# 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 [**XXXXX XXXXXXX XXXX**]: XXXXX ; H.X [**XXXXX** ]: XXXXX —*XXXXX*

## Keywords

exploratory data analysis, text mining, information visualization, digital humanities



## 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, language is a unique form of expression, and text as data is distinct from numerical and categorical data. Text has both linear and hierarchical structure, its meaning is ambiguous given its representation, it has tens of thousands or hundreds of thousands of features, and frequencies of words are distributed via a power law. 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 (including synonymy and hyponymy) and literary devices, such as metaphor, sarcasm and allusion. To illustrate, consider this 13-word "slice" of text from Shakespeare's "Romeo and Juliet", where a slice is simply a set of sentences (not necessarily consecutive, like in this example):

```
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 can be thought of as a jumping-off point to all of the other sentences in which it occurs, as well as to sentences containing other words that mean the same thing, or to other words that tend to be used near it.

- Each two-, three-, or n-word phrase can be associated with every other sentence in which it occurs

- Each grammatical relationship between words (such as "dear" being an adjective modifier of "account") can be thought of as a link to other words that enter that relation (other adjectives describing "account", or other things that are "dear".)

- Each instance of a literary device, such as the exclamation "dear account", or the imagery of debt, can be associated with all other occurrences of this phrase, or the different phrasings with which this concept surfaces in the play.

The more structure there is to text, the more kinds of associations are possible. Shakespeare plays 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 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 two case studies. The researchers driving these studies used our tool to further their projects on education and U.S-China relations. The two projects investigated

1. How college students from diverse backgrounds remember and reflect upon literacy.

2. 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. MOTIVATION

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" usually manifests itself as a summary, report, or presentation.

Pirolli and Card [3] identified "pain points" in three areas having to do with navigation and transitions between slices while studying intelligence analysts working with large collections of text-based reports:

1. **Exploring** the collection by searching and filtering. Collections were often large difficult to navigate.

    - If associated slices were easy to see and to access, exploration might become 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 should be easy 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.

- It should be easy to explore associations and to start new threads of inquiry with low overhead, and without losing their current state.

# 3. TEXT SLIDING

The twin concepts of *slices* and *views* are central to text sliding. A slice is a set of sentences, and a view is a visual representation of the data in a slice: anything from a list of the sentences in the slice, to more complex linguistic processing combined with visual analytics. A slice is like a scientific specimen, a sample of some chemical compound. and views are the different lenses, apparatuses, and tests that reveal different information about the sample.

Through the richness of language, slices are *linguistically associated* with other slices (that also contain a particular word or phrase for example). If there is metadata accompanying the text, such as Act, Scene, and Speaker in Shakespeare's plays, there are also *metadata associations*. Finally, through the wide variety of visual analytics tool available, there can be many different ways of viewing and analyzing the data in a single slice. Text sliding makes all these associations accessible. In fact, we define text sliding as:

- Getting a different view of the same slice (lateral movement)
- Opening a new view of an associated slice (drilling down, following a new thread, or 'more like this')

## 3.1 Slices

In our tool, we define a slice as *a set of sentences*, although future tools might use different units of text, such as paragraphs, documents, or phrases. We allow arbitrarily-assembled slices, but it is easiest 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 sentences spoken by Romeo in Act 1 in which he mentions "Capulet". We start with the whole collection, the collected works of Shakespeare, apply filters for `speaker = Romeo` and `act = Act 1` and and also apply a search for `'Capulet'`. This leaves us with a slice containing only sentences that match our criteria, in this case, only one sentence:

```
{'Is she a Capulet?'}
```

### 3.1.1 Linguistic Associations

Slices do not stand alone. In our database, each sentence is indexed according to the following linguistic phenomena:

- Each word in the sentence, and its part of speech (noun, verb, adjective, etc.)
- Each consecutive two-, three-, and four-word sequence in the sentence
- Each grammatical relationship in the sentence. (These are identified using a computational linguistics technology called dependency parsing, fully explained in [1]. In particular, we use the Stanford dependency parser[1] [2].

[1]Online demo at `http://nlp.stanford.edu:8080/parser/`

By traversing the index both ways, we can quickly compute the associations for a slice. From a slice, we can query for all the words, phrases, or grammatical relations in the sentences in that slice, and from there, to all the other sentences that contain each particular item.

Figure **??** illustrates the 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.

## 3.2 Views

Views are visual representations of a slice. As shown in Figure **??**, views are window-like panels, and they contain the following components:

- Breadcrumbs identifying the slice
- A visualization of the data in the slice, one of:
  - A list of sentences
  - A list of documents that match the sentences in the slice
  - An interactive word tree [4] of the most common word in the slice, or the search term, if specified
  - Charts of how many sentences in the slice are distributed across various metadata categories
  - If the slice is a full document, a simple reading interface
  - If the slice is the result of a grammatical relation query, bar charts showing how often different words in the slice that appear in the relations
- Summary statistics of:
  - How many sentences within the slice match different metadata categories
  - The most frequent nouns, verbs, adjectives and multi-word phrases

## 3.3 Text-Sliding Interactions

Our goal is to make as many different types of associations and views available without overwhelming the user.

### 3.3.1 A Different View of the Same Slice

### 3.3.2 A New View of an Associated Slice

*Metadata Associations*

*Linguistic Associations*

# 4. CASE STUDIES

## 4.1 Literacy Autobiographies

## 4.2 U.S. Perceptions of China and Japan

# 5. RELATED WORK

# 6. DISCUSSION AND FUTURE WORK

## 7. 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.

## 8. REFERENCES

[1] D. Jurafsky and J. H. Martin. Chapter 13 syntactic parsing. In *Speech and language processing*, pages 427 — 459. Pearson Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2009.

[2] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proc. 41st annual meeting of the Association for Computational Linguistics*, volume 1 of *Association for Computational Linguistics '03*, pages 423–430, Sapporo, Japan, 2003. Association for Computational Linguistics. ACM ID: 1075150.

[3] 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.

[4] M. Wattenberg and F. B. Viegas. The word tree, an interactive visual concordance. *IEEE Transasctions on Visualization and Computer Graphics*, 14(6):1221–1228, 2008.