INLP, Finite State Automata lab

Andrés F. Lamilla

November 30, 2014

Contents

1 Goal **2**

2 Code **2**

2.1 FSA 2

3 Results **3**

3.1 FSA 3

1 Goal

For this exercise we have to build a FSA to recognize different dates.

2 Code

The main code that has the FSA is called labFSA.py. This code can be run with a python 2.7 interpreter and it is necessary to have the jpbarrette-moman library.

2.1 FSA

For finding the different dates format I build several FSA.

One to find all the months format called MONTH, another to find all the years (YEAR) other for find days (DAY) other to find the separator of day, months and years (ex. [[YEAR]][[DAY]]) called SEPARATOR and finally other that cover all the alphabet but digits (ANY).

In particular the rule MONTH cover all the posible months format, ex: ['January', 'january', 'jan', 'Jan', 'may', '01', '1', '08', '9', '12']. YEAR cover all the years from 0 to 9999 and DAY all the numbers from 0 to 31. SEPARATOR cover all the possibilities of the separators '[-,.]'

With these FSA I build others that find a complete date in a line of text.

1. ANY SEPARATOR MONTH SEPARATOR DAY SEPARATOR YEAR SEPARATOR ANY
2. ANY SEPARATOR DAY SEPARATOR MONTH SEPARATOR YEAR SEPARATOR ANY
3. YEAR SEPARATOR MONTH SEPARATOR DAY
4. ANY SEPARATOR YEAR SEPARATOR ANY
5. YEAR
6. MONTH SEPARATOR DAY SEPARATOR YEAR
7. DAY SEPARATOR MONTH SEPARATOR YEAR
8. MONTH SEPARATOR DAY SEPARATOR YEAR SEPARATOR
9. MONTH SEPARATOR YEAR
10. ANY SEPARATOR YEAR SEPARATOR MONTH SEPARATOR DAY SEPARATOR ANY
11. MONTH SEPARATOR DAY
12. YEAR SEPARATOR MONTH SEPARATOR DAY ANY

13. ANY SEPARATOR MONTH SEPARATOR DAY SEPARATOR YEAR

For example the line:
[[April 20]], 1586 is covered by the rule 13.

3 Results

The file examples_birth_date.txt has a total of 12296 lines with different dates format.

7 of that lines couldn't be analyzed because they had non standard characters.

The program could recognize 11759 lines and was unable to recognize 530 lines.

The accuracy was 95.69

3.1 FSA

The number of lines recognized by each of the FSA were:

ANY SEPARATOR MONTH SEPARATOR DAY SEPARATOR YEAR SEPARATOR ANY: 7660

ANY SEPARATOR DAY SEPARATOR MONTH SEPARATOR YEAR SEPARATOR ANY: 1296

YEAR SEPARATOR MONTH SEPARATOR DAY: 1189

ANY SEPARATOR YEAR SEPARATOR ANY: 918

YEAR: 270

MONTH SEPARATOR DAY SEPARATOR YEAR: 150

DAY SEPARATOR MONTH SEPARATOR YEAR: 65

MONTH SEPARATOR DAY SEPARATOR YEAR SEPARATOR: 59

MONTH SEPARATOR YEAR: 59

ANY SEPARATOR YEAR SEPARATOR MONTH SEPARATOR DAY SEPARATOR ANY: 50

MONTH SEPARATOR DAY: 48

YEAR SEPARATOR MONTH SEPARATOR DAY ANY: 33

ANY SEPARATOR MONTH SEPARATOR DAY SEPARATOR YEAR: 11

Here we can see that the most common type of date in the file has the form MONTH DAY YEAR