

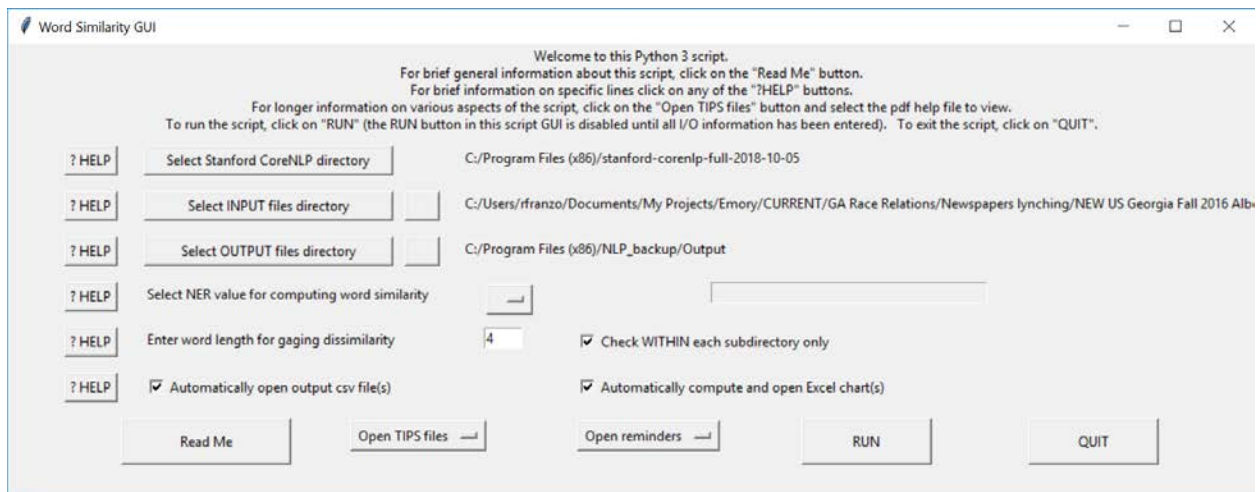
Word Similarity (Levenshtein Distance)

Table of contents

Proper noun spelling checker.....	1
Stanford CoreNLP NER (Named Entity Recognition) annotator.....	1
Levenshtein's word/edit distance.....	1
INPUT.....	2
OUTPUT.....	3
References.....	3

Proper noun spelling checker

The algorithm uses two different set of tools (Stanford CoreNLP NER, Named Entity Recognition, annotator and Levenshtein's word/edit distance) to check for possible misspellings in proper names, i.e., names that would be tagged as NNP or NNPS in the POSTAG values of the CoNLL table.



Stanford CoreNLP NER (Named Entity Recognition) annotator

The algorithm takes as input the NER values of POSTAG (Part of Speech tag) proper names (NNP and NNPS, proper noun singular and plural), i.e., City, Country, State-Or-Province, Location, Organization, Person, as computed by Stanford CoreNLP.

Levenshtein's word/edit distance

The Levenshtein's word or edit distance is defined as the smallest number of insertions, deletions, and substitutions required to change one string or tree into another. (2) A $\Theta(m \times n)$ algorithm to compute the distance between strings, where m and n are the lengths of the strings (Levenshtein 1966).

The NLP Suite algorithm uses the NLTK implementation of the edit distance (edit_distance library) among the many available (for a long list, in different programming languages, see the webpage at the US National Institute of Standards, NIS, <https://xlinux.nist.gov/dads/HTML/Levenshtein.html>).

INPUT

The algorithm can either process all the txt files in a directory (Fig. 1), including its subdirectories, computing for each file the user-selected NER values and comparing values with the NER values of every other file.





















Name	Date modified	Type	Size
 Charlotte Daily Observer_09-23-1906_1_1.txt	2/8/2020 2:42 PM	TXT File	6 KB
 Chicago Daily Tribune_09-23-1906_1_5.txt	2/8/2020 2:42 PM	TXT File	7 KB
 Chicago Daily Tribune_09-24-1906_8_4.txt	2/8/2020 2:42 PM	TXT File	3 KB
 Daily Press_09-23-1906_1_1.txt	2/8/2020 2:42 PM	TXT File	5 KB
 Dallas Morning News_09-23-1906_2_1.txt	2/8/2020 2:42 PM	TXT File	1 KB
 Dallas Morning News_09-23-1906_2_1_2.txt	2/8/2020 2:42 PM	TXT File	1 KB
 Dallas Morning News_09-23-1906_2_1_3.txt	2/8/2020 2:42 PM	TXT File	1 KB
 Dallas Morning News_09-23-1906_2_1_4.txt	2/8/2020 2:42 PM	TXT File	1 KB
 Dallas Morning News_09-23-1906_2_1_5.txt	2/8/2020 2:42 PM	TXT File	1 KB
 Dallas Morning News_09-23-1906_2_1_6.txt	2/8/2020 2:42 PM	TXT File	1 KB
 Los Angeles Time_09-23-1906_2_4.txt	2/8/2020 2:42 PM	TXT File	4 KB
 Los Angeles Times_09-24-1906_1_2.txt	2/8/2020 2:42 PM	TXT File	8 KB
 Los Angeles Times_09-24-1906_6_1.txt	2/8/2020 2:42 PM	TXT File	1 KB
 New York Times_09-23-1906_1_1.txt	2/8/2020 2:42 PM	TXT File	9 KB
 New York Times_9-24-1906_2_5.txt	2/8/2020 2:42 PM	TXT File	11 KB

Figure 1 – Document list in a directory

The algorithm can also limit the computation and computing of user-selected NER values to the files *within* subdirectories. Thus, if a user has organized the documents (e.g., newspaper articles) that describe specific events in subdirectories, each subdirectory bearing the ID of the event in a database, a specific name, or a combination of ID and/or name, the user can restrict the NER computations and comparisons to *within* each subdirectory (see Fig. 2).

Name	Date modified	Type	Size
 3	2/18/2020 10:14 PM	File folder	
 4	2/18/2020 10:15 PM	File folder	
 5	2/18/2020 10:15 PM	File folder	
 6	2/18/2020 10:15 PM	File folder	
 7	2/18/2020 10:15 PM	File folder	

Name	Date modified	Type	Size
3_Jim Cobb	2/18/2020 10:14 PM	File folder	
4_Frank Hardeman	2/18/2020 10:15 PM	File folder	
5_Palseo	2/18/2020 10:15 PM	File folder	
6_Owen Jones	2/18/2020 10:15 PM	File folder	
7_Jet Hicks	2/18/2020 10:15 PM	File folder	

Name	Date modified	Type	Size
3_Jim Cobb_1918	2/18/2020 10:14 PM	File folder	
4_Frank Hardeman_1900	2/18/2020 10:15 PM	File folder	
5_Palseo_1890	2/18/2020 10:15 PM	File folder	
6_Owen Jones_1890	2/18/2020 10:15 PM	File folder	
7_Jet Hicks_1906	2/18/2020 10:15 PM	File folder	

Figure 2 – Event subdirectories structure

OUTPUT

References

Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Doklady Akademii Nauk SSSR*, 163(4):845-848, 1965 (Russian). English translation in *Soviet Physics Doklady*, 10(8):707-710, 1966. (Doklady is Russian for "Report". Sometimes transliterated in English as Doclady or Dokladi.)