# Representing Character Sequences as Sets

## A simple and intuitive string encoding algorithm for text data cleaning

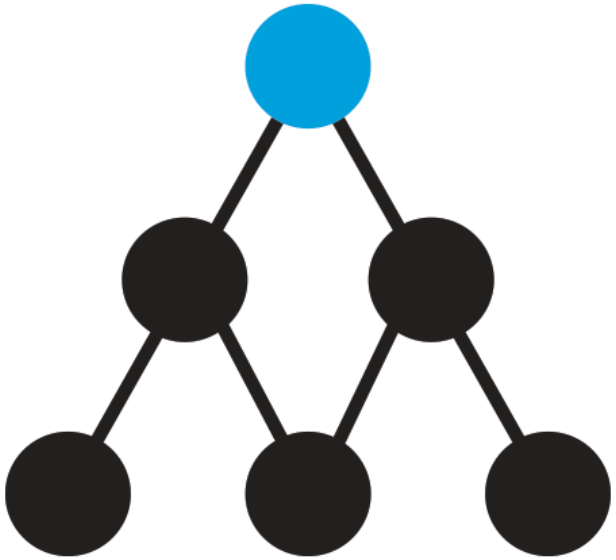
Martin Marinov – PhD student

Alexander Efremov – Associate Professor, scientific advisor

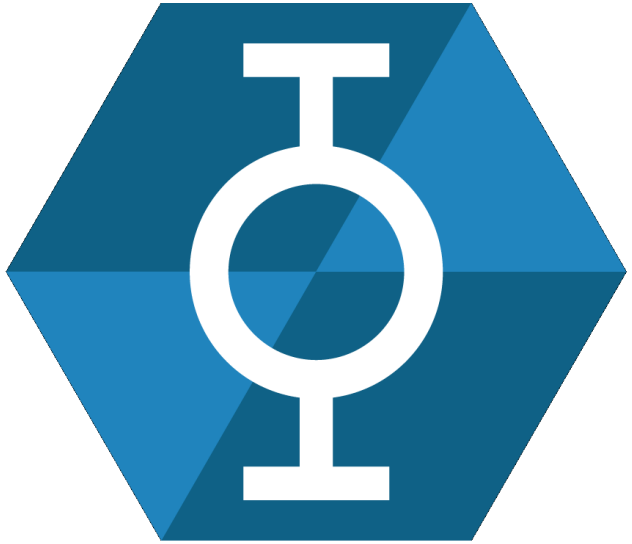
Faculty of Automation, Technical University of Sofia, Bulgaria

# Inspiration

## Numenta



## Cortical.io



# Encoding Information

## Information storage created by humans

## How nature does it - SDR

# Divergence

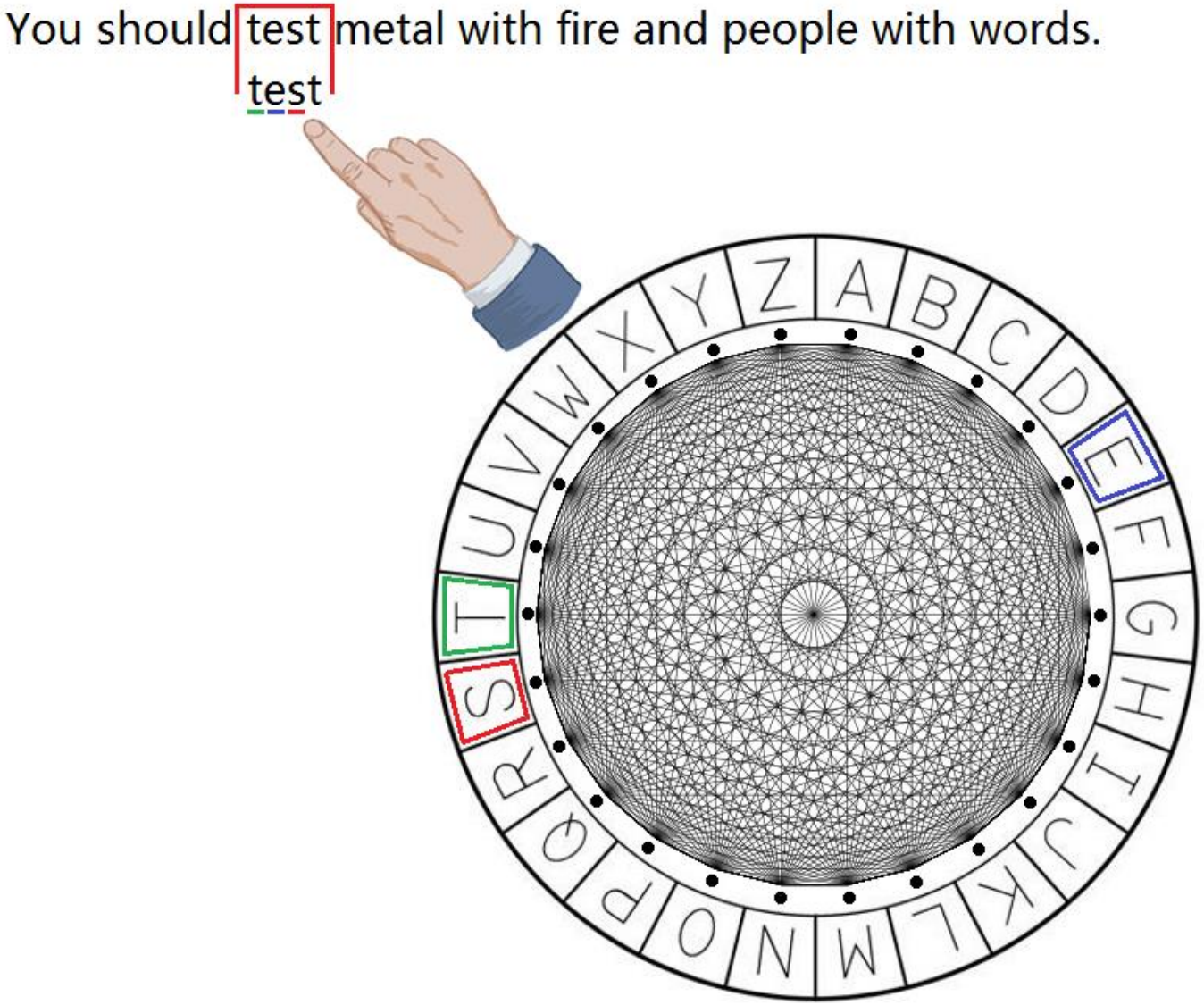
## What Numenta does:

- Streaming data/real-time prediction.
- Emphasis on forecasting/predictive modeling.
- Focuses on development of a generalized predictive framework.
- Develops weak AI, but also flirts with the idea of strong AI.

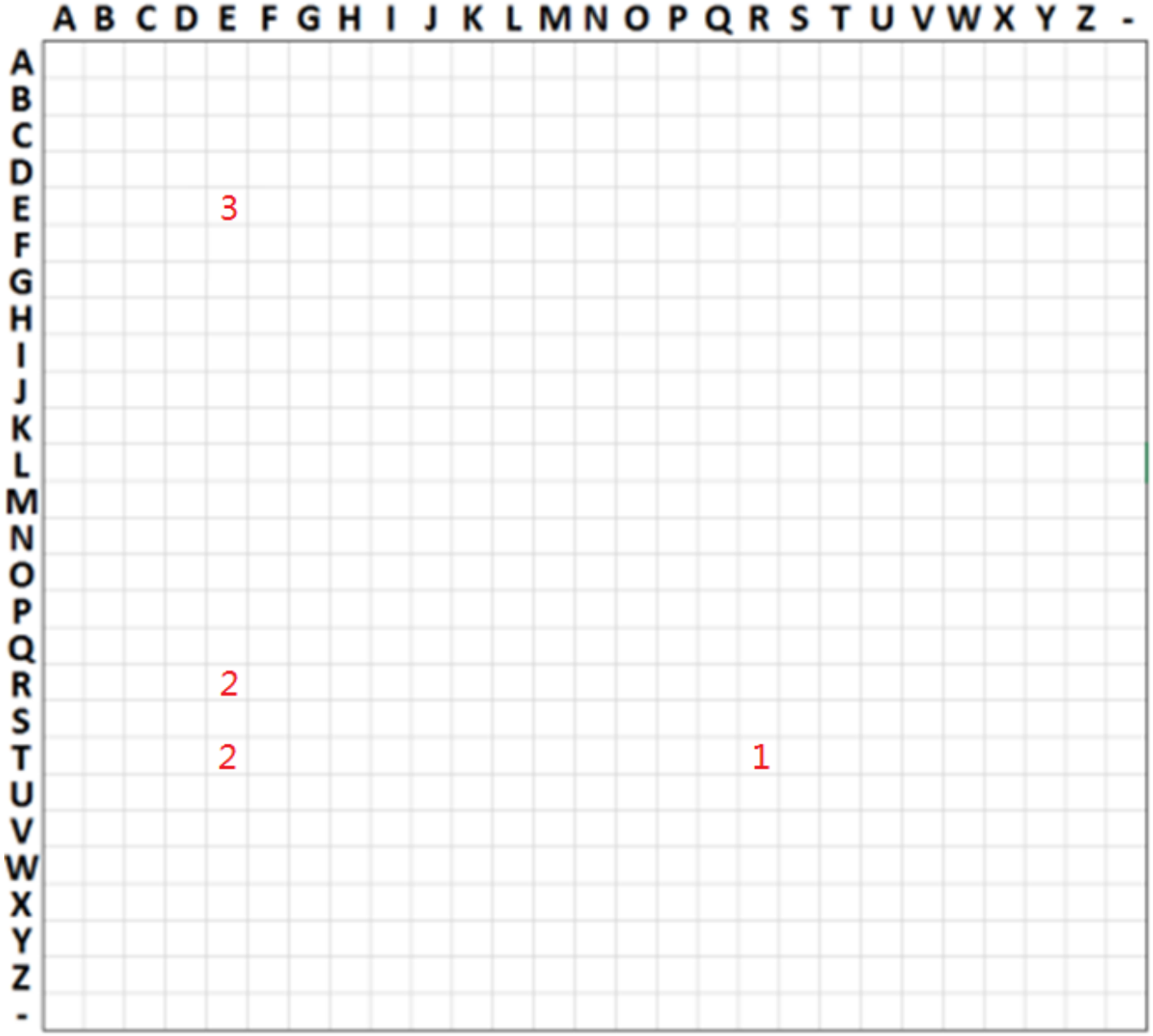
## The focus of the PhD thesis:

- Natural Language Processing.
- Emphasis on data preparation and cleaning.
- Development of a domain-specific encoding algorithm.
- Strictly focused on weak AI.

# ANN for Memorizing Character Sequences




- Sentences are tokenized and each token is scanned separately.
- The sensory organ can perceive only one letter at a time.
- Each artificial neuron corresponds to a specific Unicode character.
- As a token is scanned, connections are formed by active neurons only.
- All formed connections are saved as a group at the end of the scan.




Set representation: { tr, te, re, ee }

# Basic word encoding algorithm

Encoding the word **tree**.

Steps:

0. no_prior_char ← t
(do nothing)

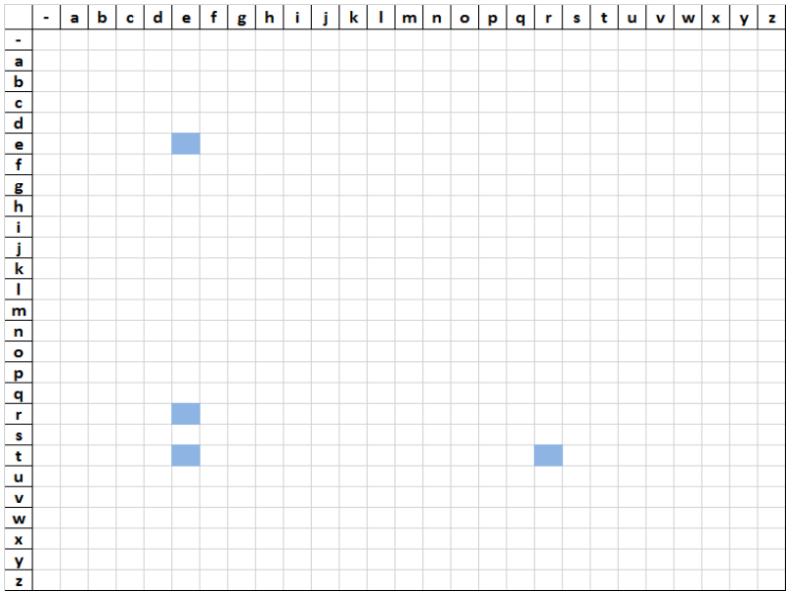
1. t ← r
(link **t** and **r**)

2. t ← e, r ← e
(form two links)

3. t ← e, r ← e , e ← e
(form three links)

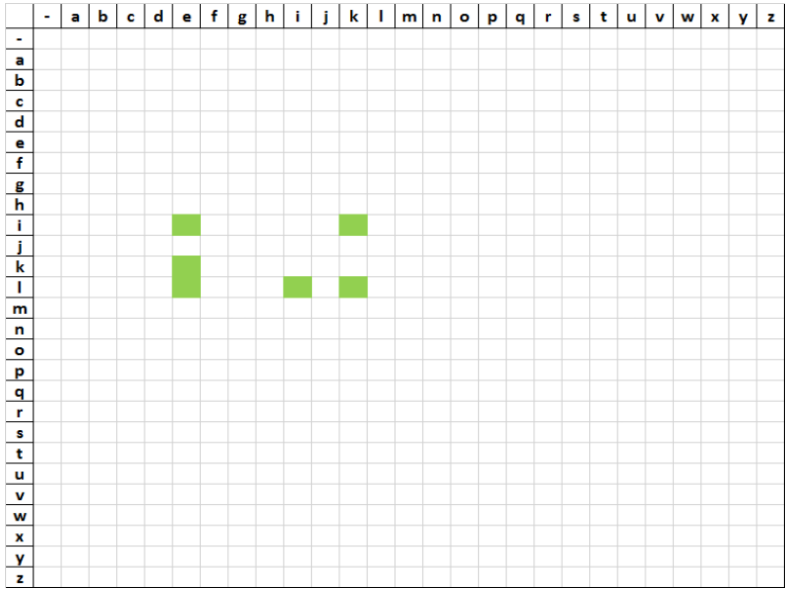
# Properties of Encoded Words

## tree



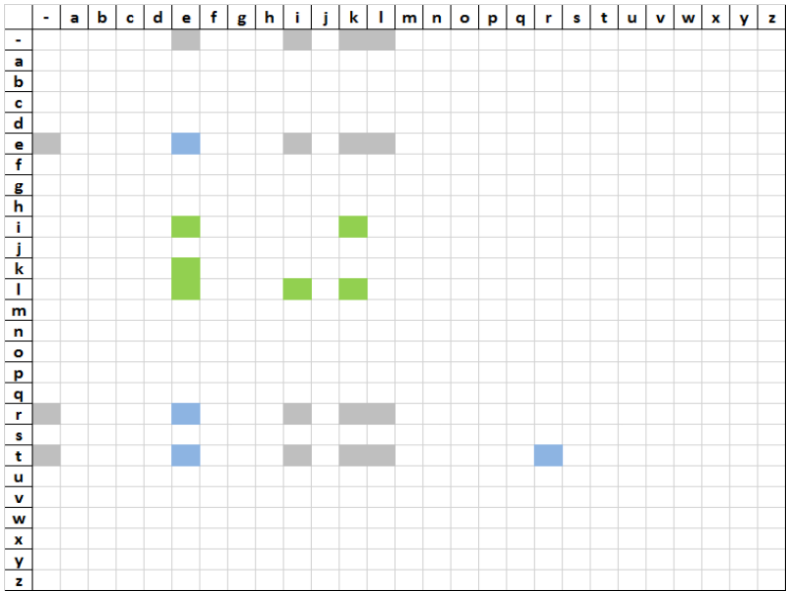
{ ee, re, te, tr }

## like



{ ie, ik, ke, le, li, lk }

## tree-like



{ -e, -i, -k, -l, e-, ee, ei, ek, el, ie, ik, ke, le, li, lk, r-, re, rl, t-, te, tr }

# Word Comparison Heuristic

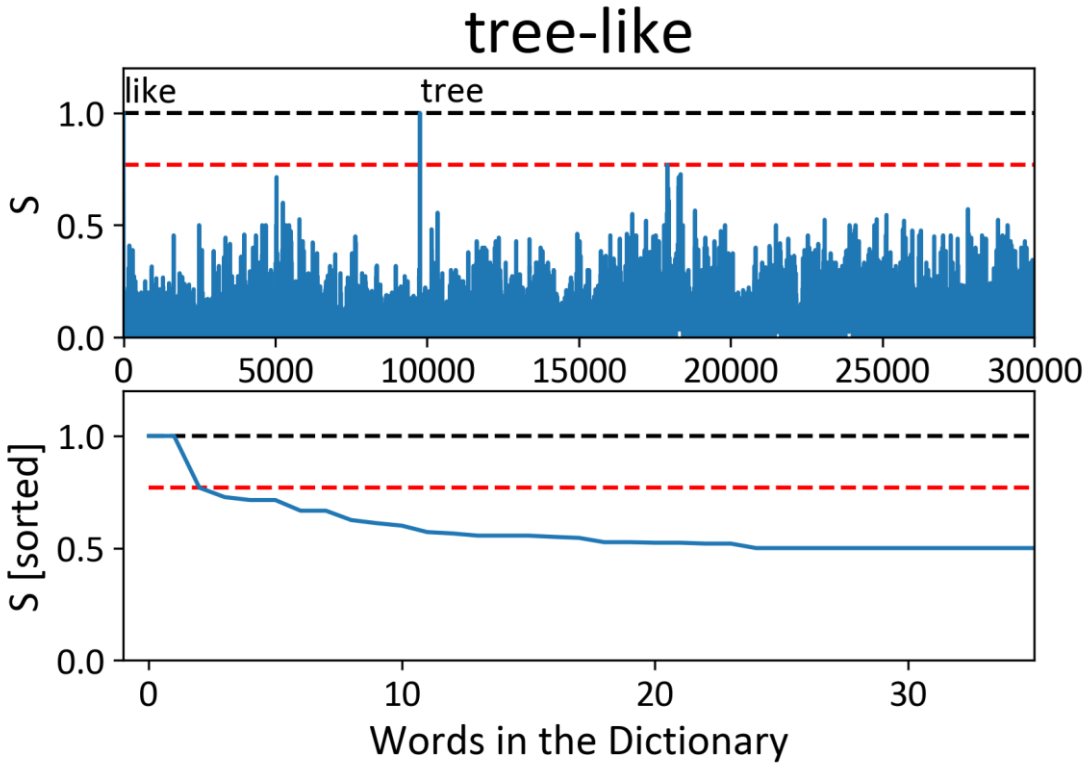
$$S = \frac{n(D_i \cap C)}{n(D_i)}$$

$S$ – Match score, a float within the interval [0, 1]. Quantifies similarity between words from the text data and a set of dictionary words.

$D_i$ – The set of character pairs, of the i-th encoded dictionary word.

$C$ – set of character pairs, of a single word from the text data being processed.

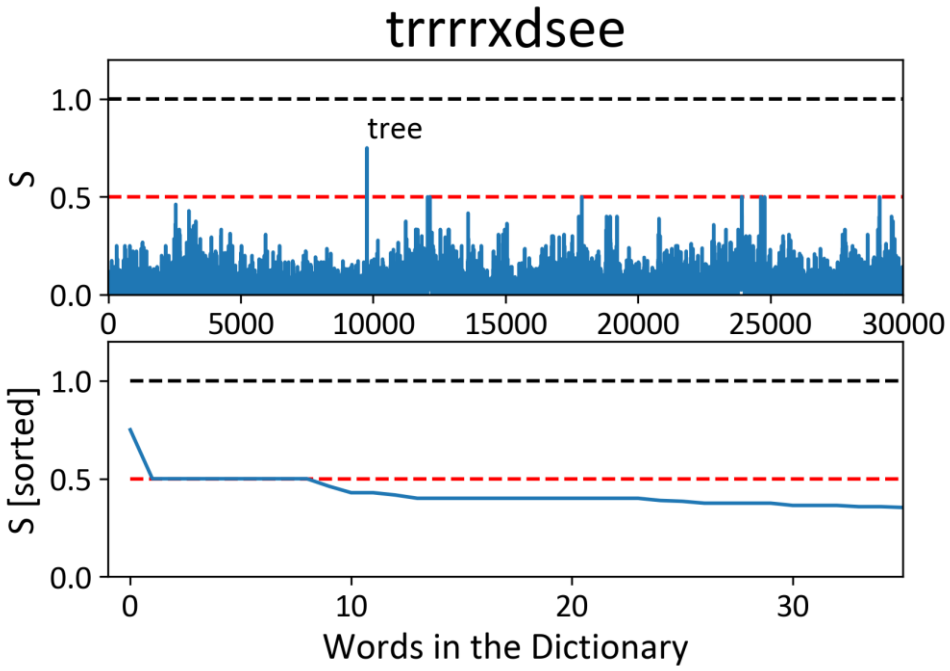
# Word Selection Heuristic



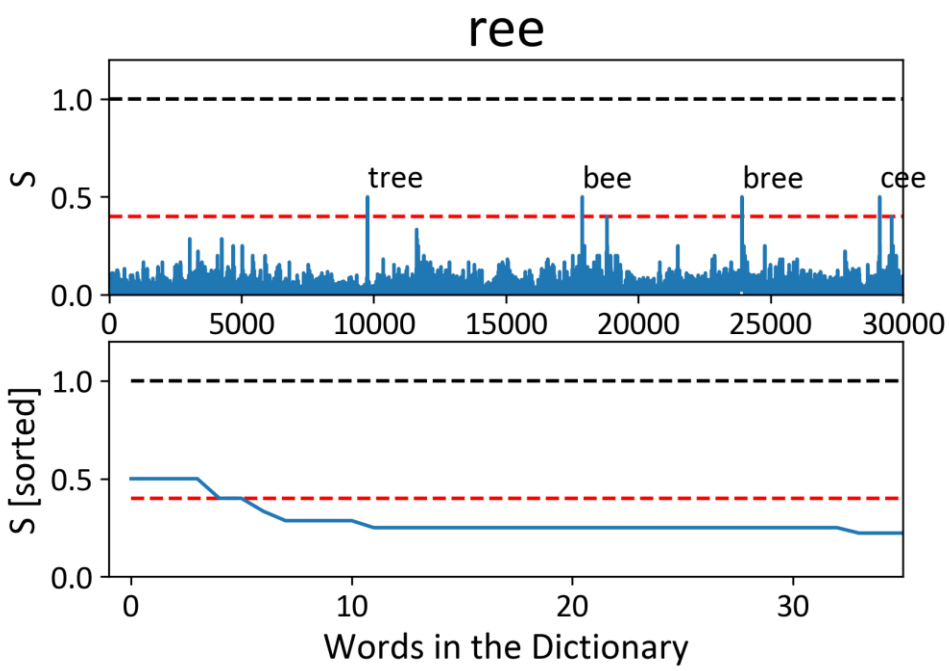

1. Sort all similarity scores in descending order.
2. Find the first score, which is lower than the mean, of the one before and the one after it.
3. Keep only dictionary entries, which have a similarity score above that threshold score.

# Interesting Examples

## Corrupted word lookup




## Handling Missing characters



# Uses of the Encoding Method and its Limitations

Strengths:

- Resilient to spelling mistakes and morphological variations in words.
- Can detect compound words or words merged by mistake into a single string.

Limitations:

- Doesn’t take into account word context, just works with character sequences.

Potential applications:

- Automated dictionary compilation from unstructured text documents.
- Automated, or perhaps automatic, data preparation for text data.
- Resolving database nomenclature inconsistencies.

# Thank you.