#metoo Group 1

Angela, Jane, Michael, & Sarah

#metoo Project Website
https://www.schlesinger-metooproject-radcliffe.org/

#metoo Digital Media Collection

- ~4 GB of saved test crawls from Archive-It
- Web sites and pages: News articles, blog posts, organizational websites
- Full text (derivative file): 370 MB

- Does not contain Twitter data

What do we want to explore?

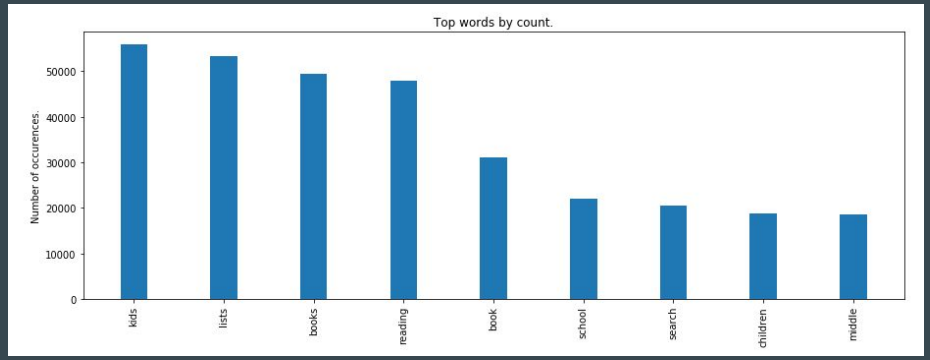
- What content is linked to from the web archive?
- Is that content (the content linked to from the crawled websites or web pages) also in the web archive?
  - Can we access the text of the content that is linked to?

What were we able to do?

- Run through the code in the Jupyter notebook!
  - Looked at top domains and tokens
  - Added additional stop words to try to get more meaningful data
    - Needed to add years, months, etc. to get rid of some of the noise
- Explore hyperlink relationships between domains in Gephi
- Try to identify content that is linked to but not crawled or accessible to users in Archive-It

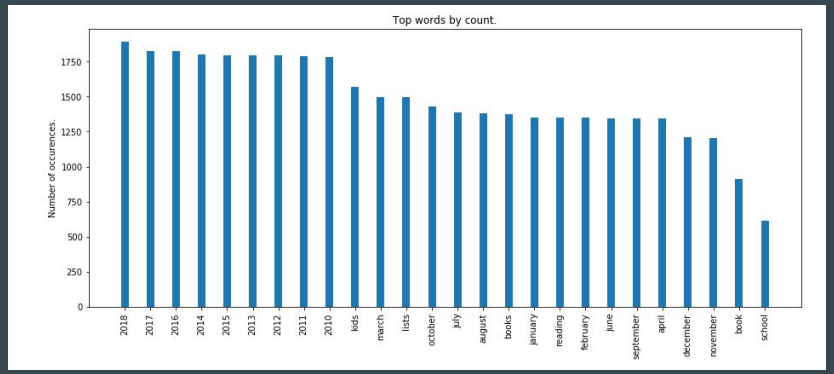
What have we learned?

- Decision-making around web crawl QA has a BIG impact on the data we're working with in AUT
  - Two large crawls of the Pragmatic Mom blog throw the data out of whack
- Data that's useful when provided through Archive-It/Wayback Machine might be problematic for this type of analysis



What have we learned?

- A lot of data cleaning is necessary to make text analysis meaningful



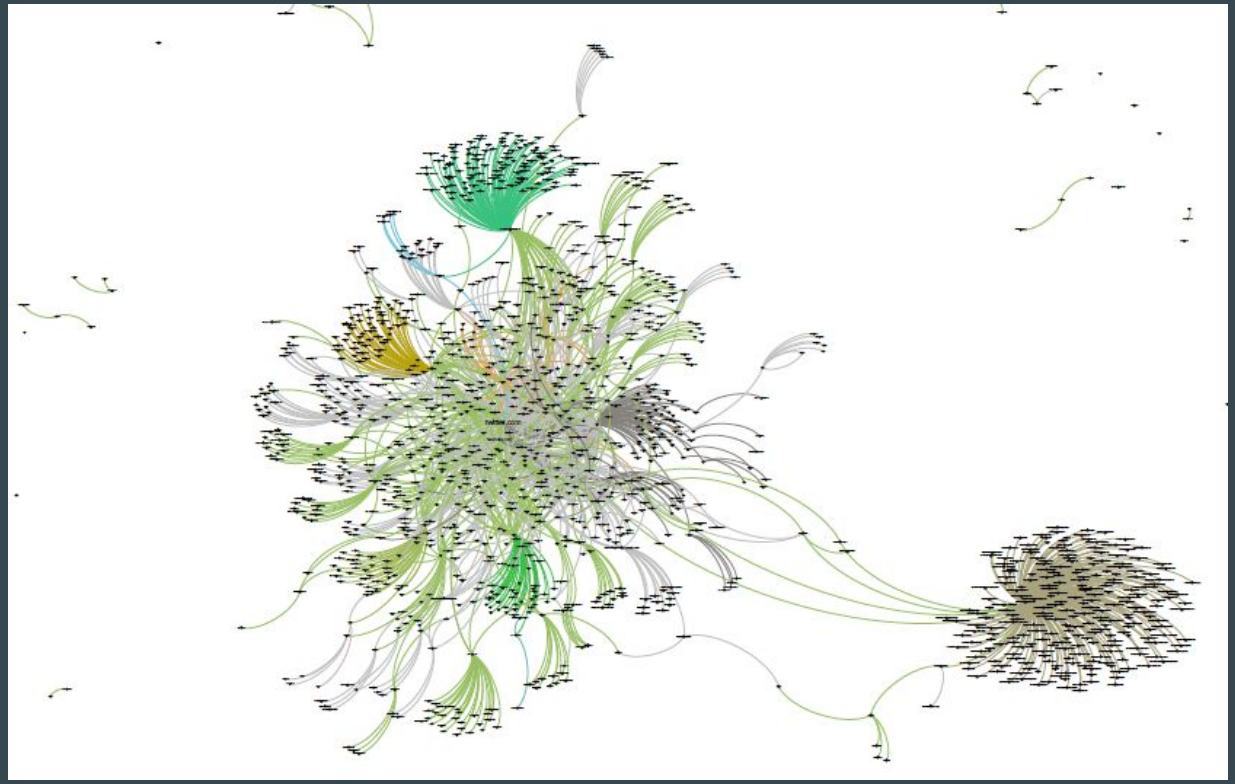

```
In [15]: # Get a list of the top tokens, separated by year.

         res = get_top_tokens_by(year, OVERALL_SHOW_TOKENS)[0:OVERALL_OUTPUT_SIZE]
         lst = []


         [(z[0], [x for y in res for x in y[1] if x[0].lower() not in STOP_WORDS]) for z in res]
```

```
Out[15]: [('2018',
           [('2018', 1834),
            ('2017', 1787),
            ('2016', 1785),
            ('2014', 1784),
            ('2015', 1783),
            ('2010', 1781),
            ('2012', 1781),
            ('2013', 1781),
            ('2011', 1780),
            ('Kids', 1554),
            ('March', 1483),
            ('Lists', 1483),
            ('October', 1410),
            ('July', 1375),
            ('August', 1362),
            ('February', 1334),
            ('January', 1334),
```

Network Analysis in Gephi



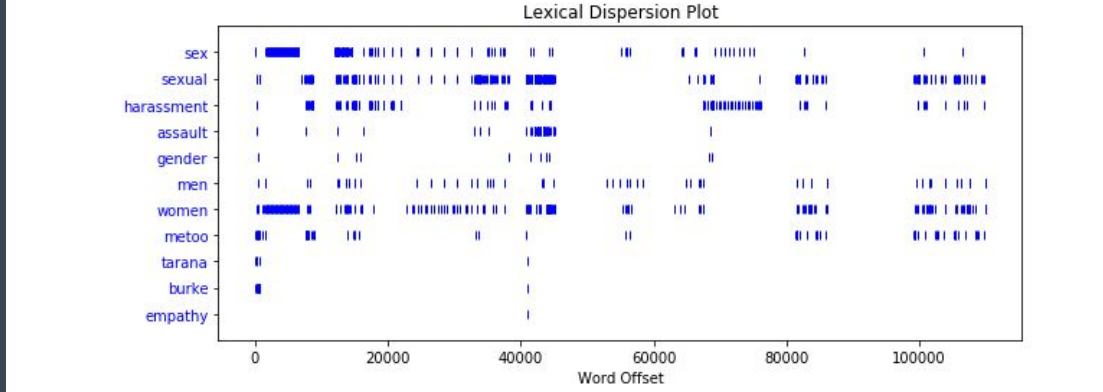
How can we glean meaningful information from this?

What have we learned?

- It looks like Tarana Burke, who originally began #metoo, is mentioned at the beginning of texts. Is she getting credit for her work and her role in this movement?

```
In [46]: # Create a dispersion plot, showing where the list of words appear in the text.

         text = get_text_tokens(1) # Need to have one to include words with fewer than 3 letters.
         dp(text, DISPERSION_PLOT_WORDS) # Uses the nltk dispersion plot library (dp).
```



What else do we want to know?

- Explore what tweets are embedded in web content?
  - Would use .warc files for this
- What can we learn from working with warc files instead of derivatives?
- Can NER help us identify more people/organizations that should be represented in this collection? Can that help point us to more content that will allow us build more inclusive or diverse collections, or will that only identify what's already here?
- What is the best way for us to continue working on this once we get home?