Lab Exercise 3 : Regular Expression

In this lab exercise, we will perform some text pre-processing tasks using Regular expression.

1. Import the necessary libraries.
2. Load the text in **labE4.txt** to a variable `text`. Here is the sample screenshot of the file:

```
WEBVTT

NOTE duration:"00:13:48.4180000"

NOTE language:en-us

NOTE Confidence: 0.96506834

c17ea77e-3314-46db-b21e-cd9b76ef1baf
00:00:11.030 --> 00:00:12.566
Welcome back today.

NOTE Confidence: 0.96506834

a5e85a51-935b-46b4-aae6-b9bf8d3d916a
00:00:12.566 --> 00:00:16.798
We'll see one more technique to reduce the

NOTE Confidence: 0.96506834

c36dd7c9-901b-4b8e-9488-f5b6162ff6d0
00:00:16.798 --> 00:00:20.128
volume of your text using lemmatization.

NOTE Confidence: 0.96506834

92bc02cb-73d8-4a26-b7d3-09206fcab0a7
00:00:20.130 --> 00:00:22.620
Previously we had discussed about stemming,
```

3. Apply **regular expression** and clean the data. The expected output:

```
Welcome back today. We'll see one more technique to reduce the volume of your text using
lemmatization. Previously we had discussed about stemming, so stemming is the process of also
reducing the text into some normal form, which is called as the stem. Today we'll see again a
normalization technique which is called as limitation, which will reduce the tokens into root word,
also called as the lemma. We'll also discuss about resource, which is used by lemmatization that is
called as Wordnet. So we'll have a discussion about that. And we'll also have a discussion of whether
to go for stemming or whether to go for limitation. So let's begin. So let's start with the
limitation. Before we start the discussion, let us first go to the results of a few words. The stem
words that are generated for the few words from these Timor algorithms. So you have Sam and some. The
base word that is generated or the stem that is generated from a stem. Morris sang for song. It is
sung itself and for sang also it is sang itself sings it is singing, so it's stripped. We know that
the suffixes are removed for singing ING is removed and you get sing. But all of us know that even
sang and sung both are derived from the word sing. So this is not done in the Stemmer algorithms.
Why? Because there are no suffixes as such that are being attached to the root word, so. To handle
such kind of problem where all the verb forms. So all these things are reduced to sing, we have to go
for limitation. So let's see. So limitation is the process of mapping all the different word forms to
its root word, not the stem. So we are going to the root word which is called as Lima. So some sang
and sings. All these words are reduced to single women. We had seen this example earlier in the
stemming algorithm. So women is reduced to women. So women, women both are reduced to woman was is
reduced to be. Settling is reduced to settle better, is reduced to good, so these are the things
which are seen in limitation. So like stemming also limitation will remove the affix is to get the
```

4. Find the regular expressions for :

- Date of form: “MM-DD-YYYY”
- Time with format: “HH:MM:SS”
- Time with format: “HH:MM:SS AM” or “HH:MM:SS PM”
- Full Name: “Mr. ……” or “Ms. ……” or “Mrs. …….”
- Enrollment number for Amrita University
- Indian Landline Number : eg :- 0476-2802017 or 022-12345678
- Any Mobile Numbers: +91-1234567890, +12-3456789000
- Indian Postal Pin code: 690 525
- Tokenize a sentence (?!.)
- finding any 4 digit in a text.
- Card Number :(0000-0000-0000-0000)
- Twitter User : (@username)
- IP Address
- IPv6 Address : eg - 2001:cdba:0000:0000:0000:0000:3257:9652