Topic 7: Word Embeddings

Charles Hendrickson

05/17/2022

This week's Rmd file here: https://github.com/MaRo406/EDS_231-text-sentiment/blob/main/topic_7.Rmd

Assignment

Download a set of pretrained vectors, GloVe, and explore them.

Grab data here:

1. Recreate the analyses in the last three chunks (find-synonyms, plot-synonyms, word-math) with the GloVe embeddings. How are they different from the embeddings created from the climbing accident data? Why do you think they are different?
2. Run the classic word math equation, "king" - "man" = ?
3. Think of three new word math equations. They can involve any words you'd like, whatever catches your interest.

```
# read in the glove data
glove_data <- fread(here("../../data/glove.6B/glove.6B.300d.txt"), header = F)

# check if the data frame has row names
has_rownames(glove_data)
```

```
## [1] FALSE
```

```
# make a column into rownames
glove_data <- glove_data %>%
     column_to_rownames(var = 'V1')

# make a matrix
glove_matrix <- data.matrix(glove_data)
```

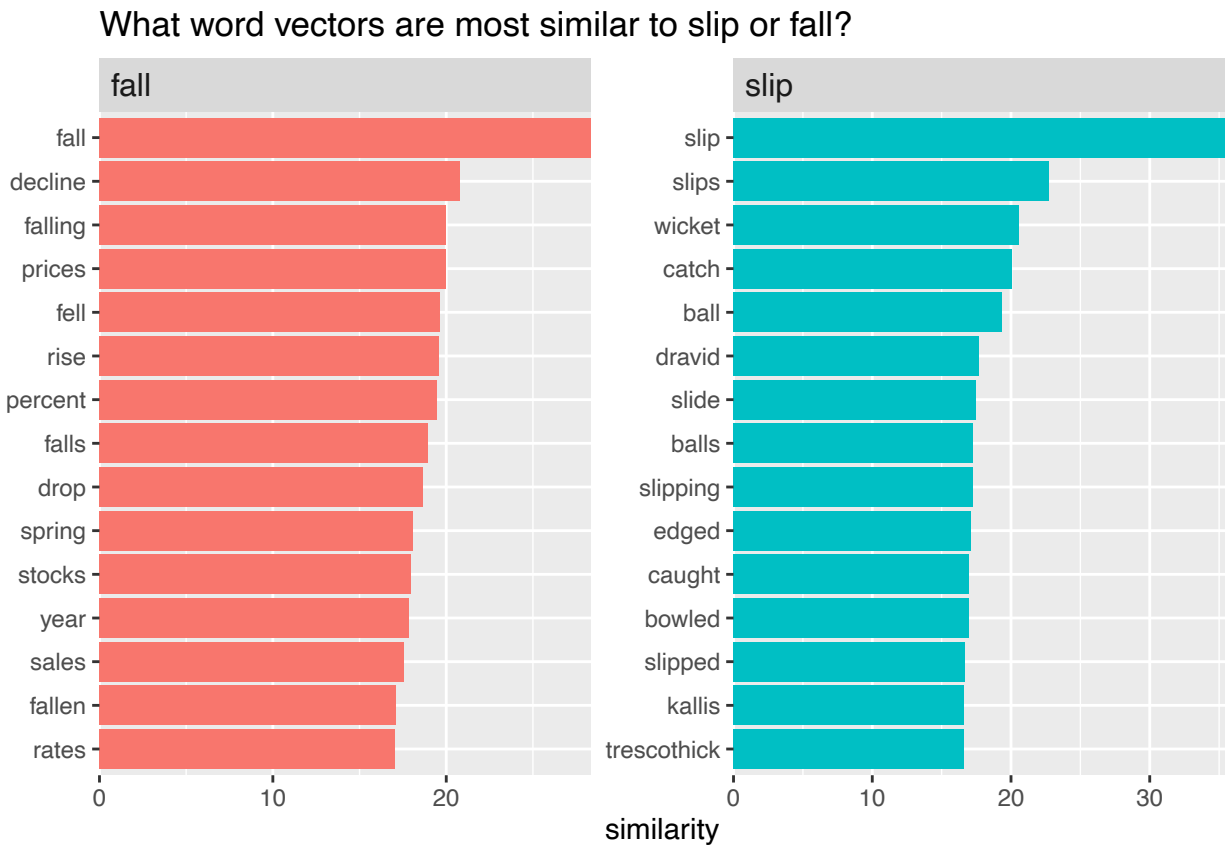
```
# create a function that searches for synonyms and produces similarity score
search_synonyms <- function(word_vectors, selected_vector) {
dat <- word_vectors %*% selected_vector

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

similarities %>%
       arrange(-similarity) %>%
        select(c(2,3))
}
```

```
# use the search synonyms function to get the similarity scores for words like 'fall' and 'slip'.
fall <- search_synonyms(glove_matrix,glove_matrix["fall",])
```

```
slip <- search_synonyms(glove_matrix,glove_matrix["slip",])
```

The similarity scores for the GloVe embeddings are much higher overall compared to the climbing accident embeddings. This difference could be due to the climbing data set containing text that exclusively uses the terms 'fall' and 'slip' to describe climbing situations instead of much generic terms such as 'decline' or 'wicket' meaning an opening like a window especially. These terms are very related to 'fall' and 'slip' when taken out of context, however the climbing data set does not use them. This difference could be due to the climbing data set being much smaller than the glove data set and also climbing.

```
slip %>%
    mutate(selected = "slip") %>%
    bind_rows(fall %>%
                  mutate(selected = "fall")) %>%
    group_by(selected) %>%
    top_n(15, similarity) %>%
    ungroup %>%
    mutate(token = reorder(token, similarity)) %>%
    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 slip or fall?")
```



```r
# word math equation, "king" - "man"
word_math1 <- glove_matrix["king",] - glove_matrix["man",]
search_synonyms(glove_matrix, word_math1)
```

```
## # 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
```

```r
# word math equation, "soldier" + "fighter"
word_math2 <- glove_matrix["soldier",] + glove_matrix["fighter",]
search_synonyms(glove_matrix, word_math2)
```

```
## # A tibble: 400,000 x 2
##    token    similarity
##    <chr>         <dbl>
##  1 fighter        68.5
##  2 soldier        64.7
##  3 soldiers       49.2
##  4 fighters       48.4
##  5 f-16           46.3
##  6 combat         46.2
##  7 wounded        45.7
##  8 army           45.3
##  9 bomber         45.3
## 10 aircraft       43.5
## # ... with 399,990 more rows
```

```r
# word math equation, "flood" + "fill"
word_math3 <- glove_matrix["flood",] + glove_matrix["fill",]
search_synonyms(glove_matrix, word_math3)
```

```
## # A tibble: 400,000 x 2
##    token     similarity
##    <chr>          <dbl>
##  1 flood           52.7
##  2 fill            40.7
##  3 flooding        39.2
##  4 floods          37.8
##  5 water           35.3
##  6 flooded         35.1
##  7 rains           32.9
##  8 dam             32.2
##  9 levees          32.0
## 10 inundated       30.0
```

```
## # ... with 399,990 more rows
```

```
# word math equation, "hunt" + "kill"
word_math4 <- glove_matrix["hunt",] + glove_matrix["kill",]
search_synonyms(glove_matrix, word_math4)
```

```
## # A tibble: 400,000 x 2
##    token      similarity
##    <chr>           <dbl>
##  1 kill             54.6
##  2 hunt             45.4
##  3 killed           38.4
##  4 hunting          36.6
##  5 killer           36.0
##  6 killing          35.4
##  7 militants        34.4
##  8 kills            34.1
##  9 terrorists       32.9
## 10 qaeda            32.6
## # ... with 399,990 more rows
```