Topic 7: Word Embeddings

Mia Forsline

2022-05-17

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

Read in the data

- wrangle to get the dataframe in the proper format to use the synonyms function

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

glove_df <- glove_data %>%
     remove_rownames() %>%
     column_to_rownames(var = 'V1') #make the first column the index
```

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 compare it to the entire matrix
#then output a 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))
}
```

Check similarity scores of words most similar to “fall” and “slip” using the GloVe data

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

fall <- search_synonyms(glove_matrix, glove_matrix["fall",])
head(fall, n = 10) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),
                latex_options = "HOLD_position")
```

token	similarity
fall	28.35289
decline	20.78131
falling	19.97644
prices	19.97596
fell	19.62625
rise	19.58406
percent	19.46760
falls	18.96819
drop	18.66136
spring	18.09208

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".

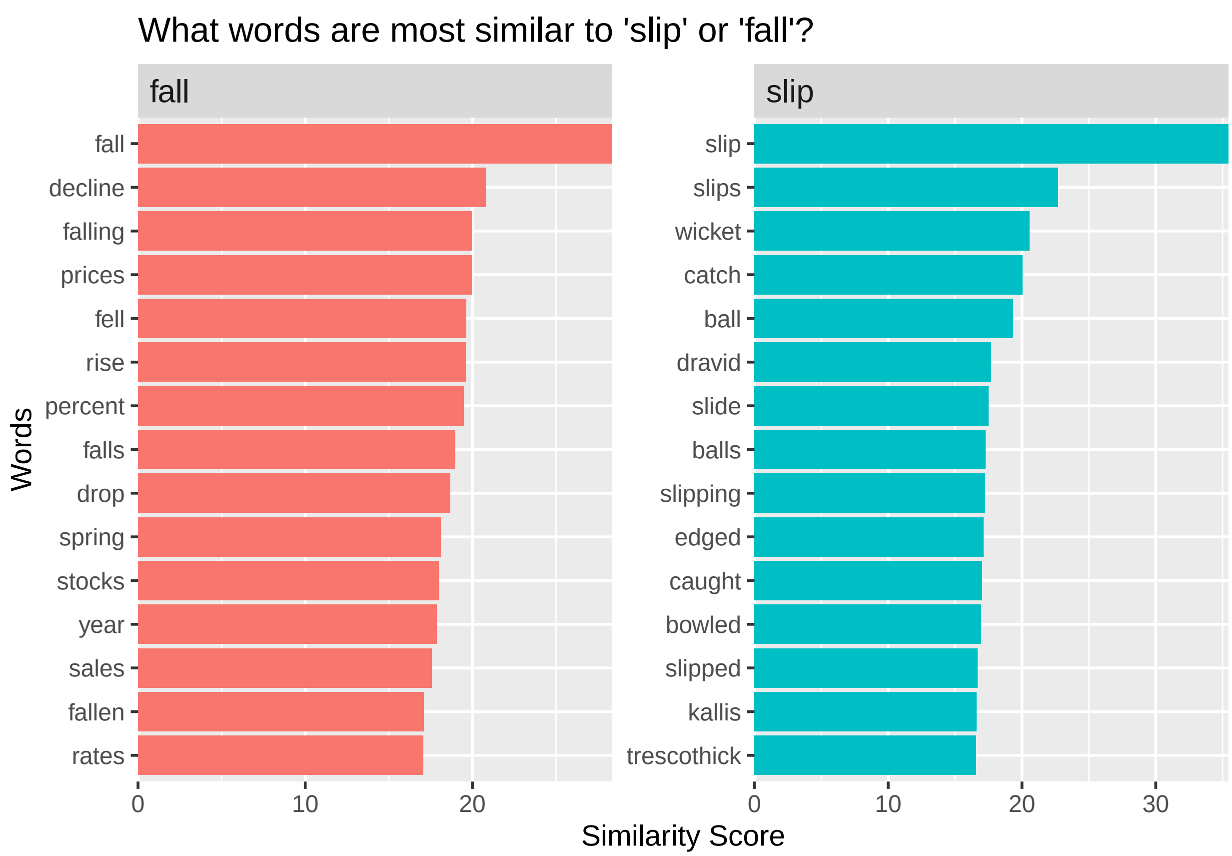
```r
slip <- search_synonyms(glove_matrix, glove_matrix["slip",])
head(slip, n = 10) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),
                latex_options = "HOLD_position")
```

token	similarity
slip	35.43341
slips	22.70521
wicket	20.55729
catch	20.05911
ball	19.33358
dravid	17.70322
slide	17.50436
balls	17.26482
slipping	17.24516
edged	17.14493

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.

```r
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) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),
                latex_options = "HOLD_position")
```

token	similarity
snow	57.58158
rain	40.56130
danger	40.46035
snowfall	34.84752
weather	34.37406
winds	33.96186
rains	33.95089
fog	33.59895
landslides	33.27340
threat	32.97454

```
no_snow_danger <- glove_matrix["danger",] - glove_matrix["snow",]
head(search_synonyms(glove_matrix, no_snow_danger), n = 10) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),              latex_options = "HOLD_positi
```

token	similarity
danger	23.31435
risks	20.22485
imminent	18.67691
dangers	17.89223
risk	17.77783
32-team	17.56241
mesdaq	17.46916
inflationary	17.42012
risking	17.20070
2001-2011	17.02498

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) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),              latex_options = "HOLD_positi
```

token	similarity
king	35.29707
kalākaua	26.82616
adulyadej	26.34680
bhumibol	25.87043
ehrenkrantz	25.45746
gyanendra	25.21709
birendra	25.20759
sigismund	25.05872
letsie	24.68315
mswati	24.00341

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) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),                latex_options = "HOLD_positi
```

token	similarity
ball	33.70580
deflected	29.30528
backhand	28.18227
header	27.80514
footed	27.66999
dribbled	27.46802
crossbar	27.45483
layup	27.42671
3-pointer	27.06143
forehand	26.94229

b) red + apple

```
red_apple <- glove_matrix["red",] + glove_matrix["apple",]
head(search_synonyms(glove_matrix, red_apple), n = 10) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),                latex_options = "HOLD_positi
```

token	similarity
apple	60.81770
red	57.36138
yellow	41.37920
blue	40.99693
orange	38.32256
pink	38.08597
green	36.50215
fruit	35.69151
juice	35.15918
ipod	34.99317

c) dog + cat

```
dog_cat <- glove_matrix["dog",] + glove_matrix["cat",]
head(search_synonyms(glove_matrix, dog_cat), n = 10) %>%
  kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"),                latex_options = "HOLD_positi
```

token	similarity
dog	73.30799
cat	68.80145
dogs	58.57034
pet	51.93694
cats	48.82777
horse	44.78663
puppy	41.73210
animal	41.61641
rabbit	39.05114
hound	38.76443