11-411/611 HW #2: Lemmatizer Setup

S22

1 Setup

For this assignment, you will be working with the library wfst4str to make your own FSTs. Due to complications with setting up this library on local machines, we highly recommend you to use Google Colab for this assignment. Colab is convenient for importing libraries and running code.

There are two files we have given you: **lemmatizer.py** and **lemmatizer.ipnyb**.

lemmatizer.ipnyb is a Jupyter Notebook which contains the instructions you will need for the assignment, examples, and the code foundation for the FSTs you will write. This is what you will import into Colab, and where you will be working.

lemmatizer.py is the file you will be submitting. This will be discussed more in section (4), "submitting your file".

2 Using Colab

Signing into your Google account, you will import **lemmatizer.ipnyb** into Colab. You will see a document which has instructions for the assignment as well as cells which can run code.

In order to run code, you must press the button at the top left corner of the cell. Note that this will only run the code in **that** particular cell. Every time you open Colab, you will generally want to run the cells from top to bottom (in particular, you must always run the cells in **Imports and Definitions**).

If you need to test your code and some parts depend on other parts, make sure that you run those cells in the correct order. However, afterwards, you do not need to run those particular cell blocks unless you have changed something. When a cell has finished running, you will see a green check mark to the left of it. Some cells take a long time to run (in particular, the data and test cells

might take up to a minute, but all other cells should finish between 0-5 seconds).

Some cells, such as the pre- and post- processing and the data, do not need to be worked on and can be hidden for your convenience. The following code at the top of a cell:

#@title [your title here]

may be used to hide a cell. To hide the cell, double click on the title which should appear on a gray bar to the right of the cell. To later unhide, click "show code".

You may also collapse an entire section of code by clicking the button to the left of the section title.

You can clear output after running a cell by clicking the X button which appears underneath the "run" button.

3 Testing

If you want to test a single wFST by itself, you may write tests into the same cell (or create a new one for your own tests). However, if you want to test your entire lemmatizer, we have provided a **batch test** cell.

After running the cell to define the batch test function, you may test the 3 datasets provided by running the cells which test **CARNEGLISH**, **INVOCAB**, and **OUTVOCAB**.

4 Submitting your file

You will be submitting a zipped version of the file **lemmatizer.py** onto Gradescope (note: not **lemmatizer.ipynb**).

We would like you to copy and paste the entire "Your wFSTS Here" cell into **lemmatizer.py**. Make sure that you do **not** rename any variables in those cells.

When you are submitting to gradescope, make sure that you do not leave any FSTs blank; this will cause all of your outputs to be empty. Instead, you may use the default code provided from the starter file for an FST which maps each character to itself.