Week 9 Tutorial: Embeddings

POP77032 Quantitative Text Analysis for Social Scientists

Tom Paskhalis

Exercise 1: Validating Embeddings

- In this exercise we will try validating embeddings using Rodriguez & Spirling (2022) approach.
- Download and check out the `RodriguezSpirling2022_CR.rds` dataset. You can find a description of it in the Assignment 3.
- Let's try to follow the steps that we would ask the human evaluators to do in this Turing-style setup.
- In groups of 2-3 decide on a few (2-3) prompt words (you can re-use the ones from the paper, but you don't have to). You can choose to draw them at random or just pick some political/social concepts that you are interested in.
- Think about 10 candidate words that are closest in meaning to each of the prompt words.
- Now fit or load a pre-trained word embedding model to find the 10 closest words to each of the prompt words using the cosine similarity.
- Do any of the candidate words match the closest words from the embedding model?
- Create a spreadsheet with a prompt word column and 2 columns representing the human and embedding model closest words.
- Exchange these spreadsheets across groups and either individually or in groups pick the best fit for each word pair.
- Calculate relative performance of the embedding model compared to human.

```r
cr <- readRDS("../data/RodriguezSpirling2022_CR.rds")
```

```r
dim(cr)
```

```
[1] 1411740      12
```

```r
table(cr$session_id)
```

```
   102    103    104    105    106    107    108    109    110    111
164559 162663 195770 140048 141402 116413 125660 119500 133178 112547
```

```r
head(cr)
```

```
    speech_id
9  1020000009
13 1020000013
23 1020000023
28 1020000028
29 1020000029
30 1020000030

speech
9
respectively advanced to the desk of the vice president the oath prescribed by law was administered to them by
the vice president and they severally subscribed to ahe oath in the official oath book
13
respectively advanced to the desk of the vice president the oath prescribed by law was administered to them by
the vice president and they severally subscribed to the oath in the official oath book
23
respectively advanced to the desk of the vice president the oath prescribed by law was administered to them by
the vice president and they severally subscribed to the oath in the official oath book
28 mr president i will momentarily suggest the absence of a quorum so that the roll will be called and a quorum
established for the purpose of beginning the proceedings of this senate but i believe it appropriate to note at
```

Here is an example with the synthetic dataset for prompt words: “soccer”, “computer” and “potato”.

```r
set.seed(123)
n <- 100

cues <- c("soccer", "computer", "potato")

models <- c("6_300", "human")

# Top 10 nearest neighbours per cue per model
# Overlap: soccer = 7/10, computer = 5/10, potato = 6/10
nn_list <- list(
  "6_300" = list(
    soccer   = c("goal", "pitch", "penalty", "player", "match", "kick", "score",
                 "midfielder", "dribble", "goalkeeper"),
    computer = c("keyboard", "software", "monitor", "program", "mouse",
                 "algorithm", "compiler", "cursor", "desktop", "processor"),
    potato   = c("fries", "mashed", "chips", "roasted", "vegetable", "boiled",
                 "starch", "tuber", "harvest", "baked")
  ),
  "human" = list(
    soccer   = c("goal", "pitch", "penalty", "player", "match", "kick", "score",
                 "football", "team", "coach"),
    computer = c("keyboard", "software", "monitor", "program", "mouse",
                 "laptop", "internet", "data", "email", "screen"),
    potato   = c("fries", "mashed", "chips", "roasted", "vegetable", "boiled",
                 "food", "carrot", "soup", "garden")
```

```
      cue left.source right.source left.word right.word left.choice
1  potato       6_300        human   roasted     boiled       FALSE
2  potato       human        6_300     chips      baked       FALSE
3  potato       human        6_300      soup  vegetable       FALSE
4 computer       6_300        human  keyboard    program        TRUE
```

```
5   potato       6_300        human    mashed     mashed       FALSE
6 computer       6_300        human   program   keyboard       FALSE
  right.choice
1         TRUE
2         TRUE
3         TRUE
4        FALSE
5         TRUE
6         TRUE
```

As this dataset was artificially generated, it's no surprise that the performance of two models is indistinguishable.

Exercise 2: Principal Component Analysis

- In this exercise we will try to use Principal Component Analysis (PCA) to find the main dimensions of variation in the word embedding space.
- We will use the same `RodriguezSpirling2022_CR.rds` dataset
- First, we will need to fit or load a pre-trained word embedding model to find the vector representations of the words in the dataset.
- Then we will apply PCA to these vector representations to find the main dimensions of variation.
- What is the variance explained by the first few principal components? Do they provide a good summary of the multi-dimensional structure?
- Try plotting a random subset of words along the first two principal components.