Linking via Fuzzy Matching

Alexey Osipov

11.12.2020

Affiliation.

SIDENIS

Swiss Re
Institute

Casualty R&D

Matching by meaningful content.

Statement of the problem

We have base of descriptions:

Establishments primarily engaged in the production of rice

We would like to link a phrase with it:

Rice

Examples

Classifying a new object.

Linking classifications.

How to link?

Algorithm.

- Preprocessing.
- Embeddings.
- Word2vec, GloVe.
- Embedding for a phrase, e.g., by averaging.
- Compare via cosine similarity.
- Other ways: doc2vec, text2vec.

Relevant R packages.

R packages.

- word2vec
- text2vec
- textTinyR

Linking by short phrases.

Statement of the problem

We have data like

BP plc

We would like to link a phrase with it:

BP Amoco plc

Examples

Linking datasets with company names.

How to link?

Algorithm.

- Preprocessing/normalization.
- Fuzzy matching by string distances:
  1. Jaccard.
  2. Jaro-Winkler.
  3. Levenshtein.
  4. Damereau-Levenshtein.
  5. Longest common substring.
- Filtering by threshold values.
- Disambiguation logic.

Relevant R packages.

R packages.

- stringdist
- hash
- megalodon

Linking with unstructured text.

Statement of the problem

We have data like

Grupo Acerinox, S. A. 257768

Cyclone Klaus

DO NOT USE Sweco Lietuva UAB

We would like to link it with the dataset with data like

Acerinox

Examples

Linking dataset with claim names with dataset with company names.

General scheme.

Scheme.

1. Classification block (distinguish noisy entries).
2. Entity extraction block.
3. Linking block

Classification and entity extraction block.

Scheme

- features from the dataset, features from the text.
- features based on presence of key words
- xgboost for classification.
- spacy ner model + regular expressions for entity extraction.
- xgboost, spacyr

Linking block.

Algorithm.

- calculate metric that can be calculated fast (TFIDF-based)
- calculate second level metrics (jaccard-based)
- choose threshold via clustering and manual validation
- disambiguation block
- dbscan

Thank you very much!