IIIT BHUBANESWAR

IIIT Bhubaneswar

Imagine, Innovate, Inspire

(A University established by Government of Odisha)

WEB SEARCH AND MINING

PROJECT REPORT

DESIGN AND IMPLEMENT HINDI SPELL CHECKER

Guided By : Dr. Rakesh Chandra Balabantaray

By Group 8 :

Preeti Dey(B118040)

Pritish Priyatosh Nayak(B118041)

Priya Pradeep Kumar(B118042)

Priyabrat Mahapatro(B118043)

ABSTRACT

Spell checker is one of the most common features that are used across text editors and search engines that are regularly used by people.

Normally a spell correction technique consists of two process :

1. **Error detection** : The first part consists of identifying the errors in the typed text. This part uses a language model which accounts for the words allowed in the language.

2. **Error Correction**: This involves rectifying the wrong spelled words using a set of possibilities from the dictionary.

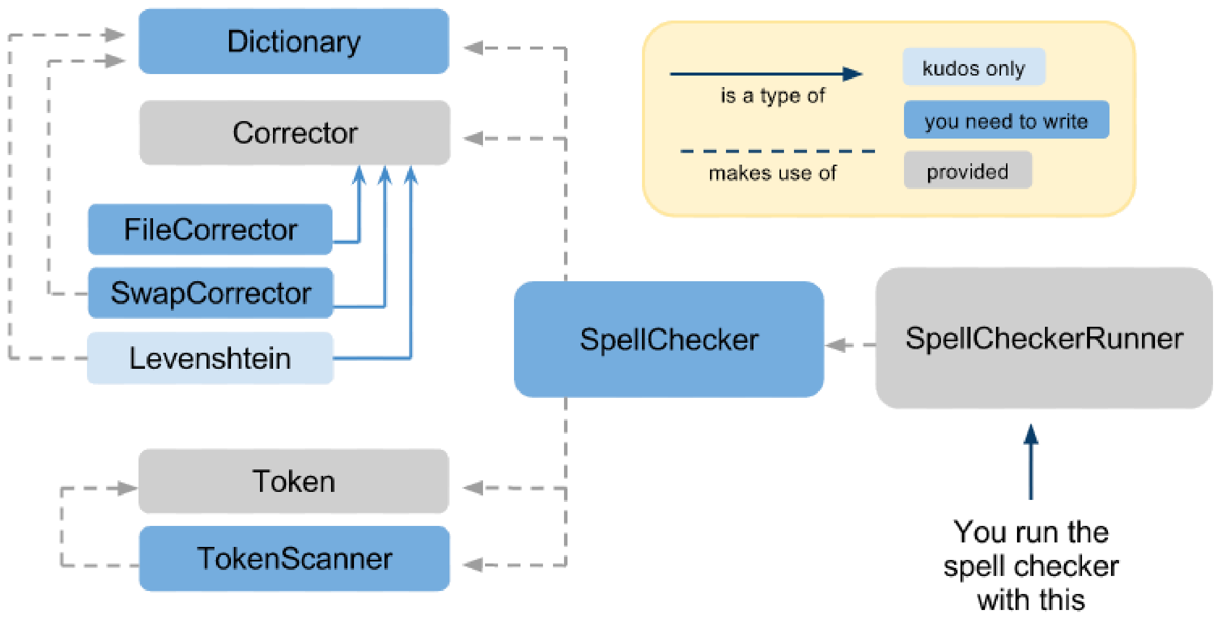


Figure 1: Diagram represting working of a simple Spell Checker.

In this project, we have built a simple spell checker for **Hindi language**.

Method used is : **Edit distance based.**
Algorithm used is : **Levenshtein distance.**

LEVENSHTEIN DISTANCE :

It is one of the most popular algorithms to calculate edit distance. It is named after the Soviet mathematician Vladimir Levenshtein, who considered this distance in 1965.
The main work of this algorithm is –given two words(strings) ,it calculates the minimum single character changes required to convert one of the word to another.

The three basic operations involved in this algorithm are:

1. Insert
2. Delete
3. Replace

It is a Dynamic Programming based algorithm whose complexity is **O(**|s1||s2|**)** ,where |s1| and |s2| denote the lengths of the two input strings.

CODE IMPLEMENTATION DETAILS:

This project uses Python 3.7.0 as the programming language.
External Libraries Used : None

1. Our corpus consists of nearly 20,000 Hindi words which were extracted from news articles. The source to the full corpus can be found : Here.

2. Our python program named spell_checker.py loads this corpus (file: hindi_corpus.txt) into memory and stores it in a global list (array). **[function used : loadCorpus()]**

3. The program then loads the input (file : input.txt) which contains the Hindi sentences which are to be checked.
   [function used : processInput()]

4. The program then checks for each word if it is present directly in the current loaded corpus. If it is present then there's no need to correct this word. Otherwise we start searching for the correct word.

5. For searching the correct word , we simply loop through all the words in the corpus and look for word with the minimum edit distance. In case of words having equal edit distance, we chose the one which is present earlier in the corpus (the corpus was pre-sorted according to the frequency of occurance of each word).
   [function used for searching correct word: getCorrectWord()]
   [function used for getting edit distance: getLevenshteinDistance()]

6. Upon finding the possible match for the correct word, we simply print the location of the mispelled word along with the corrected version of the word.

Note: Further optimization heuristics for speeding up the query time were not implemented as it was somewhat beyond the scope of the project.

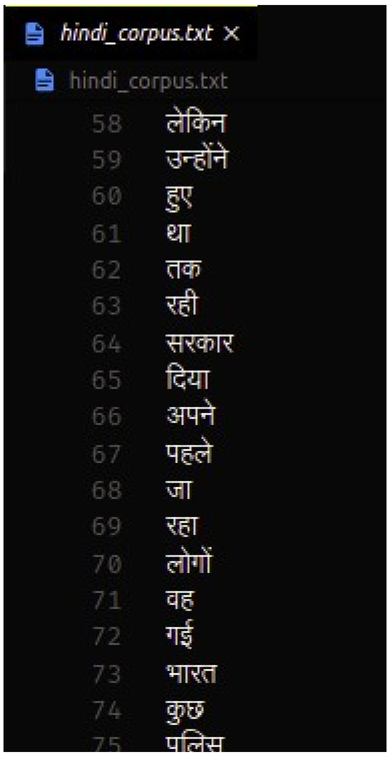
Input Format:

Paste the Hindi string into "input.txt" in the root directory of the project.

Output Format

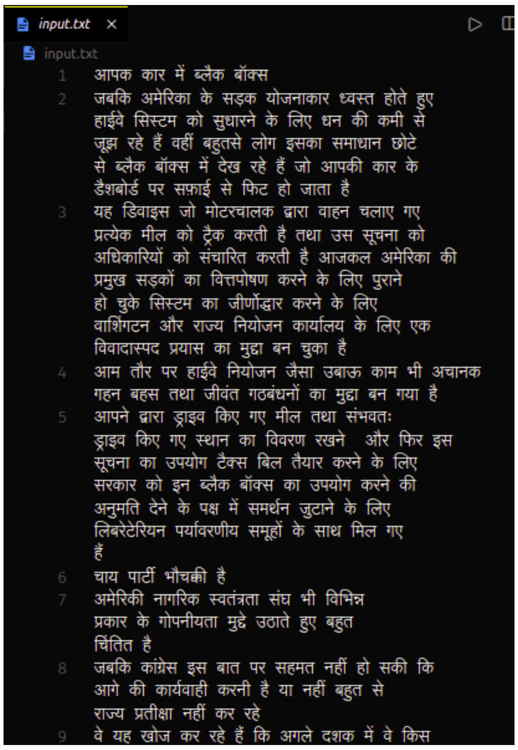
The program will create a "output.txt" file in the root directory of the project where for each wrong word a line will be printed in the format:
"At Line:" <line_number> "Word No. " <word_number> ":" <wrong_word> "->" <correct_word>

Corpus:



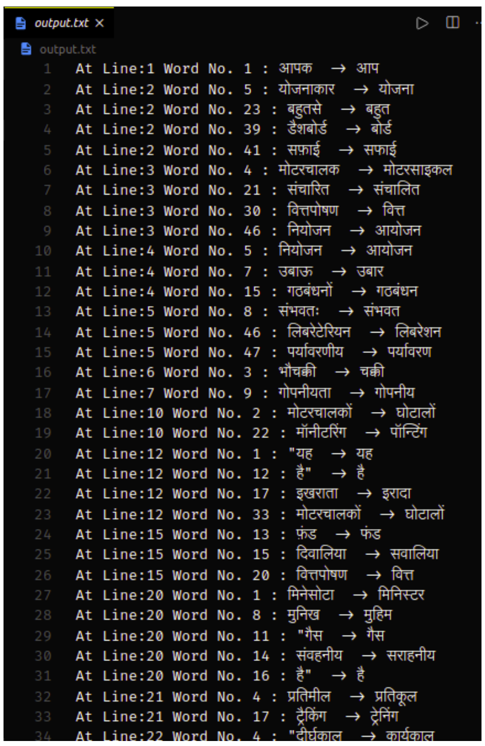
```
लेकिन
उन्होंने
हुए
था
तक
रही
सरकार
दिया
अपने
पहले
जा
रहा
लोगों
वह
गई
भारत
कुछ
पुलिस
```

Input:



```
1   आपक कार में ब्लैक बॉक्स
2   जबकि अमेरिका के सड़क योजनाकार ध्वस्त होते हुए
    हाईवे सिस्टम को सुधारने के लिए धन की कमी से
    जूझ रहे हैं वहीं बहुतसे लोग इसका समाधान छोटे
    से ब्लैक बॉक्स में देख रहे हैं जो आपकी कार के
    डैशबोर्ड पर सफ़ाई से फिट हो जाता है
3   यह डिवाइस जो मोटरचालक द्वारा वाहन चलाए गए
    प्रत्येक मील को ट्रैक करती है तथा उस सूचना को
    अधिकारियों को संचारित करती है आजकल अमेरिका की
    प्रमुख सड़कों का वित्तपोषण करने के लिए पुराने
    हो चुके सिस्टम का जीर्णोद्धार करने के लिए
    वाशिंगटन और राज्य नियोजन कार्यालय के लिए एक
    विवादास्पद प्रयास का मुद्दा बन चुका है
4   आम तौर पर हाईवे नियोजन जैसा उबाऊ काम भी अचानक
    गहन बहस तथा जीवंत गठबंधनों का मुद्दा बन गया है
5   आपने द्वारा ड्राइव किए गए मील तथा संभवतः
    ड्राइव किए गए स्थान का विवरण रखने  और फिर इस
    सूचना का उपयोग टैक्स बिल तैयार करने के लिए
    सरकार को इन ब्लैक बॉक्स का उपयोग करने की
    अनुमति देने के पक्ष में समर्थन जुटाने के लिए
    लिबरेटेरियन पर्यावरणीय समूहों के साथ मिल गए
    हैं
6   चाय पार्टी भौचक्की है
7   अमेरिकी नागरिक स्वतंत्रता संघ भी विभिन्न
    प्रकार के गोपनीयता मुद्दे उठाते हुए बहुत
    चिंतित है
8   जबकि कांग्रेस इस बात पर सहमत नहीं हो सकी कि
    आगे की कार्यवाही करनी है या नहीं बहुत से
    राज्य प्रतीक्षा नहीं कर रहे
9   वे यह खोज कर रहे हैं कि अगले दशक में वे किस
```

Output:



```
At Line:1 Word No. 1 : आपक  → आप
At Line:2 Word No. 5 : योजनाकार  → योजना
At Line:2 Word No. 23 : बहुतसे  → बहुत
At Line:2 Word No. 39 : डैशबोर्ड  → बोर्ड
At Line:2 Word No. 41 : सफ़ाई  → सफाई
At Line:3 Word No. 4 : मोटरचालक  → मोटरसाइकल
At Line:3 Word No. 21 : संचारित  → संचालित
At Line:3 Word No. 30 : वित्तपोषण  → वित्त
At Line:3 Word No. 46 : नियोजन  → आयोजन
At Line:4 Word No. 5 : नियोजन  → आयोजन
At Line:4 Word No. 7 : उबाऊ  → उबार
At Line:4 Word No. 15 : गठबंधनों  → गठबंधन
At Line:5 Word No. 8 : संभवतः  → संभवत
At Line:5 Word No. 46 : लिबरेटेरियन  → लिबरेशन
At Line:5 Word No. 47 : पर्यावरणीय  → पर्यावरण
At Line:6 Word No. 3 : भौचक्की  → चक्की
At Line:7 Word No. 9 : गोपनीयता  → गोपनीय
At Line:10 Word No. 2 : मोटरचालकों  → घोटालों
At Line:10 Word No. 22 : मॉनीटरिंग  → पॉन्टिंग
At Line:12 Word No. 1 : "यह  → यह
At Line:12 Word No. 12 : है"  → है
At Line:12 Word No. 17 : इखराता  → इरादा
At Line:12 Word No. 33 : मोटरचालकों  → घोटालों
At Line:15 Word No. 13 : फ़ंड  → फंड
At Line:15 Word No. 15 : दिवालिया  → सवालिया
At Line:15 Word No. 20 : वित्तपोषण  → वित्त
At Line:20 Word No. 1 : मिनेसोटा  → मिनिस्टर
At Line:20 Word No. 8 : मुनिख  → मुहिम
At Line:20 Word No. 11 : "गैस  → गैस
At Line:20 Word No. 14 : संवहनीय  → सराहनीय
At Line:20 Word No. 16 : है"  → है
At Line:21 Word No. 4 : प्रतिमील  → प्रतिकूल
At Line:21 Word No. 17 : ट्रैकिंग  → ट्रेनिंग
At Line:22 Word No. 4 : "दीर्घकाल  → कार्यकाल
```

Project Source Code Link :
https://github.com/pritish-n/Hindi-Spell-Checker

External Resources used :

- Theory : https://en.wikipedia.org/wiki/Levenshtein_distance
- Corpus Source : https://wortschatz.uni-leipzig.de/en/download/hindi