university of groningen

Quantitative Text Analysis – Essex Summer School

Comparing and describing documents

dr. Martijn Schoonvelde

University of Groningen

Today's class

- Describing and comparing documents
  - Document similarities or document distances
  - Document complexity
- Lab session

Describing and comparing documents

In the next days we'll learn about various approaches to categorise our texts using supervised and unsupervised methods

But it is often useful compare and describe documents at a more basic, perhaps syntactic level to help us think through meaningful categorisations

- Document similarities (distances) as well as document complexity

NB: these methods we're developed from linguistics and computer science with specific goals in mind (e.g., information retrieval) – we cannot just assume that they fit our social scientific research goals

- Validate, validate, validate (Grimmer & Stewart, 2013)

Document similarity and document distance

A frequent challenge in QTA is comparing pairs of documents and assessing how close or similar they are to one another.

- Constitutional scholars may want to know which constitutions are most alike. Communication scholars may want to know how news travels through outlets.

To this end it helps to think of documents as a vector of features as it allows us to use similarity metrics from linear algebra

- Vector space model – as document's vector is its numerical representation in a document feature matrix

```
                features
docs             so now , on this hallowed ground where just days
  2021-Biden.12  1   1 5  1    2        1      1     1    1    1
```

Euclidian distance

Let's say we have two documents with two different words: *mouse*, *cat*

	mouse	cat
Document a	12	14
Document b	5	11

In order to asses the distance between those two documents, we could calculate the euclidian distance between their feature vectors

Euclidian distance

- Euclidian distance
  - $\mathbf{a} = (12, 14)$
  - $\mathbf{b} = (5, 11)$
  - $d(\mathbf{a}, \mathbf{b}) = \sqrt{(12-5)^2 + (14-11)^2} \approx 7.6$
- This gives is some measure of distance that we can compare against distances between other documents.

Euclidian distance

- Euclidian distance
  - $\mathbf{a} = (12, 14)$
  - $\mathbf{b} = (5, 11)$
  - $d(\mathbf{a}, \mathbf{b}) = \sqrt{(12-5)^2 + (14-11)^2} \approx 7.6$
- This gives is some measure of distance that we can compare against distances between other documents.
- But it is not invariant to scaling.
  - $\mathbf{a} = (24, 28)$
  - $\mathbf{b} = (10, 22)$
  - $d(\mathbf{a}, \mathbf{b}) = \sqrt{(24-10)^2 + (10-22)^2} \approx 18.4$

Cosine similarity

A solution to this scaling issue is to calculate the angle between the two vectors

	mouse	cat
Document a	12	14
Document b	5	11

Cosine similarity between **a** and **b** can be calculated as follows (their dot product divided by their vector norms):

$$= \frac{\mathbf{a}}{\|\mathbf{a}\|} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|}$$

$$= \frac{(12,14)\cdot(5,11)}{\|12,14\|\|5,11\|}$$

$$= \frac{12\times5+14\times11}{\sqrt{12^2+14^2}\times\sqrt{5^2+11^2}} = 0.96$$

Cosine similarity

Cosine similarity is invariant to scaling

```
> library(lsa)

> a <- c(12,14)
> b <- c(5,11)
> cosine(a,b)
          [,1]
[1,] 0.9605011

> c <- c(24,28)
> d <- c(10,22)
> cosine(c,d)
          [,1]
[1,] 0.9605011
```

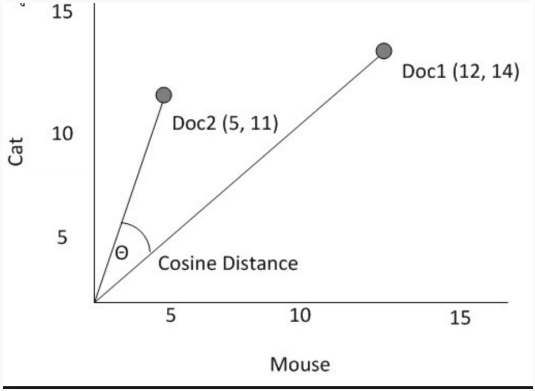

Image credit: Nick Grattan's Data Science Blog

Cosine similarity in quanteda

```
> dfmat_inaugural <- data_corpus_inaugural %>%
+     tokens(remove_punct = TRUE) %>%
+     tokens_remove(stopwords("en")) %>%
+     dfm()

> tail(textstat_simil(dfmat_inaugural,
+                     dfmat_inaugural["2021-Biden",],
+                     method = "cosine",
+                     margin = "documents"))
           2021-Biden
2001-Bush   0.5619136
2005-Bush   0.4797651
2009-Obama  0.6158540
2013-Obama  0.6061256
2017-Trump  0.5133378
```

Jaccard similarity

Jaccard similarity is another similarity measure. It's fairly easy to calculate:

- Count the number of tokens that appear in both documents
- Count the number of tokens that appear in either document
- Divide the first by the second

Jaccard similarity

```
jaccard <- function(a, b) {
  intersection = length(intersect(a, b))
  union = length(a) + length(b) - intersection
  return (intersection/union)
}

> a <- c("apple", "pear", "strawberry")
> b <- c("apple", "pineapple", "blueberry")
> jaccard(a,b)
[1] 0.2

> a <- c("apple", "apple", "apple", "pear", "strawberry")
> b <- c("apple", "apple", "apple", "pineapple", "blueberry")
> jaccard(a,b)
[1] 0.1111111
```

Jaccard similarity in quanteda

```
> tail(textstat_simil(dfmat_inaugural,
+                     dfmat_inaugural["2021-Biden",],
+                     method = "jaccard",
+                     margin = "documents"))
            2021-Biden
2001-Bush   0.1792746
2005-Bush   0.1912313
2009-Obama  0.1849711
2013-Obama  0.1913357
2017-Trump  0.1740064
2021-Biden  1.0000000
```

Comparing similarity measures

```
> cosine_inaugural <- textstat_simil(dfmat_inaugural,
+                                    dfmat_inaugural["2021-Biden",],
+                                    method = "cosine",
+                                    margin = "documents")
>
> jaccard_inaugural <- textstat_simil(dfmat_inaugural,
+                                     dfmat_inaugural["2021-Biden",],
+                                     method = "jaccard",
+                                     margin = "documents")
>
> cor(as.vector(cosine_inaugural), as.vector(jaccard_inaugural))
[1] 0.7559279
```

Similarity and distance measures in quanteda

textstat_dist options are: "euclidean" (default), "canberra", "Chisquared", "Chisquared2", "hamming", "kullback". "manhattan", "maximum", "canberra", and "minkowski".
textstat_simil options are: "correlation" (default), "cosine", "jaccard", "eJaccard", "dice", "eDice", "simple matching", "hamann", and " "faith".

Similarity in social science

Similarity and distance measures are blind to the semantic content of a text.

As social scientists we often have a specific idea in mind when we are interested in similarity between documents (e.g., the extent to which they share certain topics, or the extent to which they share a particular sentiment)

- If we want to try and measure context-specific similarity, other tools and methods are probably better
- But similarity and distance scores may help us in the conceptualization stage

Biden inaugural address

"So now, on this hallowed ground where just days ago violence sought to shake this Capitol's very foundation, we come together as one nation, under God, indivisible, to carry out the peaceful transfer of power as we have for more than two centuries." (Biden, 2021)

Boris Johnson resignation speech

"In the last few days I've tried to persuade my colleagues that it would be eccentric to change governments when we're delivering so much, when we have such a vast mandate and when we're actually only a handful of points behind in the polls." (Johnson, 2022)

Measuring textual complexity

Capturing the complexity of texts has intuitive applications for social scientists:

- Psychologists may be interested in the extent to which ideas in a text integrate various perspectives (e..g, Suedfeld & Tetlock, 1977)
- Political scientists may be interest in when politicians obfuscate when talking about certain issues (Rauh *et al.* 2020)

Measuring textual complexity

There are many different ways that complexity of a text can be captured, and – in broad terms – they differ in their focus on semantic or syntactic elements in a text

- Syntactic elements refer to the grammatical and structural elements of a text.
- Semantic elements refer to the actual content expressed in a text.

Readability scores

- Originally developed to measure 'readability' of a text by Kincaid *et al.*
- Purely based on syntax: Weighted average of word length and sentence length
  - Flesch Kincaid grade level:
  - $0.39 \times (\frac{\text{total words}}{\text{total sentences}}) + 11.8 \times \frac{\text{total syllables}}{\text{total words}} - 15.59$
  - Result is a number that corresponds with US grade level required to understand the text
  - Also: Flesch Reading Ease, $0 - 100$ scale, with 100 easiest, and 0 most difficult

Readability in quanteda

```
> corpus <- johnson_sentence + biden_sentence
> textstat_readability(corpus)
        document    Flesch
1  2022-Johnson  46.81136
2  2021-Biden  33.33884
```

Meaning of readability scores

- Topic of ongoing debate in political science

Measuring and Explaining Political Sophistication through Textual Complexity

Kenneth Benoit London School of Economics and Political Science
Kevin Munger Pennsylvania State University
Arthur Spirling New York University

Abstract: *Political scientists lack domain-specific measures for the purpose of measuring the sophistication of political communication. We systematically review the shortcomings of existing approaches, before developing a new and better method along with software tools to apply it. We use crowdsourcing to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of sophistication. This includes previously excluded features such as parts of speech and a measure of word rarity derived from dynamic term frequencies in the Google Books data set. Our technique not only shows which features are appropriate to the political domain and how, but also provides a measure easily applied and rescaled to political texts in a way that facilitates probabilistic comparisons. We reanalyze the State of the Union corpus to demonstrate how conclusions differ when using our improved approach, including the ability to compare complexity as a function of covariates.*

Lexical diversity

Type to Token Ratio (TTR)

- Ratio of different unique word stems (types) to the total number of words (tokens)
- Combines semantic and syntactic elements
  - Most informative when texts of equal length are compared as TTR tends to decrease as text length increases

Lexical diversity in quanteda

```
> tokens(corpus) %>%
+   dfm() %>%
+   textstat_lexdiv(measure = "TTR")

       document        TTR
1   2022-Jonhson  0.8409091
2  2021-Biden  0.9069767
```

Complexity in social science

Lots of work in this space has been done by computational linguists who developed many measures of the complexity of a particular text

But the research goals of social scientists are different. Not interest in absolute complexity, but in relative complexity.