Tool Demo: An R package for detecting biases in word embeddings

Chung-hong Chan[1]

[1] Mannheimer Zentrum für Europäische Sozialforschung

Author Note

The development of this package was partly supported by a research grant from the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (Bundesministerium für Familie, Senioren, Frauen und Jugend), the Federal Republic of Germany.

Correspondence concerning this article should be addressed to Chung-hong Chan, A5, 6 (section A), 68159 Mannheim, Germany. E-mail: chung-hong.chan@mzes.uni-mannheim.de

Abstract

This paper shows how the R package sweater can be used to detect biases in word embeddings. The package provides highly optimized functions to calculate the following bias metrics: mean average cosine similarity, relative norm distance, SemAxis, normalized association score, relative negative sentiment bias, embedding coherence test and word embedding association test. Using two public available word embeddings trained on media content, this paper demonstrates how sweater can be used to study implicit gender stereotypes.

Keywords: word embedding, bias, fairness, gender stereotypes

Tool Demo: An R package for detecting biases in word embeddings

Statement of need

The goal of the R package `sweater` is to detect (implicit) biases in word embeddings. 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, Olteanu, & Crawford, 2020; Packer, Mitchell, Guajardo-Céspedes, & Halpern, 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 (e.g. Kroon, Trilling, & Raats, 2020; Knoche, Popović, Lemmerich, & Strohmaier, 2019; Sales, Balby, & Veloso, 2019). 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.

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 in `sweater` are implemented in C++ and interfaced to R using the Rcpp package (Eddelbuettel, 2013). These heavily optimized methods, such as the Word Embedding Association Test (WEAT) (Caliskan, Bryson, & Narayanan, 2017), are significantly faster than the same methods implemented in interpreted languages.

In the usage section below, we demonstrated how the package can be used to detect biases and reproduce some published findings.

Usage

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. Let v_x denote a row vector of w, the word vector of the word x.

`sweater` supports two types of w. For locally trained word embeddings, word embedding outputs from the R packages *word2vec* (Wijffels, 2021), *rsparse* (Selivanov, 2020) and *text2vec* (Selivanov et al., 2020) are directly supported.[1] For pretrained word embeddings obtained online,[2] they are usually provided in the so-called "word2vec" file format and `sweater`'s function `read_word2vec` reads those files into the supported matrix format.

Query

`sweater` uses the concept of *query* (Badilla, Bravo-Marquez, & Pérez, 2020) to study the biases in w. A query contains two or more sets of seed words with at least one set of *target words* and one set of *attribute words*. `sweater` uses the $\mathcal{STAB}$ notation from Brunet, Alkalay-Houlihan, Anderson, and Zemel (2019) to form a query.

Target words are words that **should** have no bias. They 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 $\mathcal{S} = \{math, algebra, geometry, calculus, equations, computation, numbers, addition\}$ and

[1] The vignette of *text2vec* provides a guide on how to locally train word embeddings using the GLoVE algorithm (Pennington, Socher, & Manning, 2014) on a large corpus from R. https://cran.r-project.org/web/packages/text2vec/vignettes/glove.html

[2] For example, the pretrained GLoVE word embeddings provided in https://nlp.stanford.edu/projects/glove/, pretrained word2vec word embeddings provided in https://wikipedia2vec.github.io/wikipedia2vec/pretrained/ and fastText word embeddings provided in https://fasttext.cc/docs/en/english-vectors.html.

$\mathcal{T} = \{poetry, art, dance, literature, novel, symphony, drama, sculpture\}$.

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, Lim, Tsvetkov, & Black, 2019). For instance, the study of gender stereotypes by Caliskan et al. (2017) used

$\mathcal{A} = \{he, son, his, him, father, man, boy, himself, male, ...\}$ and

$\mathcal{B} = \{she, daughter, hers, her, mother, woman, girl, herself, female, ...\}$. 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, Fang, & Nguyen, 2021). Please read Antoniak and Mimno (2021) for recommendations.

Supported methods

Table 1 lists all methods supported by `sweater`. As Relative Norm Distance, SemAxis, Normalized Association Score, and Embedding Coherence Test are all distance-based measures, only Relative Norm Distance is demonstrated below.

Examples

In the following examples, the publicly available word2vec word embeddings trained on the Google News corpus is used (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). `sweater` provides a unified interface for making query, calculating effect size, and plotting result. Three functions are used:

- `query` for making query. It follows a template of `method(w, S_words, T_words, A_words, B_words, method)` for a query.
- `calculate_es` for calcuting effect size which represents the overall bias of w based on the query

- `plot` for visualizing the result of a query

The argument `method` of the function `query` determines which method to use (see Table 1 for the list of shorthands). By default, it is set to "guess", i.e. the function select the appropriate method for you based on your provided $\mathcal{S}$, $\mathcal{T}$, $\mathcal{A}$, and $\mathcal{B}$.

Mean Average Cosine Similarity

Average cosine similarity (Manzini et al., 2019) is calculated as the mean cosine similarity between the word vector of a target word v_s where $s \in \mathcal{S}$ and that of all terms in the attribute wordset $\mathcal{A}$. The same method was deployed in Kroon et al. (2020).

The average cosine similarity of many occupation words in $\mathcal{S}$ are calculated against a wordset $\mathcal{A}$ of attribute words related to male.

The code below shows how to conduct a query using the `query` function. When only $\mathcal{S}$ and $\mathcal{A}$ are provided, the function assumes that the query calls for calcuting average cosine similarity. The function returns an S3 object.

In an interactive session, printing the S3 object shows the effect size and other functions (or more accurately, S3 methods) that can further process the object (Figure 1).

For most of the functions, the returned S3 object contains a slot `P`, which stores the bias of each word (e.g. `res_mac_male$P`). The average cosine similarity values are `P` in this case. The function `plot` can be used to visualize `P` as a Cleveland Dot Plot (Figure 2).

```r
plot(res_mac_male)
```

The effect size, mean average cosine similarity, is the mean value of all average cosine similarity values.

```r
calculate_es(res_mac_male)
```

```
## [1] 0.1375856
```

Relative Norm Distance

Relative norm distance (RND) (Garg et al., 2018) is calculated with two sets of attribute words. The following analysis reproduces the calculation of "women bias" values in Garg et al. (2018). Compared with average cosine similarity, RND appears to be reflecting the underlying gender bias more accurately (Figure 3).

```r
B <- 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 = S, A_words= A, B_words = B)
plot(res_rnd_male)
```

The effect size is the sum of all `P`. As the effect size is negative, it indicates that the concept of occupation is more associated with $\mathcal{A}$, i.e. male.

```r
calculate_es(res_rnd_male)
```

```
## [1] -6.341598
```

Relative Negative Sentiment Bias

Relative negative sentiment bias (RNSB) (Sweeney & Najafian, 2020) takes the same query template as RND. But the technique is not based on a distance metric such as cosine similarity. Instead, the method trained a regularized logistic regression model on the word vectors $v_{x \in \mathcal{A} \bigcup \mathcal{B}}$ to predict the probability of x being in $\mathcal{B}$. The bias is quantified as the relative probability of the word s for being a word in the wordset $\mathcal{B}$ (Figure 4).

```r
res_rnsb_male <- query(w = googlenews, S_words = S, A_words = A,
                       B_words = B, method = "rnsb")
plot(res_rnsb_male)
```

The effect size in this case is the Kullback–Leibler divergence of P from the uniform distribution.

```r
calculate_es(res_rnsb_male)
```

```
## [1] 0.07398497
```

Word Embedding Association Test

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

In this example, a different w is used. It is the publicly available GLoVE embeddings made available by the original Stanford Team (Pennington et al., 2014). The same GLoVE embeddings were used in Caliskan et al. (2017). In the following example, the calculation of "Math. vs Arts" gender bias is reproduced. Please note that for WEAT, the returned object does not contain `P`. By default, the effect size is standardized so that it can be interpreted the same way as Cohen's D (Cohen, 2013).

```r
data(glove_math) # a subset of the original GLoVE word vectors
S <- c("math", "algebra", "geometry", "calculus", "equations", "computation",
       "numbers", "addition")
T <- c("poetry", "art", "dance", "literature", "novel", "symphony", "drama",
```

```
       "sculpture")
A <- c("male", "man", "boy", "brother", "he", "him", "his", "son")
B <- c("female", "woman", "girl", "sister", "she", "her", "hers", "daughter")
sw <- query(glove_math, S, T, A, B)
sw
```

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

The effect size can also be converted to point-biserial correlation coefficient.

```
calculate_es(sw, r = TRUE)
```

```
## [1] 0.4912066
```

One can also obtain the unstandardized effect size. In the original paper (Caliskan et al., 2017), it is referred to as “test statistic”.

```
calculate_es(sw, standardize = FALSE)
```

```
## [1] 0.02486533
```

One can also test the statistical significance of the effect size. The original paper suggests an exact test (Caliskan et al., 2017). This exact test is implemented in this package as the function `weat_exact`. But the exact test takes a long time to calculate when the number of words in $\mathcal{S}$ is larger than a few words.

Instead, we recommend the resampling approximation of the exact test. The p-value is extremely close to the reported 0.018.

```
weat_resampling(sw)
```

```
##
##  Resampling approximation of the exact test in Caliskan et al. (2017)
##
## data:  sw
## bias = 0.024865, p-value = 0.0154
## alternative hypothesis: true bias is greater than -2.784776e-05
## sample estimates:
##       bias
## 0.02486533
```

Conclusion

This paper demonstrates how `sweater` can be used to detect biases in word embeddings.

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*.

Antoniak, M., & Mimno, D. (2021). Bad seeds: Evaluating lexical methods for bias measurement. In *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)* (pp. 1889–1904).

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., & Pérez, J. (2020). WEFE: The word embeddings fairness evaluation framework. In *IJCAI* (pp. 430–436).

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. In *International conference on machine learning* (pp. 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

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Dev, S., & Phillips, J. (2019). Attenuating bias in word vectors. In *The 22nd international conference on artificial intelligence and statistics* (pp. 879–887). PMLR.

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. In *Proceedings of the 30th ACM conference on hypertext and social media* (pp. 253–257).

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

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv Preprint arXiv:1904.04047*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity. *Personality and Social Psychology Bulletin*, *31*(2), 166–180. https://doi.org/10.1177/0146167204271418

Packer, B., Mitchell, M., Guajardo-Céspedes, M., & Halpern, Y. (2018). Text embeddings contain bias. Here's why that matters. Retrieved from https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/d14-1162

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sales, A., Balby, L., & Veloso, A. (2019). Media bias characterization in brazilian presidential elections. In *Proceedings of the 30th acm conference on hypertext and social media* (pp. 231–240). https://doi.org/10.1145/3345645.3351107

Selivanov, D. (2020). *Rsparse: Statistical learning on sparse matrices*. Retrieved from https://CRAN.R-project.org/package=rsparse

Selivanov, D., Bickel, M., & Wang, Q. (2020). *Text2vec: Modern text mining framework for R*. Retrieved from https://CRAN.R-project.org/package=text2vec

Sweeney, C., & Najafian, M. (2020). Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 359–368).

Wijffels, J. (2021). *Word2vec: Distributed representations of words*. Retrieved from https://CRAN.R-project.org/package=word2vec

Table 1

All methods supported by `sweater`

Method	Target words	Attribute words	shorthand
Mean Average Cosine Similarity (Manzini et al., 2019)	$\mathcal{S}$	$\mathcal{A}$	`mac`
Relative Norm Distance (Garg, Schiebinger, Jurafsky, & Zou, 2018)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$	`rnd`
SemAxis (An, Kwak, & Ahn, 2018)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$	`semaxis`
Normalized Association Score (Caliskan et al., 2017)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$	`nas`
Embedding Coherence Test (Dev & Phillips, 2019)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$	`ect`
Relative Negative Sentiment Bias (Sweeney & Najafian, 2020)	$\mathcal{S}$	$\mathcal{A}$, $\mathcal{B}$	`rnsb`
Word Embedding Association Test (Caliskan et al., 2017)	$\mathcal{S}$, $\mathcal{T}$	$\mathcal{A}$, $\mathcal{B}$	`weat`

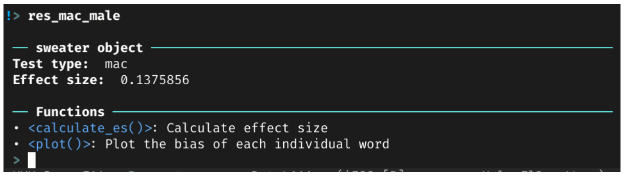

Figure 1. The printed S3 object produced with ‘query‘ in an R interactive session

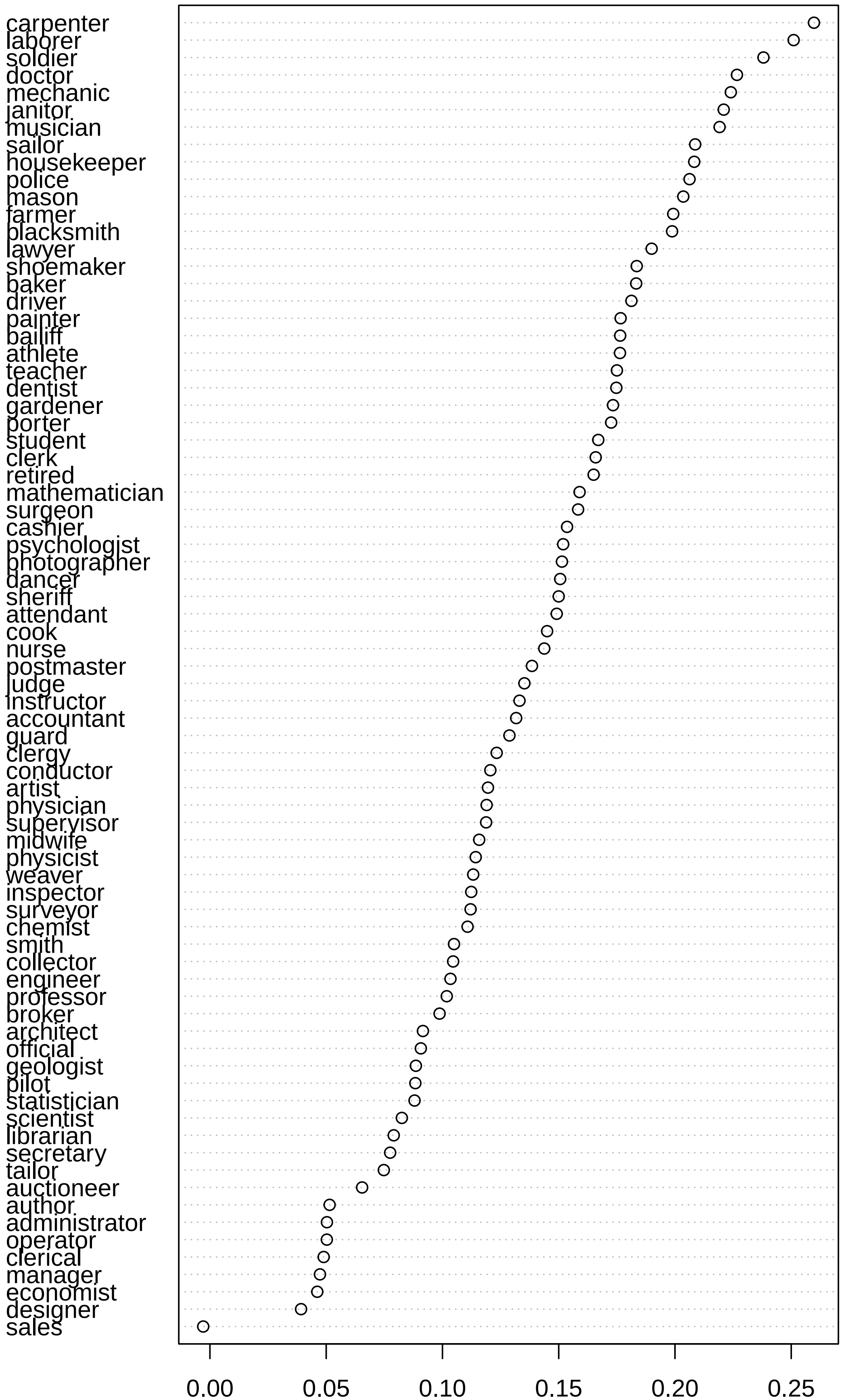

Figure 2. Bias of words in the target wordset according to average cosine similarity

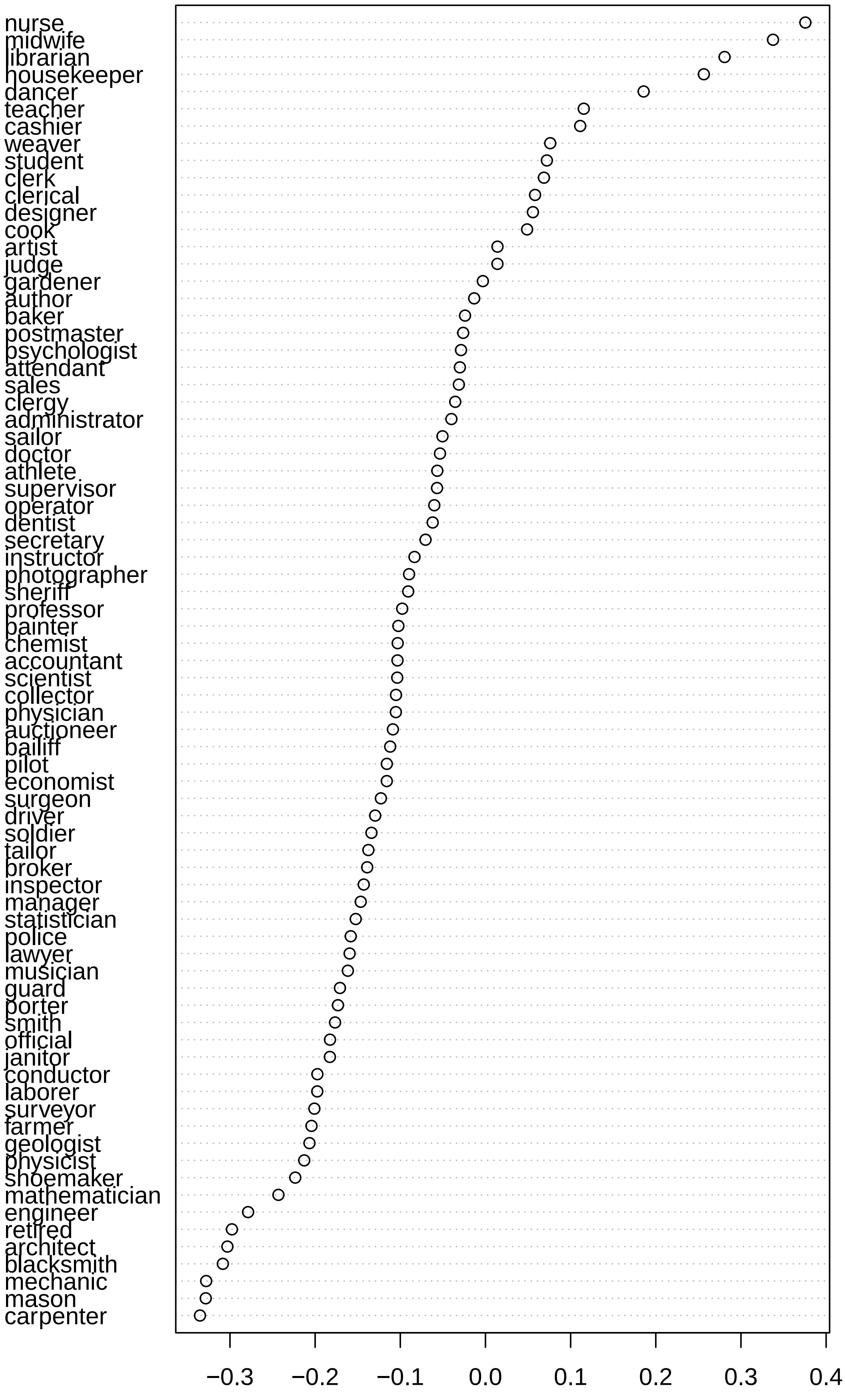

Figure 3. Bias of words in the target wordset according to relative norm distance

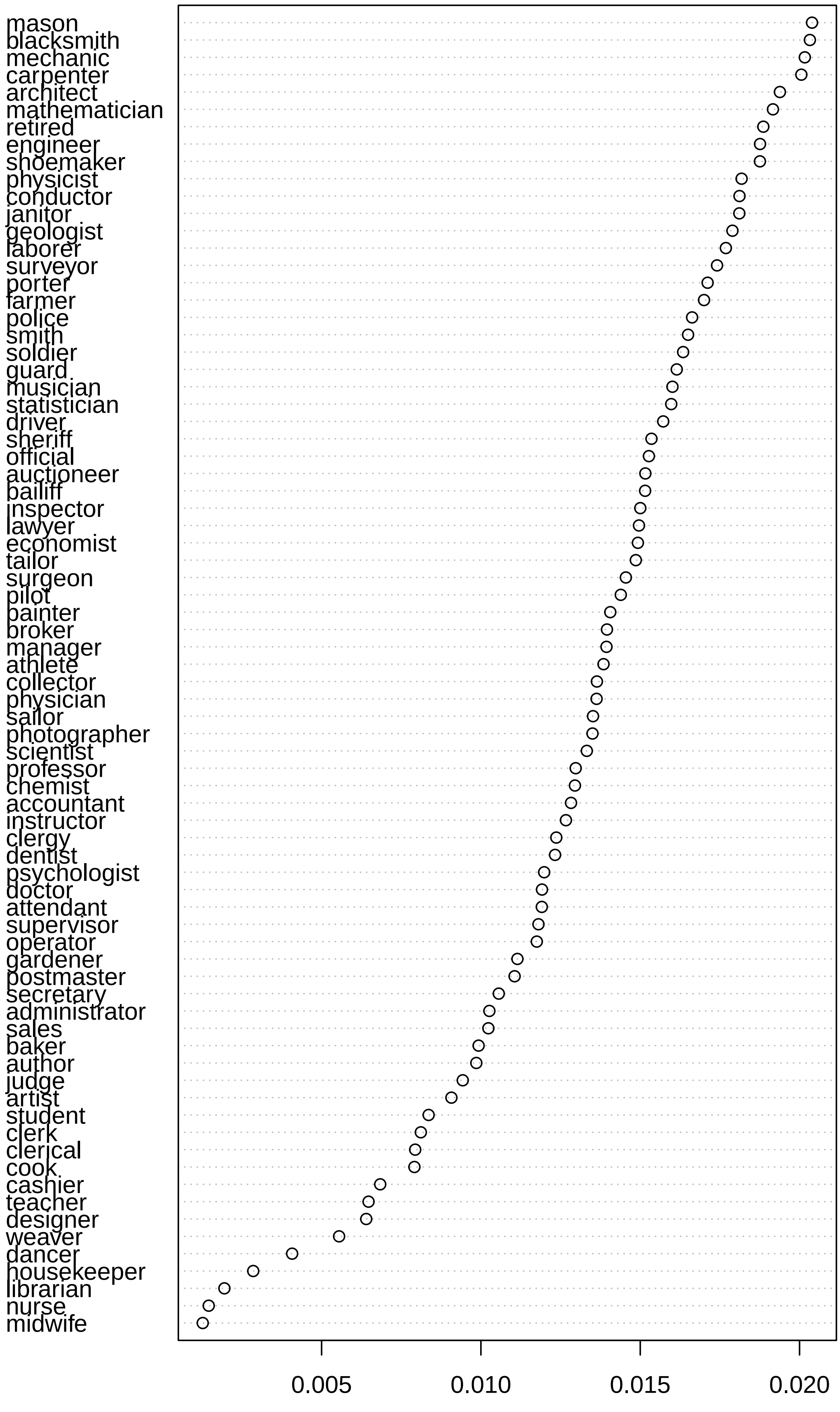

Figure 4. Bias of words in the target wordset according to relative negative sentiment bias