SAMSUNG APPATHON 2017

Pho-hinix (Phonetic library for Hindi)

Problem Statement:

Develop a library, which can match words based on similar phonetics

Description:

In multilingual natural language processing system, there are many challenges in information retrieval. One of the biggest challenge is matching two set of strings which are not exactly same but are equivalent. For languages like Hindi where it is required to handle various pronunciations by people from various communities for the same keyword, especially in rural areas. This requires some sort of approximate matching algorithms (phonetic or semantic) to improve the recall of information retrieval systems. There are many phonetic algorithms available for English and other European languages, however for many Indian languages like Hindi they do not perform well. Having the matras in Hindi also adds up to the complexities where a small variation in spelling can impact the results heavily.

Goal of this problem is to develop either phonetic or semantic matching algorithm for matching such words for Hindi language

Ex: Many English words and utterances have similar phonetics for example “John” and “Jon”. Similarly there are Hindi words, which are having similar Phonetics for example “डीपेश” & “दीपेश”.

Instruction

1. Coding Language: C, C++, Java
2. More Test cases [15] will be shared on the live coding day
3. Coding Duration: 12 Hours
4. Evaluation is done based on many corner cases which will be tested offline
5. Preferably Do Not Use Open Source Libs

Test Cases:

Word 1	Word 2	Result	Description
संतोष	संथोष	1	Should Match
शून्य	शुन्य	1	Should Match
एक	इक	1	Should Match
वास्तु	वस्तु	0	Should not Match
शूर	सुर	0	Should not Match