

# Introduction to word2vec

Claire He

2023-09-16

```
## [1] "I went on a successful date with someone I felt sympathy and connection with."
## [2] "I was happy when my son got 90% marks in his examination"
## [3] "I went to the gym this morning and did yoga."
## [4] "We had a serious talk with some friends of ours who have been flaky lately. They understood and
## [5] "I went with grandchildren to butterfly display at Crohn Conservatory\n"
## [6] "I meditated last night."
```

```
##   wid  age country gender marital parenthood
## 1   1  37.0     USA      m married          y
## 2   2  29.0     IND      m married          y
## 3   3   25     IND      m  single          n
## 4   4   32     USA      m married          y
## 5   5   29     USA      m married          y
## 6   6   35     IND      m married          y
```

*Question:* Effect of animals on happiness across gender groups.

We are going to use `word2vec` to try to investigate this question. To illustrate this we are going to try to compare use of `word2vec` and `lda`.

First convert the data to a list of characters to input into our model. Choose the dimension of the embedding (arbitrarily chose 15 but this is a tuning parameter you can try a few dimensions and see what works best for you). You can also choose either models based on `cbow` or `skipgram`.

## Overall vocabulary between happiness and animals

```
x <- tolower(cleaned_data$cleaned_hm)
cat(x[1])
```

```
## i went on a successful date with someone i felt sympathy and connection with.
```

Lemmatizing our text and using speech tag (verb, adverb, noun, adjective) will make representation easier (let's say we want to see all adjectives and nouns relative to the topic of animals).

```
## 2023-09-16 21:13:46 Annotating text fragment 1/100392
## 2023-09-16 21:13:46 Annotating text fragment 11/100392
## 2023-09-16 21:13:46 Annotating text fragment 21/100392
## 2023-09-16 21:13:46 Annotating text fragment 31/100392
## 2023-09-16 21:13:46 Annotating text fragment 41/100392
## 2023-09-16 21:13:46 Annotating text fragment 51/100392
```

[illegible]

[illegible]

[illegible]





[illegible]

[illegible]



[illegible]



[illegible]



[illegible]









[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]





[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]





















[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]





















[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]







[illegible]



[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]























[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

96



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]





















[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]



[illegible]





[illegible]

[illegible]



[illegible]





[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]





[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]



[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]







[illegible]

[illegible]



[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]



185

[illegible]

```

## 2023-09-16 21:31:27 Annotating text fragment 99961/100392
## 2023-09-16 21:31:27 Annotating text fragment 99971/100392
## 2023-09-16 21:31:28 Annotating text fragment 99981/100392
## 2023-09-16 21:31:28 Annotating text fragment 99991/100392
## 2023-09-16 21:31:28 Annotating text fragment 100001/100392
## 2023-09-16 21:31:28 Annotating text fragment 100011/100392
## 2023-09-16 21:31:28 Annotating text fragment 100021/100392
## 2023-09-16 21:31:28 Annotating text fragment 100031/100392
## 2023-09-16 21:31:28 Annotating text fragment 100041/100392
## 2023-09-16 21:31:28 Annotating text fragment 100051/100392
## 2023-09-16 21:31:28 Annotating text fragment 100061/100392
## 2023-09-16 21:31:29 Annotating text fragment 100071/100392
## 2023-09-16 21:31:29 Annotating text fragment 100081/100392
## 2023-09-16 21:31:29 Annotating text fragment 100091/100392
## 2023-09-16 21:31:29 Annotating text fragment 100101/100392
## 2023-09-16 21:31:29 Annotating text fragment 100111/100392
## 2023-09-16 21:31:29 Annotating text fragment 100121/100392
## 2023-09-16 21:31:30 Annotating text fragment 100131/100392
## 2023-09-16 21:31:30 Annotating text fragment 100141/100392
## 2023-09-16 21:31:30 Annotating text fragment 100151/100392
## 2023-09-16 21:31:30 Annotating text fragment 100161/100392
## 2023-09-16 21:31:30 Annotating text fragment 100171/100392
## 2023-09-16 21:31:30 Annotating text fragment 100181/100392
## 2023-09-16 21:31:30 Annotating text fragment 100191/100392
## 2023-09-16 21:31:31 Annotating text fragment 100201/100392
## 2023-09-16 21:31:31 Annotating text fragment 100211/100392
## 2023-09-16 21:31:31 Annotating text fragment 100221/100392
## 2023-09-16 21:31:31 Annotating text fragment 100231/100392
## 2023-09-16 21:31:31 Annotating text fragment 100241/100392
## 2023-09-16 21:31:31 Annotating text fragment 100251/100392
## 2023-09-16 21:31:31 Annotating text fragment 100261/100392
## 2023-09-16 21:31:31 Annotating text fragment 100271/100392
## 2023-09-16 21:31:32 Annotating text fragment 100281/100392
## 2023-09-16 21:31:32 Annotating text fragment 100291/100392
## 2023-09-16 21:31:32 Annotating text fragment 100301/100392
## 2023-09-16 21:31:32 Annotating text fragment 100311/100392
## 2023-09-16 21:31:32 Annotating text fragment 100321/100392
## 2023-09-16 21:31:32 Annotating text fragment 100331/100392
## 2023-09-16 21:31:32 Annotating text fragment 100341/100392
## 2023-09-16 21:31:32 Annotating text fragment 100351/100392
## 2023-09-16 21:31:32 Annotating text fragment 100361/100392
## 2023-09-16 21:31:32 Annotating text fragment 100371/100392
## 2023-09-16 21:31:33 Annotating text fragment 100381/100392
## 2023-09-16 21:31:33 Annotating text fragment 100391/100392

```

We now want to get the words most similar to animal in the embedding and we compare them to the words most similar to happy.



```
## $title
## [1] "100 most similar words to animal with word2vec - umap"
##
## attr(,"class")
## [1] "labels"
```

## Per gender

We need to combine `demo_data$gender` to `cleaned_data`, join on `wid`.

```
## 2023-09-16 21:33:18 Annotating text fragment 1/42019
## 2023-09-16 21:33:18 Annotating text fragment 11/42019
## 2023-09-16 21:33:18 Annotating text fragment 21/42019
## 2023-09-16 21:33:18 Annotating text fragment 31/42019
## 2023-09-16 21:33:18 Annotating text fragment 41/42019
## 2023-09-16 21:33:18 Annotating text fragment 51/42019
## 2023-09-16 21:33:19 Annotating text fragment 61/42019
## 2023-09-16 21:33:19 Annotating text fragment 71/42019
## 2023-09-16 21:33:19 Annotating text fragment 81/42019
## 2023-09-16 21:33:19 Annotating text fragment 91/42019
## 2023-09-16 21:33:19 Annotating text fragment 101/42019
## 2023-09-16 21:33:19 Annotating text fragment 111/42019
## 2023-09-16 21:33:19 Annotating text fragment 121/42019
## 2023-09-16 21:33:19 Annotating text fragment 131/42019
## 2023-09-16 21:33:19 Annotating text fragment 141/42019
```



190

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]









[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

















[illegible]









234

































[illegible]



[illegible]



[illegible]











[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

```
## 2023-09-16 21:40:45 Annotating text fragment 41731/42019
## 2023-09-16 21:40:46 Annotating text fragment 41741/42019
## 2023-09-16 21:40:46 Annotating text fragment 41751/42019
## 2023-09-16 21:40:46 Annotating text fragment 41761/42019
## 2023-09-16 21:40:46 Annotating text fragment 41771/42019
## 2023-09-16 21:40:46 Annotating text fragment 41781/42019
## 2023-09-16 21:40:46 Annotating text fragment 41791/42019
## 2023-09-16 21:40:46 Annotating text fragment 41801/42019
## 2023-09-16 21:40:47 Annotating text fragment 41811/42019
## 2023-09-16 21:40:47 Annotating text fragment 41821/42019
## 2023-09-16 21:40:47 Annotating text fragment 41831/42019
## 2023-09-16 21:40:47 Annotating text fragment 41841/42019
## 2023-09-16 21:40:47 Annotating text fragment 41851/42019
## 2023-09-16 21:40:47 Annotating text fragment 41861/42019
## 2023-09-16 21:40:47 Annotating text fragment 41871/42019
## 2023-09-16 21:40:47 Annotating text fragment 41881/42019
## 2023-09-16 21:40:48 Annotating text fragment 41891/42019
## 2023-09-16 21:40:48 Annotating text fragment 41901/42019
## 2023-09-16 21:40:48 Annotating text fragment 41911/42019
## 2023-09-16 21:40:48 Annotating text fragment 41921/42019
## 2023-09-16 21:40:48 Annotating text fragment 41931/42019
## 2023-09-16 21:40:48 Annotating text fragment 41941/42019
## 2023-09-16 21:40:48 Annotating text fragment 41951/42019
## 2023-09-16 21:40:48 Annotating text fragment 41961/42019
## 2023-09-16 21:40:49 Annotating text fragment 41971/42019
## 2023-09-16 21:40:49 Annotating text fragment 41981/42019
## 2023-09-16 21:40:49 Annotating text fragment 41991/42019
## 2023-09-16 21:40:49 Annotating text fragment 42001/42019
## 2023-09-16 21:40:49 Annotating text fragment 42011/42019
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



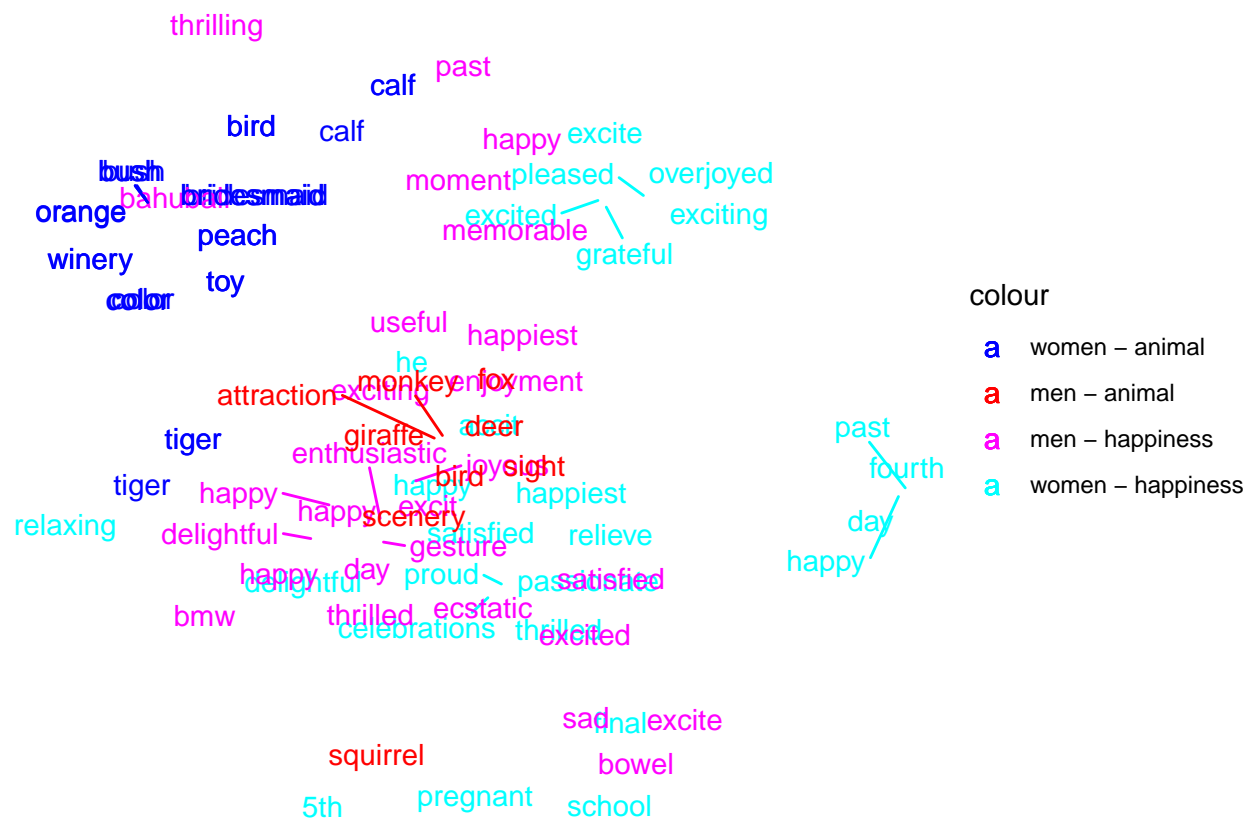
[illegible]

[illegible]

[illegible]

[illegible]

```
## 2023-09-16 21:51:18 Annotating text fragment 57481/57597
## 2023-09-16 21:51:19 Annotating text fragment 57491/57597
## 2023-09-16 21:51:19 Annotating text fragment 57501/57597
## 2023-09-16 21:51:19 Annotating text fragment 57511/57597
## 2023-09-16 21:51:19 Annotating text fragment 57521/57597
## 2023-09-16 21:51:19 Annotating text fragment 57531/57597
## 2023-09-16 21:51:19 Annotating text fragment 57541/57597
## 2023-09-16 21:51:19 Annotating text fragment 57551/57597
## 2023-09-16 21:51:20 Annotating text fragment 57561/57597
## 2023-09-16 21:51:20 Annotating text fragment 57571/57597
## 2023-09-16 21:51:20 Annotating text fragment 57581/57597
## 2023-09-16 21:51:20 Annotating text fragment 57591/57597
```



```
## $title
## [1] "Most similar words to animal with word2vec - umap"
##
## attr("class")
## [1] "labels"
```