CLIQUE

A human in China is known as a tree.
GOD give us earth to live, learn, and research.
We, human, commit errors during our stay on earth until we are back Home.

CLIQUE is a spell checker program that allows committing multiple spelling errors of various error types and error categories – like us, human : our errors are various, diverse until we find the correct path.

By: Leena Al-Hussaini
CLIQUE webpage: https://multi-error-misspellings.github.io

Motivation

- The idea of CLIQUE is to be able to commit spelling errors – multiple errors in a word.
- The idea is derived from the lack of spell checker on NASA Earth Data website.
- It is a big problem if a large dataset is to be searched by many users of different language and backgrounds, including novice users, with no spell checker.
- Spell checker helps users find the required dataset.
- Long words: long field words are prone to spelling errors, like observation, meteorological, terrestrial, environmental, visualization, etc.
- CLIQUE helps many users:
  - Any user: NASA Earth Data is open to be searched by any user – not only specialized personnel of NASA – which means inexperience users of the field terms would probably commit some sorts of misspellings.
  - Nonnative English Speakers: mind that many users are nonnative speakers of the English language so the rate of falling into spelling errors is high.
  - Experts: even experts in the field of space and earth do commit misspellings due to fast keyboard typing.

CLIQUE: The Project

- CLIQUE allows more than one error in a word, upto 5 errors. Example: "enviruunetla" for correct dictionary word "environmental".

```
Your misspell is: enviruunetla

Suggestions:
environmental
environal
environment
environ
environage
environmentally
environs
environmentalism
environmentalist
envier
```

Example of CLIQUE program

Correct/Dictionary word: e n v i r o n m e n t a l

Misspell: e n v i r u u n _ e _ t l a

substitute error insert error delete error delete error transpose error

CLIQUE: The Project

- CLIQUE uses 4 known spell checking algorithms to spell check: 1) Jaro Winkler, 2) Longest Common Subsequence, 3) Hamming, and 4) 2-Gram.
- The novelty of CLIQUE is in the use of multiple well known spell checking algorithms, algorithms of different knowledges to solve the spell checking problem.
- CLIQUE does data combination of multiple algorithms. The selected 4 algorithms are chosen after thorough experimental study which compares 12 spell checking algorithms and 2 spell checking softwares using 320 difficult misspellings of various error operations (insertion, deletion, substitution, transposition) and error categories (phonetic – *fonetik* for phonetic, misconceived – *anonomy* for anatomy , keyboard – *leatn* for learn).
- Some results of the comparison:

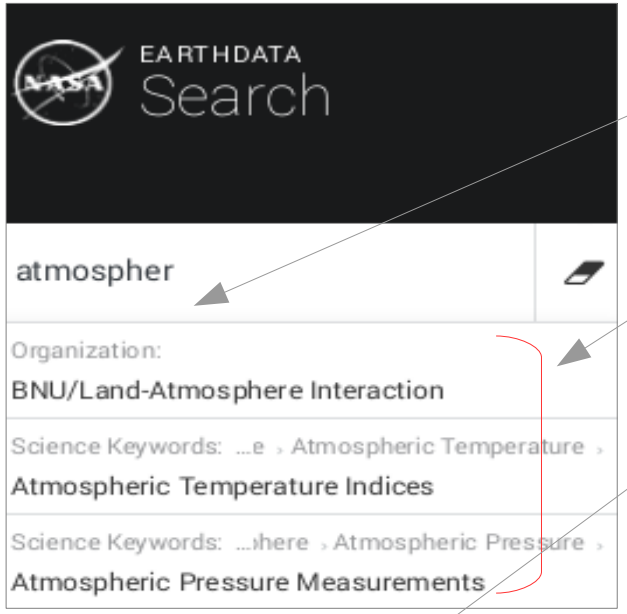
Algorithms	Phonetic Error (167)	Misconceived Error (53)	Keyboard Error (100)	Total Score (320)	Notes
CLIQUE	**90%**	**77%**	**95%**	**87%**	**Algorithm: the submitted project**
Hunspell	77%	66%	37%	60%	Software used in LibreOffice, OpenOffice, Google Chrome, etc.
Aspell	83%	50%	45%	59%	Software used in GNU OS, Gajim, Notepad, etc.
2-Gram	76%	33%	98%	69%	Algorithm.

CLIQUE Properties

- CLIQUE solves: 1) phonetic error which is usually committed by large number of users, especially dyslexic users, 2) Misconceived error which is usually committed by nonnative Speakers of English, and 3) Keyboard error which is either due to fast typing or fat-finger syndrome.
- You can add to the dictionary your own words – specialized words of Space and Earth data.
- Do NOT proceed “a space” in your entered word – to check for spelling error. Example: "<space>earthh", " earthh".
- **CLIQUE limitations**:
  - First character error is not handled by CLIQUE, like “cky” - misspelling of “sky".
- **CLIQUE weakness**:
  - If a misspelling coincides with the beginning of another word and the misspelling lacks correct aspects of the correct word, then clique fails to find the correct word. For example, the misspelling “atomsfer” for “atmosphere” contains "atom" in the beginning, so clique would guess that it is related to "atom" words. But a slight change in the misspelling by adding more correct aspects of the correct word like "atomosphere" would help Clique to find the correct word "atmosphere".

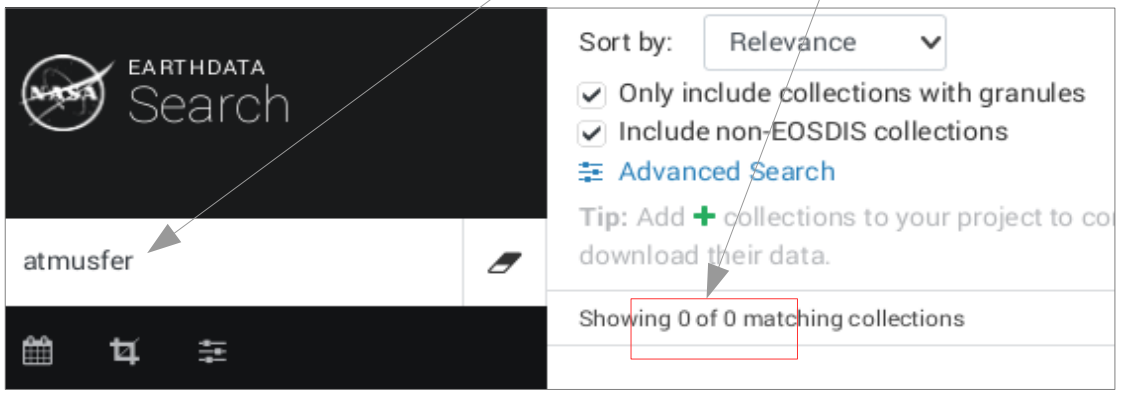
CLIQUE and NASA Earth Data Search Engine

- NASA Earth Data Website only search entered words against their pre-defined list of keywords. If a misspelling occurred in the word, it can not suggest the correct one related to their keywords.

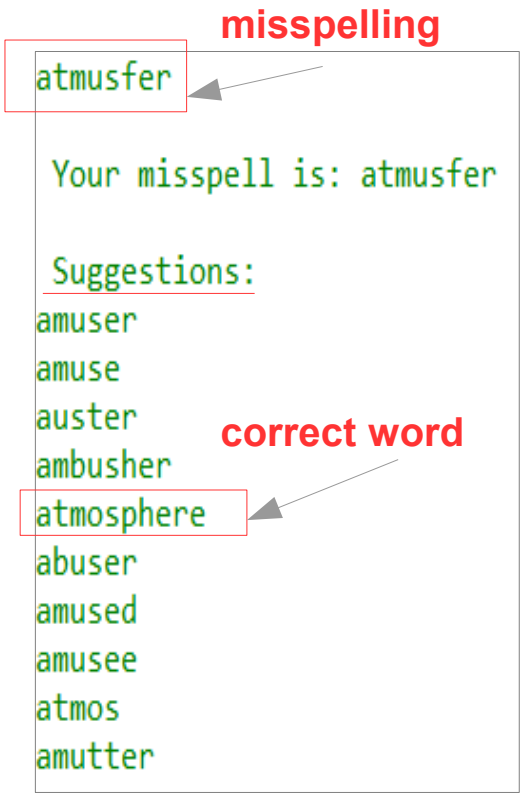


Correct word yields a list of datasets.

If a misspell of one or more errors is in the entered word, the search engine does not suggest alternatives close to the contained keywords.



CLIQUE can suggest alternatives to the misspell even with more than one error.



Conclusion

- CLIQUE is a multi-error spell checker.
- CLIQUE solves three different error categories: phonetic error, misconceived error, and keyboard error.
- CLIQUE is a combination of four spell checking algorithms with different knowledges collaborating together to solve the problem of misspelling.
- CLIQUE selected the four algorithms - Jaro Winkler, Hamming, Longest Common Subsequence, 2-Gram - after thorough study through experimental analysis which involved comparing 12 spell checking algorithms and 2 spell checking softwares.
- CLIQUE results, 89%, proves the effectiveness of combining different algorithms with different backgrounds to tackle the problem of spell checking, in general.
- CLIQUE will help many users search the NASA Earth Datasets even if they lack the knowledge of correct spelling, do fast typing, or are dyslexic users.

- Note: CLIQUE is part of Leena Al-Hussaini doctorate study. CLIQUE research paper is a paper-to-submit. If CLIQUE is useful to NASA organization, then CLIQUE paper will be shared after publication.