

Assignment 1

Spell Correction

- In this assignment, you are required to write a toy system for spelling correction.
- Both components of **channel model** and **language model** should be implemented as introduced in [Lecture 3](#).
- For the evaluation, you will be provided a text file consisting of 1,000 sentences extracted from *news articles*.
 - Each line contains one sample with three items, including sentence id, # of error words and the sentence. They are separated by tabs.
 - Out of which, 50 instances contain real word errors.

■ **testdata.txt**

- This is the text file that consists of 1,000 sample sentences extracted from *news articles*.
- Each line contains one sample with three items, including sentence id, # of error words and the sentence. They are separated by tabs.
- Out of which, 50 instances contain real word errors.

■ **ans.txt**

- This is the answer file that consists of 1,000 sentences after spelling correction.
- Each line contains one instance with two items, including sentence id and corrected sentence.

■ **vocab.txt**

- This is the dictionary for non-word error detection. Make sure you use this!

■ **eval.py**

- This is a python file including evaluation program **for your reference**. The path of the answer and result files are written inside the code. Read it.
- Run “eval.py”, it will give an accuracy number. Improve your program based on this number.
- Make sure your “ans.txt” and “result.txt” are located in the same directory with “eval.py” when running it.
- It might contain some errors. If you find any, email us!

■ **LM**

- The compiled files for SRILM.

- Generate a zip file and name it as “sid_homework-1.zip”.
- It should include a directory named *program*, an output file “result.txt” and a written report “spell correction.pdf”.
- Program: codes should be **written in python**.
- Output file: each line includes a single instance consisting of two items, namely *sentence id* and *sentence after correction*. Separate them by tab.
- Report: the report needs to be **written in English with no more than 4 pages**.

- We will mark your homework based on the three criterias:
 - Final accuracy (20%)
 - Program (30%)
 - Report (40%)
 - LM Implementation (10%)
- If you use a toolkit to estimate the language model, your maximum mark will be 90% of the full mark. If you estimate it on your own, you will have chance to get a full mark.

- Submit your homework via E-learning system.
- Deadline: **Mid-night at 2017 Oct. 18th**
- If you have any questions about this homework, send email to TA or our course mailbox.
- TA in Charge
 - 刘乾龙 (liuql2@163.com)
 - 顾云帆(13771900432@163.com)