Assignment 2

Text Generation

Question 1 (30%)

First, please visit the following website:

https://pdos.csail.mit.edu/archive/scigen/

This is a very interesting piece of work in last decade. It can generate a computer science paper for you automatically.

Go to the following github site

https://github.com/shaundsouza/scigen

and git clone the project to your local folder.

Change the permission by running

```
chmod 755 *.pl
```

Follow the instruction on the site and try to generate a paper (.pdf) with your own name.

Open the paper and you may find that the figure is oversized.

Fix the oversize problem.

Hints: You should open the source latex file (.tex) and locate the problem.

We will mark your submission by running

```
./run.sh output.pdf
```

in your folder, where output.pdf should be the output.

Question 2 (70%)

In this task, you need to train a language model (LM) and generate a sentence using greedy search in this LM.

In the previous lab, you have learnt how to use the segmentation tool to segment Chinese sentence.

Now you can follow the instructions below and learn how to train a LM with SRILM.

First, install the SRILM:

```
git clone https://github.com/gsayer/SRILM.git
cd SRILM
mkdir ~/srilm-1.7.1
tar -xvf srilm-1.7.1.tar.gz -C ~/srilm-1.7.1/
cd ~/srilm-1.7.1
export SRILM=$(pwd)
make world
vim ~/.bashrc  #add these lines as follows
            SRILM=/data/cs310/zichao/srilm-1.7.1
            Export PATH=$SRILM/bin/i686-m64:$PATH
source ~/.bashrc
```

Then, train a bigram with the SRILM commands:

```
# Usage of command
ngram-count -help
# Do the counting
ngram-count -text train.txt -order 2 -write train2gram.count
# Build the LM
ngram-count -read train2bigram.count -order -lm bigram.lm
```

In the commands above, train.txt is the training text data which has already segmented with the segmentation tool and bigram.lm is the output LM file.

Open the bigram.lm to have a look. It should be in ARPA file format. (Please refer to the lecture notes what is ARPA file format.)

You need to write a program to generate a sentence using greedy search from this LM. If the sentence you generate is too long, then just pick the first 30 characters.

The training text is placed at /home/cseadmin/asg2_input_text.

Copy this file to your local folder before your work.

We will mark your submission by running

```
./run.sh train.txt
```

in your folder, where train.txt is the training data.

Your program should output the generated sentence on the screen.

Have fun!

[Submission]

You have to create a folder named "assignment2" under your working directory at /data/cs310/XXX/. Thus, the path of your folder should be /data/cs310/XXX/assignment2/, where XXX is your login name.

Create two subfolders Q1 and Q2 under this folder.

Place the script file in the folder, then use chmod to turn off the read access of your folder so that others can't read your solutions.