Topic 7: Word Embeddings

Alex Vand

5/31/2022

1.a Recreate the analyses in the last three chunks (find-synonyms, plot-synonyms, word-math) with the GloVe embeddings.

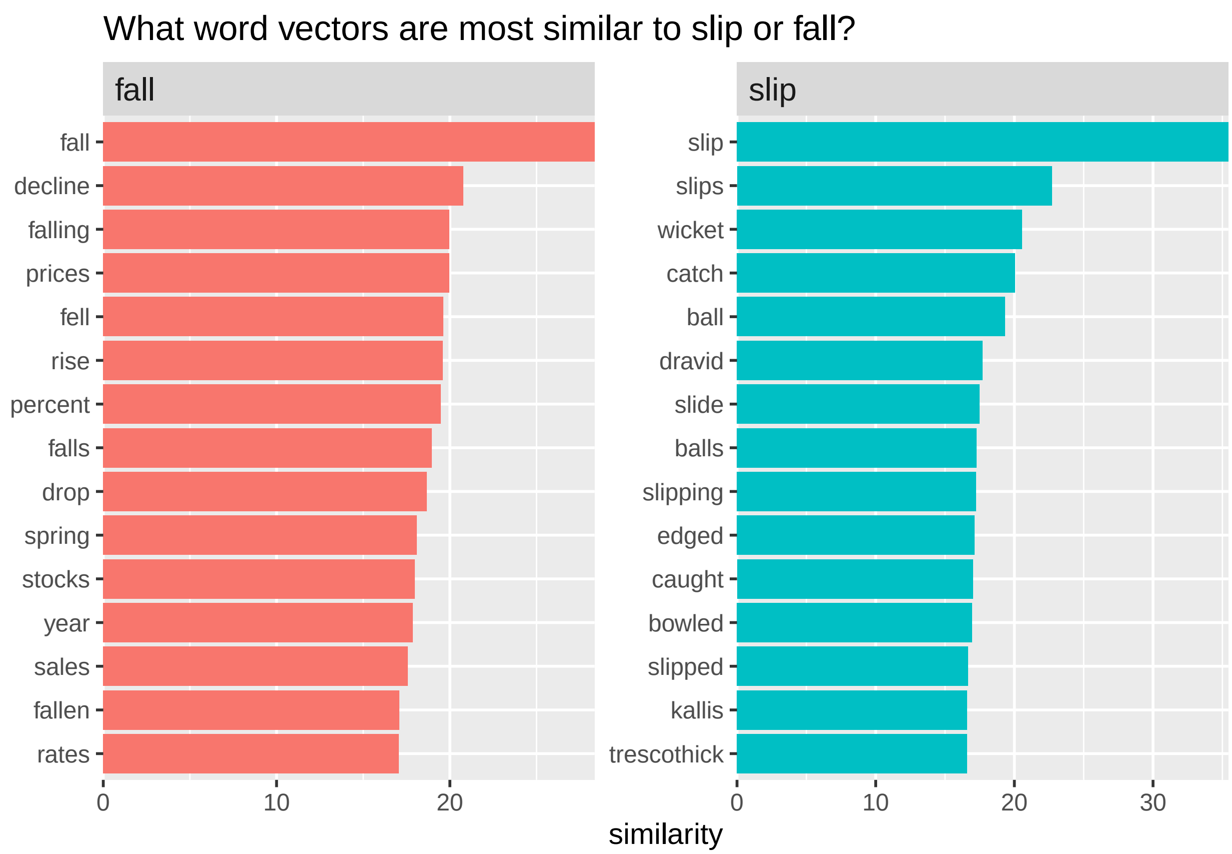
```
word_vectors <- as.matrix(data)
```

```
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))
}
```

```
fall <- search_synonyms(word_vectors,word_vectors["fall",])
slip <- search_synonyms(word_vectors,word_vectors["slip",])
```

```
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?")
```



1.b How are they different from the embeddings created from the climbing accident data? Why do you think they are different?

These are different from the embeddings created from the climbing data given the particular meaning and context of "slip" and "fall" used here. "Fall" has a financial connotation while "slip" seems to be related to sports. Further comparison should include stemming the words of interest (removing fallen, falls, etc.).

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

```
king_man <- word_vectors["king",] - word_vectors["man",]
search_synonyms(word_vectors, king_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
```

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

```r
bionic_vision <- word_vectors["bionic",] - word_vectors["vision",]
search_synonyms(word_vectors, bionic_vision)
```

```
## # A tibble: 400,000 x 2
##    token       similarity
##    <chr>            <dbl>
##  1 bionic            43.3
##  2 u.n.c.l.e.        24.3
##  3 silverbacks       23.6
##  4 forelimbs         23.2
##  5 refrigerate       22.9
##  6 republish         22.4
##  7 5.125             21.6
##  8 gmac              21.6
##  9 47-story          21.4
## 10 paratype          21.1
## # ... with 399,990 more rows
```

```r
beach_volleyball <- word_vectors["beach",] - word_vectors["volleyball",]
search_synonyms(word_vectors, beach_volleyball)
```

```
## # A tibble: 400,000 x 2
##    token     similarity
##    <chr>          <dbl>
##  1 beach          30.2
##  2 palm           27.8
##  3 fla.           26.6
##  4 boulevard      19.9
##  5 beaches        19.2
##  6 eyman          18.9
##  7 calif.         18.6
##  8 gushee         18.5
##  9 stoda          18.4
## 10 bay            18.3
## # ... with 399,990 more rows
```

```r
scuba_dive <- word_vectors["restaurant",] - word_vectors["menu",]
search_synonyms(word_vectors, scuba_dive)
```

```
## # A tibble: 400,000 x 2
##    token       similarity
##    <chr>            <dbl>
```

```
## 1 restaurant        19.6
## 2 hotel             18.4
## 3 apartment         17.8
## 4 nightclub         17.7
## 5 downtown          15.8
## 6 suburb            15.5
## 7 near              14.7
## 8 restaurants       14.4
## 9 motel             14.4
## 10 condominium      14.4
## # ... with 399,990 more rows
```