

Spell Corrector

This project develops a Python spell corrector inspired by Peter Norvig's approach. It identifies misspelled words and suggests corrections using NLP techniques like text normalization and probabilistic modeling.



Corpus and Probability Model

Corpus Usage

Uses a selected corpus for word frequency statistics.

Word Probabilities

Calculates unigram probabilities from corpus frequencies.

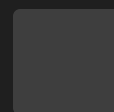
Tokenization

Extracts lowercase words using regex for analysis.

```
    null  
    => null  
    => "admin"  
    => null  
    => "info@mecanbay.com"  
    erified at" => null  
    d" => "$2y$10$1rmusskiz0M  
    e" => 1  
    le" => "Administrator"  
    => "assets/img/users/def  
    r_token" => "0dwr7SXo3pw  
    at" => "2022-01-01 22:56  
    at" => "2022-01-02 15:01
```

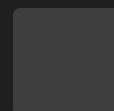


Generating Candidate Corrections



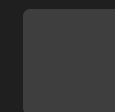
Edits1

Generates words one edit away: deletions, transpositions, replacements, insertions.



Edits2

Generates candidates two edits away by applying edits1 twice.



Custom Edit Costs

Costs assigned:
deletions 1.0
transpositions 0.8
replacements 0.9
insertions 1.2.

Candidate Selection and Scoring

Selection Steps

- Generate edits1 candidates.
- Generate edits2 candidates.
- Remove non-existent words
- Calculate score of the word
- Word with higher score = selection

Scoring Formula

Score = $\text{probability}(\text{word}) / \text{edit cost}$, prioritizing frequent and likely edits.

Theoretical Foundations

1

Probabilistic Model

Maximizes $P(c|w)$ using Bayes' theorem.

2

Language Model

Uses unigram frequency as prior probability $P(c)$.

3

Edit Distance

Focuses on Damerau-Levenshtein distance with four edit types.

4

Error Model

Incorporates edit costs to model $P(w|c)$.

Results and Demonstration

Corpus Stats

Over 1 million words, 29,157 unique, frequent words confirmed.

Correction Example

"korrektud" corrected to "corrected" with scored candidates.

Sentence Correction

Fixes multiple misspellings in sentences accurately.