The-35-Project

A project to translate English Wikipedia pages to Punjabi.

The significance of the title of the project is that in the series of Punjabi Varnmala we have total 35 alphabets. Since this project is all about the glory of the Punjabi Language, we named this project as THE 35 PROJECT.

Business Case

This project provides a platform for the people (people who speak punjabi) to easily gain the knowledge of different topics when translated in punjabi.

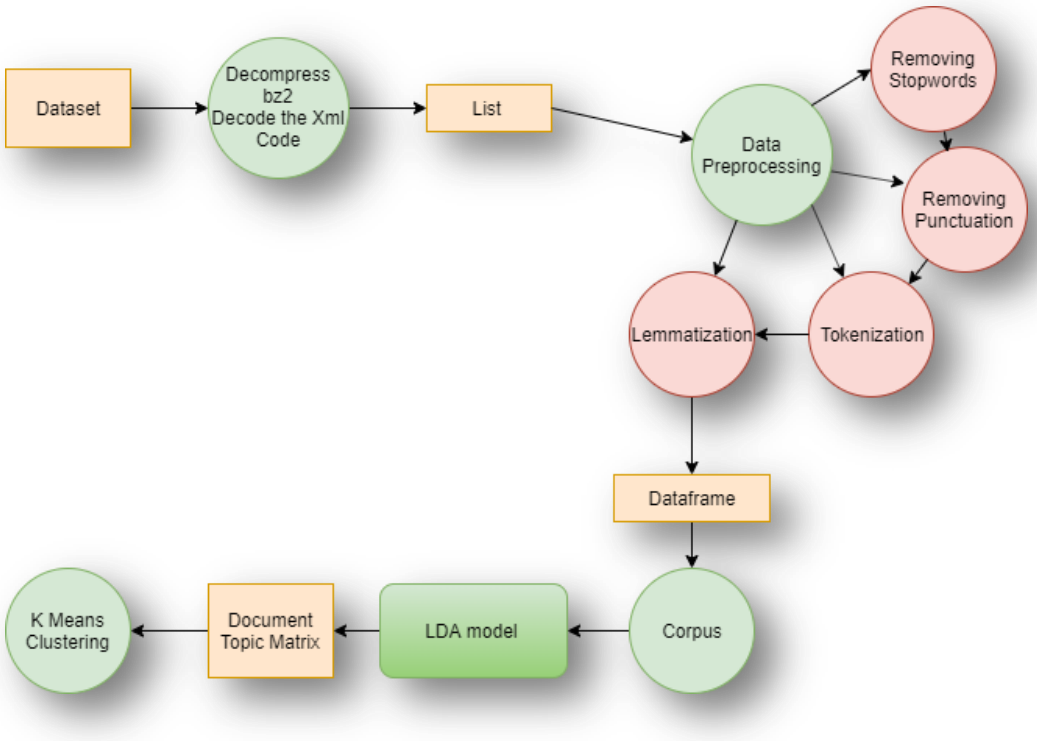
Scope of the Project

The project will affect mass all over the world to easily understand Punjabi Language which has been brought into importance by the Gurus of Sikhism.

The Existing System

The developed project api throws a random topic to the volunteers so that they can translate the article into Punjabi. After this the project can easily translated into Punjabi. The data of the project was taken from Wikimedia Dumps. The Dataset we downloaded is bz2 file containing XML code of the Wikipedia Articles.

Data Flow Diagram



Details of Code Files

The zip file contains all the code files. The code is organised into 9 .py files. Each file represents an atomic function. These files in order of their execution are:

1. import_modules.py
2. data_decode.py
3. data_cleaning.py
4. lematize.py
5. create_corpus.py
6. LDA.py
7. clustering.py
8. services.py
9. API.py

The API.py file can be used to execute the whole code.