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: 10 November 2020

Summary

Information retrieval (Baeza-Yates et al., 1999; Manning et al., 2008) and record linkage (Christen, 2012; Fellegi & Sunter, 1969; Herzog et al., 2007) 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 (Odell, 1956) invented 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 **Talisman** JavaScript library 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 (Levenshtein, 1966) or the Jaccard similarity (Jaccard, 1912), etc.
- **phonetic algorithms**: functions aiming at producing a fuzzy phonetical representation of the given strings to enable comparisons such as the *Kölner phonetik* (Postel, 1969) or the *Metaphone* (Philips, 1990), etc.
- **stemmers**: functions reducing given strings to a *stem* to ease comparisons of a word's various inflections such as the Porter stemmer (Van Rijsbergen et al., 1980), etc.
- **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, this library can also be used to perform some benchmarks of those building blocks, wrt. precision, recall etc. of the fuzzy matches, which is seldom done in the literature because of how hard it can be to find comprehensive archives aggregating many phonetic algorithms, stemmers etc.

Related works

- abydos: a python library implementing similar utilities.
- java-string-similarity, stringdistance: Java libraries implementing string distance/similarity functions.
- OpenRefine: a fully-fledged application designed to apply similar methods to typical data cleaning tasks.
- clj-fuzzy: a Clojure library which stands as an earlier version of **Talisman**

References

Baeza-Yates, R., Ribeiro-Neto, B., & others. (1999). *Modern information retrieval* (Vol. 463). ACM press New York.

Christen, P. (2012). *Data Matching*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31164-2

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Herzog, T. N., Scheuren, F., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer. https://doi.org/10.1007/0-387-69505-2

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge university press.

Odell, M. K. (1956). The profit in records management. *Systems (New York)*, *20*, 20.

Philips, L. (1990). Hanging on the metaphone. *Computer Language*, *7*(12), 39–43.

Postel, H. J. (1969). Die kölner phonetik. Ein verfahren zur identifizierung von personennamen auf der grundlage der gestaltanalyse. *IBM-Nachrichten*, *19*, 925–931.

Van Rijsbergen, C. J., Robertson, S. E., & Porter, M. F. (1980). *New models in probabilistic information retrieval*. British Library Research; Development Department London.