Proposal for CDS Project

[55 41 59 46 32]

October 21, 2019

1 Music Instrument Classifier

We plan to train a classifier which (given an audio sample) can classify which instrument it belongs to. Many similar Dataset are available publically available at http://musicinformationretrieval.wordpress.com

In the above dataset most instruments are western, however we are thinking to include some of the Indian instruments (in our dataset) like Shenai, Tabla, etc.

2 Author Identification: This article seems interesting, but who is the author?

The initial idea was to collect some editorial articles from various indian national dailies viz. Hindu, The Telegraph, The Indian Express, Business Standard. Some of the newspapers are meant for different audiences and have different usage of english vocabulary. We are planning to train a classifier which will be able to classify the article i.e. whether it is from newspaper A or B.

But we find a problem here — Given a newspaper, there are multiple authors who co-create content in the editorial sections. So their writing style should also be different.

Alternatively we can try for Given a set of articles from any other sections (apart from editorial) like Sports, Business, Politics we try to attribute it to the Newspaper. Our assumption being the Vocabulary of Hindu is comparatively rich compared to ToI.

To summarize, this idea would be a Shallow Text Analysis based technique for classification of attribution.

Source of dataset: Some of IAS coaching websites provides repository of past published articles in pdf format in a google drive. We plan to build a plain crawler to extract the text from pdf