A fuzzy string-matching algorithm capable of interpreting abbreviations, errata & whitespace

Jargon Errata Standardizing Text Ranking

Cr to our bsnss is the intprtn of sneta nces lik thes

Author: Sudhir Sekharan

SINGULARIUM TECHNOLOGIES

#306, First Floor 2nd Main, 7th Cross, Domlur Layout,
Bangalore KA 560071

What is JESTR?

The **JESTR** or **J**argon **E**rrata **S**tandardizing **T**ext **R**anking algorithm compares two different strings for similarity. This proprietary fuzzy string-matching algorithm compares complex string structures including errors in spelling, abbreviations and structural errors caused by whitespaces. It is a general string similarity algorithm that does not require training.

JESTR's fuzzy string matching **determines** that the two strings score high on similarity:

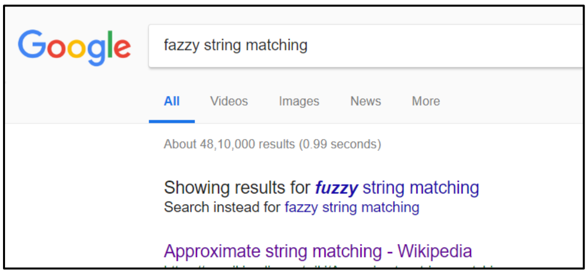
a. fyi, pickup for air port est. at 5:30
b. pick up for airport estimated at 5:30, for your information

The ubiquitous Fuzzy String-matching & it's uses!

In computer programming, a string is a sequence of characters such as "qw3rty" or "A sample string!".

Approximate / fuzzy string matching is the technique of finding strings that match an input string approximately. Fuzzy matching employs mathematical constructs such as the Levenshtein edit distance [(1)] or the Hamming distance to compare strings and generate matches when no exact matches are found.

Standard string dictionaries are used to enable fuzzy matching in search engines, chatbots, messaging services, analytics engines and all manners of text-based applications. These enhance user experiences by correcting / interpreting user input strings to return meaningful information.



fazzy string matching

About 48,10,000 results (0.99 seconds)

Showing results for ***fuzzy*** string matching
Search instead for fazzy string matching

Approximate string matching - Wikipedia

Figure 1 Google uses similarity to improve search results

What makes JESTR different?

Complex strings, say a message between people on chat, are often large strings built using individual words (separated by whitespace). They are also reduced to easy-to-type formats to reduce verbiage. JESTR introduces a fundamental change in similarity evaluation for complex strings through the following -

1. Compound transformations involving whitespace separated strings & characters v/s simple character transformation

Most fuzzy techniques operate on character level transformations, JESTR uses a combination of characters & whitespace separated strings as the base element for similarity evaluation, *i.e.*

t|h|i|s| |i|s| |a| |s|a|m|p|l|e| |s|t|r|i|n|g

current fuzzy techniques: 23 base elements

v/s

this|is|a|sample|string

JESTR: 5 base elements

2. Evaluation of abbreviations

In complex strings, however, a few characters in one string could map to many characters in the second string. JESTR evaluates potential abbreviations and their quality objectively. Simpler fuzzy techniques are incapable of the same, *for e.g.*

lol (3 chars) = laugh out loud (14 chars)

est (4 chars) = established (11 chars)

est (4 chars) = estimated (9 chars)

estmtd (6 chars) = estimated (9 chars)

3. Structural whitespace manipulation

When whitespace separated strings are used for transformations, errors in the positions of whitespaces fundamentally affect the transformations. JESTR internally optimizes whitespaces to interpret & maximize the scores between two strings.

A couple of case studies

Case I

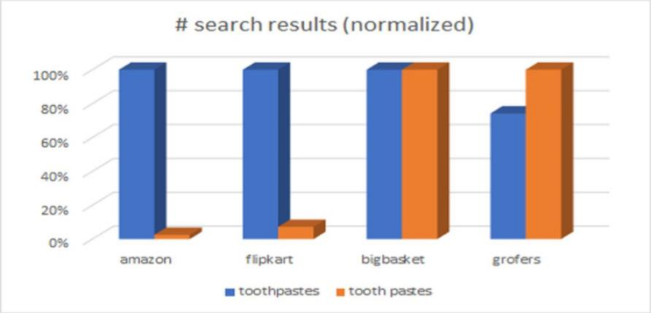
E-commerce search optimization

Simple phrases used for items of daily consumption were queried on search engines deployed on popular online marketplaces in India – Amazon [2], Flipkart [3], Big Basket [4] & Grofers [5]. The impact of commonly used non-standard inputs are evaluated by

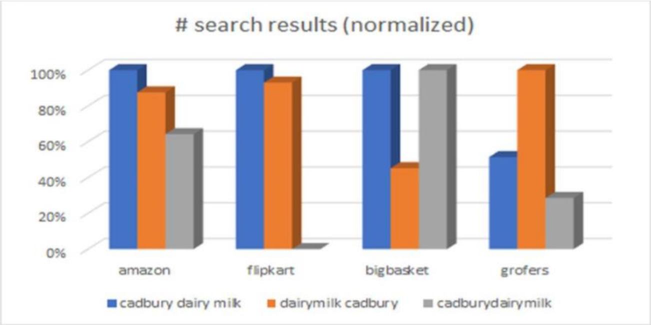
(i) the number of results returned and

(ii) the number of relevant results returned in the top 10 results, in the default search mode deployed by the engine.

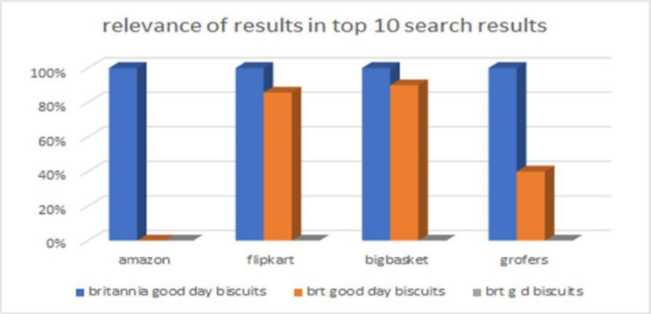
1. Variations of "toothpastes"



2. Variations of "cadbury dairy milk"



3. Variations of "britannia good day biscuits"



Changes in whitespaces or use of jargon in the search strings significantly impacted the number & quality of results across engines.

Case II

Product descriptor standardization

A study was taken up to evaluate the market share of popular snacking brands "Lays" & "Kurkure", marketed by PepsiCo. JESTR was utilized to map item names stored by retailers on their PoS systems to forty-two standardized names as per PepsiCo's catalogue.

Coupled with basic pre-processing, JESTR was used to pick out the most similar strings in two different similarity normalization modes ("strict" and "default"). JESTR demonstrated its capability in interpreting these complex strings and determining the appropriate items sharply.

A few interesting samples from this exercise are highlighted below; with the correctly interpreted answers are highlighted in green, and incorrect in red.

1. lyshtndswtchl30grs10

def: lays west indies hot & sweet chilli, 0.807

str: lays west indies hot & sweet chilli, 0.461

2. lays amrcanso 30g

def: lays american sour cream & onion potato chips, 0.890

str: lays american sour cream & onion potato chips, 0.667

3. kk pu yc 60g

def: kurkure zigzag yummy cheese, 0.848

str: kurkure puffcorn yummy cheese, 0.833

4. laysspnsh tt

def: lays spanish tomato tango potato chips, 0.972

str: lays spanish tomato tango potato chips, 0.648

5. lysclaics ltd 52g

def: lays classic salted potato chips, 0.764

str: lays classic salted potato chips, 0.459

A technical overview of the algorithm

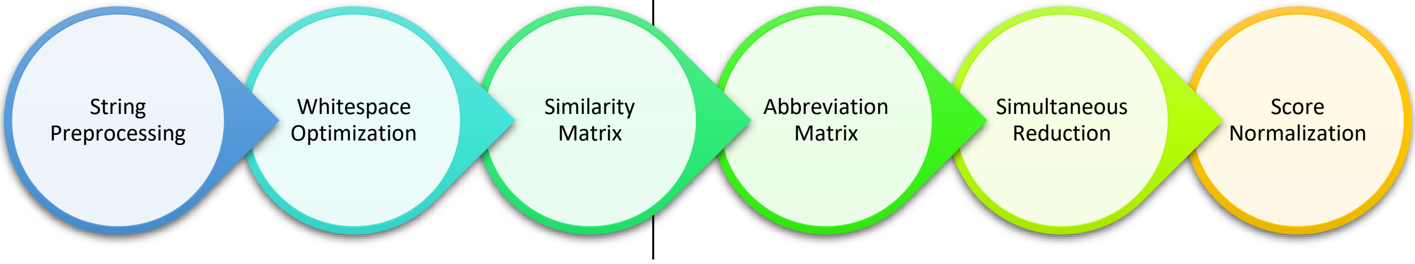


Figure 2 Process flow within the JESTR algorithm

Levenshtein and working with complex strings

Simple strings are single words - "these", "are", "a", "few", "examples" etc. Several methods evaluate the similarity between such strings using metrics such as common characters, position of such characters & total number of characters.

Popular among these is the Levenshtein distance - which is the number of deletions, insertions or substitutions required to transform the input string into a standard string. Similarity could be therefore be normalized as

$$1 - \frac{edits\ required}{total\ characters}$$

For e.g.

String 1	String 2	Similarity
whale	while	9 / 10
apple	ball	4 / 9
lol	laugh out loud	6 / 17

JESTR implements a combination of Levenshtein for evaluation of similarities between the simple strings that make up a complex string.

Additionally, a proprietary whitespace manipulation & abbreviation scoring algorithm is employed to build relationships between simple strings beyond the Levenshtein similarity method.

Process steps within JESTR

Given a non-standard input string comprising "m" whitespace separated simple strings and a standard string comprising "n" whitespace separated simple strings as the comparator, the following steps are undertaken -

1. Pre-processing the strings by removing unnecessary special characters as well as reducing all characters to their base form, i.e., "Crème" to "creme" etc.
2. A recursive tree manipulates whitespace characters to maximize the possibility of a higher final score against the standard string, i.e.
    a. "goodday" to "good day" or
    b. "tooth pastes" to "toothpastes"
3. Generates an ***m*** x ***n*** similarity score matrix using Levenshtein distance
4. Generates an ***m*** x ***n*** abbreviation score matrix that employs a proprietary scoring mechanism basis quality of abbreviation
5. Simultaneous matrix reductions extract a similarity array for the two strings.
6. The normalized score is generated in two modes:
    a. **default**

$$\frac{sum(match\ score)}{len(match\ score) + len(unmatched\ input\ string)}$$

    b. **strict**

$$\frac{sum(match\ score)}{len(match\ score) + len(unmatched\ total\ string)}$$

Implications & way forward

JESTR's greatest strength is its capability to work on data sets being evaluated for the first time. Unlike most AI/ML systems, JESTR is not a trained algorithm and yet demonstrates problem solving capability at the highest levels of accuracy. Internal benchmarking demonstrated that JESTR was in fact in certain cases able to resolve jargon & standardization puzzles better than trained industry experts.

At Singularium, JESTR was able to crash time taken to setup and train supervised ML systems to 20% of the original setup time.

On the way forward, JESTR is still a work in progress, with immense scope for optimization both in terms of the efficiency of algorithm as well as its failure in some edge cases.

It is envisaged that the techniques employed here can be extended to phonetic & semantic similarities as well to build an even more general-purpose algorithm.

Potential Applications

A few snippets of contemporary problems being tackled – a few of these are already under commercial engagement with organizations pioneering cutting edge tech (marked in *italics*)

1. Enhancing Search Engine keyword mapping
2. Address resolution-Flipkart
3. Hate speech-detecting AIs are fools for 'love'
4. *Joining independent data sources by automating the primary keys mapping*
5. *Automating competitive backend price mapping services in e-commerce businesses*
6. *Accelerating setup & training of ML/AI /Deep-learning technology systems by standardizing training sets.*
7. Interpretation of colloquial communication in conjunction with other semantic / NLP techniques by using language dictionaries instead of standardized strings.

Try out the JESTR framework for yourself in "strict" scoring method:

www.singularium.in/jestr

References

(1) Levenshtein Distance
(2) Amazon
(3) Flipkart
(4) Bigbasket
(5) Grofers
(6) Lays & Kurkure masters

Standard names for "Lays" &"Kurkure"

kurkure chilli chatka
kurkure hyderabadi hungama
kurkure monster paws funky tomateo
kurkure naughty tomato
kurkure puffcorn yummy cheese
kurkure solid masti twisteez teekha meetha khatta
kurkure corn cups yummy cheese
kurkure zigzag yummy cheese
kurkure monster paws fun cheese
kurkure crunchy rings
kurkure green chutney rajasthani style
kurkure masala munch
kurkure monster paws mad masala
kurkure puffcorn mad masala
kurkure solid masti twisteez masala
kurkure solid masti twisteez mango pickle flavour
kurkure monster smileez tom cheese
kurkure zigzag mad masala
kurkure papad o nutz chilli masala
kurkure trangles masala
kurkure trangles lime pickle
kurkure trangles mango achari
kurkure trangles buttery
kurkure multigrain curry& herbs
lays american style cream n onion potato chips
lays west indies hot n sweet chilli
lays spanish tomato tango potato chips
lays thai sweet chilli flavour potato chips
lays maxx sizzling barbeque potato chips
lays baked cream herb & onion potato chips
lays crispz herb& onion potato chips
lays twistz herb &onion potato chips
lays classic salted potato chips
lays indias magic masala potato chips
lays chile limon potato chips
lays swiss grilled cheese potato chips
lays maxx macho chilli potato chips
lays maxx hot&sour punch potato chips
lays baked original salted potato chips
lays baked sunkissed tomato potato chips
lays crispz saucy tomatina potato chips
lays twistz saucy tomatina potato chips

FOR QUERIES ON JESTR, CONTACT US AT

crm@singularium.in