Week 6 Tutorial: Unsupervised Modelling

POP77032 Quantitative Text Analysis for Social Scientists

Similarity

- It is sometimes desirable to compare two texts to see how similar they are.
- How do we do that?
- Recall the representation of a single document from the DTM is a vector
- As long as we can compare two vectors, we can compare two texts.
- Some desirable properties of a similarity measure:
  - Maximum similarity when the document is compared to itself.
  - Minimum similarity when no words in common (**orthogonal** documents)
  - Symmetry: the similarity between document A and B is the same as between B and A.

Inner Product

- One straight-forward way to compare two vectors is to calculate the **inner product**.
- For two documents $\mathbf{W}_A$ and $\mathbf{W}_B$ with J features, the inner product is:

$$\mathbf{W}_A \cdot \mathbf{W}_B = \sum_{j=1}^{J} W_{A,j} \times W_{B,j}$$

- E.g. for two documents with 3 features:

$$\begin{aligned} \mathbf{W}_A &= [1, 0, 1] \\ \mathbf{W}_B &= [0, 1, 2] \\ \mathbf{W}_A \cdot \mathbf{W}_B &= 1 \times 0 + 0 \times 1 + 1 \times 2 = 2 \end{aligned}$$

- The major downside: inner product is sensitive to vector length.

Cosine Similarity

- To alleviate this problem we can *normalise* each vector before calculating the inner product.
- We will normalise it by vector magnitude.
- Where vector magnitude is the square root of the sum of squares of all elements (alternatively, Euclidean distance between the vector and itself):

$$\|\mathbf{W}\| = \sqrt{\sum_{j=1}^{J} W_j^2}$$

- A document $\mathbf{W}$ can then be normalised by dividing it by its magnitude:

$$\frac{\mathbf{W}_A}{\|\mathbf{W}_A\|} = \left[\frac{W_{A,1}}{\|\mathbf{W}_A\|}, \frac{W_{A,2}}{\|\mathbf{W}_A\|}, \cdots, \frac{W_{A,J}}{\|\mathbf{W}_A\|}\right]$$

Cosine Similarity

- The **cosine similarity** between two documents $\mathbf{W}_A$ and $\mathbf{W}_B$ is then:

$$\cos(\mathbf{W}_A, \mathbf{W}_B) = \frac{\mathbf{W}_A \cdot \mathbf{W}_B}{\|\mathbf{W}_A\| \times \|\mathbf{W}_B\|}$$

- It is called “cosine similarity” because it is based on the size of the angle between the two vectors.

Example: Cosine Similarity

- For the two vectors $\mathbf{W}_A = [1, 0, 1]$ and $\mathbf{W}_B = [0, 1, 2]$:

$$\|\mathbf{W}_A\| = \sqrt{1^2 + 0^2 + 1^2} = \sqrt{2}$$
$$\|\mathbf{W}_B\| = \sqrt{0^2 + 1^2 + 2^2} = \sqrt{5}$$
$$\cos(\mathbf{W}_A, \mathbf{W}_B) = \frac{1 \times 0 + 0 \times 1 + 1 \times 2}{\sqrt{2} \times \sqrt{5}} = \frac{2}{\sqrt{10}} \approx 0.63$$

- In R we can easily implement this ourselves:

```r
W_A <- c(1, 0, 1)
W_B <- c(0, 1, 2)

cosine_similarity <- sum(W_A * W_B) / (sqrt(sum(W_A^2)) * sqrt(sum(W_B^2)))
cosine_similarity
```

```
[1] 0.6324555
```

Exercise 1: Modelling Text

- Let's re-visit the Federalist papers that we looked at in the lecture.
- Now we will consider the full dataset as opposed to a tiny subset.
- The essays are available as part in the `federalist.csv` file on Blackboard.
- Calculate μ for each author that includes all words that appear in the corpus.
- Use cosine similarity to compare the $\hat{\mu}$ of each author with the μ of disputed essays (those that have `NA` in the `author` column).

```r
federalist_papers <- readr::read_csv(
  "../data/federalist.csv"
)
```

```r
head(federalist_papers)
```

```
# A tibble: 6 × 4
  paper_number paper_numeric author   text
  <chr>                <dbl> <chr>    <chr>
1 No. 1                    1 hamilton "AFTER an unequivocal experience of the i…
2 No. 2                    2 jay      "WHEN the people of America reflect that …
3 No. 3                    3 jay      "IT IS not a new observation that the peo…
4 No. 4                    4 jay      "MY LAST paper assigned several reasons w…
5 No. 5                    5 jay      "QUEEN ANNE, in her letter of the 1st Jul…
6 No. 6                    6 hamilton "THE three last numbers of this paper hav…
```

```r
table(federalist_papers$author, useNA = "ifany")
```

```
hamilton      jay  madison     <NA>
      51        5       18       11
```

Exercise 2: Topic Modelling

- In this exercise we will re-visit the Irish party manifestos.
- Load the file `ireland_ge_2020-24_manifestos.csv` containing the manifestos for 2020 and 2024 General elections.
- Fit an LDA model with 3, 5 and 10 topics.
- Compare the log-likelihood, top terms and overall topics across manifestos.
- Try labelling some topics based on the top terms.

```r
manifestos <- readr::read_csv(
  "../data/ireland_ge_2020-24_manifestos.csv"
)
```

```r
str(manifestos)
```

```
spc_tbl_ [17 × 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ party: chr [1:17] "AO" "FF" "FG" "GR" ...
 $ year : num [1:17] 2024 2024 2024 2024 2024 ...
 $ text : chr [1:17] "Our\nCommon Sense\n  Manifesto 2024\n\n                Opening statement\nIn the last year
Aontú has come of ag"| __truncated__ "MOVING FORWARD. TOGETHER.\n\nAG BOGADH AR AGHAIDH. LE CHÉILE.\nGeneral
Election Manifesto 2024\n\n\n\n\n       "| __truncated__ "General Election 2024\n   M A N I F E S TO\n
1\n\nFINE GAEL | GENERAL ELECTION MANIFESTO"| __truncated__ "towards\n2030\na decade of change\nvolume II\nGreen
Party Manifesto 2024\n\n\n\n\n              greens\n       "| __truncated__ ...
 - attr(*, "spec")=
  .. cols(
  ..   party = col_character(),
  ..   year = col_double(),
  ..   text = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```