Annotation Efficient Language Identification Using Weak Labels

Shriphani Palakodety*
Onai (spalakod@onai.com)

Ashiqur KhudaBukhsh*
CMU (akhudabu@cs.cmu.edu)

Indic Text Mining:

- Noisy
- Multilingual
- Romanized

Language Identification:

- Critical first step
- SOTA weak at:
  - ***Noisy***
  - ***Romanization***

Solution:

- Support ***Noisy, Romanized***
- ***Annotation Efficient***

Method:

- Obtain monolingual clusters

  [Palakodety, KhudaBukhsh, Carbonell, 2020]
- Label 20 documents per cluster
- Assign labels to other cluster members ***(weak labels)***
- ***260 annotations***
- ***2.4 million weak labels***

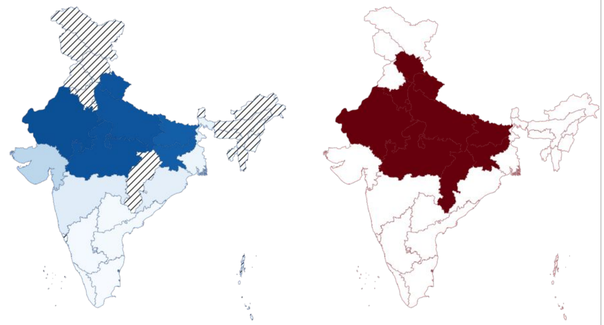
https://github.com/onai/indic-language-identification

ONAI

Carnegie Mellon University

Weak Geographic Extents:

- Online Hindi Extent Matches Hindi Belt



Bi / Multi Linguality: