NLP and Visualization for Digital Humanities

Guest Lecture: Dr. Christopher Collins
University of Ontario Institute of Technology
http://vialab.ca
Outline

• What is Digital Humanities?
• Problems and Challenges for DH Research
• Using Natural Language Processing to Solve DH Problems
• Exploratory Analysis of Alice in Wonderland with Demos
Outline

• What is Digital Humanities?

• Problems and Challenges for DH Research

• Using Natural Language Processing to Solve DH Problems

• Exploratory Analysis of Alice in Wonderland with Demos
What is Digital Humanities?
What are the humanities?

The humanities can be described as the study of how people process and document the human experience. Since humans have been able, we have used philosophy, literature, religion, art, music, history and language to understand and record our world.

These modes of expression have become some of the subjects that traditionally fall under the humanities umbrella. Knowledge of these records of human experience gives us the opportunity to feel a sense of connection to those who have come before us, as well as to our contemporaries.

http://shc.stanford.edu/what-are-the-humanities
Goals of the Humanities

• To think about the past in the present in order to take action in the future…

• To examine the past critically, thinking about what you wish to emulate, what you wish to avoid and what you wish to do that may be new…

• Establishing shared spaces beyond/ complementing our tribal/ ethnic/ religious identities…

-- Gregory Crane
What is Digital Humanities?

Digital humanities (DH) is an area of scholarly activity at the intersection of computing and the disciplines of the humanities.

The nature of this activity ranges broadly, from the practical, such as digitizing historical texts, to the philosophical, such as reflection on the nature of representation itself.

This spectrum of activities is reflected in definitions of the field that range from it being a collection of methods to being a distinct epistemology and a kind of science.

Within this variation, a distinctive feature of digital humanities is its cultivation of a two-way relationship between the humanities and the digital: the field both employs technology in the pursuit of humanities research and subjects technology to humanistic questioning and interrogation, often simultaneously.
What is Digital Humanities?

“Interdisciplinary, collaborative, design-aware, iteration-positive, oriented toward accessibility (in many senses).

Applies digital technologies alongside more conventional ones to cultural and social questions & has the potential to enable new questions through the use of those technologies.”

-- Margarita Nafpaktitis

“Work that engages with the development of digital tools or uses these same tools within the traditional scope of Humanities.”

-- Henriette Roued-Cunliffe

http://whatisdigitalhumanities.com/
What is Digital Humanities?

You might be a digital humanist if:

• You use emerging technologies to rethink what "the humanities" is all about.
• You collaborate with people across disciplines and backgrounds to explore innovations in academia
• You are willing to experiment with new pedagogies and research approaches made possible by new technologies
• You are not afraid to work closely and collaboratively with your friendly neighborhood technologist
• You are willing to take a do it yourself approach to creating the tools and technologies you need for your teaching and/research
• You use technology in a way that disregards traditional boundaries between disciplines and hierarchies

-- Patrick Murray-John
Collaborations between CS and…

• Disciplines
  • Classics
  • History
  • Linguistics
  • Literatures, Cultures & Languages
  • Music
  • Philosophy
  • Religious Studies
  • Theater & Performance Studies

• Associated
  • Journalism
  • Political Science
  • Legal Studies
Digital Humanities Collaborations

- Interdisciplinary research is very challenging
- It is crucial to have an open communication
  - ... about the used concepts and terminology
  - ... and the process of tackling the overall research question
- Divide and conquer: Smaller self-contained tasks are easier to solve
- The research question has to be:
  - clear (allows only one interpretation)
  - focused (concentrates on specific aspects)
  - concise (based on rigorous conceptualization)
  - complex (includes multiple layers)
  - arguable (addresses a research gap)
Digital Humanities Collaborations

JOHN W. TUKEY*

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.
Outline

• What is Digital Humanities?

• Problems and Challenges for DH Research

• Using Natural Language Processing to Solve DH Problems

• Exploratory Analysis of Alice in Wonderland with Demos
Unify Close and Distant Reading

Koch et al., VarifocalReader — In-Depth Visual Analysis of Large Text Documents
"Prosthetic Reading" vs. "Slow Analysis"

Mehta et al., MetaTation: Augmenting Free-Form Annotations to Bridge Close and Distant Reading of Poetry
Supporting Serendipitous and Guided Discovery

Thudt et al., The Bohemian Bookshelf: Supporting Serendipitous Book Discoveries through Information Visualization
Supporting Serendipitous and Guided Discovery

Alexander et al., Serendip: Topic Model-Driven Visual Exploration of Text Corpora
Linking Outside Resources into the Analysis

Cho et al., VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History
Aesthetic Experiences

McCurdy et al., Poemage: Visualizing the Sonic Topology of a Poem
Annotation of Resources

Faisal Cheema et al., AnnotateVis: Combining Traditional Close Reading with Visual Text Analysis
Outline

• What is Digital Humanities?
• Problems and Challenges for DH Research
• Using Natural Language Processing to Solve DH Problems
• Exploratory Analysis of Alice in Wonderland with Demos
NLP for DH

- Word counting
- POS tagging
- Stop word removal
- Keyword extraction
- Named entity recognition
- Lexical distribution
- Clustering documents in a corpus
- Comparing documents / document with corpus
- Word sense disambiguation
- Parsing
- Topic modelling
- Edit distance (variation)
- Pronunciation / sonic patterns
- Extracting sentiment and other features
- …
NLP for DH

- Word counting
- POS tagging
- Stop word removal
- Keyword extraction
- Named entity recognition
- Lexical distribution
- Clustering documents in a corpus
- Comparing documents / document with corpus
- Word sense disambiguation
- Parsing
- Topic modelling
- Edit distance (variation)
- Pronunciation / sonic patterns
- Extracting sentiment and other features
- …
Outline

• What is Digital Humanities?

• Problems and Challenges for DH Research

• Using Natural Language Processing to Solve DH Problems

• Exploratory Analysis of Alice in Wonderland with Demos
Tools – Lexical Data

- AntConc
- Wordandphrase.info
- Word Wanderer
- Text Arc
- Arc Diagram
- Word Tree
- Phrase Net

Word counts, concordances

Word relationships, Structure
Since Carroll was a mathematician at Christ Church, it has been suggested that there are many references and mathematical concepts in both this story and also in Through the Looking-Glass; examples include:

- In chapter 1, "Down the Rabbit-Hole", in the midst of shrinking, Alice waxes philosophic concerning what final size she will end up as, perhaps "going out altogether, like a candle."; this pondering reflects the concept of a limit.
- In chapter 2, "The Pool of Tears", Alice tries to perform multiplication but produces some odd results: "Let me see: four times five is twelve, and four times six is thirteen, and four times seven is—oh dear! I shall never get to twenty at that rate!" This explores the representation of numbers using different bases and positional numeral systems: \(4 \times 5 = 12\) in base 18 notation, \(4 \times 6 = 13\) in base 21 notation, and \(4 \times 7\) could be \(14\) in base 24 notation. Continuing this sequence, going up three bases each time, the result will continue to be less than \(20\) in the corresponding base notation. (After 19
Word Counts

• How many times does each (stemmed?) word appear?
Word Counts

Source: CNN, National Archives, NaNoWriMo, overviewbible.com, Open Source Shakespeare

Concordance

- Lines taken from a corpus (one book, multiple books) showing a target word in context
- Target word is always in the middle allowing skimming of contexts

1. Is Oxfam interested in disarmament?
2. But she was not interested in education for its own sake.
3. Parents are obviously interested in the various skills and ...
4. .. to change as more women become interested in outdoor pursuits.
5. In other words, he was not interested, as Singer was, in proving ...
6. .. last month said a range of other interested parties had emerged, including ..
Concordance

- Lines taken from a corpus (one book, multiple books) showing a target word in context
- Target word is always in the middle allowing skimming of contexts

1. Is Oxfam interested in disarmament?
2. But she was not interested in education for its own sake.
3. Parents are obviously interested in the various skills and ...
4. .. to change as more women become interested in outdoor pursuits.
5. In other words, he was not interested, as Singer was, in proving ...
6. .. last month said a range of other interested parties had emerged, including ..
Concordance

• Lines taken from a corpus (one book, multiple books) showing a **target word** in **context**
• Target word is always in the **middle** allowing skimming of contexts

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Is Oxfam <strong>interested</strong> in disarmament?</td>
</tr>
<tr>
<td>2</td>
<td>But she was not <strong>interested</strong> in education for its own sake.</td>
</tr>
<tr>
<td>3</td>
<td>Parents are obviously <strong>interested</strong> in the various skills and ...</td>
</tr>
<tr>
<td>4</td>
<td>.. to change as more women become <strong>interested</strong> in outdoor pursuits.</td>
</tr>
<tr>
<td>5</td>
<td>In other words, he was not <strong>interested,</strong> as Singer was, in proving ...</td>
</tr>
<tr>
<td>6</td>
<td>.. last month said a range of other <strong>interested</strong> <strong>parties had emerged,</strong> including ..</td>
</tr>
</tbody>
</table>
KWIC – Keyword in Context

- A form of permuted index

<table>
<thead>
<tr>
<th>Titel</th>
<th>Catalognummer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office 2003 Timesaving Techniques for Dummies</td>
<td>100344</td>
</tr>
<tr>
<td>Excel for Scientists and Engineers</td>
<td>2704</td>
</tr>
<tr>
<td>How To Do Everything With Microsoft Office Excel 2003</td>
<td>894</td>
</tr>
<tr>
<td>How To Do Everything With Microsoft Office PowerPoint 2003</td>
<td>900</td>
</tr>
<tr>
<td>Data Analysis Using SQL and Excel</td>
<td>12377</td>
</tr>
<tr>
<td>Automated Data Analysis Using Excel</td>
<td>3434</td>
</tr>
<tr>
<td>How To Do Everything With Microsoft Office Excel 2003</td>
<td>904</td>
</tr>
<tr>
<td>Accessing and Analyzing Data with Excel 2003</td>
<td>5</td>
</tr>
<tr>
<td>Excel 2003 Formulas</td>
<td>708</td>
</tr>
<tr>
<td>Microsoft Office Excel 2003 Inside Out</td>
<td>18785</td>
</tr>
<tr>
<td>Mastering Excel 2003 Programming with VBA</td>
<td>7655</td>
</tr>
<tr>
<td>Excel 2003 VBA Programmer’s Reference</td>
<td>3099</td>
</tr>
<tr>
<td>Creating Spreadsheets and Charts in Microsoft Office</td>
<td>203</td>
</tr>
<tr>
<td>Excel 2007</td>
<td>2704</td>
</tr>
<tr>
<td>Beginning Excel: What-if Data Analysis Tools</td>
<td>8457</td>
</tr>
<tr>
<td>Fixing Powerpoint Annoyances</td>
<td>5575</td>
</tr>
<tr>
<td>Excel 2003 Formulas</td>
<td>708</td>
</tr>
<tr>
<td>Microsoft Visio 2003 User’s Guide</td>
<td>32001</td>
</tr>
<tr>
<td>How To Do Everything With Microsoft Office Excel 2003</td>
<td>904</td>
</tr>
<tr>
<td>How To Do Everything With Microsoft Office PowerPoint</td>
<td>900</td>
</tr>
<tr>
<td>Microsoft Office Excel 2003 Inside Out</td>
<td>18785</td>
</tr>
<tr>
<td>Mastering Excel 2003 Programming with VBA</td>
<td>7655</td>
</tr>
<tr>
<td>How To Do Everything With Microsoft Office Excel 2003</td>
<td>904</td>
</tr>
</tbody>
</table>

AntConc

A freeware corpus analysis toolkit for concordancing and text analysis.

[AntConc Homepage] [Screenshots] [Help]

Downloads:

- Windows (3.4.4)
- Macintosh OS X 10.7-10.12 (3.4.4)
- Macintosh OS X 10.6 (3.4.1)
- Linux (3.4.3)
- Older versions
Keyword Plots

Queen

No. of Hits = 75
File Length (in chars) = 163817

Alice

No. of Hits = 403
File Length (in chars) = 163817

Cat

No. of Hits = 37
File Length (in chars) = 163817
Keyword Plots

• Using keyword plots, at what points in the last Republican/Democratic debates did <applause> occur?

Look Familiar?

As published in the Washington Post
N-Grams

- An “n-gram” is a sequence of words, POS, or letters of length n
- Each sequence is counted
- Used in document analysis, but also spelling and grammar checking, translation, and more

\[
\begin{align*}
N = 1 : & \text{This is a sentence} \quad \text{unigrams:} \quad \text{this, is, a, sentence} \\
N = 2 : & \text{This is a sentence} \quad \text{bigrams:} \quad \text{this is, is a, a sentence} \\
N = 3 : & \text{This is a sentence} \quad \text{trigrams:} \quad \text{this is a, is a sentence}
\end{align*}
\]
AntConc n-grams

- View top n-grams
- Search n-grams
not get dry very soon. ‘Ahem!’ said the Mouse with an important Alice in Won
icia and Northumbria—’” ‘Ugh!’ said the Lory, with a shiver. ‘I beg Alice in Won
with a shiver. ‘I beg your pardon!’ said the Mouse, frowning, but very Alice in Won
politely: ‘Did you speak? ‘Not I!’ said the Lory hastily. ‘I thought Alice in Won
hastily. ‘I thought you did,’ said the Mouse.’—I proceed. “Ed Alice in Won
it advisable—”’ ‘Found WHAT?’ said the Duck. ‘Found IT,’ the Mou Alice in Won
dell enough, when I find a thing,’ said the Duck: ‘it’s generally a fr Alice in Won
to dry me at all.’ ‘In that case,’ said the Dodo solemnly, rising to Alice in Won
tic remedies—’ ‘Speak English!’ said the Eaglet. I don’t know the Alice in Won
ublicly. ‘What I was going to say,’ said the Dodo in an offended tone Alice in Won
clined to say anything. ‘Why,’ said the Dodo, ‘the best way to Alice in Won
es asked. ‘Why, SHE, of course,’ said the Dodo, pointing to Alice in Won
have a prize herself, you know,’ said the Mouse. ‘Of course,’ the Alice in Won
Alice sadly. ‘Hand it over here,’ said the Mouse. Then they all cro Alice in Won
a bat, and that's very like a mouse, you know. But do cats eat mice? Alice in Wonderland made out that it was only a mouse that had slipped in like this. Alice in Wonderland thought Alice, 'to speak to this mouse? Everything is so out-of-way.' Alice in Wonderland. She began: 'O Mouse, do you know the way out?' Alice in Wonderland. She was very much out of her mind, as she was swimming about here, O Mouse!' (Alice thought this must be the right way of speaking to a mouse: she had never done such a thing before.) Alice in Wonderland. She had a Latin Grammar, 'A mouse--of a mouse--to a mouse--a mouse--a mouse--O mouse!' The Mouse looked at her rather inquiringly, as Alice in Wonderland. She had a rather curious French lesson, and in her French lesson-book the Mouse gave a sudden leap out of the window.

Search Term: mouse

Search Window Size: 50

Kwic Sort:
Level 1: 1R
Level 2: 2R
Level 3: 3R

Advanced Options
Collocates

• “Collocations of a given word are statements of the habitual or customary places of that word.”

• Presence of one is highly indicative of the other (high mutual information)

• Semantics of collocates have a lesser degree of compositionality:
  • “strong tea” vs *“forceful tea”
  • “broad daylight” vs *“narrow darkness”

• In the extreme, can be used to detect idioms:
  • “heard it through the grapevine”
  • “kick the bucket”
Collocates

• Generally:
  • Weakly or non-compositional meaning
    • White wine vs white hair
    • “kick the bucket”
  • Non substitutive
    • “White wine” vs. ?? “yellow wine”
    • “kick the bucket” vs. “kick the pail”
  • Non-modifiable
    • “a frog in your throat” vs ?? “an ugly frog in your throat”
### Words

![AntConc 3.4.4w (Windows) 2014](image)

**Corpus Files:** Alice in Wonderland, Through the Looking Glass

**Word Types:** 4210  
**Word Tokens:** 61053  
**Search Hits:** 0  
** Lemma Word Form(s)**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3419</td>
<td>the</td>
</tr>
<tr>
<td>2</td>
<td>1847</td>
<td>and</td>
</tr>
<tr>
<td>3</td>
<td>1546</td>
<td>to</td>
</tr>
<tr>
<td>4</td>
<td>1451</td>
<td>a</td>
</tr>
<tr>
<td>5</td>
<td>1276</td>
<td>it</td>
</tr>
<tr>
<td>6</td>
<td>1205</td>
<td>i</td>
</tr>
<tr>
<td>7</td>
<td>1118</td>
<td>of</td>
</tr>
<tr>
<td>8</td>
<td>1097</td>
<td>she</td>
</tr>
<tr>
<td>9</td>
<td>1097</td>
<td>you</td>
</tr>
<tr>
<td>10</td>
<td>935</td>
<td>said</td>
</tr>
<tr>
<td>11</td>
<td>853</td>
<td>alice</td>
</tr>
<tr>
<td>12</td>
<td>835</td>
<td>in</td>
</tr>
<tr>
<td>13</td>
<td>715</td>
<td>was</td>
</tr>
<tr>
<td>14</td>
<td>689</td>
<td>that</td>
</tr>
</tbody>
</table>

**Search Term:** over  
**Hit Location:**

**Sort by:** Invert Order
**Sort by Freq:**

**Files Processed:** [ ]

**Total No.:** 2
Stop Word Removal

Stopwords

Stopword Lists
Collection of stopword lists in many languages. Find the English stopwords below and/or follow the links to view our other language stop word lists.

English Stopwords

Default English stopwords list
This list is used in our Page Analyzer and Article Analyzer for English text, when you let it use the default stopwords list.

- a
- about
- above
- after
- again
- against
- all
- am
- an
- and
- any
- are
- aren't
- as
- at
- be
- because
- been
- before
- being
- below
- between
- both
- between

Stopwords in other languages

- Arabic
- Armenian
- Basque
- Bengali
- Brazilian
- Bulgarian
- Catalan
- Chinese
- Czech
- Danish
- Dutch
- Finnish
- French
- Galician
- German
- Greek
- Hindi
**Word Types:** 3651  **Word Tokens:** 25944  **Search Hits:** 0

<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>935</td>
<td>said</td>
</tr>
<tr>
<td>2</td>
<td>853</td>
<td>alice</td>
</tr>
<tr>
<td>3</td>
<td>260</td>
<td>queen</td>
</tr>
<tr>
<td>4</td>
<td>259</td>
<td>one</td>
</tr>
<tr>
<td>5</td>
<td>245</td>
<td>little</td>
</tr>
<tr>
<td>6</td>
<td>213</td>
<td>know</td>
</tr>
<tr>
<td>7</td>
<td>210</td>
<td>like</td>
</tr>
<tr>
<td>8</td>
<td>171</td>
<td>went</td>
</tr>
<tr>
<td>9</td>
<td>160</td>
<td>thought</td>
</tr>
<tr>
<td>10</td>
<td>153</td>
<td>see</td>
</tr>
<tr>
<td>11</td>
<td>141</td>
<td>now</td>
</tr>
<tr>
<td>12</td>
<td>141</td>
<td>time</td>
</tr>
<tr>
<td>13</td>
<td>129</td>
<td>king</td>
</tr>
<tr>
<td>14</td>
<td>126</td>
<td>first</td>
</tr>
</tbody>
</table>

**Search Term**  
- **Check:** Words  
- **Case:**  
- **Regex:**  

**Hit Location**  
- **Search Only:** 0

**Lemma list**  
- **Loaded:**  

**Sort by**  
- **Invert Order:**  
- **Freq:**
## Collocates Analysis

<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>Freq(L)</th>
<th>Freq(R)</th>
<th>Stat</th>
<th>Collocate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>7.64075</td>
<td>furiously</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6.64075</td>
<td>wringing</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>wonders</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>warning</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>vicious</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>tinkling</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>thursdays</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>stroked</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>6.64075</td>
<td>spied</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>shyly</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>6.64075</td>
<td>shrieked</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>seizing</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>scrubbing</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6.64075</td>
<td>roared</td>
</tr>
</tbody>
</table>

### Search Term

- **Search Term**: queen
- **Words**
- **Case**
- **Regex**

### Window Span

- **Same**
- **From**... SL
- **To**... SR

### Min. Collocate Frequency

- 1

### Tools

- **Start**
- **Stop**
- **Sort**
- **Sort by Stat**
- **Clone Results**
Corpus n-grams vs Collocates
Finding Keywords
Dunning Log Likelihood ($G^2$)

$$E1 = c \frac{(a + b)}{(c + d)}$$

$$E2 = d \frac{(a + b)}{(c + d)}$$

<table>
<thead>
<tr>
<th></th>
<th>Analysis Text</th>
<th>Reference Text</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count of word</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Count of other words</td>
<td>c - a</td>
<td>d - b</td>
<td>c + d - a - b</td>
</tr>
<tr>
<td>Total</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
</tbody>
</table>

$$G^2 = 2((a \cdot \ln(a / E1)) + (b \cdot \ln(b / E2)))$$

Translates approximately to significance, e.g. 6.63 = $p < 0.01$

Macbeth

Word Counts

begotten
hast
love
thou
Macbeth

Log Likelihood

wrath
blood
lady
witch
Keywords

![AntConc 3.4.4w (Windows) 2014 interface](image)

- Types Before Cut: 3651
- Types After Cut: 3312
- Search Hits: 0

<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>Keyness</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>853</td>
<td>6930.483</td>
<td>alice</td>
</tr>
<tr>
<td>2</td>
<td>935</td>
<td>2982.574</td>
<td>said</td>
</tr>
<tr>
<td>3</td>
<td>260</td>
<td>1891.394</td>
<td>queen</td>
</tr>
<tr>
<td>4</td>
<td>245</td>
<td>795.826</td>
<td>little</td>
</tr>
<tr>
<td>5</td>
<td>171</td>
<td>572.442</td>
<td>went</td>
</tr>
<tr>
<td>6</td>
<td>129</td>
<td>541.539</td>
<td>king</td>
</tr>
<tr>
<td>7</td>
<td>117</td>
<td>520.425</td>
<td>quite</td>
</tr>
<tr>
<td>8</td>
<td>160</td>
<td>491.822</td>
<td>thought</td>
</tr>
<tr>
<td>9</td>
<td>81</td>
<td>482.544</td>
<td>stone</td>
</tr>
<tr>
<td>10</td>
<td>56</td>
<td>469.858</td>
<td>hatter</td>
</tr>
<tr>
<td>11</td>
<td>76</td>
<td>468.107</td>
<td>cried</td>
</tr>
<tr>
<td>12</td>
<td>55</td>
<td>461.468</td>
<td>gryphon</td>
</tr>
<tr>
<td>13</td>
<td>53</td>
<td>444.687</td>
<td>dumpty</td>
</tr>
<tr>
<td>14</td>
<td>53</td>
<td>444.687</td>
<td>humpty</td>
</tr>
</tbody>
</table>
### Negative Keywords

The table below shows the frequency and keyness of words from the corpus file "Through the Looking Glass" from the Alice in Wonderland series. The keyness score is calculated based on the frequency and the context of the word in the text. The words with the lowest keyness score are considered the most negative.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Freq</th>
<th>Keyness</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>37.320</td>
<td>says</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>36.358</td>
<td>years</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>31.744</td>
<td>year</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>28.988</td>
<td>also</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>19.350</td>
<td>around</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>18.659</td>
<td>new</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>16.923</td>
<td>need</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>15.564</td>
<td>family</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>14.956</td>
<td>today</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>14.456</td>
<td>woman</td>
</tr>
</tbody>
</table>

**Search Term:**
- **thought**

**Hit Location:**
- Reference Corpus: Loaded

**Total No.:** 2
**Files Processed:** 2

---

The software used is AntConc 3.4.2w (Windows) 2014.
You can enter any text that you would like in the form at the top, and then see useful information about words and phrases in that text.

First, it will highlight all of the medium and lower-frequency words in the text, and it can allow you to see "what the text is about" (i.e. text meaning).

Second, you can click through the words in the text to see the word sketches (i.e. meaning into the meaning and usage of the word), re-sortable by COCA frequency.

Finally, you can do powerful searches on selected phrases. These can be likely different styles of English. For example, if you click on 'high' phrases, it will show you the frequencies of those phrases in COCA -- over 50 milllion words.

By Mark Davies
CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversations?’

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, ‘Oh dear!’

Word frequency highlighting
Investigate Words in Context

- External context: From Corpus of Contemporary American English
Word Wanderer

We are experimenting with visual ways in which we can enhance people’s engagement with language. By fusing the information we can obtain from corpus searches, concordance outputs and word clouds we are aiming to enable and encourage people to notice and wander through the words they read, write and speak.

To try an early prototype of the WordWanderer just select a text below or paste your own and take it for a walk.

- Hansel and Gretel (Brothers Grimm)
- The Yellow Wallpaper (Gillman)
- In the Penal Colony (Kafka)
- 2 B R 0 2 B (Vonnegut)
- The Library of Babel (Borges)
- Jekyll & Hyde (Stevenson)
- Your own text: paste in box →

remembering her own child-life, and the happy summer days.
THE END

By Marian Dörk and Dawn Knight; http://wordwanderer.org/
... millennium fulcrum edition 3.0 chapter i . down the rabbit-hole alice was beginning to get very tired of sitting by her...

... what is the use of a book , ' thought alice 'without pictures or conversations ? ' so she was considering...

... was nothing so very remarkable in that ; nor did alice think it so very much out of the way to...

... and looked at it , and then hurried on , alice started to her feet , for it flashed across her...

... rabbit-hole under the hedge . in another moment down went alice after it , never once considering how in the world...

... , and then dipped suddenly down , so suddenly that alice had not a moment to think about stopping herself before...

... as she fell past it . ' well ! ' thought alice to herself , ' after such a fall as this , ...

... down , i think--' ( for , you see , alice had learnt several things of this sort in her lessons ...
Alice's Adventures In Wonderland

They very soon came upon a Gryphon, lying fast asleep in the
But do cats eat bats, I wonder? And here Alice began to get...
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do. Once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, and what is the use of a book, thought Alice `without pictures or conversation?'

So she was considering in her own mind (as well as she could for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, `Oh dear! Oh dear! I shall be late!' (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural; but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-Pocket and looked at it, and then hurried on, Alice thought that very odd.)
Alice's Adventures in Wonderland

came the guests, mostly Kings and Queens, and among them Alice
Word Tree

Word Tree

http://fernandaviegas.com/wordtree.html
Alice, Mouse, Gryphon, and Mock Turtle cried.

In a shrill, passionate voice, 'Would YOU like cats if you were me?' 'Well, perhaps not,' said Alice, who was trembling down to the end of his tail. 'As if I would talk on such a subject! sharply and very angrily. 'A knot!' said Alice, always ready to make herself useful, and looking.

The Mock Turtle had just begun to repeat it, when a cry of 'The trial's be taking Alice by the hand, it hurried off, without waiting for the end of the story. 'Of course,' the Mock Turtle said: 'advance twice, set to partners--' 'change lobsters, an.

Seals, turtles, salmon, and so on; then, when you've cleared all the jelly-fish out of the Guppy, and are not afraid of being all jumping up and down in an agony of terror. 'Oh, there goes his PRECIOUS nose'; as an unusually large sausages, quite forgetting in the hurry of the moment how large she had grown in the last few minutes, and she jumped.

(she was so much surprised, that for the moment she quite forgot how to speak good English); 'now I'm opening out hastily, afraid that she had hurt the poor animal's feelings. 'I quite forgot you didn't like cats. again, for this time the Mouse was bristling all over, and she felt certain it must be really offended.

in a sorrowful tone. 'I'm afraid I've offended it again!' For the Mouse was swimming away from her as hard as it could.

'Come, there's no use in crying like that!' said Alice to herself, rather sharply. 'I advise you to leave off this minute!

'Wake up, Dormouse!' And they pinched it on both sides at once. The Dormouse slowly opened his eyes.

when they saw Alice coming. 'There's PLENTY of room!' said Alice indignantly, and she sat down in a large arm-chair at one side, 'Silence in the court!' and the King put on his spectacles and looked anxiously round, to make out who was talking. so much!' said Alice, as she swam about, trying to find her way out. 'I shall be punished for it now, I suppose, by being drowned in my own

forgot you didn't like cats. 'Not like cats!' cried the Mouse, in a shrill, voice. 'I'd LIKE cats if you were me!'

'Well, perhaps not,' said Alice soothingly; 'don't be angry about it. And yet I think it's such a dear quiet thing to me. She is such a dear quiet thing to me ever since we went on, half to her service, as she was lazily about in the pool, and she sits purring so nicely by the fire, licking her paws and washing her face—and she is such a nice soft thing to have in the house. I've a such a capital

for catching mice—oh, I beg your pardon!' cried Alice, for this time the Mouse was bristling all over, and she felt certain it must be offended. 'We won't talk about any more if you'd rather not.'

'We indeed!' cried the Mouse who was trembling down to the end of his tail. 'As if I talk down such a subject! Our family always HATED cats: nasty, vulgar things! Don't mention the name again!'
Tools – Semantic data

• Entity Resolution
  • TextRazor
• Semantic Structure
  • DocuBurst
Named Entity Recognition

• Detect people, places, companies, etc. from text
• Heuristic (rule-based) and other approaches
• Entity resolution means to link the entity with a unique (shareable) ID, e.g. DBPedia, PermID
never before seen a **rabbit** with either a waistcoat-pocket, or a watch
Mihalcea and Tarau, 2004

Wattenberg et al., 2008
DocuBurst

games → game
taken → take

absolute, noun, 10
class, noun, 2
moment, noun, 11
game, noun, 30
reality, noun, 3
take, verb, 13
represent, verb, 17

WordNet

game is activity
chair is furniture
animal
vertebrate
chordate
cordate
mammal
placental
insect
arthropod
reptile
amphibian
diapsid
anapsid
fish
aquatic vertebrate
bird
aquatic bird
salamander
diabetes
frog
mammal
diapsid
reptile
vertebrate
As wet as ever, said Alice in a melancholy tone: it doesn’t seem to dry me at all.

In that case, said the Dodo solemnly, rising to its feet, I move that the meeting adjourn, for the immediate adoption of more energetic remedies—

Speak English! said the Eaglet. I don’t know the meaning of half those long words, and, what’s more, I don’t believe you do either! And the Eaglet bent down its head to hide a smile; some of the other birds tittered audibly.

What I was going to say, said the Dodo in an offended tone, was, that the best thing to get us dry would be a Caucus-Race.

What IS a Caucus-Race? said Alice; not that she wanted much to know, but the Dodo had paused as if it thought that SOMEBODY ought to speak, and no one else seemed inclined to say anything.

Why, said the Dodo, the best way to explain it is to do it. (And, as you might like to try the thing...
Of course it is. Its called WABE, you know, because it goes a long way before it, and a long way behind it.

And a long way beyond it on each side, Alice added.

Exactly so. Well, then, MIMSY is flimny and miserable (there's another portmanteau for you). And a BOROCOVE is a thin shabby-looking bird with its feathers sticking out all round—something like a live mop.

And then MOME RATHS? said Alice. I'm afraid Im giving you a great deal of trouble.

Well, a RATH is a sort of green pig: but MOME Im not certain about. I think its short for from home—meaning that they lost their way, you know.

And what does OUTGRABE mean?

Well, OUTGRABING is something between bellowing and whistling, with a kind of sneeze in the middle: however, you'll hear it done, maybe—down in the wood yonder—and when you've...
Tools - Document Collections

• Clustering documents
  • Lexos
ALICE'S ADVENTURES IN WONDERLAND Lewis Carroll THE MILLENNIUM FULCRUM EDITION 3.0 CHAPTER I. Down the Rabbit-Hole Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had ... give tale, perhaps even with the dream of Wonderland of long ago: and how she would feel with all their simple sorrows, and find a pleasure in all their simple joys, remembering her own child-life, and the happy summer days. THE END
K-Means Clustering

- Turn documents into vectors
- Cluster $n$ items based on their similarity in $k$ groups
### K-Means

Mouseover points or rows in the table to see document labels.

| 0   | hellobarbie_lines_v2_2  |
| 0   | hellobarbie_lines_v2_4  |
| 0   | hellobarbie_lines_v2_22 |
| 0   | hellobarbie_lines_v2_23 |
| 0   | hellobarbie_lines_v2_24 |
| 0   | hellobarbie_lines_v2_29 |
| 1   | Alice_in_Wonderland_without_Gutenberg |
| 1   | Alice_in_Wonderland_without_Gutenberg |
| 1   | Alice_in_Wonderland_without_Gutenberg |
| 1   | Alice_in_Wonderland_without_Gutenberg |
| 1   | Alice_in_Wonderland_without_Gutenberg |
| 1   | Alice_in_Wonderland_without_Gutenberg |
| 1   | Alice_in_Wonderland_without_Gutenberg |
| 2   | hellobarbie_lines_v2_13  |
| 2   | hellobarbie_lines_v2_14  |
| 2   | hellobarbie_lines_v2_15  |
Comprehensive Tools

• Voyant
  • Developed by Digital Humanities Researchers
  • Provides many of the functions of previous tools online
  • Active development and user community
ALICE'S ADVENTURES IN WONDERLAND

Lewis Carroll

THE MILLENIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversations?'

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.
Embedded Tools
More Resources

DH Toychest
Digital Humanities Resources for Project Building
(curated by Alan Liu)

Guides to Digital Humanities | Tutorials | Tools | Examples | Data Collections & Datasets

http://dhresourcesforprojectbuilding.pbworks.com
Voyant 2.0 is a complete rewrite of Voyant. It provides a suite of text analysis tools that will work with most texts you can upload or find on the web. These tools are combined in skins. For documentation see Documentation for Voyant 2.0.
Text Visualization Browser
Provided by ISOVIS group

Techniques displayed: 341

Search:

Time filter:
1976  2017

Analytic Tasks

Visualization Tasks

Data

Source

http://textvis.lnu.se
Next Time

• More visualization techniques:
  • Visualizing topic modelling, document variations, sonic patterns, readability, and more

• Case Studies in Visual Text Analytics for DH