

EXPLORING NLP FUZZY MATCHING ALGORITHMS

Madhurima Nath
Data Scientist, Slalom

WHAT IS FUZZY MATCHING?

FUZZY/APPROXIMATE STRING MATCHING

- Method to find strings which match a pattern approximately.
- Identifies the likelihood/probability that two records are true match based on some parameters.

ID	Name
01	Microsoft
02	Microsoft Co.
03	Microsoft Corporation
04	Mcrosoft Corp
05	Microosoftt



Microsoft Corporation

APPLICATIONS OF FUZZY MATCHING

- Spell checker
- Deduplication of records
- Master data management
- Plagiarism detection
- Bioinformatics and DNA sequencing
- Spam filtering
- Content search

HOW DOES FUZZY MATCHING WORK?

FUZZY MATCHING ALGORITHMS

- Edit distance metric - quantifies how dissimilar two strings are by counting the minimum number of operations required to transform one string into the other
 - Levenshtein distance
 - Damerau–Levenshtein distance
- Bitap algorithm - tells whether a given text contains a substring which is "approximately equal" (defined in terms of Levenshtein distance) to a given pattern
- n-gram - predicts next item in a sequence of text (in form of a Markov model)

EDIT DISTANCE METRICS

measures the number of edits needed to transform one word into another

Levenshtein Distance

- most common metric
- techniques for editing:
 - insertion
 - deletion
 - substitution/replacement
- e.g.: $LD(\text{irks} \rightarrow \text{risk}) = 3$

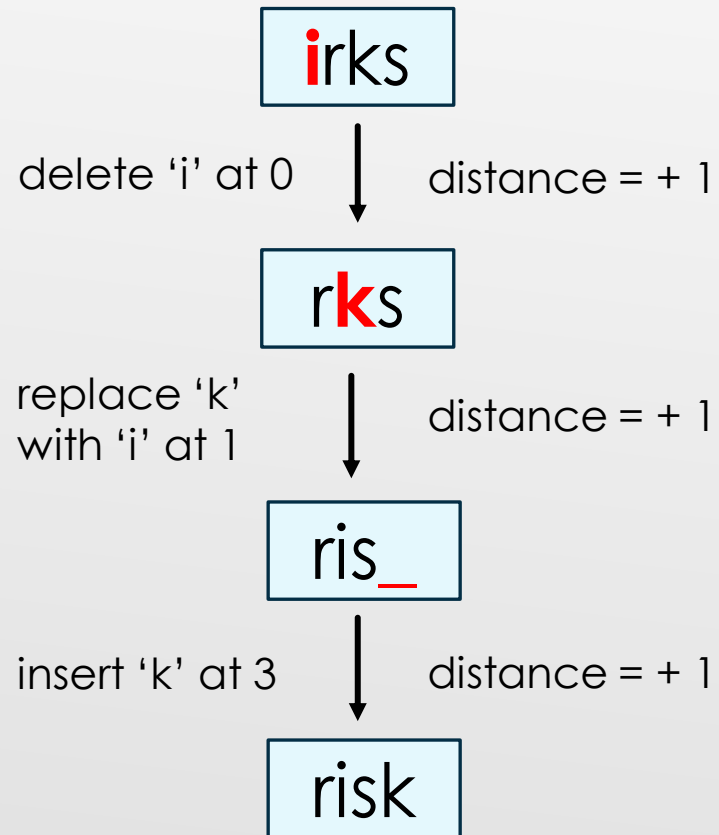
Damerau-Levenshtein Distance

- similar to Levenshtein distance
- techniques for editing:
 - insertion
 - deletion
 - substitution/replacement
 - transposition
- e.g.: $DLD(\text{irks} \rightarrow \text{risk}) = 2$

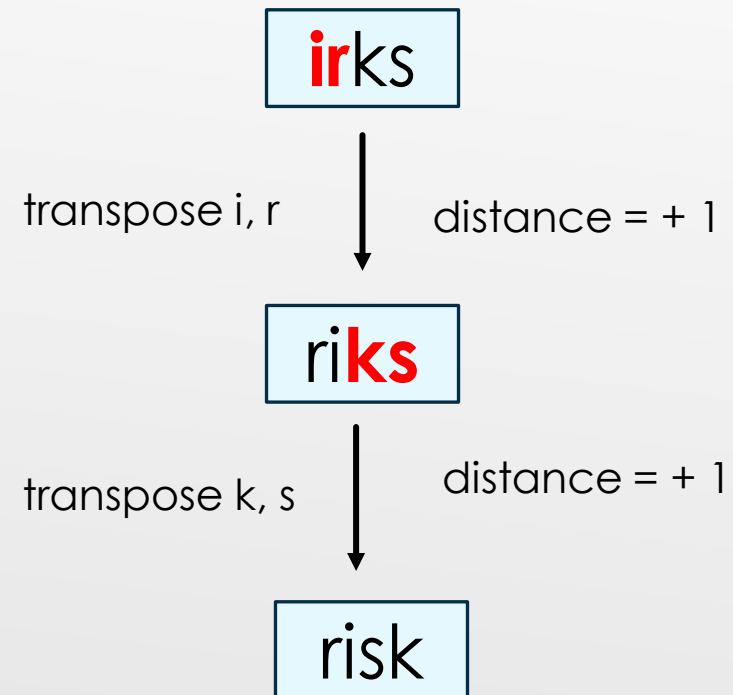
EDIT DISTANCE METRICS

measures the number of edits needed to transform one word into another

Levenshtein Distance



Demerau-Levenshtein Distance



OTHER FUZZY MATCHING ALGORITHMS

- Edit distance
 - Longest common subsequence
 - Hamming distance
 - Jaro distance
- Needleman–Wunsch algorithm
- Smith–Waterman algorithm
- BK Tree metric
- Soundex or Metaphone – phonetic algorithms