CSC 4309 Natural Language Processing
Semester 1 2019/20

Assignment 1: Finite State Transducer (Group Submission)
Due: Tuesday, 15/10/2019 (12 midnight)

English-Chinese Transliteration

Transliteration is the conversion of a text from one script to another. For example, the Arabic script بي كتا (i.e., written) can be converted to the Latin script ***kitabi*** for English. English words have been also been commonly used in a various international language as loan/borrowed words such as for technical terms (e.g., 'computer') or names of places (e.g., 'New York'). Although different languages may have different pronunciations and sound inventories, some phonetic equivalents can be applied to a target language written in the script of the target language. For example in Japanese, the word **computer** can be written as ***konpyutaa*** while in Chinese, the word **tiramisu** can be written as ***tilamisu*** '提拉米苏' (tí lā mǐ sū). Each Chinese character has a distinct meaning.

a) Given the Table 1 containing the Chinese Hanzi syllabic script, build an **fst** that receives an **English syllable as input** and **output the equivalent transliteration of the Chinese syllable**. For example, the English syllable ***vi*** in the word ***vitamin*** will output ***wei*** in Chinese:

Table 1

To use entire words as input or output symbols, enclose the word in square brackets (not in parentheses). Example: to add an arc that takes the string ***vi*** as input and returns ***wei*** when going from state 1 to 2, you should use:

```
f.add_arc('1','2',['vi'],['wei'])
```

Test your program with the following inputs: (your program will also be tested using random inputs).

input	Output
vi ta min	*wei ta ming*
la tte	*na tie*
mo cha	*mo ka*
ti ra mi su	*ti la mi su*
bun gee	*beng ji*
la ser	*lei she*
hac ker	*hei ke*

Print all input-output mappings into an output file named **Chn-trans.dat** in the following format.

vi -- > wei
min -- > ming
bun -- > beng
la -- > lei

Figure 1: Example of English-Chinese mapping of the syllables

Submit the following for your assignment:

i) a python FST program for the Chinese-Eng transliteration for loan/borrowed words
ii) an output file that prints the mappings of the transliterations as shown in Figure 1
iii) an FST construction generated by the Python program – Tkinter (image or word document)