Word Embeddings

Shale

5/17/2022

Load Data

```
# Use this file: 'glove.6B.300d.txt'
GloVe <- read_table(here('data/glove.6B.300d.txt'), col_names = FALSE)
```

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

```
glove_matrix <- GloVe %>%
    column_to_rownames(var = 'X1') %>%
    as.matrix()
```

Recreate Analyses

```
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(glove_matrix, glove_matrix["fall",])
head(fall, 15)
```

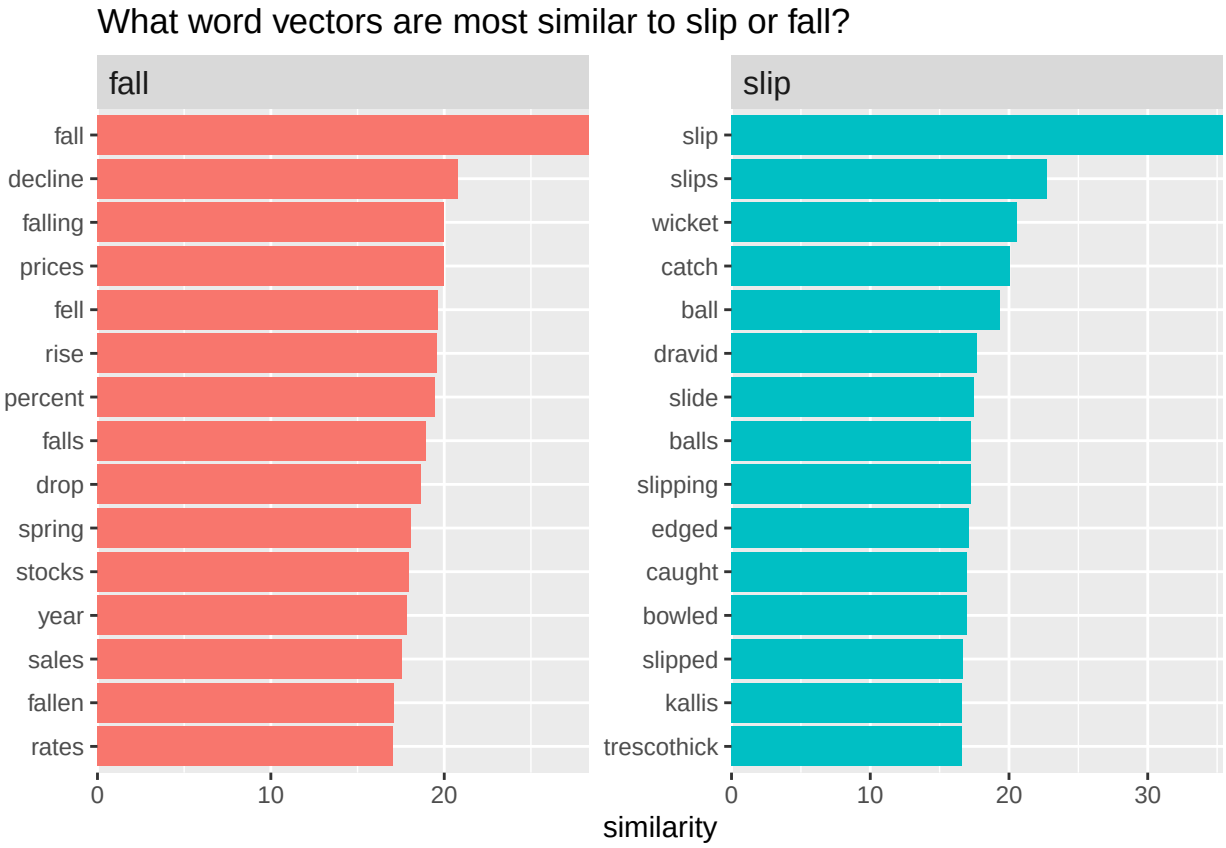
```
## # A tibble: 15 x 2
##    token   similarity
##    <chr>        <dbl>
##  1 fall          28.4
##  2 decline       20.8
##  3 falling       20.0
##  4 prices        20.0
##  5 fell          19.6
##  6 rise          19.6
##  7 percent       19.5
```

```
##  8 falls         19.0
##  9 drop          18.7
## 10 spring        18.1
## 11 stocks        18.0
## 12 year          17.9
## 13 sales         17.6
## 14 fallen        17.1
## 15 rates         17.1
```

```
slip <- search_synonyms(glove_matrix, glove_matrix["slip",])
head(slip, 15)
```

```
## # A tibble: 15 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
## 11 caught            17.0
## 12 bowled            17.0
## 13 slipped           16.7
## 14 kallis            16.6
## 15 trescothick       16.6
```

```
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?")
```



Compared to the word embeddings for `slip` and `fall` in the climbing accident data, the word embeddings from `GloVe` cover a broader range of linguistic context. This is because the texts that the embeddings are based on are more general, while the climbing accident data was specific to the context of incident reports.

```
snow_danger <- glove_matrix["snow",] + glove_matrix["danger",]
head(search_synonyms(glove_matrix, snow_danger), 15)
```

```
## # A tibble: 15 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
## 11 ice                32.8
## 12 avalanches         32.7
## 13 flooding           32.6
## 14 temperatures       32.5
## 15 mountain           32.3
```

```
no_snow_danger <- glove_matrix["danger",] - glove_matrix["snow",]
head(search_synonyms(glove_matrix, no_snow_danger), 15)
```

```
## # A tibble: 15 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
## 11 threat                17.0
## 12 extinction            16.7
## 13 incertae              16.7
## 14 peril                 16.6
## 15 dothideomycetes       16.6
```

Similar to the point mentioned above, the lists for danger with and without snow are different in `GloVe` than in the climbing accident data because the texts cover a broader range of topics. Here, danger terms associated with snow are (broadly speaking) weather-related terms, while danger terms not associated with snow are much more ecclectic.

'King' - 'Man'

```
K_M <- glove_matrix["king",] - glove_matrix["man",]
k = head(search_synonyms(glove_matrix, K_M), 15)
k
```

```
## # A tibble: 15 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
## 11 soopers           22.9
## 12 władysław         22.9
## 13 tuanku            22.8
## 14 prussia           22.7
## 15 norodom           22.6
```

```
# Or, a single answer
k[2,1]
```

```
## # A tibble: 1 x 1
##   token
##   <chr>
```

```
## 1 kalākaua
```

Exploration of Word Math

Nature - Man

```
M_W <- glove_matrix["nature",] - glove_matrix["man",]
head(search_synonyms(glove_matrix, M_W), 15)
```

```
## # A tibble: 15 x 2
##    token           similarity
##    <chr>                <dbl>
##  1 icasualties.org       23.8
##  2 nature                21.2
##  3 aonb                  19.2
##  4 computerologist       19.0
##  5 geoscience            18.7
##  6 habitats              18.4
##  7 forex.com             17.8
##  8 ecosystems            17.6
##  9 crites                17.6
## 10 iucn                  17.4
## 11 ecology               17.3
## 12 sssi                  17.3
## 13 biodiversity          16.9
## 14 xil                   16.9
## 15 ecological            16.8
```

Interesting that 'nature' has a higher similarity to icasualties.org than to itself?

Life - Love

```
L_P <- glove_matrix["life",] - glove_matrix["love",]
head(search_synonyms(glove_matrix, L_P), 15)
```

```
## # A tibble: 15 x 2
##    token               similarity
##    <chr>                    <dbl>
##  1 life                      16.0
##  2 postbellum                15.8
##  3 disability-adjusted       15.3
##  4 expectancies              14.4
##  5 imprisonment              14.4
##  6 2001-2011                 14.1
##  7 commuted                  13.9
##  8 post-football             13.6
##  9 gerst                     13.2
## 10 prison                    12.9
## 11 cambrian                  12.7
## 12 reinsurance               12.4
## 13 14,000-member             12.2
## 14 preservers                12.2
## 15 expectancy                12.2
```

Life - Pain

```
MW <- glove_matrix["life",] - glove_matrix["pain",]
head(search_synonyms(glove_matrix, MW), 15)
```

```
## # A tibble: 15 x 2
##    token            similarity
##    <chr>                 <dbl>
##  1 life                   23.4
##  2 lives                  15.2
##  3 marine                 15.1
##  4 celibate               14.9
##  5 civilization           14.9
##  6 idyllic                14.7
##  7 preservers             14.6
##  8 extraterrestrial       14.5
##  9 tycoon                 14.2
## 10 living                 13.9
## 11 post-football          13.7
## 12 expectancies           13.7
## 13 fictional              13.4
## 14 biography              13.4
## 15 herediano              13.2
```