Project Scope Document :

Text Processing Framework for HINDI

Team number 51

Mentor : Pulkit Parikh

Project Number : 2

Group Members

Utsav Chokshi, *PG1, 201505581*
Deepanshu Vijay, *UG3*, 201302093
Arushi Dogra*, UG3,201302084*

Problem Statement

To develop text processing framework for Hindi Language. Framework includes all basic text processing algorithms like Stop Word detection, Tokenisation, Sentence Breaker , POS tagging, Concept Identification, Named Entity Recognition and Categorization of documents for Hindi language.

Applications

1) Efficient retrieval of hindi language resources
2) Annotating/Categorizing hindi documents
3) Building responsive UIs in hindi language

Challenges

Developing such a framework is challenging task as hindi language is morphologically very rich and need to handle very subtle variations in writings. Also we need to develop linguistic understanding for hindi language.

Also Hindi has borrowed words from different languages like farsi, urdu, arabic. Words from different languages follow different rules for word formation. Hence, no single rule for word formation is applicable.

Major Project - Second Deliverable Details

Parsing Hindi Documents

1. Tokenization
2. Sentence breaker
3. Stop-word detection [and removal]

Identifying Variations

POS tagging for Hindi Language

4. Identifying various part of hindi sentences like noun, verb etc.
5. Identify key concepts using the same.

Major Project - Third Deliverable Details

Entity Recognition

Categorization of given documents

Tools to be used

Language : Python

Libraries : NLTK for indian languages, libindic

References

1. http://airccj.org/CSCP/vol3/csit3639.pdf [POSTagger]
2. http://talukdar.net/papers/KBCS04_HPL-1.pdf [Tokenization]
3. http://www.aclweb.org/anthology/I08-5014 [Stemming]
4. https://sanchay.co.in/papers/cpms-long-iwlc-06.pdf [Identifying Variations]
5. http://www.enggjournals.com/ijcse/doc/IJCSE12-04-05-213.pdf [Entity Recognition]