Romanagari Detection in Twitter
Hrishikesh Terdalkar - Shubhangi Agarwal

Problem:

Romanagari is writing Devanagari-script-based-language in Roman script. We often see examples of these in social media such as twitter. e.g. "kya kar rahe ho". Most NLP techniques deal with only English text, and thus in majority of twitter based studies, such tweets have to be discarded as noise. We intend to tackle this problem, by attempting detection and translation of such tweets. Targeted languages are Hindi and Marathi.

Approach:

The main task is that of "back-transliteration". That is, going back to Devnagari script from Romanagari text. After which, translation is mere formality.

Steps:

- Create dictionary of Romanagari words, by handling words at phoneme level
- Consider common inflections of Romanagari lexemes
- Collect / Obtain a twitter-dataset containing Romanagari text
- Attempt the back transliteration followed by translation
  - for this work by Barman, Utsab, et al 2013 [2] can be used as reference point

Dataset Collection:

Since we are working on twitter, dataset will be (most likely) collected with help of existing tweet-collection software. We choose python programming language. One of the notable tools available is,

- **Tweepy**: https://github.com/computermacgyver/twitter-python

Details about usage at:

References:

6. Proceedings of Social india 2014