Text-as-data

Max Callaghan

2023-09-11

Welcome

Introductions

I am a researcher at the Mercator Research Institute on Global Commons and Climate Change, where I work on applications of Natural Language Processing (NLP) in climate science.

Before that I completed a Bachelors degree in English Literature at the University of York, and a Masters of Public Policy at the Hertie School of Governance (2016). I received my PhD from the School of Earth and Environment at the University of Leeds in 2022.

Introductions

I am a researcher at the Mercator Research Institute on Global Commons and Climate Change, where I work on applications of Natural Language Processing (NLP) in climate science.

Before that I completed a Bachelors degree in English Literature at the University of York, and a Masters of Public Policy at the Hertie School of Governance (2016). I received my PhD from the School of Earth and Environment at the University of Leeds in 2022.

Over to you. Who are you? What do you already understand by "text as data"? What are you excited about covering in the course? What are you nervous about tackling?

Objectives for the course

In this course, we will learn about how, and why, to treat text as data. We'll get an overview of the main methods for doing this, and practical experience applying these to texts that are interesting to social scientists.

This will involve learning new quantitative techniques, but the emphasis will be on applications over theory. By the end of the course you should be

- Able to acquire and process text data from various sources
- comfortable producing a range of different analyses of text data
- comfortable using programming creatively as a tool to solve problems
- aware of what research questions are possible to address with text data, and able to understand what a good research question is
- able to *critically* engage with this type of analysis
- able to report the results of your work, and show your working

Organisation

R and Python

```r
square_x <- function(x) {
  return(x**2)
}
square_x(2)
```

```
## [1] 4
```

Everything in the course will be taught with R by default.

```python
def square_x(x):
  return x**2
square_x(2)
```

```
## 4
```

Python is more common in many areas working with text, so I'll provide examples of python workflows too.

It's up to you how you complete your assignments: assessments will be made on the basis of whether the code does what it is supposed to in a clear and transparent way.

RMarkdown and Github

I use:

- git for version control of projects,
- Github to host projects, and
- Rmarkdown to produce documents with integrated code and output

For example, all materials for this course will be contained in the Github project. This presentation is available here, and the Rmarkdown file used to create the presentation is available here.

You can download all the materials for yourself by "forking" or "cloning" the repository. We will also use the GitHub repository to **collectively** troubleshoot. If you have a problem, open an issue. If you can help, post an answer. I *want* you to help each other, but do this transparently!

A basic familiarity with git and github is *extremely useful*. Here is an intro.

Expectations

- Turn up to every class with a laptop with R or Python installed and ready to run code (let me know if you do not have a laptop or cannot do this for any other reason).
- Stop me if you don't understand something
- Treat each other with respect in class (and let me know if something has made you uncomfortable)
- Hand **something** in with **some** answer for every question for every assignment
- Write to me (callaghan@mcc-berlin.net) if you need something
- Office hours: arrange a meeting with me on Mondays (other days potentially possible online).

Course Mechanics

Classes

Classes will consist of an input from me along with examples of code, short coding exercises, and interactive tasks.

Assignments

Assessment will comprise

- 3 short programming exercises (due 25.09, 02.10, 09.10) (30%)
- A data analysis exercise, set 3 November - due on November 17 (30%)
- Group presentation of a research project, detailed instructions November 3, presentations on December 1 (40%)

Introduction to text as data

Why texts?

- "Politics and political conflict often occur in the written and spoken word" [Grimmer and Stewart, 2013]

Why texts?

- "Politics and political conflict often occur in the written and spoken word" [Grimmer and Stewart, 2013]
- Many other processes – science, culture, consumer behaviour, corporate behaviour – are observable through texts.

Why texts?

- "Politics and political conflict often occur in the written and spoken word" [Grimmer and Stewart, 2013]
- Many other processes – science, culture, consumer behaviour, corporate behaviour – are observable through texts.
- *Reading* texts allows us to understand these domains and processes

Why texts?

- "Politics and political conflict often occur in the written and spoken word" [Grimmer and Stewart, 2013]
- Many other processes – science, culture, consumer behaviour, corporate behaviour – are observable through texts.
- *Reading* texts allows us to understand these domains and processes
- Research that reads texts can result in useful findings about the world

Why texts?

- "Politics and political conflict often occur in the written and spoken word" [Grimmer and Stewart, 2013]
- Many other processes – science, culture, consumer behaviour, corporate behaviour – are observable through texts.
- *Reading* texts allows us to understand these domains and processes
- Research that reads texts can result in useful findings about the world

---

What are some examples of texts we can learn from?

Why treat text as data?

- Reading texts takes time

Why treat text as data?

- Reading texts takes time
- Digitized collections of texts are increasingly available at increasingly large scales

Why treat text as data?

- Reading texts takes time
- Digitized collections of texts are increasingly available at increasingly large scales
- How computers can “read” texts has advanced in leaps and bounds.

How can we treat text as data?

Text is incredibly complex. Turning this into data is not easy.

> *"Time flies like an arrow. Fruit flies like a banana."*

To turn a text into data we need a way of representing its "features'' in numerical form.

How can we treat text as data?

Text is incredibly complex. Turning this into data is not easy.

> *"Time flies like an arrow. Fruit flies like a banana."*

To turn a text into data we need a way of representing its "features'' in numerical form.

Which words are used, how frequently, in what order, in what context?

What questions can we ask?

- What is this text about?
- What texts is this text similar to?
- What emotions/opinions/political stances does this text express?

What do we need to consider when treating text as data?

- All models of text are wrong, some are useful, some of the time

What do we need to consider when treating text as data?

- All models of text are wrong, some are useful, some of the time
  - There is no single best method

What do we need to consider when treating text as data?

- All models of text are wrong, some are useful, some of the time
  - There is no single best method
  - Validation is key to understanding what is useful, when, to do what

What do we need to consider when treating text as data?

- All models of text are wrong, some are useful, some of the time
  - There is no single best method
  - Validation is key to understanding what is useful, when, to do what
- Models of text can be a "dangerous supplement" link - they are better complements to than substitutes for human reading.

Outlook

What will we cover this term?

Part 1 (weeks 1-5) - Foundations

In the first part of the course we will encounter a range of different texts, and get comfortable working with these in R

Part 2 (weeks 6-11) - Analysis

In the following 5 weeks, we will learn different approaches to analysing texts, how to carry these out in R, and how to use and interpret the results

Wrapup (week 12)

In the final session, you will present your group research projects.

Appendix

Rapidly declining remarkability

- What do the authors do with the texts?
- What techniques do they employ?
- How do they validate these? How does that validation affect their/our confidence in the model?

[Moore et al., 2019]

Plato's Pharmacy

back

The Father of Logos

The story begins like this:

Socrates: Very well. I heard, then, that at Naucratis in Egypt there lived one of the old gods of that country, the one whose sacred bird is called the ibis; and the name of the divinity was Theuth. It was he who first invented numbers and calculation, geometry and astronomy, not to speak of draughts and dice, and above all writing (*grammata*). Now the King of all Egypt at that time was Thamus who lived in the great city of the upper region which the Greeks call the Egyptian Thebes; the god himself they call Ammon. Theuth came to him and exhibited his arts and declared that they ought to be imparted to the other Egyptians. And Thamus questioned him about the usefulness of each one; and as Theuth enumerated, the King blamed or praised what he thought were the good or bad points in the explanation. Now Thamus is said to have had a good deal to remark on both sides of the question about every single art (it would take too long to repeat it here); but when it came to writing, Theuth said, "This discipline (*to mathēma*), my King, will make the Egyptians wiser and will improve their memories (*sophōterous kai mnēmonikōterous*): my invention is a recipe (*pharmakon*) for both memory and wisdom." But the King said . . . etc. (274*c*–*e*)

But the king said, "Theuth, my master of arts (*Ōtekhnikōtate Theuth*), to one man it is given to create the elements of an art, to another to judge the extent of harm and usefulness it will have for those who are going to employ it. And now, since you are father of written letters (*patēr ōn grammatōn*), your paternal goodwill has led you to pronounce the very opposite (*tounantion*) of what is their real power. The fact is that this invention will produce forgetfulness in the souls of those who have learned it because they will not need to exercise their memories (*lēthēn men en psuchais parexei mnēmēs ameletēsiai*), being able to rely on what is written, using the stimulus of external marks that are alien to themselves (*dia pistin graphēs exōthen hup' allotriōn tupōn*) rather than, from within, their own unaided powers to call things to mind (*ouk endothen autous huph' hautōn anamimnēskomenous*). So it's not a remedy for memory, but for reminding, that you have discovered (*oukoun mnēmēs, alla hupomnēseōs, pharmakon hēures*). And as for wisdom (*sophias de*), you're equipping your pupils with only a semblance (*doxan*) of it, not with truth (*alētheian*). Thanks to you and your invention, your pupils will be widely read without benefit of a teacher's instruction; in consequence, they'll entertain the delusion that they have wide knowledge, while they are, in fact, for the most part incapable of real judgment. They will also be difficult to get on with since they will be men filled with the conceit of wisdom (*doxosophoi*), not men of wisdom (*anti sophōn*)." (274*e*–275*b*)

Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mps028.

Frances C. Moore, Nick Obradovich, Flavio Lehner, and Patrick Baylis. Rapidly declining remarkability of temperature anomalies may obscure public perception of climate change. *Proceedings of the National Academy of Sciences*, 116(11):4905–4910, March 2019. doi: 10.1073/pnas.1816541116.