Auto-Spelling Checker using Natural Language Processing

Chinmay Patil[1], Rexson Rodrigues[2], Reuben Ron[3]

[1]Chinmay Patil Xavier Institute of Engineering, Mumbai, Maharashtra, India
[2]Second Xavier Institute of Engineering, Mumbai, Maharashtra, India
[3]Ruben Ron, Xavier Institute of Engineering, Mumbai, Maharashtra, India

---***---

Abstract - The spell checker is the basic requirement for any documentation in any language. The spell checker is software that suggests the incorrect word and provides its most possible correct word. This paper describes the techniques. Natural language processing (NLP) is a field which deals with the interactions between computers and human.

Key Words: Auto correction, spell checker.

1. INTRODUCTION

The Spell-checking is the process of detecting and suggesting incorrect spelled words in a paragraph. Spell checking system first detects the incorrect words and suggests correct answers. Spell checking system is a combo of standard rules of the languages for which spell checking system is to be created and a dictionary that contains the accurate spellings of various words. Better rules and a large dictionary of words help to improve the rate of error detection otherwise all the errors cannot be detected. After wrong or misspelled words, the various suggestions are given. There are many systems available for detecting and correcting the text. The system is made to check the spellings from the list of words in a text file.

The primary objective of the project is to make the user more reliable by facilitating with approximate facilities to opt for this platform. The project presents a simple and easy way to detect the spelling errors made by a user which makes work easier to perform.

1.1 Flow of the project

1) Input and Output of the project

2) Analysis of words

3) Scanning from different set of words in text file4) Giving the answer whether it is right or wrong.

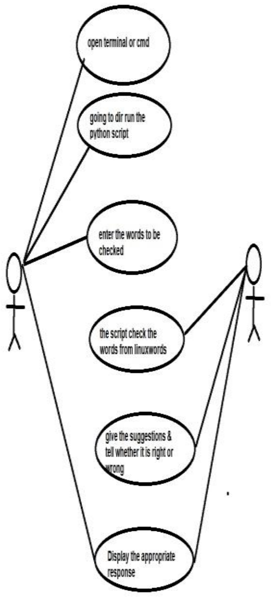
Finally, we complete our project about the suggestion of wrong words.

1.2 Scope of the project

The entire project will help us to identify incorrect spelling of the words and help us to develop a small environment where user can check their words without a specialised software. Addition to this it will be more secure as your data remains in tacked in your own device only.

2. Workflow

Taking input from the user.


open terminal or cmd
going to dir run the python script
enter the words to be checked
the script check the words from linuxwords
give the suggestions & tell whether it is right or wrong
Display the appropriate response

The spell checker compares every word typed with a list of thousands of correctly spelled, words and then uses algorithms to determine the correct spellings. If a word (e.g., a name), is spelled correctly, you can add it to the program's exceptions list so it will not be flagged as misspelled in the future.

You can run this on any platform, it has wide support of the program regardless we are working on windows, mac or Linux the only requirement is python script and text file which contains big kind of words.

3. Existing System

1. R-ruled based technique system: This rule-based system is used to check the spelling this system has a collection of rules which captures the most common spelling and typographical errors and these are applied to the misspelled words. These rules are the outcome of common errors therefore each word taken here is choosen as a solution.

2. Dictionary Lookup Technique: - Dictionary-lookup is a method, which has a huge significance in the Spellchecked technique. By checking strings of any language word in the dictionary or the database, it tells if the given word is correct or not. If it is found in the database the given word is true. If the string of word does not appear in the database, it is the incorrect word. Due to this it gives very high accuracy and is considered to be very dependable.

3. N Gram technique: An n-gram is considered to be a collection of ensuing characters of length N. N-gram analysis is a process to detect whether the entered words in the document are wrong spelled or not. Here we do not compare each word with the dictionary instead we use the N-gram method. If the place is said to be empty or deficient n-gram is found, It is assumed as an incorrect, or it is assumed to be correct. If N is found to be 1 then it is a unigram, if N is 2 then it is a Bigram, if N is 3 then it is a trigram etc. The n-grams algorithm is also referred as or a "neutral string-matching algorithm".

4. General Working of error detection

Spelling errors can be generally categorized into two types, Real word errors and Non word errors. Real word errors are those error words that are acceptable words within the dictionary. The non-word errors on the other hand are the errors that cannot be found in the dictionary. Further the errors can be classified as

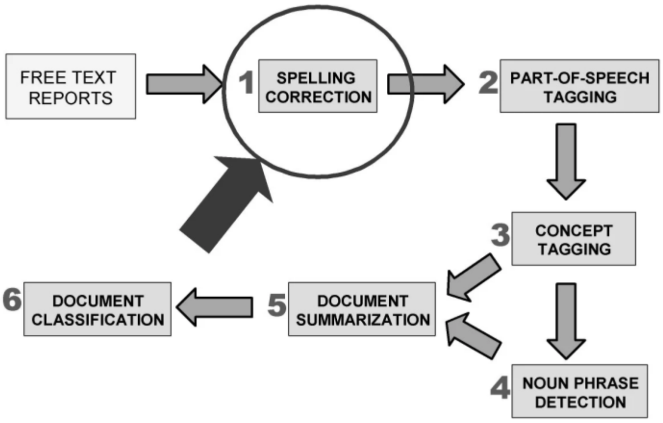
1. **Cognitive Errors**: The previous two types of errors result not from ignorance of a word or its correct spelling. Cognitive errors can occur due to those factors. The words *piece* and *peace* are homophones (sound the same). So, you are not sure which one is which. Sometimes *your* damn sure about your spellings despite a couple of grammar nazis claim *you're* not.

2. **Short forms/Slang/Lingo**: These are possibly not even spelling errors. May be u r just being kewl. Or you **try** hard **to suit** in everything within a text message or a tweet and must commit a spelling sin. We mention them here for the sake of completeness.

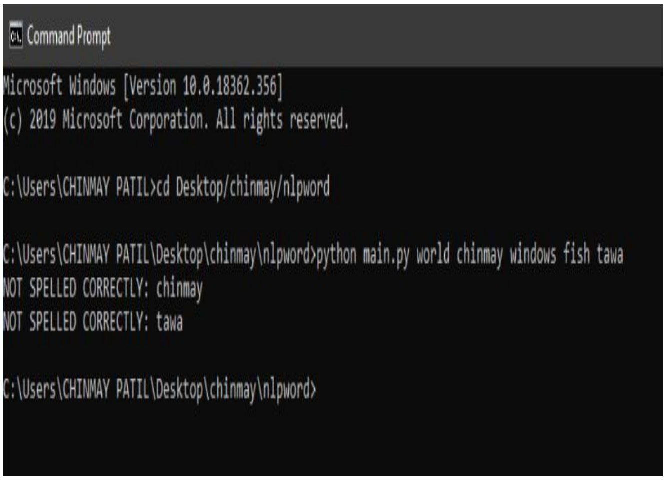
3. **Intentional Typos**: Well, because you are clever. You type in teh and pwned and zomg carefully and frown if they get autocorrected. It could be a marketing trick, one that probably even backfired.

5. Diagrams

5.1 UML diagram



5.2 System Execution



```
Microsoft Windows [Version 10.0.18362.356]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\CHINMAY PATIL>cd Desktop/chinmay/nlpword

C:\Users\CHINMAY PATIL\Desktop\chinmay\nlpword>python main.py world chinmay windows fish tawa
NOT SPELLED CORRECTLY: chinmay
NOT SPELLED CORRECTLY: tawa

C:\Users\CHINMAY PATIL\Desktop\chinmay\nlpword>
```

6. CONCLUSIONS

While autocorrect, tools do have a mind of their own, spellchecking and autocorrection may be a well addressed problem for English and other European languages. However, spell correction features a great distance to travel for other languages, especially Indian languages.

One of the major challenges in building error models for languages other than English include lack of datasets like the Birbeck corpus. There are also syntax related challenges. Indian languages are phonetic, and it's not clear what sorts of spelling error patterns exist. This is an area that needs to be studied a lot more.

Logs that are collected from applications that are multilingual, like input tool, multilingual search, and localization APIs might give us some insight into error patterns. Also, user generated content from social media and forums is another useful source for building the error model.

REFERENCES

[1] D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[2] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[3] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] K. Elissa, "Title of paper if known," unpublished.