Handling Diacritics

A **diacritic** also called a **diacritic** or **diacritical mark**, **point**, or **sign**, is a small sign added to a letter to alter pronunciation or to distinguish between similar words. In codiing we often come across them while reading input from webpages.

There might be some API's available to handle the task, but we are choosing hard way to do that to save time to explore them and then test them. We need to perform the following steps in order to find and replace them with similar sounding words.

1. Download the page which contains diacritics eg: http://geoiptool.com/en/?IP=192.42.43.22 contains Neuchatel with 'a' as a diacritic.
2. From a UNIX machine get the dump of the page using command line tool 'xxd' and grep the word so that you get hex dump of the alphabet required eg:

   ```
   xxd index.html?IP=192.42.43.22 | grep Neuch
   ```

   a. The output would be something like :

      ```
      0001d00: 6c64 223e 4e65 7563 68e2 7465 6c3c 2f74  ld">Neuch?tel</t
      ```

   b. Every two char at left represent one char at right and its starts after the colan":" eg 6c represents l and 64 represents d
3. Count the char to find the missing alphabet which "e2" in our case.
4. Replace the alphabet using the pattern matching for hex by \x

   ```
   if($city =~ m/\xe2/){
         $city  =\~ s/\xe2/a/g;
    }
   ```