Exploratory Text Analysis and The Middle Distance

Marti A. Hearst
U.C Berkeley

Joint Work with Aditi Muralidharan
Collaborators: Bryan Wagner, Chris Fan, Rex Ganding
Sponsored by NEH HK50011
In common with most of van der Weyden's male portraits, Antoine is shown half profile, staring aloofly into the middle distance.
Outline

• Close Reading

• Distant Reading / Culturomics / Text Mining

• Exploratory Text Analysis: What is Needed & Related Work

• WordSeer: Case Studies

• What Remains to be Done
Foreground: The Close Read
“Close reading describes, in literary criticism, the careful, sustained interpretation of a brief passage of text. Such a reading places great emphasis on the particular over the general, paying close attention to individual words, syntax, and the order in which sentences and ideas unfold as they are read.”

-English Wikipedia, 6/4/2012
“The subtleties of the use of ‘you’ and ‘thou’ that have emerged … can seem, at worst, random or, at best, unfathomable. …
Text Mining in Moretti, “Distant Reading”, 2013

Shrinking Book Title Lengths Across the Years

Length of titles

Number of words

1720 1740 1760 1780 1800 1820 1840 1860
0 5 10 15 20 25 30 35 40

mean
median
Comparing Uses of Adjectives in Book Titles

Semantic fields in very short titles: the role of adjectives

- Blue bars: titles without adjective
- Red bars: titles with adjective

Categories:
- Fakeer, pirate, vampyre etc.
- Mother, daughter, father etc.

Text Mining in Moretti, “Distant Reading”, 2013
“Distant Read” Example: Ngram Viewer applied to *middle distance* vs. *middle ground*
She drank a cup of coffee.
They usually drink coke after work.
Sharon drinks champagne to celebrate.
Text Mining:
Statistical Text Analysis since at least 1990

She drank a **cup of coffee**.
They usually drink **coke** after work.
Sharon drinks **champagne** to celebrate.
She drank a cup of coffee.  
They usually drink coke after work.  
Sharon drinks champagne to celebrate.  
He drank concrete with dinner.

Word association norms, mutual information, and lexicography  
KW Church, P Hanks - Computational linguistics, 1990
Text Mining:
Statistical Text Analysis since at least 1990

She drank a cup of coffee. They usually drink coke after work. Sharon drinks champagne to celebrate. 
X He drank concrete with dinner.

Word association norms, mutual information, and lexicography
KW Church, P Hanks - Computational linguistics, 1990
Text Mining:
Statistical Text Analysis since at least 1990

Kenneth Church and Patrick Hanks

Table 5. What Can You Drink?

<table>
<thead>
<tr>
<th>Verb</th>
<th>Object</th>
<th>Mutual Info</th>
<th>Joint Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>drink/V</td>
<td>martinis/O</td>
<td>12.6</td>
<td>3</td>
</tr>
<tr>
<td>drink/V</td>
<td>cup_water/O</td>
<td>11.6</td>
<td>3</td>
</tr>
<tr>
<td>drink/V</td>
<td>champagne/O</td>
<td>10.9</td>
<td>3</td>
</tr>
<tr>
<td>drink/V</td>
<td>beverage/O</td>
<td>10.8</td>
<td>8</td>
</tr>
<tr>
<td>drink/V</td>
<td>cup_coffee/O</td>
<td>10.6</td>
<td>2</td>
</tr>
<tr>
<td>drink/V</td>
<td>cognac/O</td>
<td>10.6</td>
<td>2</td>
</tr>
<tr>
<td>drink/V</td>
<td>beer/O</td>
<td>9.9</td>
<td>29</td>
</tr>
<tr>
<td>drink/V</td>
<td>cup/O</td>
<td>9.7</td>
<td>6</td>
</tr>
<tr>
<td>drink/V</td>
<td>coffee/O</td>
<td>9.7</td>
<td>12</td>
</tr>
<tr>
<td>drink/V</td>
<td>toast/O</td>
<td>9.6</td>
<td>4</td>
</tr>
<tr>
<td>drink/V</td>
<td>alcohol/O</td>
<td>9.4</td>
<td>20</td>
</tr>
<tr>
<td>drink/V</td>
<td>wine/O</td>
<td>9.3</td>
<td>10</td>
</tr>
<tr>
<td>drink/V</td>
<td>fluid/O</td>
<td>9.0</td>
<td>5</td>
</tr>
<tr>
<td>drink/V</td>
<td>liquor/O</td>
<td>8.9</td>
<td>4</td>
</tr>
<tr>
<td>drink/V</td>
<td>tea/O</td>
<td>8.9</td>
<td>5</td>
</tr>
<tr>
<td>drink/V</td>
<td>milk/O</td>
<td>8.7</td>
<td>8</td>
</tr>
<tr>
<td>drink/V</td>
<td>juice/O</td>
<td>8.3</td>
<td>4</td>
</tr>
<tr>
<td>drink/V</td>
<td>water/O</td>
<td>7.2</td>
<td>43</td>
</tr>
<tr>
<td>drink/V</td>
<td>quantity/O</td>
<td>7.1</td>
<td>4</td>
</tr>
</tbody>
</table>

Word association norms, mutual information, and lexicography
KW Church, P Hanks - Computational linguistics, 1990
Text Mining: Semantic Relation Detection

• **Goal:** automatically augment a lexical database

• **Many potential relation types:**
  
  • ISA (hyponymy/hypernymy)
  
  • Part-Of (meronymy)

• **Idea:** find unambiguous contexts which (nearly) always indicate the relation of interest
Text Mining: Semantic Relation Detection

(S1) Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

(1a) \( NP_0 \) such as \( NP_1 \), \( NP_2 \) ... , (and | or) \( NP_i \) \( i \geq 1 \)

are such that they imply

(1b) for all \( NP_i, i \geq 1 \), hyponym(\( NP_i \), \( NP_0 \))

Thus from sentence (S1) we conclude

\( \text{hyponym}(\text{"Gelidium"}, \text{"red algae"}). \)
Lexico-Syntactic Patterns

(2) such NP as \{NP,} \ast \{(or | and)\} NP

... works by such authors as Herrick, Goldsmith, and Shakespeare.

\(\Rightarrow\) hyponym(“author”, “Herrick”),
hyponym(“author”, “Goldsmith”),
hyponym(“author”, “Shakespeare”)

(3) NP \{, NP\} \ast \{,\} or other NP

Bruises, ..., broken bones or other injuries ...

\(\Rightarrow\) hyponym(“bruise”, “injury”),
hyponym(“broken bone”, “injury”)
Text Mining:
Adding a New Relation to a Lexicon
Middle Distance: Exploratory Text Analysis
In life, we tend to focus on the endpoints, not the middle.

Chess:

Opening

Endgame

Middlegame

?
Ngram Viewer on the Chess Metaphor
WordSeer: Exploratory Text Analysis at the Middle Distance

• Goal: help scholars analyze literary text.

• Method: combine natural language processing, information visualization, and search user interface design.

• Support the “middle game”

• Midway between close read and distant statistics.

• Help with hypothesis formulation, verification, and refinement.

• New Goal: Help with Qualitative Coding!
WordSeer Motivation: What do Literary PhDs Want to Do?

- Interviewed 12 PhD students in literature and history departments
- Some key requirements:
  - Annotating while reading (12/12)
  - Looking for something specific in the text (12/12)
    - Different ways in which a concept is discussed (7/12), attractive, slim, tall
  - How common a concept is (3/12)
    “I’ve been trying to find out if anyone has ever been positive towards the ‘new woman’, being approving of her. I would love to see how often that phrase shows up during the 1960s” – P12
  - Syntactic patterns and regularities (2/12)
    “I look for grammatical patterns, clauses, or any kind of list. Like, “I went to the post office, the supermarket, and church”. – P1
Intelligence Analysts’ SenseMaking Loop
(Pirolli, Russell, Card)
Digital Humanities SenseMaking Loop
WordSeer Interaction Paradigm: “Text Sliding”

• A **slice** is a set of sentences, and

• A **view** is a visual representation of the data in a slice
  • Including a list of sentences, a visualization, data charts

• **Text sliding** moves from one view of a slice to another, including:
  • **Showing** a different view of the same slice, or
  • **Narrowing** (by selecting metadata or other filters), or
  • **Broadening** (by removing filters), or

• **Creating** a new slice (moving laterally) including:
  • Slicing on one of the words from the context
  • Slicing on a related word from a word in the context.
  • Slicing based on grammatical relations of a word in the context.
WordSeer Video Demo

Shakespeare's Works

Frequent Words

PHRASES
- good lord (391)
- sir john (158)
- come come (143)
- let see (124)
- fare well (119)
- give leave (91)
- enter king (90)
- know st (88)
- good master (82)
- give hand (81)
- tis true (78)
- " say (67)
- come let (64)
- king henry (59)
- god save (57)
- enter messenger (57)
- come sir (56)
- make haste (54)
- well said (53)
- draw sword (51)

NOUNS
- lord (2089)
- thy (1904)
- man (1809)
- sir (1724)
- thee (1614)
- thou (1344)
- love (1249)
- king (1225)
- father (972)
- heart (921)
- time (908)
- death (823)
- art (810)
- life (797)
- hand (764)
- exit (694)
- day (664)
- ti (636)
- ay (635)
- lady (625)

Metadata

ACT
- Act 4 (37)
- Act 3 (37)
- Act 2 (37)
- Act 1 (37)
- Act 5 (37)
- Induction (1)

LINES

SCENE
- Act 2, Scene 1 (37)
- Act 3, Scene 2 (36)
- Act 5, Scene 1 (37)
- Act 1, Scene 1 (37)
- Act 1, Scene 2 (35)
- Act 3, Scene 1 (37)
- Act 4, Scene 1 (37)
- Act 5, Scene 2 (31)
- Act 2, Scene 2 (34)
- Act 4, Scene 2 (35)
- Act 4, Scene 3 (31)
- Act 2, Scene 3 (29)

Sets

PHRASE SET
- {beauty words} (3935)

SENTENCE SET
- {Canines in Shakespeare}
Related Work:
Exploratory Text Analysis Tools

- EDA on text has different demands than on DBMS’s
  - Unstructured information
  - Very high dimensionality of text.
- It was big in the ‘90’s, but is less active now.
  - Most systems focus on recognizing entities and showing relations among them
  - Very few focus on the details of language behavior.
Feldman et al. PAKM’98

Figure 2. Document Explorer architecture.
Feldman et al. PAKM98
Feldman et al. PAKM98
Figure 12 – Context Graph (Companies in Context of “Joint Venture”)
Figure 13 - Category Graph (relationship between People, Brokerage Houses and Computer Companies in Context of “merger”)
Stasko et al., Jigsaw, Information Visualization 2008
Stasko et al., Jigsaw, Information Visualization 2008

Figure 2

(A) The Control Panel. (B) The List View. Selected entities are shown in yellow and connected entities are indicated by the joining diagonal lines and the orange shading. Darker shading represents stronger connections to the selected entities. (C) The Graph View. Documents are white rectangles and entities are circles colored by type. Edges connect documents to the entities they contain. (D) The Scatter Plot View. Each axis enumerates a list of entities. Diamonds in the center indicate documents containing particular pairs of entities, one from each axis at the relative $x$, $y$ position.

For some tasks it is useful to consider all entities of a specific type, for example to find out what place has the highest frequency in a document collection. The button 'Add all' allows the user to add all entities of the selected type to a list. If a list of entities is too long for all the items to fit in the view, scrollbars appear to aid navigation. In this situation, many connected items may not be visible in the view. Thus, the List View also provides a mode in which all selected and connected entities are automatically moved to the top of the list.

Graph View

The Graph View, illustrated in Figure 2(C), represents documents and their entities in a traditional node–link graph/network visualization common in many other systems. Entities are depicted as labeled circles colored according to their type. Documents are represented as white rectangles. The edges from documents to the entities they contain are shown as well.

Following the query-based exploration approach of Jigsaw, the Graph View does not automatically draw all the documents and entities as one large network. We thought that a layout of such a large network would be overwhelming and difficult to understand, and thus would not be as helpful to the analyst in our context. Instead, Jigsaw's view is incremental. Display events (triggered from a query or another view) place documents and entities on the display, and then mouse clicks on items can expand or collapse their connections.
Figure 3

(A) The Document View. The document discussed in the second section is displayed with entities highlighted and colored by type. (B) The Calendar View. Small diamonds represent documents or entities connected to the date in which they appear. (C) The Document Cluster View. Each small rectangle represents a document in the collection, and can be colored and grouped based on entities that do or do not appear in it. (D) The Shoebox. The group, hypothesis, sentences, and link feature are used to organize evidence.

The automated entity identification process in Jigsaw typically is not perfect. Some entities may not be identified at all, some may have an incorrect entity type assigned, and some identified simply may not be entities. The Document View provides functionality to correct such errors. Identified entities can easily be removed or their assigned type changed to another existing type or to a new entity type by right clicking on the identified entity and selecting the relevant command from a pop-up menu. If an entity was missed during the identification process, it can be interactively designated as an entity by selecting the text and assigning an entity type. All these operations to modify entities can be applied to a single document or across all documents in the collection. This allows the user to make local corrections (e.g. the entity 'Paris' is always identified as a location but in one document it is a name) as well as global changes (‘BMW’ is not identified as an entity and should be added as a new entity type ‘car’ in all documents).
This section shows examples of using the search operators. The **AND** operator

From the Facet Navigation pane (Figure 5-17), we select the **Product** facet and the keyword **pine juice**, and then we click the **AND** operator. The results are limited to show only the pine juice-related documents.

**Figure 5-17  Selecting the Product facet and pine juice and clicking the AND operator**
The text miner application provides the following views to assist in text mining. These views are generated dynamically based on your query and facet selections.

- **Documents view**: Shows a list of documents that match your query.
- **Facets view**: Shows a list of keywords for a selected facet.
- **Time Series view**: Shows the frequency change over time.
- **Trends view**: Shows sharp and unexpected increases in frequency over time.
- **Deviations view**: Shows the deviation of keywords for a given time period.
- **Facet Pairs view**: Shows the correlation of keywords from two selected facets.
- **Connections view**: Shows the correlation of keywords from two selected facets in a graphical way.
- **Dashboard view**: Shows a configured dashboard layout with one or more graphs or tables in a single view.

*Figure 6-1  All views available in the text miner application*
The Connections view shows a graphical view of the relationship between keywords or subfacets within selected facet pairs. This view is another representation of the correlation of keywords within the selected facet pairs. The keyword is represented by a node, and the link between the nodes shows the correlation value between the two keywords.

Figure 6-20 shows an example of the Connections view when you select the Product facet and the Verb facet.

In the Connections view, you can determine the highly correlated keyword pairs by focusing on the size of the node, link color, and link length:

- **Node size** represents the frequency of the keyword or subfacet. The larger node size represents a higher frequency count in the entire document corpus.

**Key**

- Blue = Product
- Orange = Verb

**Frequency**

- Black = High
- Gray = Low

**Correlation Amount:**

- Low
- High

Depending on the browser window size or your operation, the representation in the Connections view can vary even though the same facet pairs are selected.
Mueller, WordHoard, Shakespeare, 2008
Sinclair & Rockwell, Voyant/Voyeur, TLLO 2009
Cohen et al., Voyeur, criminalintent.org 2011
(only works on one text collection)
Cohen et al., Voyeur, criminalintent.org 2011
Don et al., FeatureLens, CIKM 2007

Applied to Gertrude Stein’s *The Making of Americans*
What WordSeer Supports that is Missing from Existing Tools

- Seeing analyzed text in context
- Fluidly building rich queries from context
- Easily create new classes of concepts from the text
- Flexible, context-based syntactic analysis
- Flexible, multi-way comparisons
Case Studies

1. Literary scholar studying American literature’s reaction to China’s rise

2. English Composition Educator analyzing student essays

3. My colleagues and my analysis of use of a MOOC forum
China Scholar: Concept Trend

• Interested in: How U.S. perceptions of China and Japan responded to China’s rise over the last 30 years.

• Created a wordset with growth terms
confirmed his intuitions about China’s rise by checking whether growth-related words became more frequent over time in editorials about China.
China Scholar: Compare Trends

- Interested in: how China talked about over time: cold war rhetoric

Figure 8: Word Sets in WordSeer.

Figure 9: The \( \text{growth} \) Word Set used as a Word Frequencies query.

5. CASE STUDY: U.S. PERCEPTIONS OF CHINA JAPAN

As mentioned above, C.F. is a literary scholar at [university’s] English department. His work on American literature’s reactions to China’s rise draws upon a set of historical observations about the “rise of China” that are broadly accepted by historians and cultural historians. Literary scholars typically allow their claims to rest on observations made by field experts like historians and sociologists, or on their own inductive reasoning, but C.F. wanted to verify some of those observations by gathering as much empirical evidence for them as possible.

Using the Lexis/Nexis database, C.F. collected New York Times editorials from 1980 to 2012, limiting his collection to editorials tagged with subjects “China” or “Japan.” This left a set of 5,715 editorials, which we imported to WordSeer. Each editorial was associated with three pieces of metadata: year, month, and country (either China or Japan).

After a few weeks of face-to-face meetings, C.F. gradually became comfortable using WordSeer on his own. He used WordSeer to find evidence for discussion of the rise of China as described above, and for three other hypotheses, two of which are described below.

5.1 1980’s: China Insignificant Except for Cold War Strategy

While comparing the most frequent adjectives for the three decades, C.F noticed a drop in cold war words. “Soviet” went from a count of 1029 in the 1980’s to only 80 in the 2000’s, and “communist” went from 284 to 112. To investigate the drop in more detail, he used the Word Menu to quickly open up word frequency plots for these two words over time, and filtered them to just the editorials about China. Figure 10 shows a plot of these words together:

Figure 10: The frequencies of ‘communist’ (blue) and ‘soviet’ (orange) in editorials about China over time.

Plotting the terms together (Figure 10) added depth to his initial calculations. The plot shows the dominance, and equally dramatic drop, in cold war mentions over this time period. In the early 1980’s, almost 11% – more than 1 in 10 sentences in those editorials – mentions one of these the cold war terms. By the 1990’s, however, this association is down to a trickle.

An exploration of the “grammatical neighborhood” around ‘China’ helped C.F. find yet more evidence for this idea. This time, it was through a distinctive rhetorical device. As Figure 6 shows, clicking on the word “China” opens up a menu with search options. These options act as quick previews of different grammatically-related slices: they show the frequencies with which “China” occurs in different grammatical positions in the text.
China Scholar: Mid-90s Onward: conjunction(China, U.S.)

- China joined the WTO in 2001; this is when China-US relations are thought to have become more inter-dependent.

- Idea: use a grammatical search to find interdependence: conjunction, e.g.:

  “The **United States**, the world’s top energy guzzler, **and China**, with the world’s fastest-growing energy thirst . . .” (April 2006). “
China Scholar: Mid-90s Onward: conjunction(China,U.S.)

- Conjunctions much more frequent after 1994
China Scholar: Explore the “Grammatical Neighborhood”

- Interested in: how China is talked about: anything interesting or unexpected?

Grammatical relationships are identified using the Stanford dependency parser\[9\] which extracts many kinds of relationships; some of the more easily understood ones include noun compound, where two nouns come together to signify a new concept, adjective modifier where an adjective describes another word, and direct subject in which a word is the agent of a verb.

Each view automatically presents the most common nouns, verbs, adjectives, and phrases (Figure 5), along with their counts (providing a query preview [5] in a panel at the bottom.

Figure 5: The most frequent phrases, nouns, verbs, and adjectives in the New York Times editorials for Case Study 1.

Individual words are jumping-off points. They can be acted upon wherever they appear via the Word Menu (Figure 6) which enables lateral movement. Any time a user sees a word, they can follow up on it by examining the grammatical relations in which it occurs, seeing related words, and creating visualizations of the slice of sentences that contains the word, as well as the slices containing various relationships to other words.

Figure 6: The word menu for ‘China’. After selecting China ▷ Search ▷ noun compound modifier, the noun-compound relationship “China card” stood out to C.F. The related words option in the Word Menu shows the nouns, verbs, and adjectives that co-occur most frequently with the clicked-on word. For example, if we click on ‘Japan’ and open up the related words (Figure 7) the pop-up shows the words that co-occur most frequently with ‘Japan’ in this collection. Each of these related words can be clicked in turn, opening up a new Word Menu. These menus have the additional option to ‘See co-occurences’, as shown in the new Word Menu for ‘exports’. Selecting that option opens up a new view showing just those sentences in which the two words appear together (in this case, ‘Japan’ and ‘exports’).

The word menu reduces friction in both discovery and search. It only takes one menu click to discover that ‘exports’ occurs frequently with ‘Japan’, and only one more to see all the sentences in which ‘Japan’ and ‘exports’ are mentioned together.

4.4 Custom Slices with Sets

Searches and filters are useful, but cannot always express specific analysis goals. WordSeer therefore allows users to construct custom slices through Word-, Sentence-, and Document Sets. These custom slices behave like any other slices, which means that they can be summarized in views, analyzed, filtered and searched. But they are more powerful that other slices because they also behave like metadata, transforming them into categorical filters.

Word Sets are well-illustrated by an example from Case Study 1. One concept of interest in this study was “growth”. The scholar wished to confirm his intuitions about China’s rise by checking whether growth-related words became more frequent over time in editorials about China. First, he created a new Word Set and typed in some growth-related words “growing, develop, developing, grow, rise, rising” (Figure 8(a)). The result was a Word Set representing a new slice of sentences, those containing at least one of those words. After the Word Set is created, the entire user interface responds to its presence. The search box now shows a drop-down option for the set (Figure 8(b)). The Word Menu shows the option to add a new words (Figure 8(c)) and the metadata overviews (Figure 8(d)), previously restricted to pre-defined categories, now show this new “category”, and allows C.F. to filter based on it.

To verify that China was indeed described as rising, C.F. selected the \{growth\} Word Set as his search query, and opened a Word Frequencies view with the country = China filter. The resulting visualization is Figure 9, which shows almost a doubling of the frequency of these words in editorials about China over the 30-year period from 1980 to 2012. Satisfied that WordSeer was capable of reproducing this widely accepted fact, he was able to move on to deeper questions.

For sentences and documents, the idea is the same. Users can hand-pick collections of sentences from the reading view, and from search results view, or collections of documents from the document search results view. Once created, all these sets can be overviewed and analyzed like any other slice, and additionally used as filters.

This section has introduced the core text sliding features
Literacy Essays

Figure 2 – Most Frequent Adjectives

Course data

Own Able
Literacy Essays

Although  Though  While  However
Forum Analysis
I haven't looked into this but I will venture to say that berkeleyX is not part UC berkeley officially - as edX is a non-profit I assume that berkeleyX is as well and o/c UC is not - in any case the certificate is unofficial - it is issued by the professors personally and not by UC anyone?

I don't think anyone is "under the illusion" that they are receiving college credit for this course work nor do I think many of us care.

This is a new medium for education so we'll all get smarter with experience (including how to post "general" questions ;-) How's this one, "Has anyone found a good alternative to Wine emulator for getting Kindles to run on Linux?"

IMHO, such information is more appropriately placed in the help FAQ for the edX site.

The information that Mr. Jenze and Kravishy have provided is nowhere to be found in edx.org help section, nor is it easily found anywhere else on the "minimalistic" site.

@Mr-Jenze Please read post with information from my direct e-mail correspondence with edx.

as a rule we should try not to reveal the answers to assignments but as this particular quiz has unlimited
What’s Missing?

• Integration of thesauri / taxonomies

• More sophisticated syntactic pattern matching

• Tools to learn from examples

• Tools to check “the rest” of the hypothesis
Conclusions

• “Big data” is often not explored in a sophisticated way
  • Instead, perhaps we should focus on medium-sized, motivated slices to gain more insight.

• Interfaces and analysis should support the “middle game”
  • Midway between close read and distant statistics.
  • Help with hypothesis formulation, verification, and refinement.
Thank you!
Exploratory Text Analysis and The Middle Distance

Marti A. Hearst
U.C Berkeley

Joint Work with Aditi Muralidharan
Collaborators: Bryan Wagner, Chris Fan, Rex Ganding
Sponsored by NEH HK50011