sweater: Speedy Word Embedding Association Test and Extras Using R

Chung-hong Chan

Mannheimer Zentrum für Europäische Sozialforschung, Universität Mannheim

DOI: 10.21105/joss.04036

Statement of need

The goal of this R package is to detect associations among words in word embedding spaces. 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 for “PlayStation”, as measured by a distance measure such as cosine distance. Word embeddings can also be used to study associations among words that are otherwise difficult to detect. For instance, Jing & Ahn (2021) used word embeddings to study how Democrats and Republicans are divided along party lines about COVID-19.

The same technique can also be used to detect unwanted implicit associations, or biases. For example, Kroon, Trilling, & Raats (2020) measure how close the word vectors for various ethnic group names (e.g. “Dutch”, “Belgian”, and “Syrian”) are to those for 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, 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 (Knoche, Popović, Lemmerich, & Strohmaier, 2019; e.g. Kroon et al., 2020; Sales, Balby, & Veloso, 2019).

Previously, the software of these methods is only available piecemeal 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, sweater provides several methods that are implemented in C++ for speed and interfaced to R using the Rcpp package (Eddelbuettel, 2013).

Related work

The R package cbn (https://github.com/conjugateprior/cbn) by Will Lowe provides tools for replicating the study by Caliskan, Bryson, & Narayanan (2017). The Python package wefe

\footnote{Compared 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.}
(Badilla, Bravo-Marquez, & P’erez, 2020) provides several methods for bias evaluation in a unified (Python) interface.

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, Bickel, & Wang, 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, Sutskever, Chen, Corrado, & Dean, 2013), glove_math (Pennington et al., 2014), and small_reddit (An, Kwak, & Ahn, 2018).

Query

sweater uses the concept of a query (Badilla et al., 2020) to study associations in \( w \) and the \( STAB \) notation from Brunet, Alkalay-Houlihan, Anderson, & Zemel (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.

In the situation of bias detection, target words are words that should have no bias and usually represent the concept one would like to probe for biases. For instance, Garg, Schiebinger, Jurafsky, & Zou (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 \( S \) and \( T \). All methods require \( S \) while \( T \) is only required for Word Embedding Association Test (WEAT). For instance, the study of gender stereotypes in academic pursuits by Caliskan et al. (2017) used \( S = \{math, algebra, geometry, calculus, equations, ...\} \) and \( T = \{poetry, art, dance, literature, novel, ...\} \).

In the situation of bias detection, attribute words are words that have known properties in relation to the bias. They are denoted as wordsets \( A \) and \( B \). 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 \( A = \{he, son, his, him, ...\} \) and \( B = \{she, daughter, hers, her, ...\} \). In some applications, popular off-the-shelf sentiment dictionaries can also be used as \( A \) and \( B \) (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.

\(^2\)The vignette of text2vec provides a guide on how to locally train word embeddings using the GLoVE algorithm (Pennington, Socher, & Manning, 2014). https://cran.r-project.org/web/packages/text2vec/vignettes/glove.html

\(^3\)For example, the pretrained GLoVE word embeddings, pretrained word2vec word embeddings and pretrained fastText word embeddings.


**Supported methods**

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

**Table 1: All methods supported by sweater**

<table>
<thead>
<tr>
<th>Method</th>
<th>Target words</th>
<th>Attribute words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Average Cosine Similarity (Manzini et al., 2019)</td>
<td>( S )</td>
<td>( A )</td>
</tr>
<tr>
<td>Relative Norm Distance (Garg et al., 2018)</td>
<td>( S )</td>
<td>( A, B )</td>
</tr>
<tr>
<td>Relative Negative Sentiment Bias (Sweeney &amp; Najafian, 2020)</td>
<td>( S )</td>
<td>( A, B )</td>
</tr>
<tr>
<td>SemAxis (An et al., 2018)</td>
<td>( S )</td>
<td>( A, B )</td>
</tr>
<tr>
<td>Normalized Association Score (Caliskan et al., 2017)</td>
<td>( S )</td>
<td>( A, B )</td>
</tr>
<tr>
<td>Embedding Coherence Test (Dev &amp; Phillips, 2019)</td>
<td>( S )</td>
<td>( A, B )</td>
</tr>
<tr>
<td>Word Embedding Association Test (Caliskan et al., 2017)</td>
<td>( S, T )</td>
<td>( A, B )</td>
</tr>
</tbody>
</table>

**Example 1**

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). The publicly available word2vec word embeddings trained on the Google News corpus is used (Mikolov et al., 2013). Words such as “nurse”, “midwife” and “librarian” are more associated with female, as indicated by the positive relative norm distance (Figure 1).

```r
library(sweater)


A1 <- c("he", "son", "his", "him", "father", "man", "boy", "himself", "male", "brother", "sons", "fathers", "men", "boys", "males", "experimental support for quanteda dictionaries (Benoit et al., 2018) is current available for this method. The support will be expanded to all methods later.

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

https://doi.org/10.21105/joss.04036.
Example 2

Word Embedding Association Test (WEAT) (Caliskan et al., 2017) requires all four wordsets of $S$, $T$, $A$, and $B$. The method is modeled after the Implicit Association Test (IAT) (Nosek, Greenwald, & Banaji, 2005) and it measures the relative strength of $S$’s association with $A$ to $B$ against the same of $T$. 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, the publicly available GLoVE embeddings made available by the original Stanford Team (Pennington et al., 2014) were used. In the following example, the calculation of “Math vs Arts” gender bias in Caliskan et al. (2017) is reproduced. In this example, the positive effect size indicates the words in the wordset $S$ are more associated with males than are the words in wordset $T$.

```r
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
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_resampling`.

```r
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).

References


