Search Engine for Hindi Literature and Arts

Abstract
In this project, a stemmer is going to be implemented with a search engine which indexes documents containing rich text of Hindi Language. We try to develop different rules for the Hindi Language stemmer and evaluate those rules, finding out the ones which turn out to be okay and the ones which don't.

Key Concepts of IR
In order to develop a search engine in Hindi, we need to develop several different rules. The various rules involved in development of engines include stop words removal, normalization, ranking system, zones, fields, etc. However, the key concepts of IR that will be focused in our project are Stemming and Lemmatization. Stemming is the process of clipping off the affixes from the input word to obtain the respective root word. To make the search engine more meaningful and genuine, we will also try to implement Lemmatization. It is the process by which we crave out the lemma from the given word and can also add additional rules to make the clipped word a proper stem. The main purpose of stemming and lemmatization is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. This root form of the words can be indexed to provide better results for complicated queries.

Impact of this project
Hindi, the national language of India is widely spoken in the country and is the most preferred language after English. Hindi not only has one of the richest vocabularies but it has an equally rich script, grammar, word collection, and parts of speech. Our search engine will focus on the major problems of Hindi text searching over the Hindi literature. This will help literature critics and reviewers to appreciate and contribute towards it. Literature search provides not only an opportunity to learn more about a given topic but provides insight on how the topic was studied by previous analysts. It helps to interpret ideas, detect shortcomings and recognise opportunities.

Team Members:
Anubhav Ujjawal (20160010005)
Anurag Gupta (20160010006)
Garvit Kataria (20160010028)