Topic 07 - Word Embeddings

Julia Parish

2022-05-11

Word Embedding

This text sentiment analysis was completed as an assignment for the course, Environmental Data Science 231: Text and Sentiment Analysis for Environmental Problems. The data was sourced from: Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. The dataset used is Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 300d vectors, 822 MB download). For more details on the data and unsupervised learning algorithm, navigate here

Original assignment instructions can be found here

Load Libraries

```
if (!require(librarian)){
  install.packages("librarian")
  library(librarian)
}

librarian::shelf(broom,
                 here,
                 irlba,
                 kableExtra,
                 textdata,
                 tidytext,
                 tidyverse,
                 widyr)
```

Load in the Data

```
glovevecs <- read_table(here('assignments/HW07_WordEmbeddings/data/glove.6B.300d.txt'), col_names = FALS
  column_to_rownames(., var = "X1")
```

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?

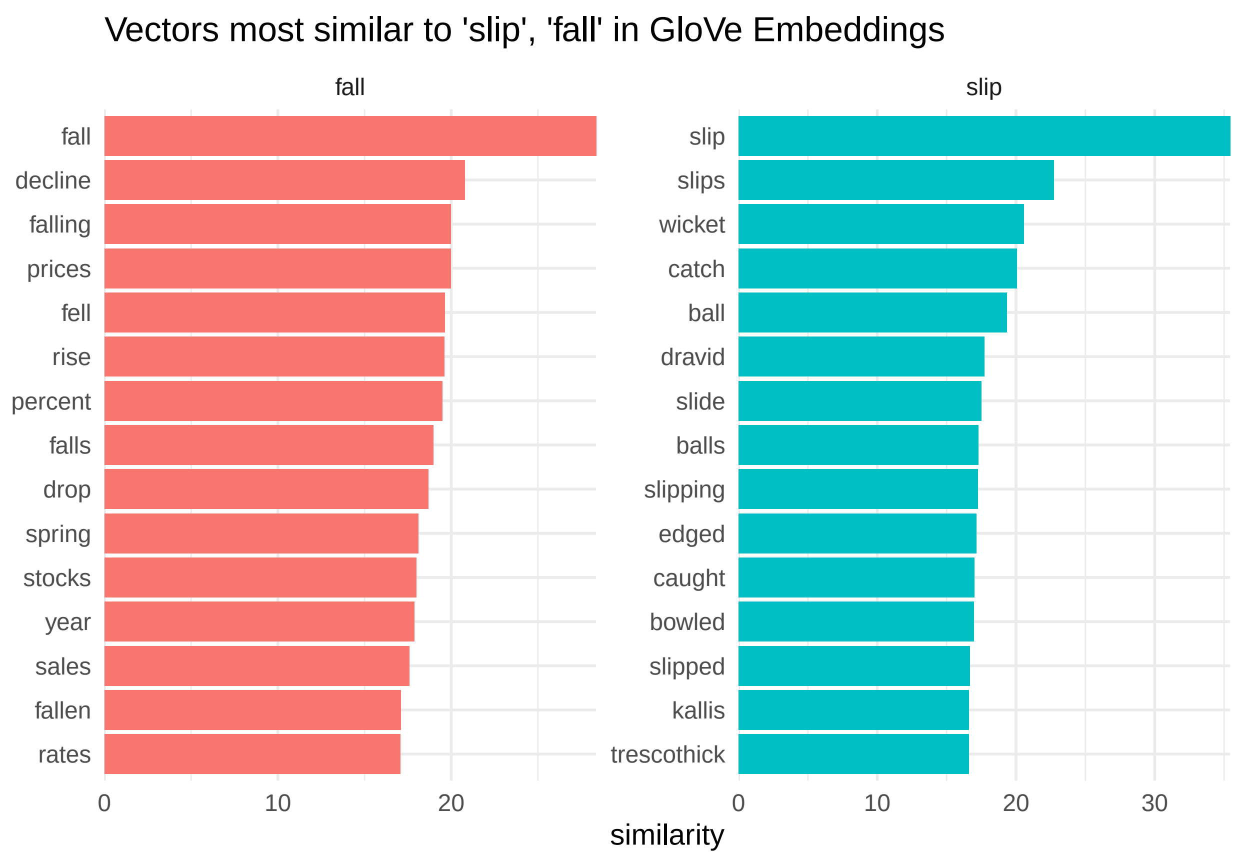
```
# function to create similarity score

search_synonyms <- function(glovevecs, selected_vector) {
```

```r
dat <- glovevecs %*% selected_vector

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

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

```r
# convert dataframe to a matrix
glove_matrix <- data.matrix(glovevecs)
```

```r
# use function: give  all word vectors (model) and the word " " to calculate similarities
fall2 <- search_synonyms(glove_matrix,glove_matrix["fall",])
slip2 <- search_synonyms(glove_matrix,glove_matrix["slip",])
```

```r
glove_plot <- slip2 %>%
  mutate(selected = "slip") %>%
  bind_rows(fall2 %>%
              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 = "Vectors most similar to 'slip', 'fall' in GloVe Embeddings") +
  theme_minimal()

glove_plot
```



In the climbing data, the top 10 most similar words to `fall` are: fall, rock, ice, accident, foot, avalanche, climber, injuries, ground, rope. The top 10 most similar words to `slip` are: fall, rope, line, short, lead, coley, gentzel, meter, operation, dome.

In the GloVe data, the top 10 most similar words to `fall` are: fall, decline, falling, prices, fell, rise, percent, falls, drop, spring. The top 10 most similar words to `slip` are: slip, slips, wicket, catch, ball, dravid, slide, balls, slipping, edged.

The GloVe data contains general words related to `fall` and `slip` across various categories like economics, sports, non-rock climbing accidents. The GloVe data is sourced from Wikipedia, so it makes sense that the climbing data tokens are more precisely aligned with rock climbing than the GloVe data.

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

```
k_minus_m <- glove_matrix["king",] - glove_matrix["man",]

km_df <- as.data.frame(search_synonyms(glove_matrix, k_minus_m))

head(km_df, n = 20) %>%
  knitr::kable(caption = "Top 20 Tokens Most Similar to King - Man") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

I am thrilled to see that the second most similar token is King Kalakaua, and another Hawaiian King is in the top 20!

Table 1: Top 20 Tokens Most Similar to King - Man

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
soopers	22.86619
władysław	22.85730
tuanku	22.79580
prussia	22.70036
norodom	22.59436
throne	22.54447
æthelred	22.44941
kamehameha	22.33307
jagiellon	22.31369
ahom	22.29553

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

```
invspe <- glove_matrix["invasive",] + glove_matrix["species",]

invspe_df <- as.data.frame(search_synonyms(glove_matrix, invspe))

head(invspe_df, n = 20) %>%
  knitr::kable(caption = "Top 20 Tokens Most Similar to Invasive + Species") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

```
cc <- glove_matrix["climate",] + glove_matrix["change",]

cc_df <- as.data.frame(search_synonyms(glove_matrix, cc))

head(cc_df, n = 20) %>%
  knitr::kable(caption = "Top 20 Tokens Most Similar to Climate + Change ") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

```
alien <- glove_matrix["illegal",] - glove_matrix["alien",]

alien_df <- as.data.frame(search_synonyms(glove_matrix, alien))

head(alien_df, n = 20) %>%
  knitr::kable(caption = "Top 20 Tokens Most Similar to Illegal - Alien") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

Table 2: Top 20 Tokens Most Similar to Invasive + Species

token	similarity
species	96.37069
invasive	84.90362
genus	76.14424
endemic	58.50159
subspecies	58.43185
endangered	58.07282
mammals	57.63567
insects	56.81169
habitat	56.45319
genera	56.20591
birds	54.86900
habitats	54.70227
larvae	52.32993
iucn	51.80757
organisms	51.63745
plants	51.00643
aquatic	49.28001
non-native	48.89803
fauna	48.48481
vegetation	48.46436

Table 3: Top 20 Tokens Most Similar to Climate + Change

token	similarity
climate	81.31213
change	57.26723
warming	56.60032
global	50.45895
emissions	46.92509
environment	46.46936
changes	46.06386
greenhouse	45.06848
environmental	43.70108
economic	43.36176
weather	43.28188
climatic	43.19583
policy	41.37874
changing	41.18879
temperature	40.18571
köppen	40.01835
temperatures	39.86019
conditions	39.29165
pollution	39.03022
biodiversity	38.78233

Table 4: Top 20 Tokens Most Similar to Illegal - Alien

token	similarity
illegal	35.35655
illegally	23.07498
illicit	22.57470
crackdown	22.50884
smuggling	21.98535
trafficking	21.02692
cocaine	19.78626
heroin	19.33944
prostitution	19.33404
ban	19.00298
banned	18.99601
laundering	18.95541
hashish	18.89176
kickbacks	18.64754
traffickers	18.58311
bribes	18.48122
arrests	18.46146
worldsources	18.32251
immigrants	18.27058
drugs	18.25602