

Text Sliding: Information Discovery with Intensely Integrated Text Analysis

Firstname LastName
Address
Address
Address
email@address.edu

Firstname LastName
Address
Address
Address
email@address.edu

Firstname LastName
Address
Address
Address
email@address.edu

Firstname LastName
Address
Address
Address
email@address.edu

ABSTRACT

There are numerous information analysis and discovery tools that allow users to view multidimensional data from different angles by selecting subsets of data, viewing those as visualizations, moving laterally to view other subsets of data, slicing into another view, expanding the viewed data by relaxing constraints, and so on. However, these tools operate over numerical and categorical data, but do not seamlessly operate over raw textual information with the same flexibility. In this paper we describe a text analysis and discovery tool that allows for highly flexible “slicing and dicing” (hence “sliding”) across a text collection. The goal of the tool is to help scholars and analysts discover patterns and formulate and test hypotheses about the contents of their text collections, midway between what traditional humanities scholars call a “close read” and the digital humanities “distant read” or “culturomics” approach. We illustrate the text sliding capabilities of the tool with two real-world case studies from the humanities and social sciences – the practice of literacy education, and U.S. perceptions of China and Japan over the last 30 years – showing how the tool has enabled scholars with no technical background to make new, important discoveries in these text collections.

Categories and Subject Descriptors

H.X [XXXXXX XXXXXXXX XXXX]: XXXXX; H.X [XXXXXX]: XXXXX — XXXXX

General Terms

Theory

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

Keywords

ACM proceedings, L^AT_EX, text tagging

1. INTRODUCTION

Current text analysis systems put computational linguistics and data visualization into the hands of users. Tools such as Jigsaw, IBM TextMiner, and SAS Text Analytics take techniques like text classification, named-entity recognition, sentiment analysis and summarization and build systems to automate their use. To make the results of these computational analyses interpretable, they display the outputs with interactive data visualization. These systems have made advanced text mining and visualization algorithms available to users without expertise in those areas.

However, text is distinct from other types of data because it is inherently meaningful and rich with associations. Even a small fragment of text does not stand alone, but is densely interconnected with other text through the *linguistic phenomena* that it contains. These are units of meaning that take surface form in text, such as words, phrases, relations between words, and literary devices, such as metaphor, sarcasm and allusion. To illustrate, consider this 13-word “slice” of text from Shakespeare’s “Romeo and Juliet”:

ROMEO:

Is she a Capulet?
O dear account! my life is my foe’s debt.

Surrounding this tiny slice is a swarm of other slices of text, each associated with the different linguistic phenomena in this slice. For instance:

- Each of the 11 distinct words in the above excerpt is a jumping-off point to all the sentences in which it occurs, to other words that mean the same thing, or to other words used near it.
- Each two-, three-, or n-word phrase can be associated with all the sentences in which it occurs
- Each grammatical relationship between words (such as “dear” being an adjective modifier of “account”) can be a link to other words that enter that relation (other adjectives describing “account”, or other things that are “dear”).

- For every instance of a literary device, such as the exclamation “dear account”, or the imagery of debt, can be associated with all other occurrences of it, or the different phrasings with which it surfaces in the play.

The more structure there is to text, the more kinds of associations are possible. In Shakespeare, we have metadata, such as speaker, act, and scene, so associations based on these dimensions also exist:

- The speaker, Romeo, can be associated with all the other speeches by him.
- The location within the play: Act 1, Scene 4 can be associated with the other scenes in that act, or the other speakers in that scene.

This network of associated slices is immediately apparent to us as humans. And to an analyst trying to make sense of an idea, some associations may be extremely meaningful. In fact, transitions and associations are central to the analytical process. People seeking knowledge from text are engaging in *sensemaking*. They do not follow a straight path from data input to analysis output, but meander between analysis, interpretation, exploration and understanding on different sub-collections of data. It is therefore important for text analysis systems to support not just algorithmic and visual analysis, but the transitions: slicing, filtering and exploration that lead from analysis to analysis, visualization to visualization, and finally to insight.

In this paper, we describe a text analysis tool that supports such transitions. It allows highly flexible slicing and dicing, as well as frictionless transitions (hence “sliding”) between visual analyses, drill-downs, lateral explorations and overviews of slices in a text collection. Our tool uses computational linguistics, information retrieval and data visualization, and enables scholars with no technical background to conduct analyses yielding concrete, useful and otherwise inaccessible knowledge.

1.1 Humanities and Social Sciences

The humanities and social sciences (HASS) are our motivation. In these fields, it is common for scholars to have hundreds, even thousands of text-based source documents of interest from which they extract evidence for complex arguments about society and culture. These collections (such as the set of all New York Times editorials about China, the complete works of Shakespeare, or the set of all 18th C. American novels) are difficult to make sense of and navigate. Unlike numerical data, they cannot be condensed, overviewed, and summarized in an automated fashion without losing significant information. And the metadata that accompanies the documents – often from library records – does not capture the varied content of the text within.

HASS scholars are an important area of focus for tool builders for another reason: low uptake. A 2012 study of computational tool use among these scholars showed that adoption was low despite an abundance of tools. The main culprit? Poor interface design due to lack of involvement with end users. In our research, we have tried to avoid this problem by taking an iterative, user-centered design approach. We collaborate with active HASS scholars working on text analysis problems of existing professional interest to

them, and let their needs, behaviors, and observations drive tool development.

We introduce the text sliding capabilities of our tool using three case studies. The researchers driving these studies used our tool to further their projects on literary history, education, and U.S-China relations. The projects investigated 1) How certain words dropped religious connotations and gained secular tones over the 18th Century in America, 2) How college students from diverse backgrounds remember and reflect upon literacy, and 3) How U.S. perceptions China and Japan responded to China’s rise over the last 30 years.

This paper is structured as follows: in the next section, we describe results from the field of sensemaking that motivate the need for “sliding” interactions between slices. After that, we explain the main ideas behind text sliding and show it in action with extended examples from case studies. Then, we describe related systems, and finally conclude with a discussion our results and future work.

2. ‘PAIN POINTS’ IN SENSEMAKING

Observational studies from the literature on sensemaking describe many problems analysts encounter while trying to make sense of text collections. These studies typically watch professionals such as government intelligence analysts [?], business analysts [?] market researchers [?] and academics [?] at work, attempt to categorize the actions they perform and identify common sequences of actions. Several models of the sensemaking process have emerged [?] These models attempt to explain what one would observe when watching analysts distill understanding from raw data, where “understanding” manifests itself as a summary, report, or presentation.

Pirolli and Card [1] identified “pain points” in three areas having to do with navigation and transitions between slices of text:

1. **Exploring** the collection by searching and filtering. Collections could be large difficult to navigate.
 - If associated slices were easier to see and to access, it might make exploration easier.
2. **Enriching**, which is the process of collecting a narrower set of items for analysis. This was a time consuming process involving going through documents returned by results, reading them to determine whether they were relevant or not, and placing them into groups.
 - It needs to be easier to select documents matching a term, quickly skim the text to determine relevance, and to collect the relevant text into a slice for later analysis.
3. **Exploiting**, which is the process of analyzing the collected information by manual schematizing, computational analysis, or visualization. Follow-up actions, such as drilling down to a finer set, noticing something interesting and starting a different analysis, or re-framing the question had a high cost.
 - This suggests that analysts should be able to explore associations and start new threads of inquiry with low overhead, and without losing their current state.

3. TEXT SLIDING

Central to the idea of text sliding are the twin concepts of *slices* and *views*.

3.1 Slices

In our tool, a slice is *a set of sentences* (future tools might define it differently: perhaps a set of documents, or a set of words, or some user-specified unit). We allow arbitrary slices, but for now, it is easier to think of slices as combinations of searches and filters. Slices are conceptually illustrated in Figure ?? . In that figure, we are assembling a slice corresponding to all of Romeo's mentions of the word "Capulet" in Act 1. We start with the whole collection, the collected works of Shakespeare, apply filters for **speaker = Romeo** and **act = Act 1** and a search for 'Capulet'. This leaves us with a slice containing only sentences that matches our criteria.

Slices do not stand alone. As described in the introduction, text is densely interconnected and associative. In our database, each sentence, and therefore each set of sentences, is associated with the following linguistic phenomena:

- Words
- Consecutive two-, three-, and four-word sequences
- Grammatical relationships, identified using a computational linguistics technology called dependency parsing, fully explained in []. In particular, we use the Stanford dependency parser¹ [].

Figure ?? illustrates some other slices associated with the slice from figure ?? . That slice only contains one sentence, 'Oh dear account!', but in our tool, it is associated with the following other slices through the words 'is', 'she', 'a' and 'capulet', grammatical relationships between those three words and other words in the collection, and the phrases 'Is she', 'she a', 'a capulet', 'is she a', and so on.

In our database, each sentences is associated with these different types of information

3.2 Case Studies

3.2.1 *Analyzing Literacy Autobiographies Written By Students*

3.2.2 *U.S. Perceptions of China and Japan 1980-2012*

3.2.3 *Secular vs. Religious 'Belief' in 18th C. American Fiction*

4. RELATED WORK

5. DISCUSSION AND FUTURE WORK

6. ACKNOWLEDGEMENTS

We sincerely thank [Firstname Lastname, Firstname Lastname, and Firstname Lastname] for their helpful feedback and comments on the use of our system in Case Study 3. We are also grateful to [Firstname Lastname] for his helpful advice and thought-provoking discussions throughout.

This work is supported by NEH grant HK-50011-12.

7. REFERENCES

- [1] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. International Conference on Intelligence Analysis*, volume 1, pages 2–4, MacLean, VA, USA, 2005.

¹Online demo at <http://nlp.stanford.edu:8080/parser/>