DS 3500

Advanced Programming with Data

Rachlin
NEU

Text Analysis Framework*

- Analyze individual texts
- Compare multiple texts.
- Text agnostic
- Extensibility — Be able to add features over time.
- The DS2500 Presidents: Pools would be about 10 lines of code!

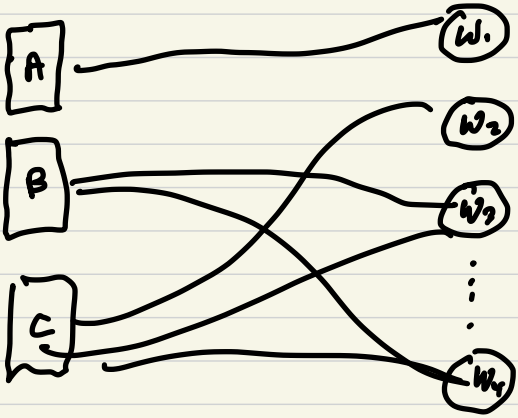
1. Possible Data Sources:
   - Gutenberg Texts
   - Political Speech
   - Tweet compilations (Biden/Trump)
   - Corporate Filings
   - Philosophical Treatises
   - Letters/Journals/Diaries
     - Civil war? North-vs-South
   - Blogs.
   - News articles

(Specific!)*

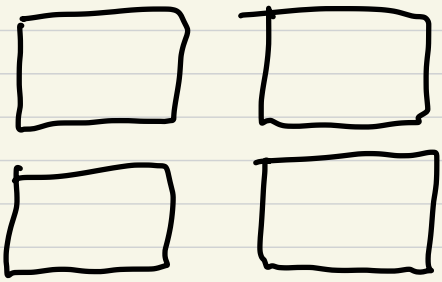
* In principle, the library should work on *any collection of texts.

2. Visualization Goals

1. Sankey Text → common word (or bigram) map.
   Top k words for each text OR provide a list of thematic words.

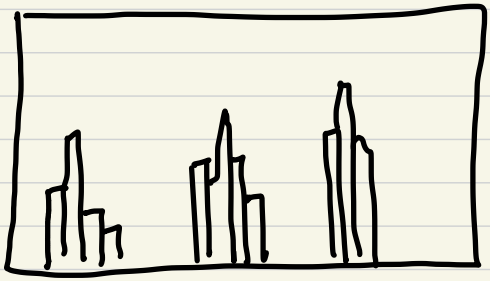


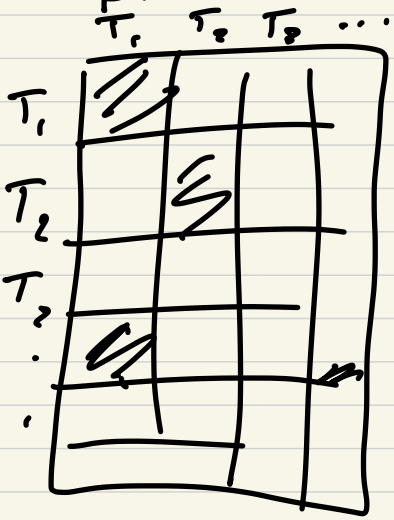
2. At least one visualization that uses subplots.



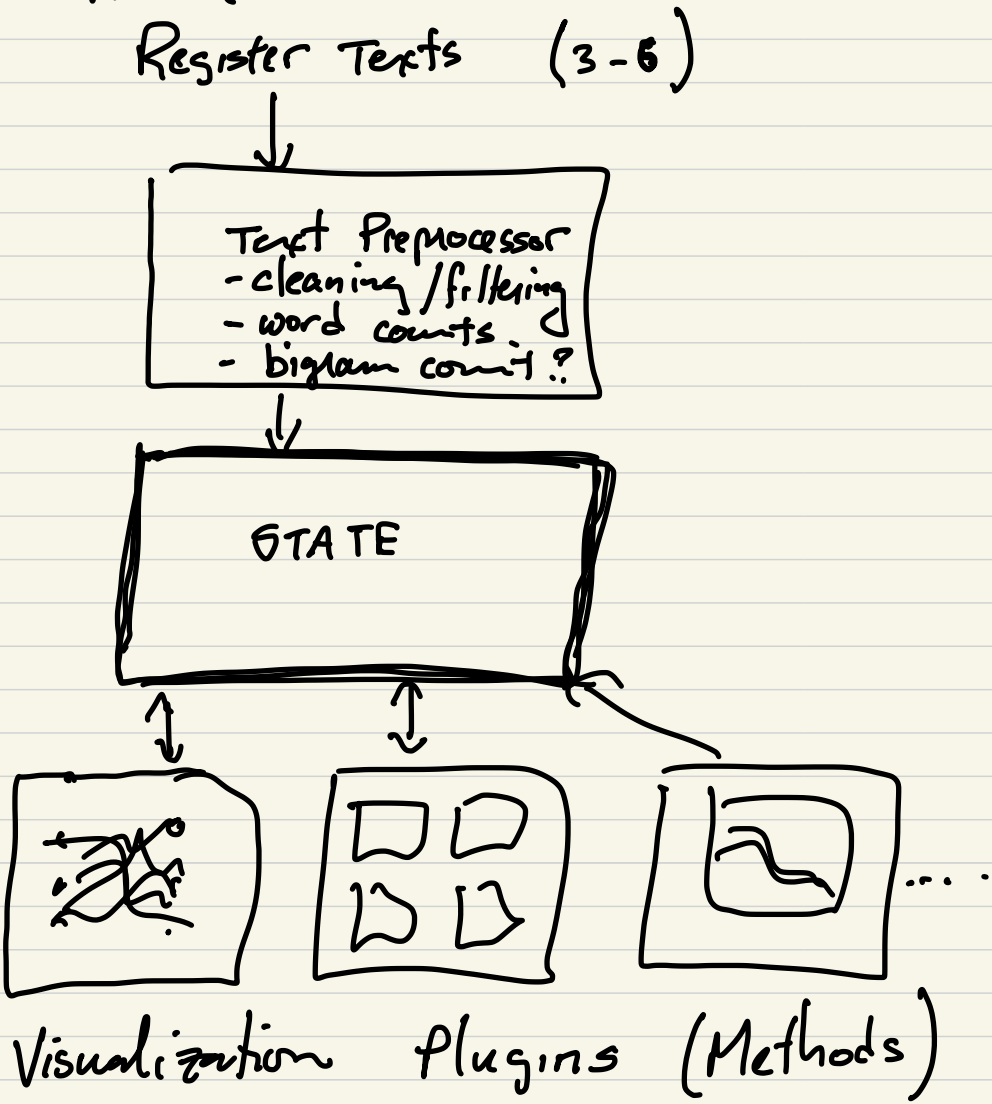
?

3. At least one visualization that does a comparative analysis on one plot.





3. Architecture.



Methods

- register_stop_words(filename)
- add_text(filename, label=...)
- common_words(label, k=10)
- text_word_sankey(word_list=[], k=20)
- zipf_analysis()
- word_clouds()