Week 8 Tutorial: Beyond Bag-of-Words

POP77032 Quantitative Text Analysis for Social Scientists

Exercise 1: Keyword in Context (KWIC)

- In the lecture we calculated a sparse vector representation of the word ‘taoiseach’ from the speeches in the 33rd Dáil.
- That vector was of length $V - N_{w=0}$, where V is the vocabulary size and $N_{w=0}$ is the number of words in the vocabulary that never occurred in the context of our word of interest.
- In this exercise calculate the sparse vector representation of the words ‘taoiseach’, ‘tánaiste’ and ‘minister’ each of length V.
- Try writing a function that does this for any given word and context window.
- Try using different context windows (e.g. 2 and 5).
- Calculate the cosine similarity between each pair of words.

```r
dail_33 <- read.csv("../data/dail_33_small.csv.gz")
```

```r
dail_33_toks <- dail_33$text |>
  quanteda::tokens(remove_punct = TRUE) |>
  quanteda::tokens_tolower() |>
  quanteda::tokens_remove(quanteda::stopwords("en"))
```

```r
V <- dail_33_toks |>
  quanteda::dfm() |>
  quanteda::featnames()
```

```r
length(V)
```

```
[1] 157414
```

```r
head(V)
```

```
[1] "seanad"   "éireann"  "accepted" "finance"  "bill"     "2024"
```

Exercise 2: Word Embeddings

- In this exercise we will take another look at dense vectors as word embeddings.
- Let's start by inspecting a pre-trained GloVe model.
- Load in the file `glove.6B.50d.txt` available on Blackboard or here.
- This is a GloVe model with 400K vocabulary in 50 dimensions trained on 2014 dump of English-language Wikipedia.
- One of the common features of word embeddings is the ability to do substantively meaningful vector arithmetic.
- Calculate the sum vectors for words 'candidate' and 'elected'.
- The calculate the cosine similarity between the resultant vector and all other words in the vocabulary.
- What are the most similar word?
- Fit a GloVe model to the Dáil 33 dataset using the `text2vec` package.
- What are the most similar words using this model?

```r
library("text2vec")
```

```r
glove_6b_50d <- as.matrix(read.csv(
  "../data/glove.6B.50d.txt",
  header = FALSE,
  sep = " ",
  quote = "",
  row.names = 1
))
```

```r
head(glove_6b_50d, 5)
```

```
         V2        V3       V4       V5      V6        V7       V8       V9
the 0.418000  0.249680 -0.41242  0.12170 0.34527 -0.044457 -0.49688 -0.17862
,   0.013441  0.236820 -0.16899  0.40951 0.63812  0.477090 -0.42852 -0.55641
.   0.151640  0.301770 -0.16763  0.17684 0.31719  0.339730 -0.43478 -0.31086
of  0.708530  0.570880 -0.47160  0.18048 0.54449  0.726030  0.18157 -0.52393
to  0.680470 -0.039263  0.30186 -0.17792 0.42962  0.032246 -0.41376  0.13228
           V10       V11      V12       V13      V14      V15        V16
the -0.00066023 -0.656600 0.278430 -0.147670 -0.55677  0.14658 -0.0095095
,   -0.36400000 -0.239380 0.130010 -0.063734 -0.39575 -0.48162  0.2329100
.   -0.44999000 -0.294860 0.166080  0.119630 -0.41328 -0.42353  0.5986800
of   0.10381000 -0.175660 0.078852 -0.362160 -0.11829 -0.83336  0.1191700
to  -0.29847000 -0.085253 0.171180  0.224190 -0.10046 -0.43653  0.3341800
          V17       V18       V19      V20       V21       V22       V23
the  0.011658  0.102040 -0.127920 -0.84430 -0.121810 -0.016801 -0.332790
,    0.090201 -0.133240  0.078639 -0.41634 -0.154280  0.100680  0.488910
.    0.288250 -0.115470 -0.041848 -0.67989 -0.250630  0.184720  0.086876
of  -0.166050  0.061555 -0.012719 -0.56623  0.013616  0.228510 -0.143960
to   0.678460  0.057204 -0.344480 -0.42785 -0.432750  0.559630  0.100320
         V24       V25       V26     V27      V28       V29       V30      V31
the -0.155200 -0.231310 -0.191810 -1.8823 -0.76746  0.099051 -0.421250 -0.19526
```

```
1 glove_6b_50d["taoiseach", ]
```

```
       V2         V3         V4         V5         V6         V7         V8
-0.9553300  0.0503800 -0.4326100 -0.1338000 -0.2815100  0.6699200 -0.0664980
       V9        V10        V11        V12        V13        V14        V15
 1.5273000 -2.7075000 -0.3199000  0.3120700  1.5754000 -0.0195480  0.9067300
      V16        V17        V18        V19        V20        V21        V22
 0.1745300  0.2038600  0.8665900 -0.3187000  0.8908700 -0.2384000  1.2805000
      V23        V24        V25        V26        V27        V28        V29
-0.6975700  0.0966820 -0.0507870  0.0019422  0.4988300 -0.5806300 -0.1049400
      V30        V31        V32        V33        V34        V35        V36
-1.1567000  0.8314200 -0.6842000 -0.4221800 -0.2066200  0.2958500  0.6492400
      V37        V38        V39        V40        V41        V42        V43
 0.0777630  0.1621100 -0.1688600  1.3416000  0.3502900 -0.5979900  1.5899000
      V44        V45        V46        V47        V48        V49        V50
-1.3567000 -1.6923000 -1.8363000 -0.0849650 -1.0890000  0.4024600  1.0683000
      V51
-0.1342900
```

```
1 glove_6b_50d["tánaiste", ]
```

```
       V2         V3         V4         V5         V6         V7         V8
-1.2960000 -0.0094144 -1.0139000  0.4761000 -0.4040000  1.0021000  0.0156320
       V9        V10        V11        V12        V13        V14        V15
 0.7018100 -0.8355800 -1.6115000  0.1048400  1.2923000  0.5174500 -0.2383000
      V16        V17        V18        V19        V20        V21        V22
 0.1962900 -0.2857200 -0.0734380  0.5677900  1.9701000 -0.0132660  0.1257100
      V23        V24        V25        V26        V27        V28        V29
-0.1101900 -0.2120200 -0.2315300  0.7386500  1.4157000  1.2749000 -0.2974800
      V30        V31        V32        V33        V34        V35        V36
-0.2968300  1.3279000 -0.5765700 -0.1477100  0.4168400  0.5120100 -0.1891300
      V37        V38        V39        V40        V41        V42        V43
 1.2648000  0.1510800 -0.1816800 -0.5313000 -0.8476400 -0.3042800 -0.1208400
      V44        V45        V46        V47        V48        V49        V50
-1.0131000 -0.9637900 -0.0609840 -0.4185800 -0.9815400 -0.9460000  0.4664800
```

```
      V51
0.2365500
```

```
1 glove_6b_50d["minister", ]
```

```
      V2        V3        V4        V5        V6        V7        V8        V9
0.039931  0.507250 -0.186250  0.322680  0.654270  0.202860 -0.496600  0.133830
     V10       V11       V12       V13       V14       V15       V16       V17
-1.711200 -1.094300  0.109880  0.747130 -0.327610  0.663330  0.606020  0.362980
     V18       V19       V20       V21       V22       V23       V24       V25
0.547330 -0.800940  1.263300  0.818980  0.497940  0.580070 -0.243260 -0.726330
     V26       V27       V28       V29       V30       V31       V32       V33
0.829710 -1.874400  1.255400 -0.242060 -1.128300  1.833700  2.896700  0.067548
     V34       V35       V36       V37       V38       V39       V40       V41
-0.972450  0.143730 -0.180870  0.586340  0.527650  0.463890 -0.031852  0.909330
     V42       V43       V44       V45       V46       V47       V48       V49
-0.397490  1.665400 -0.288790 -1.961500  0.893560  0.158350 -2.187700  0.931920
     V50       V51
1.906600 -0.158900
```