Topic 7: Word Embeddings

Mia Forsline

2022-05-17

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

Read in the data

```
glove_data <- fread(here("data", "glove.6B.300d.txt"), header = FALSE)

glove_df <- glove_data %>%
     remove_rownames() %>%
     column_to_rownames(var = 'V1')
```

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?

Create the synonyms function

```
#take a single word from the word_vectors then compares it to the entire matrix, then outputs a similar
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))
}
```

Check similarity scores of words most similar to “fall” and “slip”

```
glove_matrix <- as.matrix(glove_df)

fall <- search_synonyms(glove_matrix, glove_matrix["fall",])
head(fall, n = 10)
```

```
## # A tibble: 10 x 2
##    token   similarity
##    <chr>        <dbl>
##  1 fall          28.4
##  2 decline       20.8
##  3 falling       20.0
##  4 prices        20.0
##  5 fell          19.6
##  6 rise          19.6
##  7 percent       19.5
##  8 falls         19.0
##  9 drop          18.7
## 10 spring        18.1
```

Compared to the in-class demo, the similarity scores are much higher. Since we are using a different dataset, the word tokens are also no longer specifically climbing related. Instead, the words in the GloVe dataset seem much more general and intuitive. For example, "decline" has the second highest similarity score after the word "fall" itself when being compared to the key word of "fall".

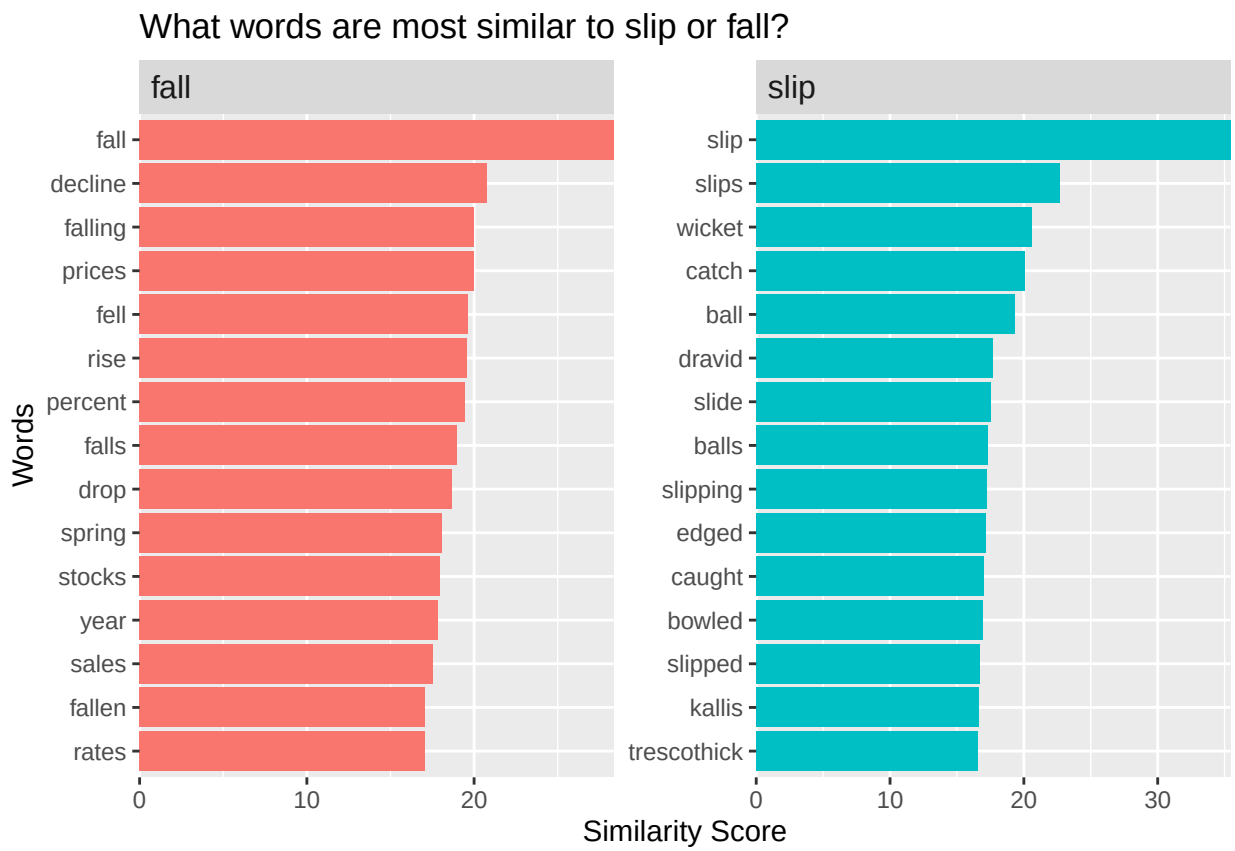
```
slip <- search_synonyms(glove_matrix, glove_matrix["slip",])
head(slip, n = 10)
```

```
## # A tibble: 10 x 2
##    token    similarity
##    <chr>         <dbl>
##  1 slip           35.4
##  2 slips          22.7
##  3 wicket         20.6
##  4 catch          20.1
##  5 ball           19.3
##  6 dravid         17.7
##  7 slide          17.5
##  8 balls          17.3
##  9 slipping       17.2
## 10 edged          17.1
```

For "slip," many of the words seem related to the sport of cricket such as "wicket." Highly scored words also include the surnames of famous cricket players such as "dravid." These differences are likely due to us using a completely different set of words compared to the analysis we performed in class.

```
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 = "Words",
         y = "Similarity Score",
         title = "What words are most similar to slip or fall?")
```



Word math: "snow" and "danger" example

```
snow_danger <- glove_matrix["snow",] + glove_matrix["danger",]
head(search_synonyms(glove_matrix, snow_danger), n= 10)
```

```
## # A tibble: 10 x 2
##    token      similarity
##    <chr>           <dbl>
##  1 snow             57.6
##  2 rain             40.6
##  3 danger           40.5
##  4 snowfall         34.8
##  5 weather          34.4
##  6 winds            34.0
##  7 rains            34.0
##  8 fog              33.6
##  9 landslides       33.3
## 10 threat           33.0
```

```
no_snow_danger <- glove_matrix["danger",] - glove_matrix["snow",]
head(search_synonyms(glove_matrix, no_snow_danger), n = 10)
```

```
## # A tibble: 10 x 2
##    token        similarity
##    <chr>             <dbl>
##  1 danger             23.3
##  2 risks              20.2
##  3 imminent           18.7
##  4 dangers            17.9
##  5 risk               17.8
##  6 32-team            17.6
##  7 mesdaq             17.5
##  8 inflationary       17.4
##  9 risking            17.2
## 10 2001-2011          17.0
```

2. Run the classic word math equation, "king" - "man" = ?

```
king_man <- glove_matrix["king",] - glove_matrix["man",]
head(search_synonyms(glove_matrix, king_man), n = 10)
```

```
## # A tibble: 10 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
```

3. Think of three new word math equations. They can involve any words you'd like, whatever catches you interest.

a) ball - cricket

```
no_ball_cricket <- glove_matrix["ball",] - glove_matrix["cricket",]
head(search_synonyms(glove_matrix, no_ball_cricket), n = 10)
```

```
## # A tibble: 10 x 2
##    token     similarity
```

```
##    <chr>          <dbl>
##  1 ball            33.7
##  2 deflected       29.3
##  3 backhand        28.2
##  4 header          27.8
##  5 footed          27.7
##  6 dribbled        27.5
##  7 crossbar        27.5
##  8 layup           27.4
##  9 3-pointer       27.1
## 10 forehand        26.9
```

b) red + apple

```
red_apple <- glove_matrix["red",] + glove_matrix["apple",]
head(search_synonyms(glove_matrix, red_apple), n = 10)
```

```
## # A tibble: 10 x 2
##    token  similarity
##    <chr>       <dbl>
##  1 apple        60.8
##  2 red          57.4
##  3 yellow       41.4
##  4 blue         41.0
##  5 orange       38.3
##  6 pink         38.1
##  7 green        36.5
##  8 fruit        35.7
##  9 juice        35.2
## 10 ipod         35.0
```

c) dog + cat

```
dog_cat <- glove_matrix["dog",] + glove_matrix["cat",]
head(search_synonyms(glove_matrix, dog_cat), n = 10)
```

```
## # A tibble: 10 x 2
##    token  similarity
##    <chr>       <dbl>
##  1 dog          73.3
##  2 cat          68.8
##  3 dogs         58.6
##  4 pet          51.9
##  5 cats         48.8
##  6 horse        44.8
##  7 puppy        41.7
##  8 animal       41.6
##  9 rabbit       39.1
## 10 hound        38.8
```