The Journal of Open Source Software

Talisman: a JavaScript archive of fuzzy matching, information retrieval and record linkage building blocks

Guillaume Plique[1]

1 médialab, SciencesPo Paris

DOI: 10.21105/joss.02405

Software

- Review
- Repository
- Archive

Editor: Kakia Chatsiou

Reviewers:

- @Fil
- @atanikan

Submitted: 11 June 2020
Published: 29 June 2020

Summary

Information retrieval and record linkage have always relied on crafty and heuristical routines aimed at implementing what is often called *fuzzy matching*. Indeed, even if fuzzy logic feels natural to humans, one needs to find various strategies to coerce computers into acknowledging that strings, for instance, are not always strictly delimited. But if some of those techniques, such as the Soundex phonetic algorithm invented by Robert Russell and Margaret King Odell at the beginning of the 20th century, are still well known and used, a lot of them were unfortunately lost to time.

As such, the JavaScript library **Talisman** aims at being an archive of a wide variety of techniques that have been used throughout computer sciences' history to perform fuzzy comparisons between words, names, sentences etc. Thus, even if **Talisman** obviously provides state-of-the-art functions that are still being used in an industrial context, it also aims at being a safe harbor for less known or clunkier techniques, for historical and archival purposes.

The library therefore compiles a large array of implementations of the following techniques:

- **keyers**: functions used to normalize strings in order to drop or simplify artifacts that could impair comparisons.
- **similarity metrics**: functions used to compute a similarity or distance between two strings, such as the Levenshtein distance, for instance.
- **phonetic algorithms**: functions aiming at producing a fuzzy phonetical representation of the given strings to enable comparisons.
- **stemmers**: functions reducing given strings to a *stem* to ease comparisons of a word's various inflections.
- **tokenizers**: functions used to cut strings into relevant pieces such as words, sentences etc.

Those building blocks can then be used to perform and improve the following tasks:

- Building more relevant search engines through fuzzy matching and indexing
- Clustering string by similarity
- Record linkage, entity resolution etc.
- Natural language processing

Finally, one should note that, being a code library, **Talisman** is able to archive a standardized way to implement some functions and algorithms whose descriptions are known to be somehow unclear, or imprecise, sometimes by consecrating typical implementation used in an industrial context and sometimes by chosing to respect the spirit of the original paper, against a faulty explanation of how the algorithm should behave.

References