Text in research context

William Lowe

Hertie School

3rd December 2020

Where to put it?

Simple lexical analysis

→ C_{ij} as an *indicator* or an effect of something non-textual

Where to put it?

Simple lexical analysis

→ C_{ij} as an *indicator* or an effect of something non-textual

We've focused more on text analysis as a *measurement problem*:

→ (Documents + assumptions) $\Longrightarrow \hat{\theta}$

Where does this fit in the larger research picture?

→ θ as an independent variable
→ θ as an dependent variable
→ θ as a confounder

Where to put it?

Simple lexical analysis

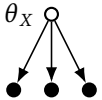
→ C_{ij} as an *indicator* or an effect of something non-textual

We've focused more on text analysis as a *measurement problem*:

→ (Documents + assumptions) $\Longrightarrow \hat{\theta}$

Where does this fit in the larger research picture?

→ θ as an independent variable
→ θ as an dependent variable
→ θ as a confounder



Remember, there's still measurement error, even if there isn't bias

Where to put it?

Simple lexical analysis

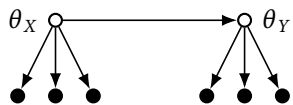
→ C_{ij} as an *indicator* or an effect of something non-textual

We've focused more on text analysis as a *measurement problem*:

→ (Documents + assumptions) $\Longrightarrow \hat{\theta}$

Where does this fit in the larger research picture?

→ θ as an independent variable
→ θ as an dependent variable
→ θ as a confounder



Remember, there's still measurement error, even if there isn't bias

Where to put it?

Simple lexical analysis

→ C_{ij} as an *indicator* or an effect of something non-textual

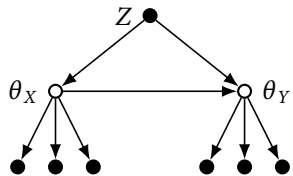
We've focused more on text analysis as a *measurement problem*:

→ (Documents + assumptions) $\Longrightarrow \hat{\theta}$

Where does this fit in the larger research picture?

→ θ as an independent variable
→ θ as an dependent variable
→ θ as a confounder

Remember, there's still measurement error, even if there isn't bias



Strategy

We can also think of a text analysis

1. θ as a large scale terrain map / sample stratifier
2. θ as a generalization check

Examples of 1 and 2:

→ Classify / scale / topic model 10,000 news stories and use $\hat{\theta}$ to see which ones to read more closely

→ Work up a small dictionary on 30 documents and apply to the 10,000 news stories

Examples of iteration:

→ Work up a small dictionary on 30 stories

→ Apply to the 10,000 stories to see macro trends

→ Sample interesting, extreme, of randomly based on θ to check the model

Tactics

Sampling?

- → Who or what is the population?
- → Down-sampled data means you iterate models faster (and risk missing something)
- → Thoughtful stratification will help you draw more robust conclusions

Model checking?

- → How you would *check* the model, e.g. for stability of the $\theta \longrightarrow W$ mapping or the βs

Tools

- → Much text analysis is inherently mechanical / automated: be very instrumental about packages and tools
- → Don't be afraid to ask for help: e.g. me, the Data Science Lab's research consulting service

Hints and Tips

Always try the Kartoffelpuffer (ideally with apple sauce)

→ Very unhealthy, but quite yummy

Hints and Tips

Always try the Kartoffelpuffer (ideally with apple sauce)

→ Very unhealthy, but quite yummy

For those of you currently outside Germany

Never eat Kartoffelpuffer (particularly with apple sauce)

→ Terrible. Especially with Glühwein.
→ You're totally not missing anything.


That's all Folks!