TEXT MINING TOOLS IN THE HUMANITIES
AN ANALYSIS FRAMEWORK

John Simpson (University of Alberta); Geoffrey Rockwell (University of Alberta); Ryan Chartier (University of Alberta);
Stéfan Sinclair (McGill University); Susan Brown (University of Alberta/University of Guelph);
Amy Dyrbye (University of Alberta); Kirsten Uszkalo (University of Alberta)

NEED FOR TOOL REVIEWS

There are many tools and methods. Most how-tos and reviews are directed at the sciences

Humanist sources for tool reviews mostly dried up in the 80s and 90s.
A new journal may help overcome this gap.

TEXTMIN
ING&VISU
ALIZATION
FORLITERA
RYHISTORY

inke

TOOL ANALYSIS FRAMEWORK

1. TEXTS

a. Standardized. The texts used must be the same across all the tools being analyzed to ensure that comparisons between tools are possible.

b. Open. The texts must be easily accessible and (ideally) free so that others can acquire them should they wish to carry out their own analysis.

c. Varied. More than one type of text should be included and these should be from a range of sources generally of interest within and across the humanities.

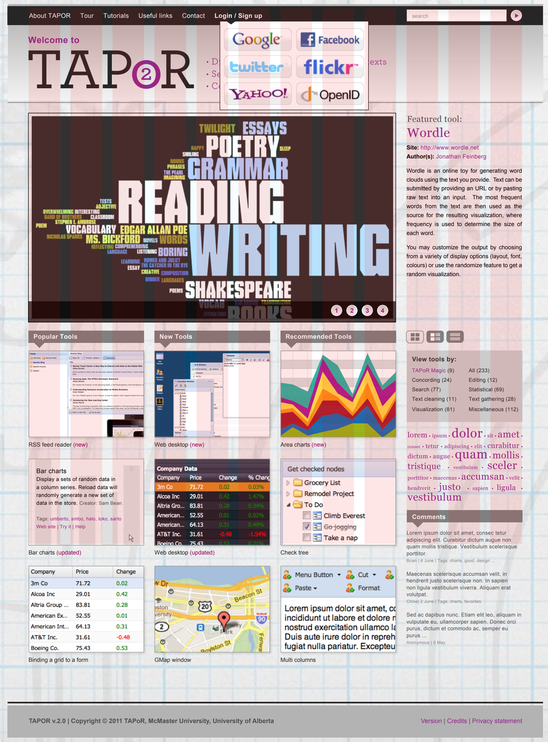
2. ANALYSIS

a. Standardized. The methods used should be the same across all texts, to the point where it should be (almost) automatic.

b. Replicable. The methods used should be available to anyone who wants them as recipes including both instructions and scripts.

c. Varied. More than one type of analysis should be included to ensure that a wide user base across the humanities will be interested in the results.

TAPoR 2.0



CONCLUSIONS

	Classification	Vocabulary & Visualization	Topic Extraction	Clustering	Network Analysis
Mallet	★★	★	★★★	★★	x
Python NLTK	★★	★★	★★	★★	★★
R	★★	★★★	★★	★★	★★
Gephi	x	x	x	x	★★★
Weka	★★★	★★	★★	★★★	x
Voyant	x	★★★	x	★	★★

★★★ Preferred Tool ★★ Capable
★ Capable, but with difficulty x No

LESSONS LEARNED

1. Text mining is an iterative process involving repeated visitation to three separate stages: Preprocessing; Analysis; and Clean-up.

2. More than one tool is often required to complete a text-mining task.

3. There is typically an inverse relation between the power of the tool and the graphical nature of the interface: to maximize performance get comfortable on the command line.

4. One benefit of command line tools is that the methods used are easily replicable. If you use a graphical interface take detailed notes of your process.

5. Don't go it alone. Find or build a community of like-minded text-miners that can support each other in discovering and exploring new tools.