram so that `$order` shall be 2.

What I've done

Mapping
I do the mapping with a python program `mapping.py`. In this program, I simply used a dictionary to store every (ZhuYin - array of Big5) pair, and output them.
Something notable is that the `big5` encoding in python is not suitable for the given data. As an alternative, `big5-hkscs` works.

Reading srlim API
I spent a lot of time reading the API (both source code and the man page). Also, I read srilm's `disambig.cc` to figure out the correct usage of the classes. Following are the header files and their contents that I took advantage of during my implementation of `mydisambig.cpp`:

- option.h
  - Opt_Parse()
- File.h
  - file.read(), file.write()
  - file.getline()
  - file.close()
- Ngram.h
  - ngram.read(fp)
  - ngram.wordProb(word, history)
- Vocab.h
  - vocab.getIndex(word)
  - vocab.getWord(index)
  - VocabWord Vocab_Unknown
  - VocabIndex Vocab_None
- VocabMap.h
  - VocabMap(vocab1, vocab2)
  - VocabMapIter(vocab_map, index)
- Prob.h
  - LogP

Implement Viterbi algorithm

After having a rough impression of the APIs, the implementation became simpler. However, there are still some things worth mentioning:

- **3 vocabulary**
  There are 3 `Vocab` class used in the program: `ZhuYin`, `Big5` from the map and `vocab` from the bigram language model. When using `getWord()` and `getIndex()`, should think twice about which vocabulary the index/word is from.
- **OOV problem**
  Some words in `Big5` do not exist in `vocab`. These words should be marked as unknown (use the default index `Vocab_Unknown`) or there would be a `Map_noKey` error.
- **<s> and </s>**
  These are tags for the beginning and the finish of a sentence, also the first and the last node of Viterbi. However, they do not exist in `Big5`. I added them in the beginning so that Viterbi algorithm can perform smoothly.