CS 106 Winter 2017

Assignment 08: Text Processing

Due: Friday, March 22nd, 11:59pm

Soundex Algorithm

The Soundex algorithm allows us to find names that sound like other names. For example, Smith sounds like Smythe. This is done by determining a 4 digit soundex code for each name, and if those soundex codes are the same then we say the names sound alike. For example, Harrigan becomes "H625" and Harrighan becomes "H625" so we say Harrigan sounds like Harrighan. Smith and Smythe are both "S530" so we say they sound alike. If you google the internet you will see various algorithms for soundex.

In this assignment you will write a function to implement the soundex algorithm and then use that function to find names that sound like other names.

The version of the soundex algorithm we will use is as follows:

a) Retain the first letter. If the name is "Jones" then retain "J". If the name is "Trump" retain the "T".
b) Replace each subsequent letter with a number using the following:
   - a, e, i, o, u, y, h, w → 0
   - b, f, p, v → 1
   - c, g, j, k, q, s, x, z → 2
   - d, t → 3
   - l → 4
   - m, n → 5
   - r → 6
c) At this time, we have a long soundex code including zeroes. For example, Harrigan at this point is the String: "H0660205". At this point the length of the original name "Harrigan" and the soundex code "H0660205" are always the same length (length of 8 in this example name). Now, for any adjacent pairs, you delete the second occurrence. In this example we remove the second "6" and thus "H0660205" becomes "H060205".
d) You then remove the zeroes. "H060205" becomes "H625".
e) If the soundex code has a length less than 4, then you pad it with zeroes on the right. For example, the name "Munn" becomes "M5" after the previous step, and it is then padded with zeroes to the right to have a total length of 4, and thus it becomes "M500".
f) If the soundex code has a length greater than 4, then you truncate it. For example, the name "Delaportilla" becomes "D41634" after step (d) above, then you truncate it to a length of 4, and thus it becomes "M416".

g) The soundex code of length four is returned by the function soundex().

This assignment is broken into two Steps.

Step 1:

You are to write a function called soundex with the following header line:

String soundex(String s)

Given a string s of any length, it returns a String of length 4 which is the soundex code.

For example, soundex(“Harrigan”) returns “H625”.

You must follow the steps in the soundex algorithm shown. There are other soundex algorithms available on the internet but you must use the algorithm above.

Step 2:

We can build a "soundex dictionary"—a sketch that lets us type in a name, and automatically find all other names that sound like it.

Proceed as follows:

This step has similarities to the demo code: “ArraySpellChecker.pde”

1) Provided to you is a long list of family names in a file "LastName.txt". Load it with code like the following:

```
surNames = loadStrings( "LastName.txt" );
```

2) Use ControlP5 to allow the user to type in a name. Save this name in a String named something like “lookupName”.
3) Determine the soundex code for lookupName.
4) Loop over the array of surNames and find any that have the same soundex code as lookupName. Whenever you find a match, use println() to print it out.

Requirements and Grading

QUESTION ONE: (12 marks)

[8 marks] Correctness

- The sketch must meet the following requirements:
    - Correctly determines the soundex code: [4 marks]
    - Correctly finds and uses println() to display all surNames that match lookupName. [4 marks]

[2 marks] Coding Style

- Comment your code appropriately. Avoid superfluous comments.
- Correctly and consistently indent your code blocks.
- Use correct inline spacing in function calls, function definitions, and variable declaration and assignment.
- Use good line spacing to chunk sections of your code.
- Pay special attention to inline spacing for your conditional statements
- One or more marks may be deducted for solutions that have obvious inefficiencies.
- Variables that are declared or assigned, but not used.
- Unnecessarily repeating the same code in multiple places.

[2 marks] Visual Design and Creativity

- Higher marks will be given to sketches with extra details for creativity and artistic appeal.

Submitting

Create a folder “A08_username”, but replace “username” with your UW id. So if your email is “jac926@edu.uwaterloo.ca” you would create a folder “A08_ jac926”.

SAVE your sketches in that folder as “A08Q01_username”. Again, replace username with your UW id.

Zip your “A08_username” folder (with “username” replaced by your UW id) and submit it the correct assignment dropbox.

It is your responsibility to submit to the correct dropbox with the correct files before the deadline. Otherwise you will have marks deducted.

Academic Integrity

All assignments in CS106 are done individually. Group work and sharing of code is not allowed.

Detecting Plagiarism:

- We monitor Reddit, File Trading Sites, past year CS106 assignments, etc.
- We use Measure Of Software Similarity (MOSS)
    - automatic system for determining the similarity of code

Discipline

Discipline (Policy 71)

- https://uwaterloo.ca/secretariat-general-counsel/policies-procedures-guidelines/policy-71