Mining the Landscape
Unified Thinking in an Uncertain Age

Robert Blades
HIST 2809 (Historian’s Craft)
Dr. Shawn Graham

“Alas! that man's stern spirit e'er should mar/A scene so pure—so exquisite as this.”[1] This is a poetic recollection of Susanna Moodie’s arrival in Canada which she writes about in *Roughing it in the Bush* (1852). I chose to data mine and analyse Susanna Moodie’s *Roughing it in the Bush* with Henry David Thoreau’s *An Excursion to Canada* (1853) because they take place in relatively the same geographical area in Canada and around the same time period. Moodie and Thoreau may have differed in background and lifestyle but the Canadian landscape unified them. Using Voyant (formerly Voyeur) tools to data mine these texts, patterns in the writing, cultural & social impacts and, specifically, the view of the Canadian landscape were revealed to present an interesting view of public history.

Data mining is a tool for public history that leaves judgment to the viewer. Unlike a museum or website, the data is presented impartially and thus eliminates any aspects of presentation bias – it offers the solution to remain as objective as possible. Many problems can, and often do, arise with presenting history in the public sphere. Traditionally, websites, documentaries, and museums, to use a few examples, have had many problems with this. The largest problem is that of bias which, unfortunately, cannot be avoided. For example, a museum curator will always struggle as to what the best way of setting up an exhibit is (what to show and what to leave out). They will also struggle with the possibility of offending people with what they choose to show, or leave out of, the exhibit. A website can be much the same, although it has the ability to show all viewpoints and artefacts. Data mining presents the information from a textual document as it is. Therefore the viewer could only be offended or misunderstand as a result of solely their bias. Although data mining allows for misinterpretation, it is reduced greatly compared to the examples of public history above. Of course it is not free of problems, and new ones are realized. One technological drawback is that it only handles textual documents which

[1] “Roughing it in the Bush - Project Gutenberg,” http://www.gutenberg.org/cache/epub/4389/pg4389.html.

need to be in a digital format that can be easily read (i.e. html files). Many projects – such as Google Books and Project Gutenberg – are solving this problem through mass book scanning.[2] Although online and available to everyone with internet access, data mining tools have the ability to turn people off to using them. It looks technical, complicated, and possibly 'boring' to common person. For them, it seems an inefficient way to learn about history. To scholars this new way of academic reading also leads many to be skeptical about the conclusions.[3] The main challenge is convincing the use of these tools.

Procedural rhetoric "is a technique for making arguments with computational systems and for unpacking computational arguments others have created."[4] In other words, it is a way to express processes and to persuade their use.[5] Data mining tools, like Voyant do this by searching textual documents – from one to millions of words – to reveal patterns that may have been hidden before. It is a new way of scholarly academic reading, and dense and difficult texts that were once left alone can now be read and understood easily.[6] The advantage is instead of sampling certain texts to read, one can now read the whole text from a 'distance'.[7] In data mining, *distance reading* uses quantitative methods to search for patterns in textual documents, representing the text in a single image and unifying it to bring things into perspective.[8] For example, the word frequency in Voyant shows trends for one or many words in a simple graph.[9] Patterns that one might have subconsciously disregarded or missed before will now be in view.[10]

[2] Matthew G Kirschenbaum, "The Remaking of Reading: Data Mining and the Digital Humanities," 2.
[3] Ibid, 1.
[4] Ian Bogost, *Persuasive Games: The Expressive Power of Videogames* (Massachusetts Institute of Technology, 2007), 3.
[5] Ibid, 2.
[6] Kirschenbaum, "The Remaking of Reading," 2, 4.
[7] "Mining the Dispatch," http://dsl.richmond.edu/dispatch/pages/intro.
[8] Kirschenbaum, "The Remaking of Reading," 2, 4.
[9] Sinclair, Stéfan and Geoffrey Rockwell. "Word Trends." Voyant. 14 Nov. 2011 <http://beta.voyant-tools.org/tool/TypeFrequenciesChart/>
[10] Kirschenbaum, "The Remaking of Reading," 3.

Other tools that ‘offer’ the same procedural rhetoric like Voyant are FeatureLens and MALLET. FeatureLens looks at texts on several different levels to search for patterns and, like Voyant, shows the frequency of words and their trend throughout a text.[11] Data mining tools are in many ways the same across the board, though some, like MALLET, involve a bit of programming which creates a problem for those without a programming background.[12] The advantage is that one can change the program for specific needs.

The data mining results for the selected texts by Thoreau and Moodie revealed the similar word usage and frequencies, despite the differing backgrounds of the two authors. The top word for Moodie – and one of the top words for Thoreau – is *old*. The mid 1800s was a time when modernity was catching up with the rest of the world yet much of the vast Canadian landscape remained untouched. In such an unsure and new time of speed, industry, and science, the landscape would have seemed *old* to them. Hence, one reason the two authors often refer to the “old country” (meaning England; the way it used to be). Seen in context, for Thoreau the word *old* was used to describe the landscape as well as people; for Moodie, *old* is used to describe people as much as objects, but not much in regards to the landscape. On the contrary, Moodie uses *young* to describe naivety and hopefulness, not only as a property for people. Word usage like this reveals that Moodie uses poetry more so than Thoreau (though Thoreau had a way with words to combine facts and metaphors).[13] Moodie was only able to describe the land as an incomprehensible wilderness and her words were not as refined as Thoreau.[14] When something is difficult to describe, such as a new landscape, people resort to familiarizing (hence the similar

[11] Anthony Don et al.,“Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization” (in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Portugal, Association for Computing Machinery, 2007), 1-2.

[12] “MAchine Learning for Language Toolkit,” http://mallet.cs.umass.edu/.

[13] John S. Pipkin, “Hiding Places: Thoreau’s Geographies,” *Annals of the Association of American Geographers* 91, no. 3 (2001): 528.

[14] Elizabeth Thompson, “Roughing It in the Bush: Patterns of Emigration and Settlement in Susanna Moodie’s Poetry,” http://uwo.ca/english/canadianpoetry/cpjrn/vol40/thompson.htm.

words). Thoreau's similes and metaphors are also revealed by the word frequency. The word *like,* is used frequently by Thoreau and when seen in context it is apparent that the word is being used for metaphorical purposes: "Soon the city of Montreal was discovered with its tin roofs shining afar. Their reflections fell on the eye *like* [my italics] a clash of cymbals on the ear."[15] Moodie also uses the word *like* frequently in use for similes and metaphors. Thoreau's factual and scientific side is revealed by uses of words like *feet* to describe the size of many things.[16] He became much more scientific in his descriptions around this time in his life.[17] While it is true that a select group of scholars may know of Thoreau's metaphors and scientific influence, most people would not and they might not pick up on that through regular reading. Water is a main aspect of the Canadian landscape and therefore both texts. Thoreau describes the landscape with *falls*, *lakes*, and *rivers*. His most frequently used word was *St.* because he was near the St. Lawrence while in Canada. In fact, Moodie even describes the St. Lawrence as a great artery of the *heart* in Canada.[18] Along with the St. Lawrence, Quebec is mentioned a lot by both authors probably because the railroad was recently opened there and Thoreau's American home was close to the new railroad that ran through both countries.[19] Since he rarely left his home area, this shows the effect of modernism: that travel was relatively easy now and would uproot him from his homeland.[20]

Data mining tools have already been utilised by scholars and many projects use tools like Voyant. One project in particular, *Data Mining with Criminal Intent* looked at a site called *Old*

[15] "An Excursion to Canada – Wikisource," http://en.wikisource.org/wiki/An_Excursion_to_Canada.

[16] John S. Pipkin, "Hiding Places: Thoreau's Geographies," *Annals of the Association of American Geographers* 91, no. 3 (2001): 542-543.

[17] Pipkin, "Hiding Places," 528, 542-543.

[18] "Roughing it in the Bush - Project Gutenberg," http://www.gutenberg.org/cache/epub/4389/pg4389.html.

[19] "oldrailhistory.com," http://oldrailhistory.com/index.php?option=com_content&task=view&id=52&Itemid=80.

[20] Pipkin, "Hiding Places," 527, 528; "Roughing It in the Bush: Patterns of Emigration and Settlement in Susanna Moodie's Poetry."

Bailey which contained years of trial documents from 1674 to 1913, or 127 million words.[21] They carried out procedures similar this project, but their methods were more advanced (having a larger and technically trained team while data mining is important but not necessary). The team first used Zotero to allocate and organize their information and data mined the 127 million words using Voyant tools.[22] They also did experiments with historians and other scholars and had them use Voyant, finding that education on data mining tools is needed for it to be successful.[23] Their conclusions found interesting changes in social aspect of life and they revealed much of the millions of words and stories that could have not been done before in such time and with such carefulness.[24] Of course their conclusions may be different than mine but the impact is similar – that both projects recognize important, often unseen, patterns among the texts.

The similarity in writing between the two authors shows the feeling Canada presented to them. Their backgrounds and states of mind were much different, yet they wrote similarly and used many of the same words. Thoreau lived in America and spent much of his time there, he would be familiar with the landscape but Moodie was an immigrant and the landscape would be new to her.[25] Since Moodie was there working the land and was having trouble since emigrating, she held a more negative view[*] of Canada whereas Thoreau was there as an observer, describing the landscape and geography.[26] . The word usage revealed that Moodie's work exhibits 'bipolar' tendencies of emotion whereas Thoreau is more unified and refined. Given their backgrounds, this idea seems true. Moodie's description of the hard life of an immigrant tells a lot about the

[21] Dan Cohen et al., *Data Mining with Criminal Intent*, White Paper, 2011, 2, 3.
[22] Ibid, 5, 8.
[23] Ibid, 20-21.
[24] Ibid, 23-24, 25-26.
[25] Pipkin, "Hiding Places," 528.
[26] Laura Groening, "The Journals of Susanna Moodie: A Twentieth-Century Look at a Nineteenth-Century Life," *Studies in Canadian Literature*, http://lib.unb.ca/Texts/SCL/bin/get.cgi?directory=vol8_2/&filename=Groening.htm.

landscape and how tough it was to work in Canadian conditions. She would have had a relationship as close to the landscape as Thoreau, just not as refined in her words.

The presentation bias that is eliminated through data mining tools allows for the most objective view of public history that exists today, even though the scope may be narrow. The need for context and secondary sources, as with all history, is still very important. Voyant – and other data mining tools – may show us patterns but it does not reveal the author completely, as judgement is left to the viewer. Voyant revealed the many metaphors and a unified view of the Canadian landscape during the time period from different sources. The similarity in use and frequency of words between Thoreau and Moodie reveal patterns that show the strong roots to a cultural story, like a tree firmly planted in the Canadian landscape.

Bibliography

Books, Journals, and Papers

Bogost, Ian. *Persuasive Games: The Expressive Power of Videogames*. Massachusetts Institute of Technology, 2007.

Cohen, Dan, Frederick Gibbs, Tim Hitchcock, Geoffrey Rockwell, Jorg Sander, Robert Shoemaker, Stéfan Sinclair, et al. *Data Mining with Criminal Intent*. White Paper, 2011. http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final.pdf.

Don, Anthony, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. "Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization." In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. Portugal: Association for Computing Machinery, 2007.

Kirschenbaum, Matthew G. "The Remaking of Reading: Data Mining and the Digital Humanities." http://www.csee.umbc.edu/~hillol/NGDM07/abstracts/talks/MKirschenbaum.pdf.

Pipkin, John S. "Hiding Places: Thoreau's Geographies." *Annals of the Association of American Geographers* 91, no. 3 (2001): 527-545.

Voyant

Sinclair, S. and G. Rockwell (2011). Cirrus. Voyant. Retrieved November 14, 2011 from http://beta.voyant-tools.org/tool/Cirrus/

Sinclair, S. and G. Rockwell (2011). Words in the Entire Corpus. Voyant. Retrieved November 14, 2011 from http://beta.voyant-tools.org/tool/CorpusTypeFrequenciesGrid/

Sinclair, S. and G. Rockwell (2011). Word Trends. Voyant. Retrieved November 14, 2011 from http://beta.voyant-tools.org/tool/TypeFrequenciesChart/

Websites

"An Excursion to Canada – Wikisource." http://en.wikisource.org/wiki/An_Excursion_to_Canada.

Groening, Laura. "The Journals of Susanna Moodie: A Twentieth-Century Look at a Nineteenth-Century Life." *Studies in Canadian Literature*. http://lib.unb.ca/Texts/SCL/bin/get.cgi?directory=vol8_2/&filename=Groening.htm.

"MAchine Learning for LanguagE Toolkit." http://mallet.cs.umass.edu/.

"Mining the Dispatch." http://dsl.richmond.edu/dispatch/pages/intro.

"oldrailhistory.com."

http://oldrailhistory.com/index.php?option=com_content&task=view&id=52&Itemid=80.

"Roughing it in the Bush - Project Gutenberg."
http://www.gutenberg.org/cache/epub/4389/pg4389.html.

Thompson, Elizabeth. "Roughing It in the Bush: Patterns of Emigration and Settlement in Susanna Moodie's Poetry."
http://uwo.ca/english/canadianpoetry/cpjrn/vol40/thompson.htm.

[*] Moodie often describes the *night* in Canada and always says that it is cold and desolate. It is clear that Moodie was not prepared for such a landscape. Her work is less focussed on the land compared to Thoreau and her struggle is with her immigration.