

sweater: Speedy Word Embedding Association Test and Extras Using R

Chung-hong Chan¹

¹ Mannheimer Zentrum für Europäische Sozialforschung, Universität Mannheim

DOI: [10.21105/joss.04036](https://doi.org/10.21105/joss.04036)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: [Sebastian Benthall](#) ↗

Reviewers:

- [@kbenoit](#)
- [@cmaimone](#)

Submitted: 20 November 2021

Published: 03 February 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Statement of need

The goal of this R package is to detect (implicit) biases in word embeddings. Word embeddings can capture how similar or different two words are in terms of implicit and explicit meanings. Using the example in [Collobert et al. \(2011\)](#), the word vector for “XBox” is close to that of “PlayStation,” as measured by a distance measure such as cosine distance. The same technique can also be used to detect biases. In the situation of racial bias detection, for example, [Kroon et al. \(2020\)](#) measure how close the word vectors for various ethnic group names (e.g. “Dutch,” “Belgian” , and “Syrian”) to that of various nouns related to threats (e.g. “terrorist,” “murderer,” and “gangster”). These biases in word embedding can be understood through the implicit social cognition model of media priming ([Arendt, 2013](#)). In this model, implicit stereotypes are defined as the “strength of the automatic association between a group concept (e.g., minority group) and an attribute (e.g., criminal).” ([Arendt, 2013, p. 832](#)) All of these bias detection methods are based on the strength of association between a concept (or a target) and an attribute in embedding spaces.

The importance of detecting biases in word embeddings is twofold. First, pretrained, biased word embeddings deployed in real-life machine learning systems can pose fairness concerns ([Boyarskaya et al., 2020](#); [Packer et al., 2018](#)). Second, biases in word embeddings reflect the biases in the original training material. Social scientists, communication researchers included, have exploited these methods to quantify (implicit) media biases by extracting biases from word embeddings locally trained on large text corpora ([Knoche et al., 2019](#); e.g. [Kroon et al., 2020](#); [Sales et al., 2019](#)).

Previously, the software of these methods is only scatteredly available as the addendum of the original papers and was implemented in different languages (Java, Python, etc.). *sweater* provides several of these bias detection methods in one unified package with a consistent R interface ([R Core Team, 2021](#)). Also, some provided methods are implemented in C++ for speed and interfaced to R using the Rcpp package ([Eddelbuettel, 2013](#)).¹

Related work

The R package *cbn* (<https://github.com/conjugateprior/cbn>) by Will Lowe provides tools for replicating the study by [Caliskan et al. \(2017\)](#). The Python package *wefe* ([Badilla et al., 2020](#)) provides several methods for bias evaluation in a unified (Python) interface.

¹Compared with a pure R implementation, the C++ implementation of Word Embedding Association Test in *sweater* is at least 7 times faster. See the benchmark [here](#).

Usage

In this section, I demonstrate how the package can be used to detect biases and reproduce some published findings.

Word Embeddings

The input word embedding w is a dense $m \times n$ matrix, where m is the total size of the vocabulary in the training corpus and n is the vector dimension size.

`sweater` supports input word embeddings, w , in several formats. For locally trained word embeddings, output from the R packages `word2vec` (Wijffels, 2021), `rsparse` (Selivanov, 2020) and `text2vec` (Selivanov et al., 2020) can be used directly with the packages primary functions, such as `query`.² Pretrained word embeddings in the so-called “word2vec” file format, such as those obtained online,³ can be converted to the dense numeric matrix format required with the `read_word2vec` function.

The package also provides three trimmed word embeddings for experimentation: `googlenews` (Mikolov et al., 2013), `glove_math` (Pennington et al., 2014), and `small_reddit` (An et al., 2018).

Query

`sweater` uses the concept of a *query* (Badilla et al., 2020) to study the biases in w and the *STAB* notation from Brunet et al. (2019) to form a query. A query contains two or more sets of seed words (wordsets selected by the individual administering the test, sometimes called “seed lexicons” or “dictionaries”). Among these seed wordsets, there should be at least one set of *target words* and one set of *attribute words*.

Target words are words that **should** have no bias and usually represent the concept one would like to probe for biases. For instance, Garg et al. (2018) investigated the “women bias” of occupation-related words and their target words contain “nurse,” “mathematician,” and “blacksmith.” These words can be used as target words because in an ideal world with no “women bias” associated with occupations, these occupation-related words should have no gender association.

Target words are denoted as wordsets \mathcal{S} and \mathcal{T} . All methods require \mathcal{S} while \mathcal{T} is only required for WEAT. For instance, the study of gender stereotypes in academic pursuits by Caliskan et al. (2017) used $\mathcal{S} = \{\text{math, algebra, geometry, calculus, equations, ...}\}$ and $\mathcal{T} = \{\text{poetry, art, dance, literature, novel, ...}\}$.

Attribute words are words that have known properties in relation to the bias. They are denoted as wordsets \mathcal{A} and \mathcal{B} . All methods require both wordsets except Mean Average Cosine Similarity (Manzini et al., 2019). For instance, the study of gender stereotypes by Caliskan et al. (2017) used $\mathcal{A} = \{\text{he, son, his, him, ...}\}$ and $\mathcal{B} = \{\text{she, daughter, hers, her, ...}\}$. In some applications, popular off-the-shelf sentiment dictionaries can also be used as \mathcal{A} and \mathcal{B} (e.g. Sweeney & Najafian, 2020). That being said, it is up to the researchers to select and derive these seed words in a query. However, the selection of seed words has been shown to be the most consequential part of the entire analysis (Antoniak & Mimno, 2021; Du et al., 2021). Please read Antoniak & Mimno (2021) for recommendations.

²The vignette of `text2vec` provides a guide on how to locally train word embeddings using the GloVe algorithm (Pennington et al., 2014). <https://cran.r-project.org/web/packages/text2vec/vignettes/glove.html>

³For example, the pretrained GloVe word embeddings, pretrained word2vec word embeddings and pretrained fastText word embeddings.

75 Supported methods

76 Table 1 lists all methods supported by sweater. The function query is used to conduct a
 77 query. The function calculate_es can be used for some methods to calculate the effect size
 78 representing the overall bias of w from the query.

Table 1: All methods supported by sweater

Method	Target words	Attribute words
Mean Average Cosine Similarity (Manzini et al., 2019)	\mathcal{S}	\mathcal{A}
Relative Norm Distance (Garg et al., 2018)	\mathcal{S}	\mathcal{A}, \mathcal{B}
Relative Negative Sentiment Bias (Sweeney & Najafian, 2020)	\mathcal{S}	\mathcal{A}, \mathcal{B}
SemAxis (An et al., 2018)	\mathcal{S}	\mathcal{A}, \mathcal{B}
Normalized Association Score (Caliskan et al., 2017)	\mathcal{S}	\mathcal{A}, \mathcal{B}
Embedding Coherence Test (Dev & Phillips, 2019)	\mathcal{S}	\mathcal{A}, \mathcal{B}
Word Embedding Association Test (Caliskan et al., 2017)	\mathcal{S}, \mathcal{T}	\mathcal{A}, \mathcal{B}

79 Example 1

80 Relative Norm Distance (RND) ([Garg et al., 2018](#)) is calculated with two sets of attribute
 81 words. The following analysis reproduces the calculation of “women bias” values in [Garg et al. \(2018\)](#). The publicly available word2vec word embeddings trained on the Google News
 82 corpus is used ([Mikolov et al., 2013](#)). Words such as “nurse,” “midwife” and “librarian” are
 83 more associated with female, as indicated by the positive relative norm distance (Figure 1).
 84

```
library(sweater)

data(googlenews)
S1 <- c("janitor", "statistician", "midwife", "bailiff", "auctioneer",
        "photographer", "geologist", "shoemaker", "athlete", "cashier",
        "dancer", "housekeeper", "accountant", "physicist", "gardener",
        "dentist", "weaver", "blacksmith", "psychologist", "supervisor",
        "mathematician", "surveyor", "tailor", "designer", "economist",
        "mechanic", "laborer", "postmaster", "broker", "chemist",
        "librarian", "attendant", "clerical", "musician", "porter",
        "scientist", "carpenter", "sailor", "instructor", "sheriff",
        "pilot", "inspector", "mason", "baker", "administrator",
        "architect", "collector", "operator", "surgeon", "driver",
        "painter", "conductor", "nurse", "cook", "engineer", "retired",
        "sales", "lawyer", "clergy", "physician", "farmer", "clerk",
        "manager", "guard", "artist", "smith", "official", "police",
        "doctor", "professor", "student", "judge", "teacher", "author",
        "secretary", "soldier")
A1 <- c("he", "son", "his", "him", "father", "man", "boy", "himself",
        "male", "brother", "sons", "fathers", "men", "boys", "males",
        "brothers", "uncle", "uncles", "nephew", "nephews")
B1 <- c("she", "daughter", "hers", "her", "mother", "woman", "girl",
        "herself", "female", "sister", "daughters", "mothers", "women",
```

```
"girls", "females", "sisters", "aunt", "aunts", "niece", "nieces")
res_rnd_male <- query(w = googlenews, S_words = S1,
                     A_words = A1, B_words= B1,
                     method = "rnd")
plot(res_rnd_male)
```

DRAFT

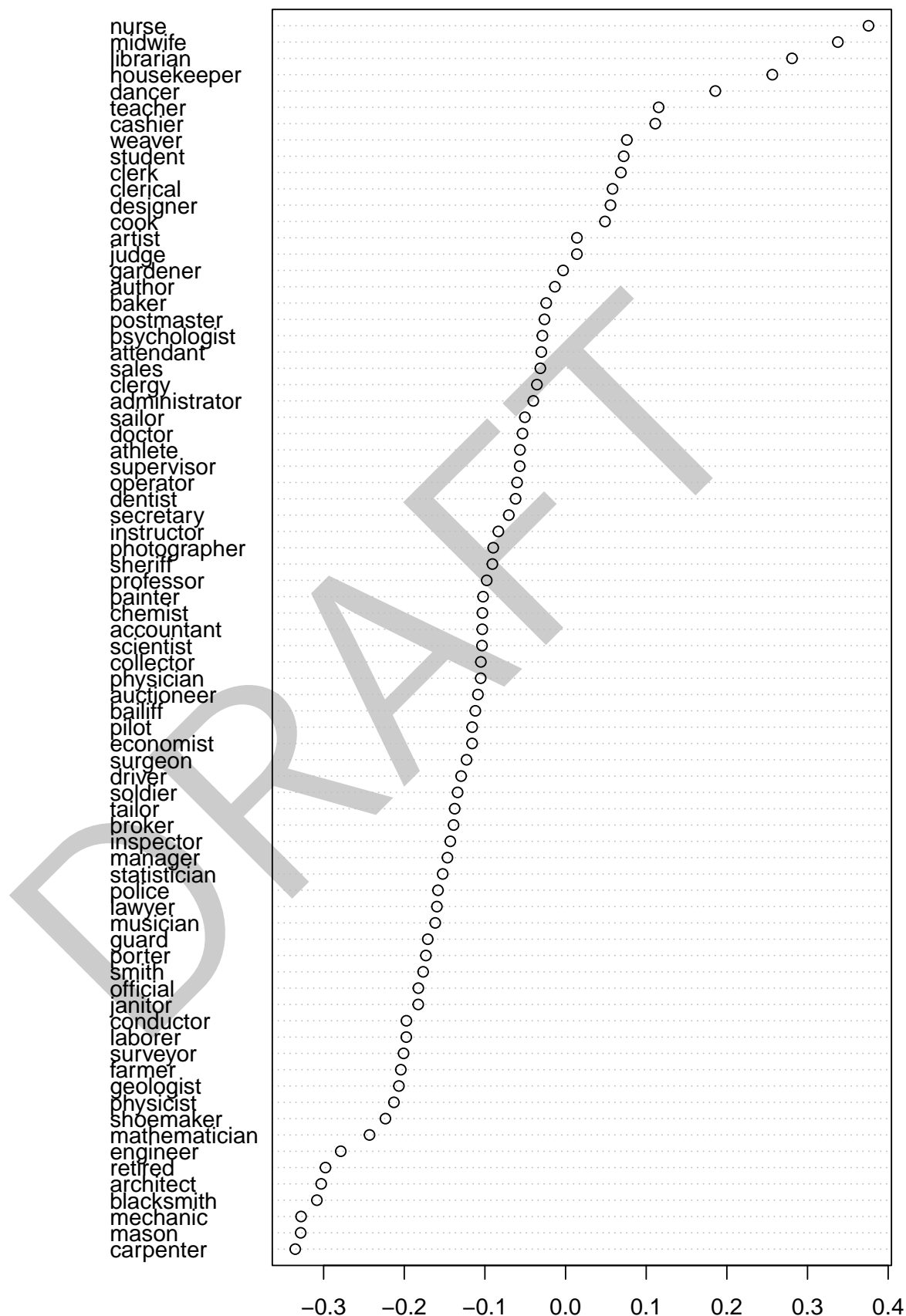


Figure 1: Bias of words in the target wordset according to relative norm distance

85 Example 2

86 Word Embedding Association Test (WEAT) (Caliskan et al., 2017) requires all four wordsets
 87 of S , T , A , and B . The method is modeled after the Implicit Association Test (IAT) (Nosek
 88 et al., 2005) and it measures the relative strength of S 's association with A to B against
 89 the same of T . The effect sizes calculated from a large corpus, as shown by Caliskan et al.
 90 (2017), are comparable to the published IAT effect sizes obtained from volunteers.

91 In this example, the publicly available GloVe embeddings made available by the original
 92 Stanford Team (Pennington et al., 2014) were used. In the following example, the calculation
 93 of "Math. vs Arts" gender bias in Caliskan et al. (2017) is reproduced. In this example, the
 94 positive effect size indicates the words in the wordset S are more associated with males than
 95 T associated with males.

```

data(glove_math) # a subset of the original GloVe word vectors
S2 <- c("math", "algebra", "geometry", "calculus", "equations",
        "computation", "numbers", "addition")
T2 <- c("poetry", "art", "dance", "literature", "novel", "symphony",
        "drama", "sculpture")
A2 <- c("male", "man", "boy", "brother", "he", "him", "his", "son")
B2 <- c("female", "woman", "girl", "sister", "she", "her", "hers",
        "daughter")
sw <- query(w = glove_math,
            S_words = S2, T_words = T2,
            A_words = A2, B_words = B2)

sw

##

## -- sweater object -----

## Test type: weat
## Effect size: 1.055015

##

## -- Functions -----

## * `calculate_es()`: Calculate effect size

## * `weat_resampling()`: Conduct statistical test

The statistical significance of the effect size can be evaluated using the function weat_resa
mpling.

weat_resampling(sw)

##
## Resampling approximation of the exact test in Caliskan et al. (2017)
##
## data: sw
## bias = 0.024865, p-value = 0.0171
## alternative hypothesis: true bias is greater than 7.245425e-05
## sample estimates:
## bias
## 0.02486533

```

Acknowledgements

The development of this package was supported by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (*Bundesministerium für Familie, Senioren, Frauen und Jugend*), the Federal Republic of Germany – Research project: “*Erfahrungen von Alltagsrassismus und medienvermittelter Rassismus in der (politischen) Öffentlichkeit.*”

References

- An, J., Kwak, H., & Ahn, Y.-Y. (2018). Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. *arXiv Preprint arXiv:1806.05521*. <https://doi.org/10.18653/v1/p18-1228>
- Antoniak, M., & Mimno, D. (2021). Bad seeds: Evaluating lexical methods for bias measurement. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1889–1904. <https://doi.org/10.18653/v1/2021.acl-long.148>
- Arendt, F. (2013). Dose-dependent media priming effects of stereotypic newspaper articles on implicit and explicit stereotypes. *Journal of Communication*, 63(5), 830–851. <https://doi.org/10.1111/jcom.12056>
- Badilla, P., Bravo-Marquez, F., & P'erez, J. (2020). WEFE: The word embeddings fairness evaluation framework. *IJCAI*, 430–436. <https://doi.org/10.24963/ijcai.2020/60>
- Boyarskaya, M., Olteanu, A., & Crawford, K. (2020). Overcoming Failures of Imagination in AI Infused System Development and Deployment. *arXiv Preprint arXiv:2011.13416*.
- Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. *International Conference on Machine Learning*, 803–811.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10.1126/science.aal4230>
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(ARTICLE), 2493–2537.
- Dev, S., & Phillips, J. (2019). Attenuating bias in word vectors. *The 22nd International Conference on Artificial Intelligence and Statistics*, 879–887.
- Du, Y., Fang, Q., & Nguyen, D. (2021). Assessing the reliability of word embedding gender bias measures. *arXiv Preprint arXiv:2109.04732*.
- Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. <https://doi.org/10.1007/978-1-4614-6868-4>
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644. <https://doi.org/10.1073/pnas.1720347115>
- Knoche, M., Popović, R., Lemmerich, F., & Strohmaier, M. (2019). Identifying biases in politically biased wikis through word embeddings. *Proceedings of the 30th ACM conference on hypertext and social media*, 253–257. <https://doi.org/10.1145/3342220.3343658>
- Kroon, A. C., Trilling, D., & Raats, T. (2020). Guilty by association: Using word embeddings to measure ethnic stereotypes in news coverage. *Journalism & Mass Communication Quarterly*, 1077699020932304. <https://doi.org/10.1177/1077699020932304>

- 159 Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as
160 caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv*
161 *Preprint arXiv:1904.04047*. <https://doi.org/10.18653/v1/n19-1062>
- 162 Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed represen-
163 tations of words and phrases and their compositionality. *Advances in Neural Information*
164 *Processing Systems*, 3111–3119.
- 165 Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and Using the
166 Implicit Association Test: II. Method Variables and Construct Validity. *Personality and*
167 *Social Psychology Bulletin*, 31(2), 166–180. <https://doi.org/10.1177/0146167204271418>
- 168 Packer, B., Mitchell, M., Guajardo-C'espedes, M., & Halpern, Y. (2018). *Text embeddings*
169 *contain bias. Here's why that matters*. [https://developers.googleblog.com/2018/04/](https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html)
170 [text-embedding-models-contain-bias.html](https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html)
- 171 Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word represen-
172 tation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*
173 *Processing (EMNLP)*. <https://doi.org/10.3115/v1/d14-1162>
- 174 R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation
175 for Statistical Computing. <https://www.R-project.org/>
- 176 Sales, A., Balby, L., & Veloso, A. (2019). Media bias characterization in brazilian presidential
177 elections. *Proceedings of the 30th ACM Conference on Hypertext and Social Media*,
178 231–240. <https://doi.org/10.1145/3345645.3351107>
- 179 Selivanov, D. (2020). *Rsparse: Statistical learning on sparse matrices*. [https://CRAN.](https://CRAN.R-project.org/package=rsparse)
180 [R-project.org/package=rsparse](https://CRAN.R-project.org/package=rsparse)
- 181 Selivanov, D., Bickel, M., & Wang, Q. (2020). *text2vec: Modern text mining framework for*
182 *R*. <https://CRAN.R-project.org/package=text2vec>
- 183 Sweeney, C., & Najafian, M. (2020). Reducing sentiment polarity for demographic attributes
184 in word embeddings using adversarial learning. *Proceedings of the 2020 Conference on*
185 *Fairness, Accountability, and Transparency*, 359–368. [https://doi.org/10.1145/3351095.](https://doi.org/10.1145/3351095.3372837)
186 [3372837](https://doi.org/10.1145/3351095.3372837)
- 187 Wijffels, J. (2021). *word2vec: Distributed representations of words*. [https://CRAN.R-project.](https://CRAN.R-project.org/package=word2vec)
188 [org/package=word2vec](https://CRAN.R-project.org/package=word2vec)