Topic 7: Word Embeddings

Steven Cognac

2022-05-11

What are word Embeddings?

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. In Natural Language Processing for text analysis, they typically come in the form of real-valued vectors that encode the meaning of words such that the words closer together are expected to be similar in meaning.

For this exercise, we'll use pre-trained word vector from the GloVe: Global Vectors for Word Representation project. Specifically, we'll use the Wiki2014 + Gigword5 300d vector file.

Import data

I downloaded the data and saved it to a local folder

```
glov_df <- read_table(here('07_word_embeddings/data/glove.6B.300d.txt'),
                      col_names = FALSE) %>%
  column_to_rownames(., var = "X1")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   X1 = col_character()
## )
## i Use `spec()` for the full column specifications.
```

Create similarity function

```
search_synonyms <- function(glov_df, selected_vector) {

  dat <- glov_df %*% selected_vector

  similarities <- dat %>%
        tibble(token = rownames(dat),
               similarity = dat[,1])

  similarities %>%
       arrange(-similarity) %>%
```

```
        select(c(2,3))
}
```

Word Synonyms

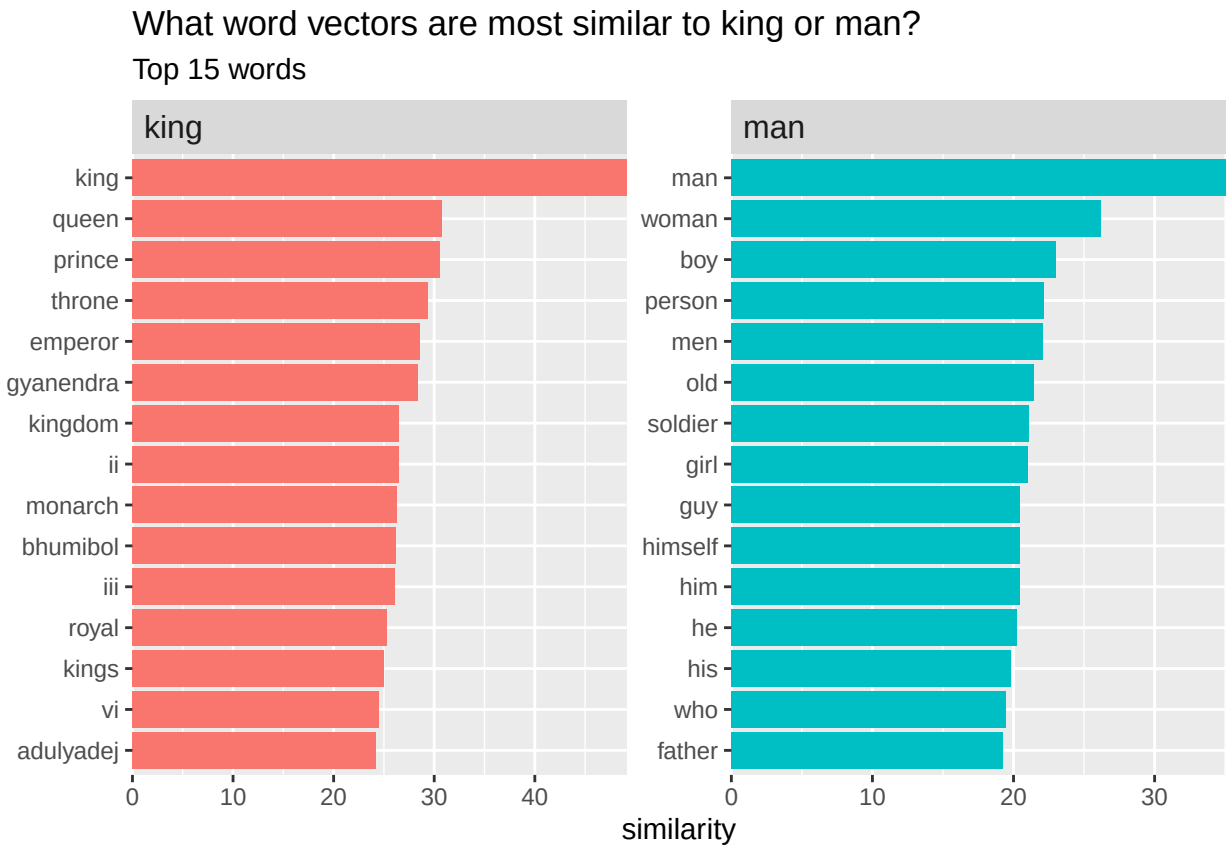
First we'll convert the dataframe to a matrix

```
glov_matrix <- data.matrix(glov_df)
```

You can search word synonyms with the `search_synonyms()` function. We can then plot those synonyms.

```
king <- search_synonyms(glov_matrix, glov_matrix["king",])
man <- search_synonyms(glov_matrix, glov_matrix["man",])
```

```
king_man_plot <- king %>%
  mutate(selected = "king") %>%
  bind_rows(man %>%
              mutate(selected = "man")) %>%
  group_by(selected) %>%
  top_n(15, similarity) %>%
  ungroup %>%
  mutate(token = reorder(token, similarity)) %>%

  # plot setup
  ggplot(aes(token, similarity, fill = selected)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~selected, scales = "free") +
  coord_flip() +
  theme(strip.text=element_text(hjust=0, size=12)) +
  scale_y_continuous(expand = c(0,0)) +
  labs(x = NULL,
       title = "What word vectors are most similar to king or man?",
       subtitle = "Top 15 words")

king_man_plot
```



In our word embedding plots, we see `king` is highly similar to other terms of royalty. These are words that have a high similarity to and in some cases, be swapped for that word. The word `man` on the other hand is more similar to other words that relate to people.

Word Math

You can use simple arithmetic to search for synonyms of multiple words "king + man" or synonyms of one word while excluding the other "king - man".

Here we'll try a couple examples

```
king_no_man <- glov_matrix["king",] - glov_matrix["man",]
search_synonyms(glov_matrix, king_no_man)
```

```
## # A tibble: 400,000 x 2
##    token       similarity
##    <chr>            <dbl>
##  1 king              35.3
##  2 kalākaua          26.8
##  3 adulyadej         26.3
##  4 bhumibol          25.9
##  5 ehrenkrantz       25.5
##  6 gyanendra         25.2
##  7 birendra          25.2
##  8 sigismund         25.1
```

```
##  9 letsie            24.7
## 10 mswati            24.0
## # ... with 399,990 more rows
```

```
# love and sad
love <- search_synonyms(glov_matrix, glov_matrix["love",])
sad <- search_synonyms(glov_matrix, glov_matrix["sad",])


love_no_sad <- glov_matrix["love",] - glov_matrix["sad",]
search_synonyms(glov_matrix, love_no_sad)
```

```
## # A tibble: 400,000 x 2
##    token    similarity
##    <chr>         <dbl>
##  1 love           22.9
##  2 passion        15.5
##  3 undying        15.4
##  4 starring       15.0
##  5 her            14.4
##  6 romance        14.0
##  7 sex            14.0
##  8 tora-san       13.6
##  9 vagner         13.5
## 10 marry          13.3
## # ... with 399,990 more rows
```

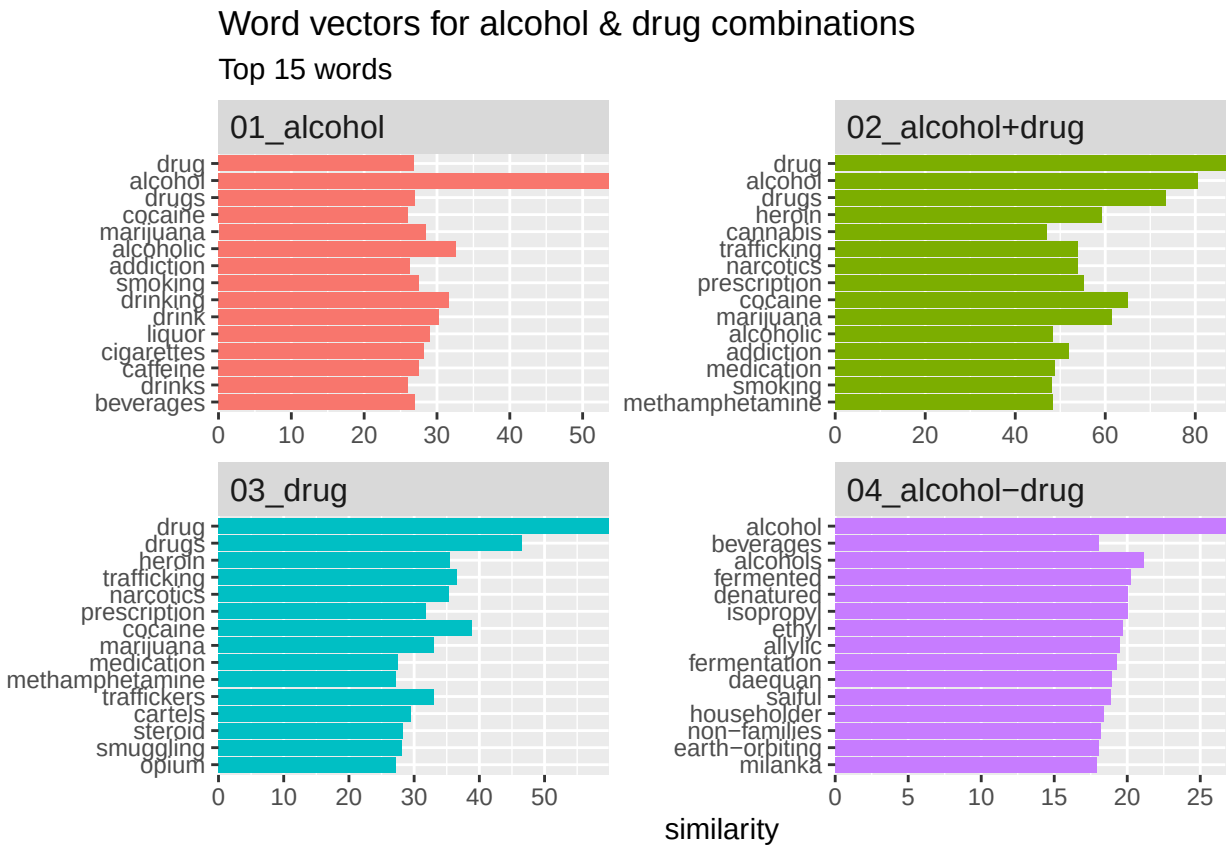
```
love_sad <- glov_matrix["love",] + glov_matrix["sad",]
search_synonyms(glov_matrix, love_sad)
```

```
## # A tibble: 400,000 x 2
##    token   similarity
##    <chr>        <dbl>
##  1 sad           53.9
##  2 love          52.4
##  3 i             41.3
##  4 'm            40.9
##  5 me            39.7
##  6 my            38.8
##  7 feel          37.9
##  8 you           37.5
##  9 tragic        37.2
## 10 feeling       37.0
## # ... with 399,990 more rows
```

```
# alcohol and drug
alcohol <- search_synonyms(glov_matrix, glov_matrix["alcohol",])
drug <- search_synonyms(glov_matrix, glov_matrix["drug",])

alcohol_no_drug <- glov_matrix["alcohol",] - glov_matrix["drug",]
alcohol_no_drug <- search_synonyms(glov_matrix, alcohol_no_drug)
```

```r
alcohol_drug <- glov_matrix["alcohol",] + glov_matrix["drug",]
alcohol_drug <- search_synonyms(glov_matrix, alcohol_drug)
```

```r
alcohol_drug_plot <- alcohol_no_drug %>%
  mutate(selected = "04_alcohol-drug") %>%
  bind_rows(alcohol %>%
              mutate(selected = "01_alcohol"),
            drug %>%
              mutate(selected = "03_drug"),
            alcohol_drug %>%
              mutate(selected = "02_alcohol+drug")) %>%
  group_by(selected) %>%
  top_n(15, similarity) %>%
  ungroup %>%
  mutate(token = reorder(token, similarity)) %>%

  # plot setup
  ggplot(aes(token, similarity, fill = selected)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~selected, scales = "free") +
  coord_flip() +
  theme(strip.text=element_text(hjust=0, size=12)) +
  scale_y_continuous(expand = c(0,0)) +
  labs(x = NULL,
       title = "Word vectors for alcohol & drug combinations",
       subtitle = "Top 15 words")

alcohol_drug_plot
```


Word vectors for alcohol & drug combinations
Top 15 words
01_alcohol
drug
alcohol
drugs
cocaine
marijuana
alcoholic
addiction
smoking
drinking
drink
liquor
cigarettes
caffeine
drinks
beverages
0
10
20
30
40
50
02_alcohol+drug
drug
alcohol
drugs
heroin
cannabis
trafficking
narcotics
prescription
cocaine
marijuana
alcoholic
addiction
medication
smoking
methamphetamine
0
20
40
60
80
03_drug
drug
drugs
heroin
trafficking
narcotics
prescription
cocaine
marijuana
medication
methamphetamine
traffickers
cartels
steroid
smuggling
opium
0
10
20
30
40
50
04_alcohol−drug
alcohol
beverages
alcohols
fermented
denatured
isopropyl
ethyl
allylic
fermentation
daequan
saiful
householder
non−families
earth−orbiting
milanka
0
5
10
15
20
25
similarity