Applied Data Science - Analysis of Philosophy Texts

- Name: Shubham Laddha
- UNI: sl4983

The dataset comprises of philosophy texts (downloaded from Kaggle). It contains over 300,000 sentences from over 50 texts spanning 10 major schools of philosophy. The represented schools are: Plato, Aristotle, Rationalism, Empiricism, German Idealism, Communism, Capitalism, Phenomenology, Continental Philosophy, and Analytic Philosophy.

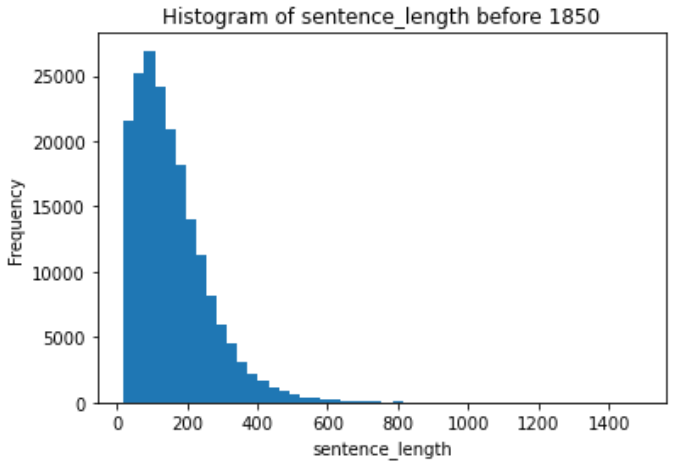
1. Exploratory Data Analysis

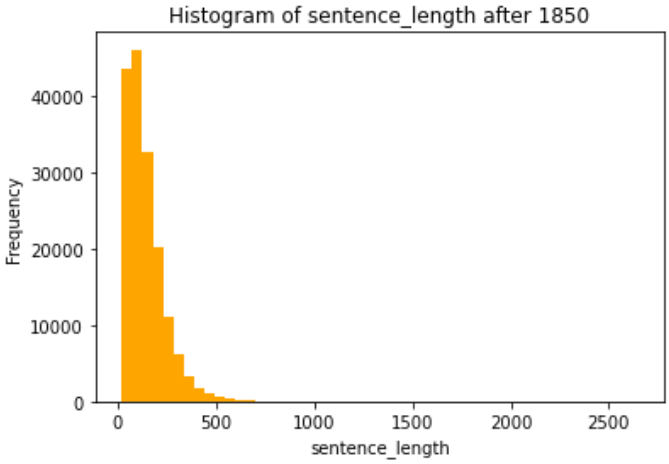
	title	author	school	sentence_spacy	sentence_str	original_publication_date	corpus_edition_date	sentence_length	sent
0	Plato - Complete Works	Plato	plato	What's new, Socrates, to make you leave your ...	What's new, Socrates, to make you leave your ...	-350	1997	125	soc yo
1	Plato - Complete Works	Plato	plato	Surely you are not prosecuting anyone before t...	Surely you are not prosecuting anyone before t...	-350	1997	69	sur an
2	Plato - Complete Works	Plato	plato	The Athenians do not call this a prosecution b...	The Athenians do not call this a prosecution b...	-350	1997	74	th p
3	Plato - Complete Works	Plato	plato	What is this you say?	What is this you say?	-350	1997	21	v
4	Plato - Complete Works	Plato	plato	Someone must have indicted you, for you are no...	Someone must have indicted you, for you are no...	-350	1997	101	hav f

```
******************************
Authors related to each School
******************************
plato: ['Plato']
aristotle: ['Aristotle']
empiricism: ['Locke' 'Hume' 'Berkeley']
rationalism: ['Spinoza' 'Leibniz' 'Descartes' 'Malebranche']
analytic: ['Russell' 'Moore' 'Wittgenstein' 'Lewis' 'Quine' 'Popper' 'Kripke']
continental: ['Foucault' 'Derrida' 'Deleuze']
phenomenology: ['Merleau-Ponty' 'Husserl' 'Heidegger']
german_idealism: ['Kant' 'Fichte' 'Hegel']
communism: ['Marx' 'Lenin']
capitalism: ['Smith' 'Ricardo' 'Keynes']
stoicism: ['Epictetus' 'Marcus Aurelius']
nietzsche: ['Nietzsche']
feminism: ['Wollstonecraft' 'Beauvoir' 'Davis']

************************************
Publication Year of texts across Schools
************************************
plato: [-350]
aristotle: [-320]
empiricism: [1689 1739 1779 1713 1710]
rationalism: [1677 1710 1637 1641 1674]
```

```
analytic: [1921 1912 1910 1953 1985 1950 1959 1972 1975]
continental: [1963 1961 1966 1967 1968 1972]
phenomenology: [1945 1936 1907 1927 1950]
german_idealism: [1788 1790 1781 1798 1817 1807 1820]
communism: [1883 1848 1862]
capitalism: [1776 1817 1936]
stoicism: [125 170]
nietzsche: [1888 1886 1887]
feminism: [1792 1949 1981]

****************************************
List of Authors and their publication years
****************************************
Plato: [-350]
Aristotle: [-320]
Locke: [1689]
Hume: [1739 1779]
Berkeley: [1713 1710]
Spinoza: [1677]
Leibniz: [1710]
Descartes: [1637 1641]
Malebranche: [1674]
Russell: [1921 1912]
Moore: [1910]
Wittgenstein: [1953 1921 1950]
Lewis: [1985]
Quine: [1950]
Popper: [1959]
Kripke: [1972 1975]
Foucault: [1963 1961 1966]
Derrida: [1967]
Deleuze: [1968 1972]
Merleau-Ponty: [1945]
Husserl: [1936 1907]
Heidegger: [1927 1950]
Kant: [1788 1790 1781]
Fichte: [1798]
Hegel: [1817 1807 1820]
Marx: [1883 1848]
Lenin: [1862]
Smith: [1776]
Ricardo: [1817]
Keynes: [1936]
Epictetus: [125]
Marcus Aurelius: [170]
Nietzsche: [1888 1886 1887]
Wollstonecraft: [1792]
Beauvoir: [1949]
Davis: [1981]
```

1.1 Histogram of Sentence Length over the years



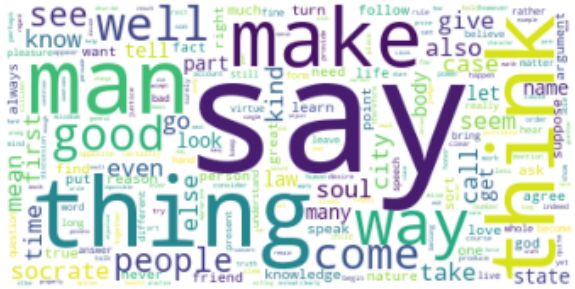



```
Average sentence length before 1850: 158.7
Average sentence length after 1850: 141.75
```

- The histogram above indicates that authors have preferred shorter sentences after 1850. However, there is an outlier after 1850 which has sentence_length=2500. The average sentence length supports our prior observation.

1.2 WordClouds per School

plato WordCloud:

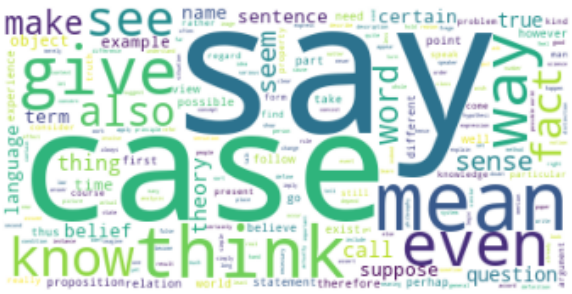


aristotle WordCloud:

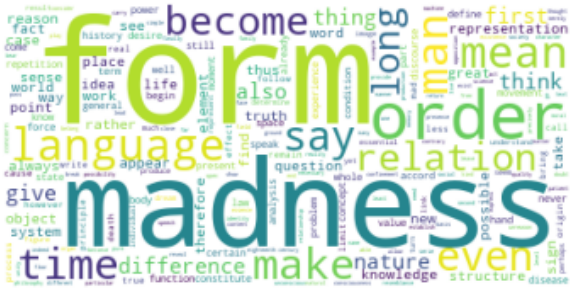
empiricism WordCloud:

rationalism WordCloud:

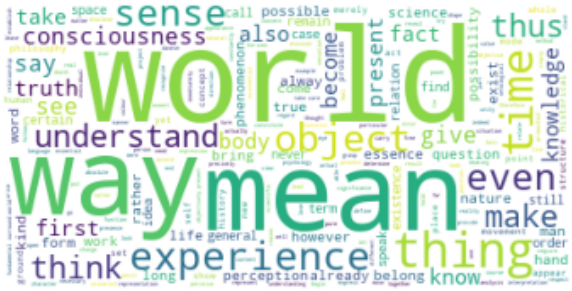
analytic WordCloud:



continental WordCloud:



phenomenology WordCloud:



german_idealism WordCloud:

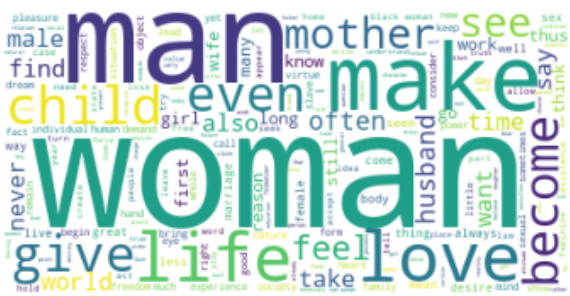
communism WordCloud:

capitalism WordCloud:

stoicism WordCloud:

nietzsche WordCloud:

feminism WordCloud:



The wordclouds per school provides us with a brief idea of the words that appear most often.

- Empiricism and Rationalism Philosophy have common frequently occuring words in wordcloud: idea, make, mind, man, reason.

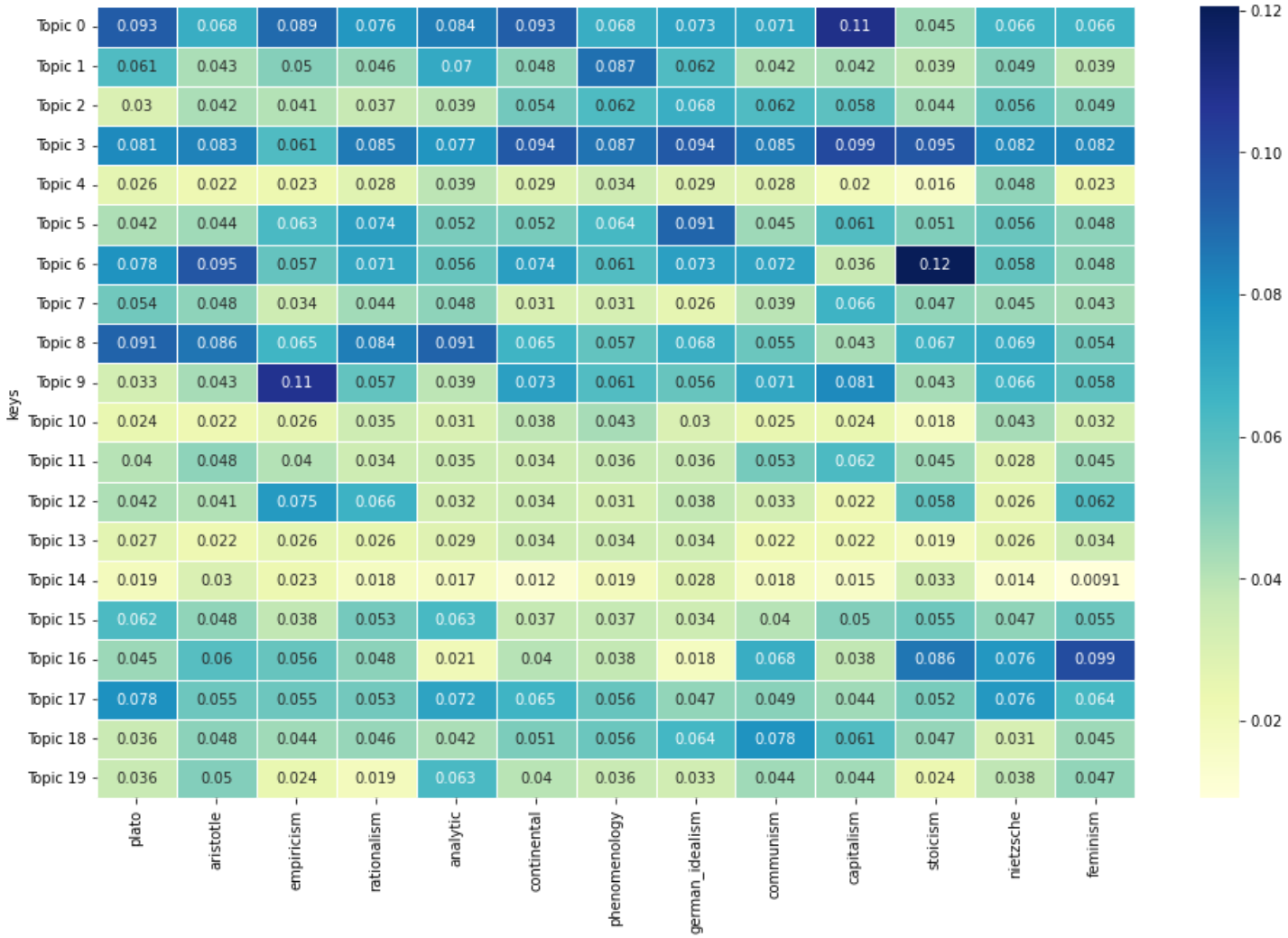
2. Topic Modeling - On entire dataset

Based on the texts, the idea is to find topics in an unsupervised manner (LDA model) on the entire dataset and create relative

frequency heatmap to identify common topic words. Before we build the model, we will:

- Transform tokenized text data.
- Remove Stop Words.
- Create Bigrams.
- Lemmatize text data.
- Create Corpus.

2.1 Heatmap of Relative Frequencies



2.2 Topic Words

```
Topic 0: ['make', 'think', 'people', 'clearly', 'maintain', 'like', 'charge', 'perhaps', 'difference', 'afr
aid']
Topic 1: ['way', 'believe', 'put', 'present', 'hear', 'answer', 'strange', 'appearance', 'obvious', 'surel
y']
Topic 2: ['become', 'consider', 'already', 'important', 'indeed', 'hard', 'superior', 'regard', 'mine', 'pu
pil']
Topic 3: ['give', 'first', 'desire', 'good', 'care', 'speak', 'public', 'later', 'possible', 'die']
Topic 4: ['call', 'old', 'truth', 'teach']
Topic 5: ['however', 'accept', 'far', 'reason', 'understand', 'full', 'quite', 'opinion', 'agree', 'would_s
eem']
```

```
Topic 6: ['many', 'thing', 'form', 'tell', 'really']
Topic 7: ['great', 'young', 'friend', 'true', 'ask', 'city', 'happen', 'source', 'obviously', 'take_care']
Topic 8: ['say', 'right', 'cause', 'death', 'escape', 'wish', 'back', 'matter', 'pursue', 'relate']
Topic 9: ['long', 'new', 'much', 'leave', 'idea', 'war', 'instead', 'spend', 'belong', 'battle']
Topic 10: ['even', 'father', 'knowledge', 'majority']
Topic 11: ['want', 'need', 'part', 'bring', 'meet', 'head', 'justice']
Topic 12: ['mother', 'other', 'turn', 'yet', 'mind', 'realize', 'son', 'similar', 'ignorance', 'wise']
Topic 13: ['find', 'try', 'rather', 'talk', 'weak']
Topic 14: ['kind', 'action', 'relative', 'opposite']
Topic 15: ['see', 'well', 'seem', 'let', 'share', 'watch', 'notice']
Topic 16: ['man', 'hand', 'kill', 'do', 'throw', 'foot', 'send', 'bind']
Topic 17: ['go', 'know', 'come', 'write', 'keep', 'one', 'easily', 'sign', 'certainly']
Topic 18: ['time', 'law', 'act', 'fear', 'serious', 'clear', 'observe', 'intend']
Topic 19: ['case', 'fight', 'look', 'heart', 'use', 'carry', 'attempt', 'story', 'start', 'dead']
```

A few observations from the relative frequency heatmap are:

- Topic 18 has higher relative frequency in Communism. Words like fear, serious, law and act are associated with it.
- Capitalism has considerably high relative frequency in topic 9, which has words like new, idea, war, battle, and spend. Surprisingly, words like money and price have not appeared in the topics.
- Empiricism is defined as knowledge based on experience & experimentation, and Rationalism is defined as knowledge based on reason or logic. Both schools have high relative frequency in Topic 5, which comprises of words like accept, reason, understand, opinion, and agree.
- Topic 3 seems to have significant relative frequency across almost all schools.

```
Perplexity:  -25.810532119509247
```

- Lower Perplexity indicates better model. We can also compute Coherence Score to measure interpretability of topics.

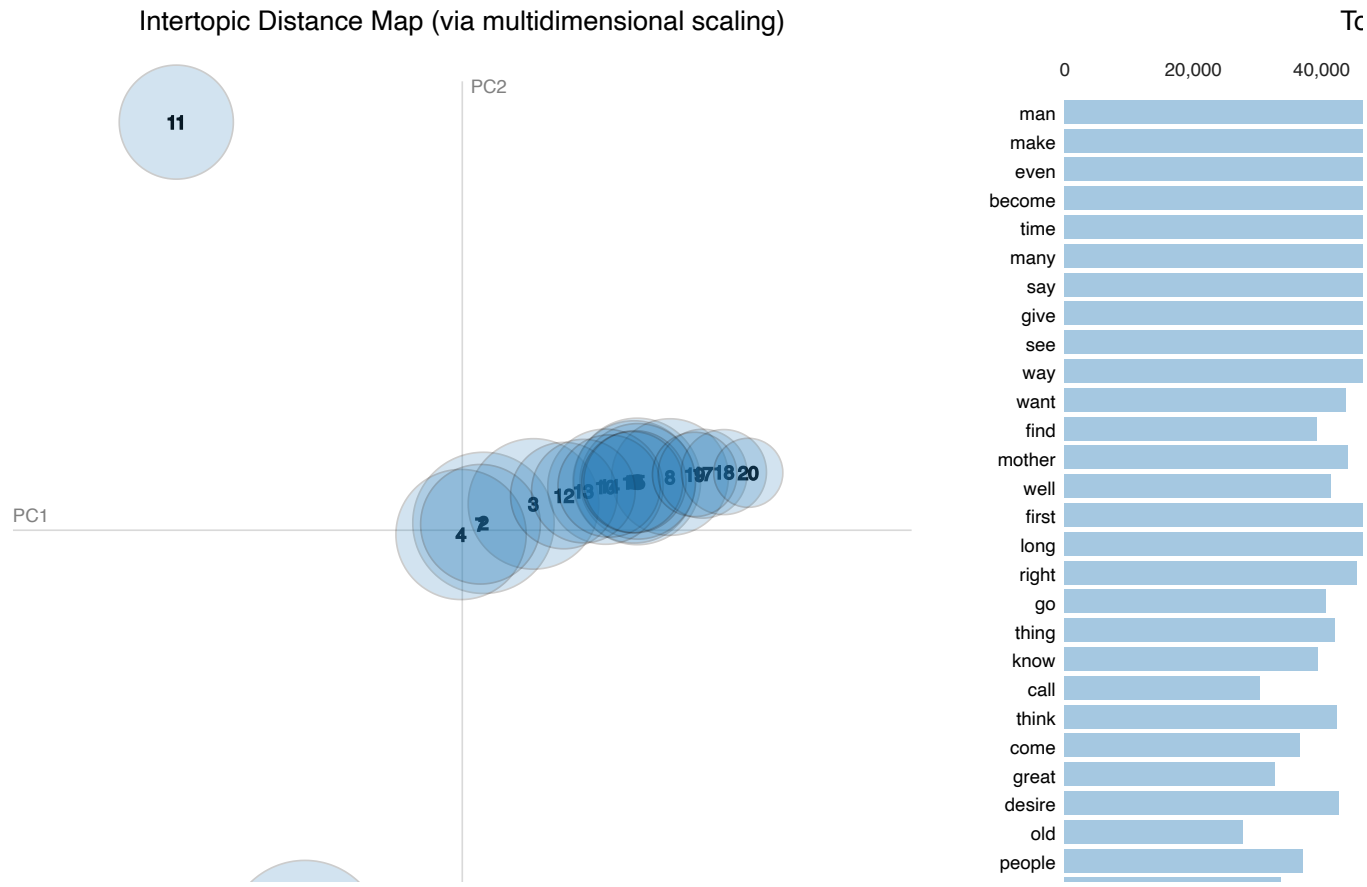
2.3 Interactive Visualization of topics

```
/Users/shubhamladdha/opt/anaconda3/lib/python3.8/site-packages/pyLDAvis/_prepare.py:246: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  default_term_info = default_term_info.sort_values(
```

Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric (2)

λ = 1



2.4 Further Exploration of School - Capitalism

We will create a small topic model on the capitalism texts to identify if words like money are prevalent in the texts. The steps to create the LDA model remain the same as above.

```
Topic 0: ['interest', 'mean', 'however', 'marginal_efficiency', 'volume', 'argument', 'exist', 'far', 'diff
erent', 'period']
Topic 1: ['level', 'make', 'great', 'effect', 'case', 'country', 'proportion', 'whole', 'take', 'reason']
Topic 2: ['wage', 'thus', 'factor', 'high', 'present', 'work', 'sense', 'possible', 'current', 'motive']
Topic 3: ['increase', 'change', 'investment', 'employment', 'quantity', 'labour', 'consumption', 'demand',
'consume', 'much']
Topic 4: ['rate', 'price', 'value', 'term', 'rise', 'state', 'commodity', 'depend', 'determine', 'equal']
Topic 5: ['theory', 'production', 'condition', 'expectation', 'reduction', 'cause', 'bring', 'correspond',
'think', 'long']
Topic 6: ['capital', 'income', 'stock', 'way', 'employ', 'first', 'advantage', 'provide', 'expect', 'reac
h']
Topic 7: ['money', 'give', 'fall', 'time', 'point', 'even', 'cost', 'good', 'amount', 'say']
```

- As indicated in the "capitalism" wordcloud above, we observe words like money and investment here. Majority of these words were absent from our topic model, when applied on the entire dataset.

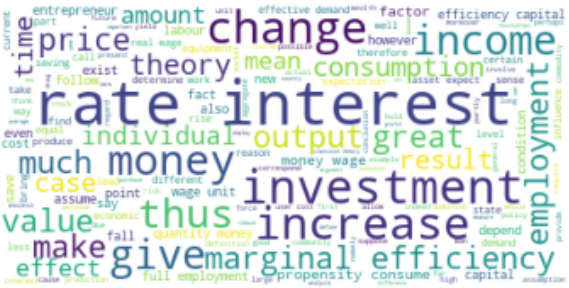
3. Analysis of Authors within same school - Capitalism

- Three authors have texts on Capitalism: Keynes (1936), Ricardo (1817), and Smith (1776). I would like to explore how the texts on capitalism have changed over 3 centuries.

```
Smith WordCloud:
```

Ricardo WordCloud:

Keynes WordCloud:



The three wordclouds have very different distribution of words. This emphasizes how "Capitalism" has evolved over time.

- Smith (1776): country, great and time are the most frequently occuring words.
- Ricardo (1817): tax, produce, country and value.
- Keynes (1936): interest, investment, income and rate.

4. Conclusion

1. Empiricism and Rationalism schools are quite similar in terms of topic word distribution & wordclouds.
2. Capitalism texts have evolved over time as indicated by their wordclouds acorss 3 centuries.
3. Sentence length of philosophy texts are shorter on average post 1850s compared to prior years.

5. Further Work

The above analysis gives an overview of a few schools based on their topic words and wordclouds. It would be interesting to further analyze sentiment of texts across schools and over time in the same school.