Software Requirements Specification

for

Developing a corpus of sentiment analysis of Hindi text

Prepared by Tarun Gupta, Aryan Verma, Kartik Garg and Srijan Saini

Indian Institute of Technology Indore
Software Engineering Project
Discipline of Computer Science and Engineering

Friday, 14. February 2020

Contents

Revision History

Revision	Date	Author(s)	Description
1.0	14.02.2020	Kartik Garg, Tarun Gupta, Srijan Saini, Aryan Verma	Initial Draft of SRS
2.0	29.05.2020	Kartik Garg, Tarun Gupta, Srijan Saini, Aryan Verma	Updated Document Conventions, Product Scope, Assumptions and Dependencies

Chapter 1

Introduction

1.1 Purpose

The purpose of this project is to a create database for sentiment analysis of Hindi text. There exists a wide variety of databases of English text for sentiment analysis such as Amazon product review database, however datasets available for Hindi language are very limited or not up to the mark for some specific purposes. In this project we aim to create one of the largest Hindi text database for sentiment analysis and natural language processing. We further aim to create a user interface to predict polarity of a movie review by creating and using a NLP model trained on our database.

1.2 Document Conventions

This document uses following conventions:

- ML: Machine Learning.
- NLP: Natural Language Processing.
- JSON: Javascript Object Notation.
- CSV: Comma separated value.
- ULMFIT: Universal Language Model Fine-Tuning
- iNLTK: Natural Language Toolkit for Indic Languages
- CLTK: Classical Language Toolkit
- HTML: Hypertext Markup Language
- CSS: Cascading Style Sheets

1.3 Intended Audience and Reading Suggestions

Intended Audience:

- Researchers doing work on Hindi Text.
- Students.

Reading Suggestions:

- Manning, Christopher D., Christopher D. Manning, and Hinrich Schütze. Foundations of statistical natural language processing. MIT press, 1999.
- Jain, Leena, and Prateek Agrawal. "Text independent root word identification in Hindi language using natural language processing." International Journal of Advanced Intelligence Paradigms 7.3-4 (2015): 240-249.
- Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." arXiv preprint arXiv:1801.06146 (2018).

1.4 Product Scope

Sentiment analysis models detect polarity within a text (e.g. a positive or negative opinion), whether it's a whole document, paragraph, sentence, or clause. Understanding people's emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before. However to do sentiment analysis the primary requirement is a large database to work on. However, for Hindi language the databases for sentiment analysis available as of today are very small in size. We aim to provide the largest Hindi text database for researchers who wish to apply methods and techniques of natural language processing on Hindi text.

1.5 References

- Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019.
- Abadi, Martín, et al. "Tensorflow: A system for large-scale machine learning." 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). 2016.
- Chollet, François. "Keras documentation." keras. io (2015).
- McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012.
- Richardson, Leonard. "Beautiful soup documentation." April (2007).

Chapter 2

Overall Description

2.1 Product Perspective

The database will be made freely available to any person requiring it for non-commercial purposes after they fill an End-User License Agreement (EULA) form. Further the code for natural language processing is supposed to be an open source, under the GNU general Public License.

2.2 Product Functions

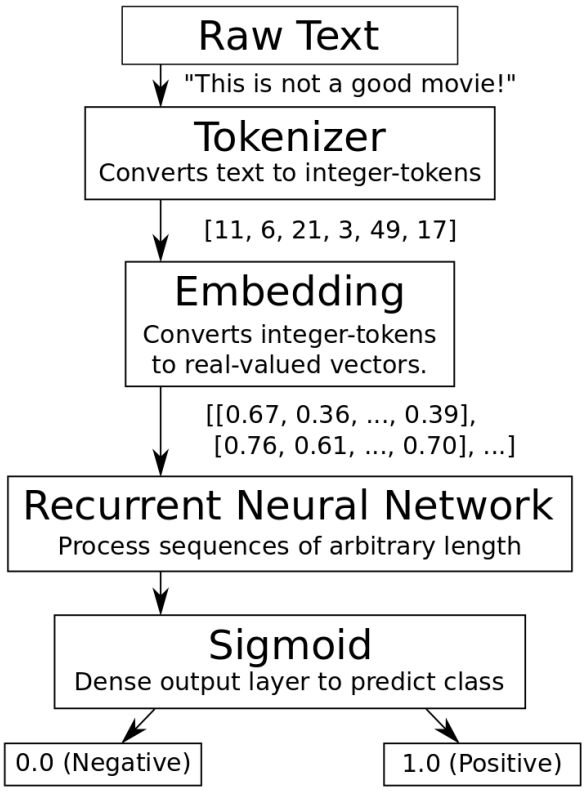

Figure 2.1: Basic NLP pipeline

The application stores project data in both CSV and JSON format to enable easy integration with 3rd party applications.

After creating the database we further aim to apply natural language processing techniques to do sentiment analysis of the extracted Hindi Text. Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural language generation.

As seen in Fig. 2.1, we first convert raw Hindi Text to tokens using tokenizer from Classical Language Toolkit. This tool can break a sentence into its constituent words. It simply splits the text into tokens of words and punctuation's. Then we can simply use Recurrent Neural Network models of Keras or Tensorflow (GRU, LSTM, etc.) to build our machine learning model.

2.3 User Classes and Characteristics

The product caters to the needs of users who happen to be machine learning enthusiasts, students, researchers and anyone who wishes to explore the scope of natural language processing and wishes to conduct research experiments on Hindi text. It is considered that the user do have the basic knowledge of reading and manipulating CSV and JSON files. Further to use the database for NLP the user should have basic knowledge of some high level language like Python, MATLAB etc. and its machine learning libraries.

2.4 Operating Environment

The operating environment for using our Hindi text corpus is listed as below.

- Operating Environment: Windows(7 or higher)/ Linux/ macOS
- any application for reading JSON or CSV files is required.

The operating environment for using our NLP model of sentimental analysis using our Dataset is listed as below.

- Operating Environment: Windows(7 or higher), Linux, macOS
- Python (3.6 or higher) required.

2.5 Design and Implementation Constraints

2.6 User Documentation

The data consists of Bollywood movie names, their ratings as given by critics or users out of 5, their reviews in Hindi language as well as their polarity (0 or 1) scrapped from Hindi websites. This data from different websites is assembled into a CSV file to get the desired dataset.

2.7 Assumptions and Dependencies

The PyTorch model used is `1.5.0+cu101`, and fastai model used is `1.0.61`. These can be installed easily to reproduce the results using the following commands:

- pip install torch==1.5.0+cu101 -f https://download.pytorch.org/whl/torch_stable.html
- pip install fastai==1.0.61

Dependencies -

- BeautifulSoup - It is a library that makes it easy to scrape information from web pages by pulling out data from HTML and XML files.
- Requests - It is a python module used to send all kinds of HTTP requests which helps in opening URLs.
- Pandas - It is used for data manipulation and analysis as it offers data structures and operations for manipulating numerical tables and time series.
- NLTK - The Natural Language Toolkit is a suite of libraries and programs for symbolic and statistical natural language processing in the Python progamming language.
- Numpy - It is a library which adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- TensorFlow - It is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.
- Keras - It is a powerful and easy to use free open source python library for developing and evaluating deep learning models which wraps the efficient numerical computation libraries Theano and TensorFlow.
- JSON - It is an open-standard file format or data interchange format that uses human-readable text to transmit data objects consisting of attribute-value pairs and array data types.
- CSV - It is a delimited text file that uses a comma to seperate values and allows data to be saved in a tabular format.
- ULMFIT - It is a transfer learning technique which can be applied to almost any NLP task.
- iNLTK - It aims to provide out of the box support for various NLP tasks that an application developer might need for Indic languages.
- CLTK - It offers natural language processing support for the languages of Ancient, Classical and Medieval Eurasia.
- Flask - It is a lightweight WSGI web application framework based on Python.
- HTML - It is the standard markup language for documents designed to be displayed in a web browser.
- CSS - It is a style sheet language used for describing the presentation of a document written in a markup language like HTML.

Chapter 3

External Interface Requirements

3.1 User Interfaces

- Our Hindi Text Corpus is a CSV file. It can be appropriately viewed with any CSV reader.
- We have provided a python script for our NLP model for sentimental analysis of Hindi Text. The user can hence interact with it via any IDE with python language and run it on the terminal.

3.2 Hardware Interfaces

- Windows(7 or higher)/ Linux/ macOS.
- Hardware should support python(3.6 or higher).

3.3 Software Interfaces

Software Used	Description
Programming Language	We have used python for creating our NLP script as it is one of the most popular language used in this field and it has pre-defined well structured libraries for most of our sub-tasks.
Operating System	As we are using python script, it can be easily used in most of the operating systems(Windows/Linux/macOS)
Corpus	We have used both CSV as well as JSON to store our database as these are most popular options used for storing tabular data

Chapter 4

Other Nonfunctional Requirements

4.1 Safety Requirements

If there is extensive damage to a wide portion of the database due to catastrophic failure, such as a disk crash, the recovery method restores a past copy of the database that was backed up to archival storage (typically tape) and reconstructs a more current state by reapplying or redoing the operations of committed transactions from the backed up log, up to the time of failure.

4.2 Security Requirements

Security systems need database storage just like many other applications. However, the special requirements of the security market mean that vendors must choose their database partner carefully.