April - Wed 28

To be discussed:

- Data Acquisition
- Data Processing
- Baseline model

Rough thoughts

- *What is the minimal thing that I can do to get going?*
  - *Read more.*
  - *Get data NOW.*
  - *Is it realistic to get TabFact-like model working fast?*
  - *If not, what should my baseline be?*
  - *Figure out metrics to log, informative diagnostics.*

Data Acquisition

- Data Acquisition
  - ONS API is available, HTTP, open and unrestricted / can use get requests
  - Would also need to build a scraper, to get the "bulletin" text on the landing page. I have an idea how to do this.
  - Is there another source I am missing?

Data Processing

- Generating false examples

  - Can we establish a method of simplifying the "query"? (subject, predicate, object)? If so we can use this triplet to our advantage and modify?
  - For numeric statements, this can be to add noise. (*future*: we might want to allow some slack though, think about a threshold/interval)
  - For declarative statements, we can use a [MASK]? What about swapping of alternative words via POS tags?
  - ...
- Figure out how the mapping to true or false should be handled

  - Is it reallistic to do both SQL/Query matching and NLP approach?

Baseline model

- What is our baseline?
  - Should I be implementing the techniques described in TabFact/ OpenFact?