|  |  |
| --- | --- |
| |  | | --- | |  | |
| **Natural Language Processing** |
|  |
|  |

**Group Members**

DATE NORMALIZATION

**S. Abu Owais Bin Nasar B-10101086**

**M. Umair Khan B-10101067**

**Anus Ejaz B-10101104**

**Hasan Moiz B-10101037**

**Section B**

Class

VIII Semester BSCS-(Morning)

Course Instructor

Sir Badar Sami

Department

Computer Science

UNIVERSITY OF KARACHI

**“DATE NORMALIZATION”**

**NORMALIZATION**

When extracting nonlinguistic entities, the entities encountered are normalized to group like entities according to predefined formats.

**DATES IN NORMALIZATION**

By default dates in an English template are recognized in the American style date format; that is: month, date, year. Other countrie has their own formats.

Many countries use the ISO YYYY-MM-DD date format generally, and they have the advantage of being unambiguous. (YYYY means four-digit year, MM means two-digit month, and DD means two-digit day.) Local conventions can vary, and for the commonly used Gregorian calendar include formats like DD-MM-YYYY, MM-DD-YYYY, YYYY.

Basic components of a calendar date for the most common calendar systems:

Y – Year

M – Month

D – Day

(year, month, day), e.g. 1996-04-22

(day, month, year), e.g. 22.04.96 or 22/04/96 or 22 April 1996

(month, day, year), e.g. 04/22/96 or April 22, 1996

Separators of the components

"/" – Slash

"." – Dots or full stops

"-" – Hyphens or dashes

Some Countries Date Format

|  |  |  |
| --- | --- | --- |
| **COUNTRY** | | **DATE FORMAT** |
| Canada  China  India  Iran | | YYYY MM DD, DD MM YYYY AND MM DD YYYY  YYYY MM DD  DD MM YYYY  DD MM YYYY and YYYY MM DD |
| Iraq  Pakistan  Saudi Arabia  United Kingdom  United State of America | DD MM YYYY  DD MM YYYY  DD MM YYYY  DD MM YYYY and MM DD YYYY  YYYY MM DD | |

Mostly Countries used the DD MM YY Date Format.

**WORKING OF PROJECT**

This project is developed in Visual Studio and language is used C#.

As this project can extract all the date format which are given. It reads all the text from the file and extract all formats of the date.

We are using Regular Expression to extract all the formats of date.

These are given below

**YY DD MM and YY MM DD**

"\d{4}[- /.]([1-9]|0[1-9]|1[012])[- /.]([1-9]|0[1-9]|[12][0-9]|3[01

"\d{4}[- /.](January|February|March|April|May|June|July|August|September|October|November|December| JANUARY|FEBRUARY|MARCH|APRIL|MAY|JUNE|JULY|AUGUST|SEPTEMBER|OCTOBER|NOVEMBER|DECEMBER| january|february|march|april|may|june|july|august|september|october|november|december| JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC| jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec| Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)[- /.]([1-9]|0[1-9]|[12][0-9]|3[01

"\d{4}[- /.](0[1-9]|[12][0-9]|3[01])[- /.]([1-9]|0[1-9]|1[2])"

"\d{4}[- /.](0[1-9]|[12][0-9]|3[01])[- /.](January|February|March|April|May|June|July|August|September|October|November|December| JANUARY|FEBRUARY|MARCH|APRIL|MAY|JUNE|JULY|AUGUST|SEPTEMBER|OCTOBER|NOVEMBER|DECEMBER| january|february|march|april|may|june|july|august|september|october|november|december| JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC| jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec| Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)"

**DD YY MM and DD MM YY**

"\d{1,2}(?:[- / .])(January|February|March|April|May|June|July|August|September|October|November|December| JANUARY|FEBRUARY|MARCH|APRIL|MAY|JUNE|JULY|AUGUST|SEPTEMBER|OCTOBER|NOVEMBER|DECEMBER| january|february|march|april|may|june|july|august|september|october|november|december| JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC| jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec| Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)(?:[- / .])\d{1,4}"

"\d{1,2}(st|nd|rd|th|ST|ND|RD|TH)( January | February | March | April | May | June | July | August | September | October | November | December | JANUARY | FEBRUARY | MARCH | APRIL | MAY | JUNE | JULY | AUGUST | SEPTEMBER | OCTOBER | NOVEMBER | DECEMBER | january | february | march | april | may | june | july | august | september | october | november | december | JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG | SEP | OCT | NOV | DEC | jan | feb | mar | apr | may | jun | jul | aug | sep | oct | nov | dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec)\d{1,4}"

"(0[1-9]|[12][0-9]|3[01])[- /.]([1-9]|0[1-9]|1[2])[- /.]\d{4}"

"\d{1,2}(?:[- / .])\d{1,4}(?:[- / .])(January|February|March|April|May|June|July|August|September|October|November|December| JANUARY|FEBRUARY|MARCH|APRIL|MAY|JUNE|JULY|AUGUST|SEPTEMBER|OCTOBER|NOVEMBER|DECEMBER| january|february|march|april|may|june|july|august|september|october|november|december| JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC| jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec| Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)"

"\d{1,2}(st |nd |rd |th |ST |ND |RD |TH )\d{1,4}( January| February| March| April| May| June| July| August| September| October| November| December| JANUARY| FEBRUARY| MARCH| APRIL| MAY| JUNE| JULY| AUGUST| SEPTEMBER| OCTOBER| NOVEMBER| DECEMBER| january| february| march| april| may| june| july| august| september| october| november| december| JAN| FEB| MAR| APR| MAY| JUN| JUL| AUG| SEP| OCT| NOV| DEC| jan| feb| mar| apr| may| jun| jul| aug| sep| oct| nov| dec| Jan| Feb| Mar| Apr| May| Jun| Jul| Aug| Sep| Oct| Nov| Dec)"

"(0[1-9]|[12][0-9]|3[01])[- /.]\d{4}[- /.]([1-9]|0[1-9]|1[2])"

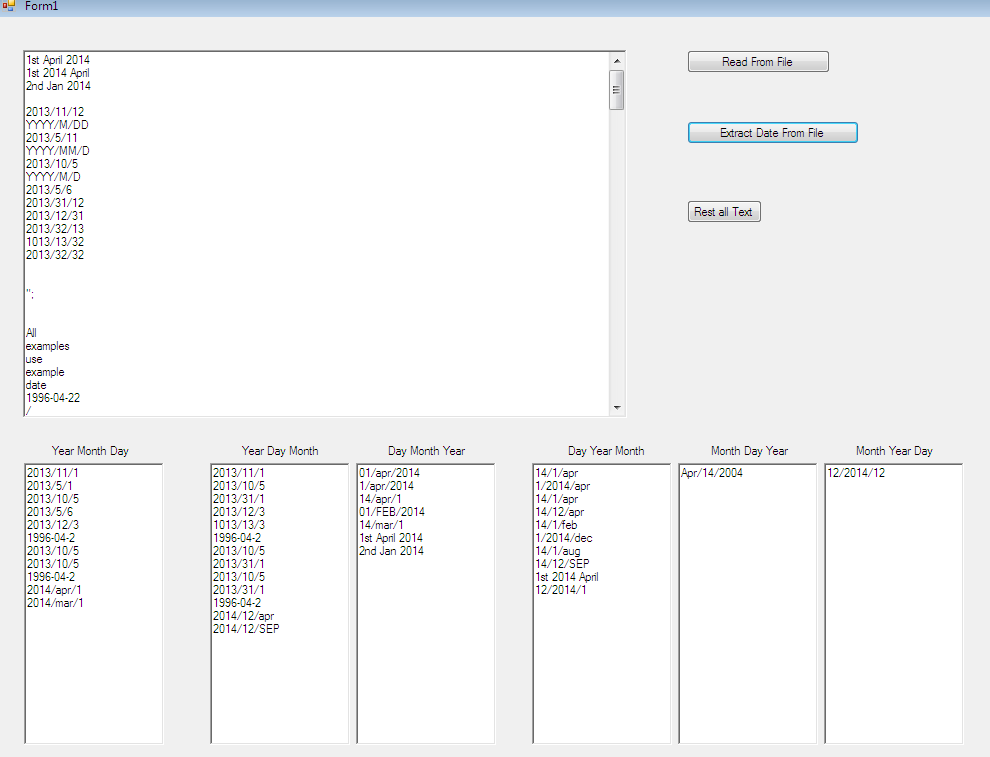
**MM DD YY and MM YY DD**

"(January|February|March|April|May|June|July|August|September|October|November|December| JANUARY|FEBRUARY|MARCH|APRIL|MAY|JUNE|JULY|AUGUST|SEPTEMBER|OCTOBER|NOVEMBER|DECEMBER| january|february|march|april|may|june|july|august|september|october|november|december| JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC| jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec| Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)(?:[- / .])\d{1,2}(?:[- / .])\d{1,4}");

"([1-9]|0[1-9]|1[2])[- /.](0[1-9]|[12][0-9]|3[01])[- /.]\d{4}"

"(January|February|March|April|May|June|July|August|September|October|November|December| JANUARY|FEBRUARY|MARCH|APRIL|MAY|JUNE|JULY|AUGUST|SEPTEMBER|OCTOBER|NOVEMBER|DECEMBER| january|february|march|april|may|june|july|august|september|october|november|december| JAN|FEB|MAR|APR|MAY|JUN|JUL|AUG|SEP|OCT|NOV|DEC| jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec| Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)(?:[- / .])\d{1,4}(?:[- / .])\d{1,2}"

"([1-9]|0[1-9]|1[2])[- /.]\d{4}[- /.](0[1-9]|[12][0-9]|3[01])"



Conclusion:

The goal of this project is to get all the formats of date. Many countries has their own formats of date. By using this project, it extract all the formats of the date.