Classification of Hindi Literature according to Author Writing Style

Dhruv Anand
11251

Srijan Shetty
11727

Motivation

- Document Fraud Detection
- Classifying works from unknown authors
- From a Literary perspective
  - Repeating trends of authors
  - Adopting styles of popular authors

Previous Work

- ➔ Extensive work done on Author Attribution for English (using domain-specific datasets like blogs, emails, forum posts, short stories and novels)
- ➔ No work has been done on Hindi datasets
- ➔ Various lexical and syntactic features have been tried by researchers in this field

Challenges

- ➔ Non-uniform data for Hindi
- ➔ Variance of writing style markers in Hindi Literature
- ➔ Multiple derivative words that must be aggregated without any pre-programmed tool for lemmatization. (The language is morphologically rich.)

Problem Statement

- Apply known methods of Author Attribution to a Hindi dataset
- Analyse difference in effectiveness of various methods between English and Hindi
- Exploring new types of lexical and syntactic features to give better results for Hindi Literature

Methodology

Proposed Features

- Word n-grams
  - Stemmed/non-stemmed unigrams
  - Collocations (bigrams)
- Character n-grams
- Sentence length distribution
- Word length distribution
- Feature word frequency distribution

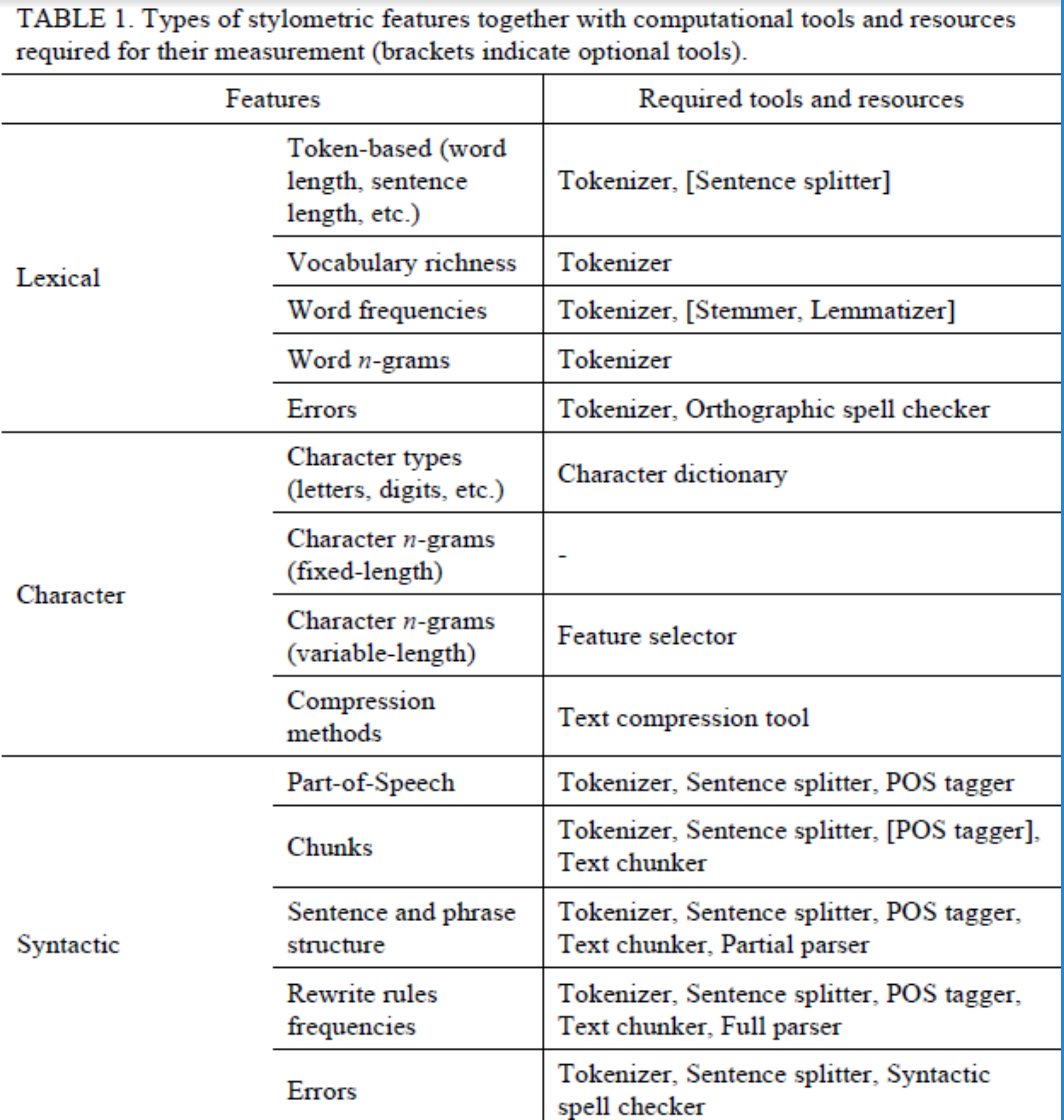


TABLE 1. Types of stylometric features together with computational tools and resources required for their measurement (brackets indicate optional tools).

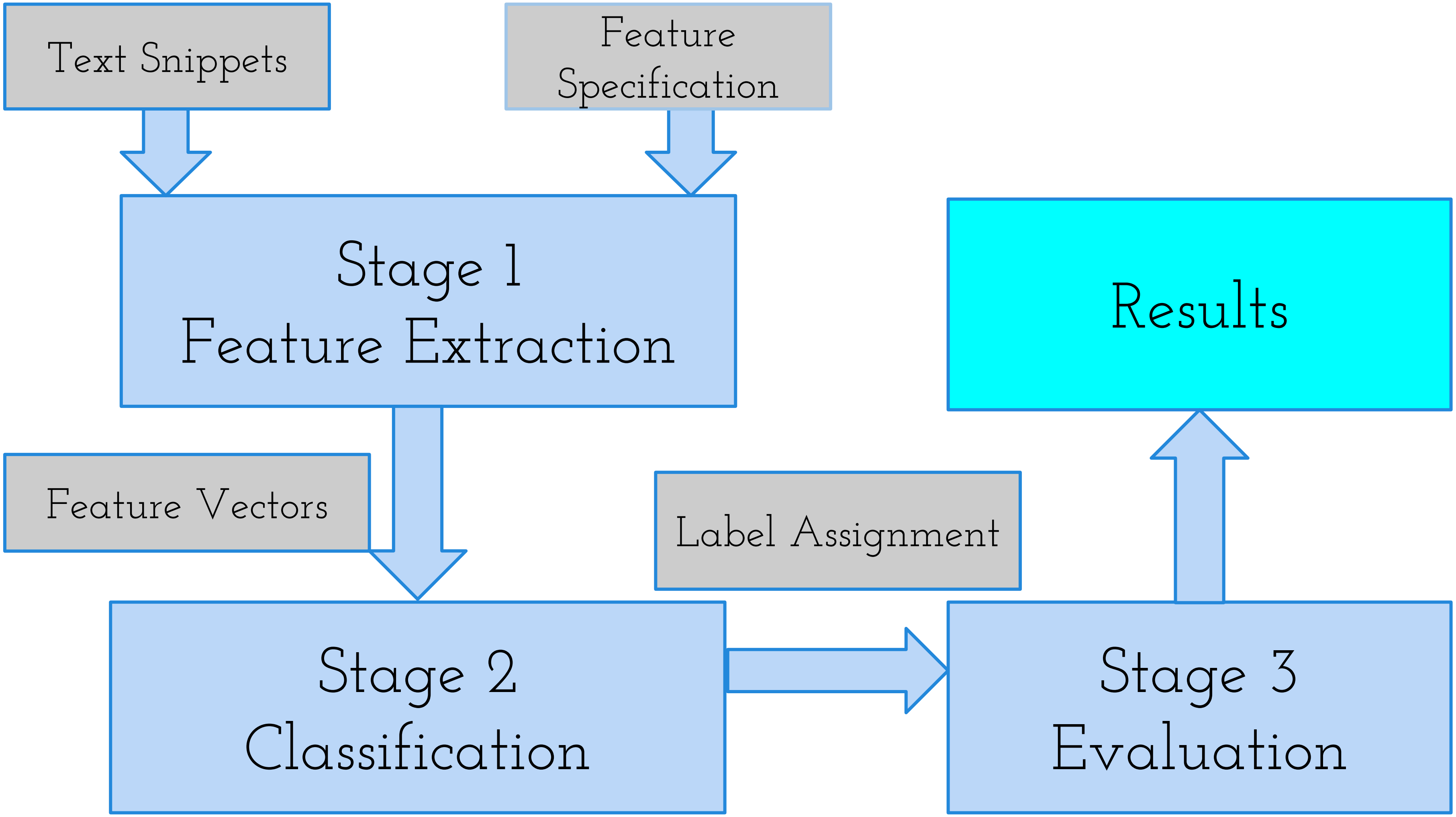
Features		Required tools and resources
Lexical	Token-based (word length, sentence length, etc.)	Tokenizer, [Sentence splitter]
	Vocabulary richness	Tokenizer
	Word frequencies	Tokenizer, [Stemmer, Lemmatizer]
	Word *n*-grams	Tokenizer
	Errors	Tokenizer, Orthographic spell checker
Character	Character types (letters, digits, etc.)	Character dictionary
	Character *n*-grams (fixed-length)	-
	Character *n*-grams (variable-length)	Feature selector
	Compression methods	Text compression tool
Syntactic	Part-of-Speech	Tokenizer, Sentence splitter, POS tagger
	Chunks	Tokenizer, Sentence splitter, [POS tagger], Text chunker
	Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser
	Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser
	Errors	Tokenizer, Sentence splitter, Syntactic spell checker

***image from [Sta09]**

Classification

- Supervised
  - SVMs
  - Bayesian Multinomial Regression (BMR)
- Unsupervised
  - K-means clustering

Framework



Text Snippets

Feature Specification

Stage 1
Feature Extraction

Results

Feature Vectors

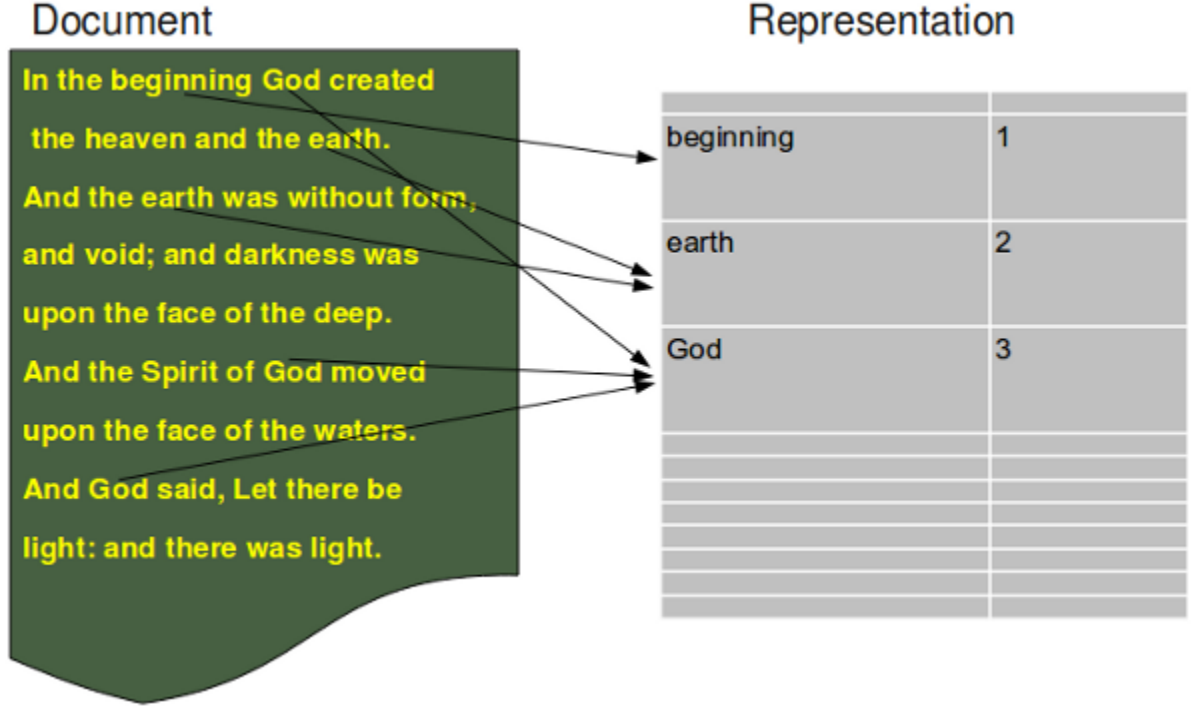
Label Assignment

Stage 2
Classification

Stage 3
Evaluation

A bit of theory

Bag of Words

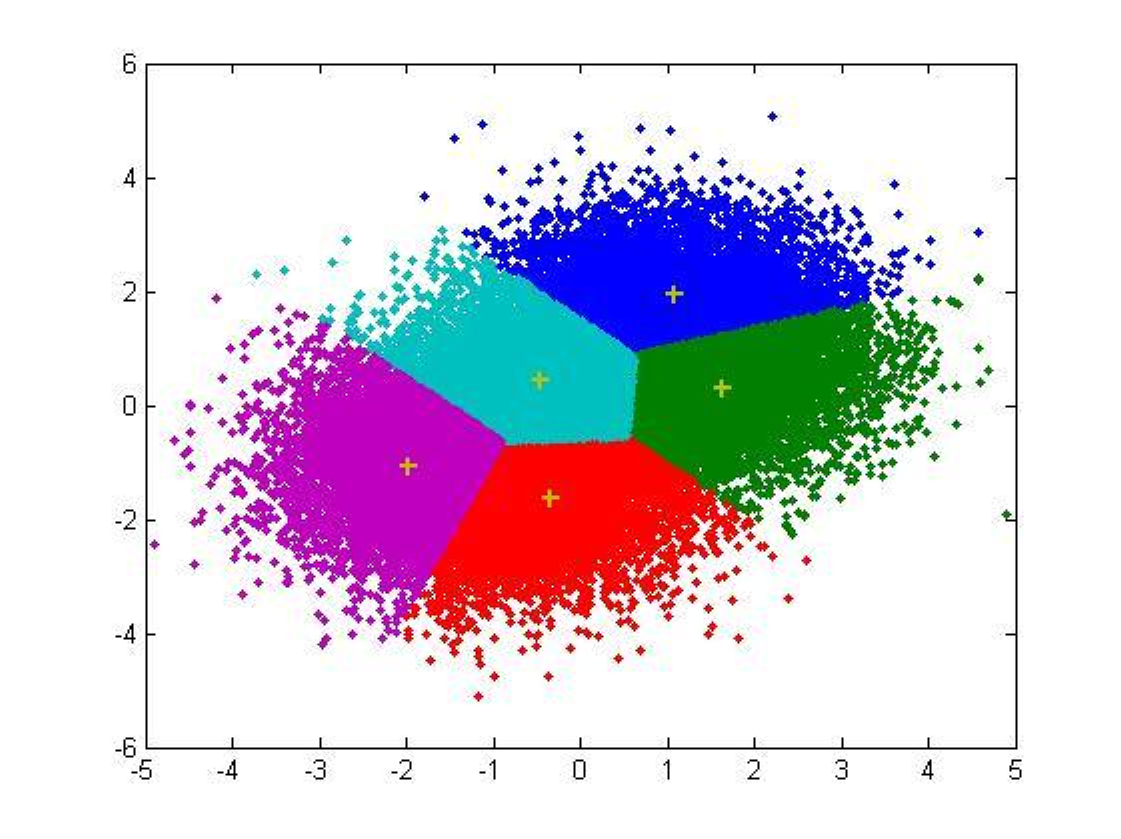


In the beginning God created
the heaven and the earth.
And the earth was without form,
and void; and darkness was
upon the face of the deep.
And the Spirit of God moved
upon the face of the waters.
And God said, Let there be
light: and there was light.

beginning	1
earth	2
God	3

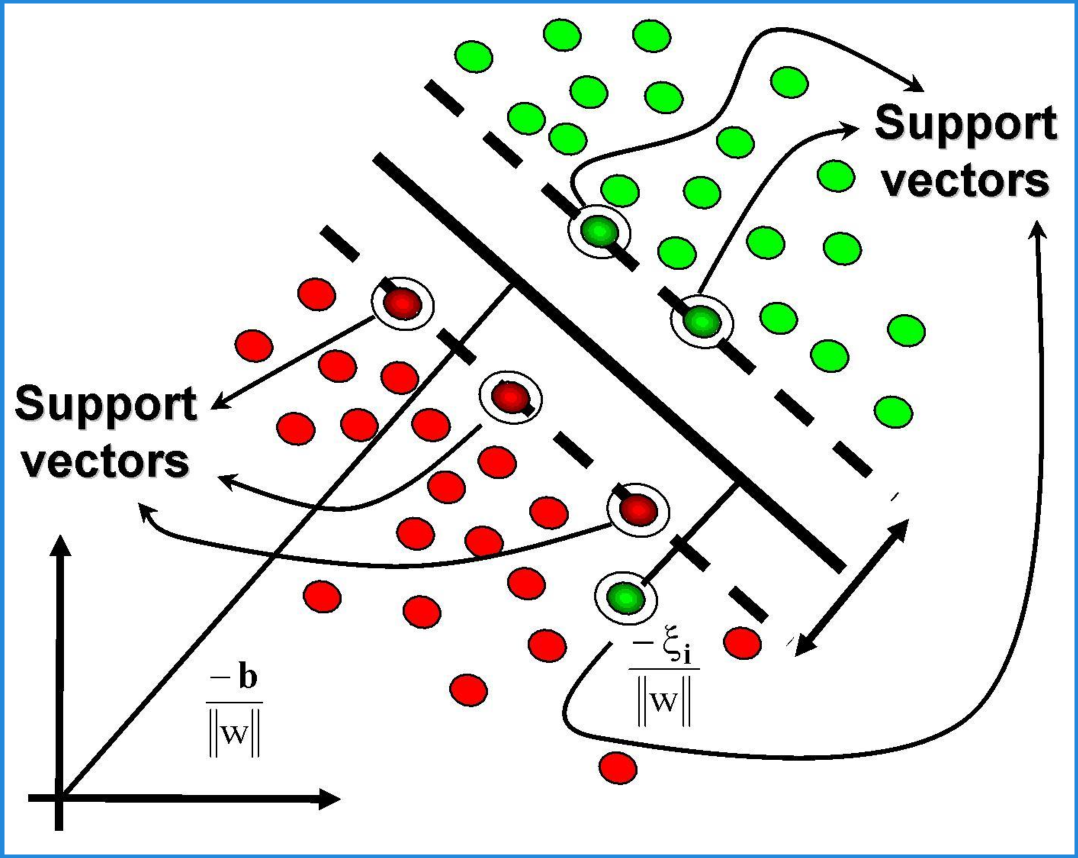
http://www.python-course.eu/images/document_representation.png

K Means



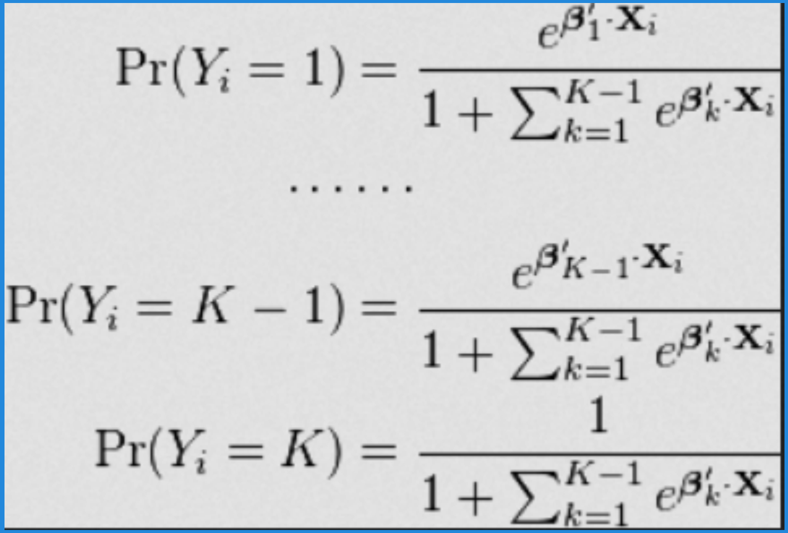
(http://www.mathworks.com/matlabcentral/fileexchange/screenshots/2240/original.jpg)

SVM



http://www.thebookmyproject.com/wp-content/uploads/Intrusion-Detection-Technique-by-using-K-means-Fuzzy-Neural-Network-and-SVM-classifiers.jpg

BMR



$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}'_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}'_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

$$\Pr(Y_i = K - 1) = \frac{e^{\boldsymbol{\beta}'_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}'_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}'_k \cdot \mathbf{X}_i}}$$

http://upload.wikimedia.org/math/2/e/e/2eeac600b65d77080381284f530f37d4.png

Where do we stand

Dataset Compilation

- No standard dataset for classical/contemporary hindi authors (novels and stories)
- Scraped HindiSamay.com manually to build a database of Classical Hindi literature.
    - 5 authors
    - 2-4 lakh words per author
- Each author's work has been divided into multiple snippets of 500 words.

Unigrams

- Belief: Authors repeat the same set of words
- Stemming: BOW using all tokens and BOW using 4500 most frequent words (>20 frequency in the entire corpus)
- Classification: K-means on 3 classes (RNT, Premchand, V.N.Rai) and on 5 classes.
- Results for 3 classes:
  - Average Precision: 50% (v/s baseline of 33%)
  - Average Recall: 48% (v/s baseline of 33%)

Results with 5 authors

	0	1	2	3	4	Snippets	Precision	Recall
RNT	111	14	20	0	6	151	22.65%	73.5%
Prem	108	21	58	0	211	398	71.77%	53.01%
Dharamvir	11	24	14	150	2	201	100%	74.6%
Sarat	142	332	3	0	65	542	82.19%	61.25%
VN	118	13	277	0	10	418	74.46%	66.26%

Insights

- Corpus has mostly stories for Rabindranath Tagore, both recall and precision for him are low indicating that across multiple works frequent words used by author change.
- Corpus contained only novels for Premchand and so both recall and precision for him were high > 70%
- The corpus contained essays by V.N.Rai, indicating high amount of content words.

Future Work

In the coming weeks

- Use collocations (bigrams) to as a feature.
- Analyzing sentence structure:
  - Sentence lengths
  - Number of subjects, verbs, objects in a sentence (instead of POS tagging we will lookup common words from HindiWordNet)
- Reducing dimensionality using *PCA*.
- Training on multiple features together (using *multivariate discriminant analysis*)
- Improving results by tuning snippet length and parameters used in classification.

In the future

- Exploring the possibility of using a morphological tagger to get more accurate style measures for authors.
- Extending the method to Hindi tweets, forum comments and messages to compare accuracy.

References

Literature

1. **[KSA09]** Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol., 60(1):9-26, January 2009.
2. **[KSA11]** Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. Lang.Resour. Eval., 45(1):83-94, March 2011.
3. **[Sta09]** Efstathios Stamatatos. A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol., 60(3):538-556, March 2009.

Tools Used

- ZSH
- Python Modules
  - indicngram
  - nltk, scipy, scikit-learn
- Snippets of code have been taken from
  - http://www.csc.villanova.edu/˜matuszek/spring2012/snippets.html

*www.python.org

THANK YOU!

Questions?