June 28th

The dataset

I've re-scrapped the ONS website because I saw that I was limitting the number of reports per subject.

- Now 395 unique bulletins that are usable (with main-points and datasets)
- with 2011 unique main-points
- I follow what Patrick was doing in the "Unsupervised Question Answering by Cloze Translation" paper to create our Clozes

Out[20]:

	bulletin	type	point	data
0	businessindustryandtrade/business/businessserv...	date_and_percent	In 2019, approximate gross value added at basi...	[/businessindustryandtrade/business/businessse...
1	businessindustryandtrade/business/businessserv...	date_and_percent	The non-financial services sector, which accou...	[/businessindustryandtrade/business/businessse...
2	businessindustryandtrade/business/businessserv...	date_and_percent	Total turnover and purchases of the UK non-fin...	[/businessindustryandtrade/business/businessse...
3	businessindustryandtrade/business/businessserv...	date_and_percent	Out of the 12 UK regions, 8 regions experience...	[/businessindustryandtrade/business/businessse...
4	businessindustryandtrade/business/businessserv...	date_and_percent	West Midlands, Yorkshire and The Humber, Scotl...	[/businessindustryandtrade/business/businessse...

Generating CLOZES

Here I show a few samples from our Cloze generation process:

```
ORIGINAL TEXT:
In 2019, approximate gross value added at basic prices (aGVA) of the UK non-financial business economy was estimated to be
£1,313.9 billion; an increase of £42.8 billion (3.4%) compared with 2018.



CLOZES:
0. In TEMPORALMASK, approximate gross value added at basic prices (aGVA) of the UK non-financial business economy was estim
ated to be £1,313.9 billion; an increase of £42.8 billion (3.4%) compared with 2018.

1. In 2019, approximate gross value added at basic prices (aGVA) of the UK non-financial business economy was estimated to
be NUMERICMASK; an increase of £42.8 billion (3.4%) compared with 2018.

2. In 2019, approximate gross value added at basic prices (aGVA) of the UK non-financial business economy was estimated to
be £1,313.9 billion; an increase of NUMERICMASK (3.4%) compared with 2018.

3. In 2019, approximate gross value added at basic prices (aGVA) of the UK non-financial business economy was estimated to
be £1,313.9 billion; an increase of £42.8 billion (NUMERICMASK) compared with 2018.

4. In 2019, approximate gross value added at basic prices (aGVA) of the UK non-financial business economy was estimated to
be £1,313.9 billion; an increase of £42.8 billion (3.4%) compared with TEMPORALMASK.


ORIGINAL TEXT:
The biggest component of services imported into NUTS1 areas was travel, as it received 28% (£50.6 billion) of UK total impo
rts of services (£180.9 billion).



CLOZES:
0. The biggest component of services imported into NUTS1 areas was travel, as it received NUMERICMASK (£50.6 billion) of UK
total imports of services (£180.9 billion).

1. The biggest component of services imported into NUTS1 areas was travel, as it received 28% (NUMERICMASK) of UK total imp
orts of services (£180.9 billion).

2. The biggest component of services imported into NUTS1 areas was travel, as it received 28% (£50.6 billion) of UK total i
mports of services (NUMERICMASK).


ORIGINAL TEXT:
In the UK, 14.1% of people reported struggling to make ends meet in 2017, below the EU-28 average of 21.6%, and one-fifth r
eported that they were “very satisfied” with their household income in 2018, above the EU-28 average.



CLOZES:
0. In the UK, NUMERICMASK of people reported struggling to make ends meet in 2017, below the EU-28 average of 21.6%, and on
e-fifth reported that they were “very satisfied” with their household income in 2018, above the EU-28 average.

1. In the UK, 14.1% of people reported struggling to make ends meet in TEMPORALMASK, below the EU-28 average of 21.6%, and
one-fifth reported that they were “very satisfied” with their household income in 2018, above the EU-28 average.

2. In the UK, 14.1% of people reported struggling to make ends meet in 2017, below the EU-28 average of NUMERICMASK, and on
e-fifth reported that they were “very satisfied” with their household income in 2018, above the EU-28 average.

3. In the UK, 14.1% of people reported struggling to make ends meet in 2017, below the EU-28 average of 21.6%, and NUMERICM
ASK reported that they were “very satisfied” with their household income in 2018, above the EU-28 average.

4. In the UK, 14.1% of people reported struggling to make ends meet in 2017, below the EU-28 average of 21.6%, and one-fift
h reported that they were “very satisfied” with their household income in TEMPORALMASK, above the EU-28 average.


ORIGINAL TEXT:
General government saw a decrease in its net borrowing position to £59.7 billion in Quarter 4 which equates to 11.0% of GDP
compared to 12.9% in Quarter 3.



CLOZES:
0. General government saw a decrease in its net borrowing position to NUMERICMASK in Quarter 4 which equates to 11.0% of GD
P compared to 12.9% in Quarter 3.

1. General government saw a decrease in its net borrowing position to £59.7 billion in TEMPORALMASK which equates to 11.0%
of GDP compared to 12.9% in Quarter 3.

2. General government saw a decrease in its net borrowing position to £59.7 billion in Quarter 4 which equates to NUMERICMA
SK of GDP compared to 12.9% in Quarter 3.

3. General government saw a decrease in its net borrowing position to £59.7 billion in Quarter 4 which equates to 11.0% of
GDP compared to NUMERICMASK in Quarter 3.

4. General government saw a decrease in its net borrowing position to £59.7 billion in Quarter 4 which equates to 11.0% of
GDP compared to 12.9% in TEMPORALMASK.
```

Discussion/ Questions

- The Clozes generated are quite similar, why wouldn't our model learn to treat them all the same?
- How can we force the model to USE the data, and not learn to regurgitate some numbers back?
- What language models should I be looking into? distilBert, and bert-like?
- I need to better understand how to handle datasets, since each dataset has 3-4 sheets/pages within it.
  - Would it be smart to build a "detection" system to identify the most appropriate sheet?
- Is it time to rethink what the dissertation's main tasks should be? Pasquale suggested this can be an attempt to make a new dataset?