CS 517 - Natural Language Processing

Programming Assignment #2

Due date: 04/21/14 16:00 hrs

Statement

You must have noticed many of the state of the art search engines automatically suggest you correct spellings for search terms. Here you have to implement a very rudimentary spell correcting system using edit-distance model.

Hint: The corpus provided in the **data** directory will be used as a dictionary. That means, any word appearing in the corpus will be assumed to be a valid English word.

Code

Download pa2.zip where the basic framework for `SpellCorrector.java` has been provided. You only have to implement `SpellCorrector` class with *two* public methods as follows:

- `public void learn(String word)`
  - This will be called for every word found in the given corpus
  - You should be able to build required data-structure to hold learnt model within the class
- `String correct(String misspelled_word)`
  - This method will return a corrected word as suggestion to the given word

Assumptions:

- All words are English and use just ASCII character set. That means, only a-z are possible characters.
- All the words passed to this class will be in lowercase.

Code Structure

`data/` -> This directory contains some works from Shakespeare, which will be used as corpus for this assignment.
`java/` -> This directory contains all the JAVA code. You only need to modify `SpellCorrector.java`.
`run.py` -> A tool to compile, run & test your code.

Extract and Run

Download the file pa2.zip. Extract it. Let's say you extracted into 'pa2' directory.
From terminal:

```
$cd pa2
```

```
# To run and evaluate your code
$./run.py
```

Compatibility

Note that, all assignments (including this one) will be tested under Linux environment with Python and Oracle Java is installed. Given code might work on other platforms (like Windows, etc.) but has not been tested. Hence, it is encouraged to develop and test your code in a Linux based environment.

Submission

You should only modify and upload `SpellCorrector.java` to Blackboard. **Any change in the filename will not be accepted and you will not be evaluated in that case**.

Evaluation

There is some held out data-set against which your code will be tested and evaluated. Hence, it is encouraged to write generic code that work in most of the cases. You are expected to employ classic edit-distance model up to a distance 2. To give an idea on accuracy in given data set, a model with edit distance of 1 produces ~15% correctness and with edit distance of 2, it produces 17%. You are free to read more techniques and use them to improve your accuracy.

Honor Code

I encourage students to discuss the programming assignments including specific algorithms and data structures required for the assignments. However, students should not share any source code for solution.

One of the most important principles of modern natural language processing is the use of a true unseen test set. Therefore, whenever we give a problem with training data, development data, and test data, you may not tamper with the submission process or scripts so as to attempt to look at the test data.

Code exists on the web for many problems including some that we may pose in problem sets or assignments. Students are expected to come up with the answers on their own, rather than extracting them from code on the web. This also means that we ask that you do not share your solutions to any of the homework - programming assignments, or problem sets - with any other students. This includes any sort of sharing, whether face-to-face, by email, uploading onto public sites, etc. Doing so will drastically detract from the learning experience of your fellow students.