DIGITAL SCHOLARSHIP WORKSHOP

INTRO TO TEXT MINING

WITH DS LIBRARIAN MEGAN MARTINSSEN
WHAT IS TEXT MINING?
WHY SHOULD I DO IT?
The process or practice of examining large collections of written resources in order to generate new information, typically using specialized computer software.
## Significant Themes in 19th Century Literature by Matthew L. Jockers (University of Nebraska) David Mimno (Princeton University)

<table>
<thead>
<tr>
<th>Label</th>
<th>Male-Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Fashion</td>
<td>-0.2015</td>
</tr>
<tr>
<td>Flowers And Natural Beauty</td>
<td>-0.1698</td>
</tr>
<tr>
<td>Tears And Sorrow</td>
<td>-0.1619</td>
</tr>
<tr>
<td>Drawing Rooms</td>
<td>-0.16</td>
</tr>
<tr>
<td>Drink As In Liquor And Beer And Tobacco</td>
<td>0.1489</td>
</tr>
<tr>
<td>Governesses And Education of Children</td>
<td>-0.1469</td>
</tr>
<tr>
<td>Nurses For Children</td>
<td>-0.1467</td>
</tr>
</tbody>
</table>
JK Rowling unmasked as author of bestselling crime novel

The detective novel penned by an unknown first-time author was “perfectly good”, “certainly well written” but “not something I fell in love with”.

COMPARISON

“Rowling and “Galbraith”: an authorial analysis by Peter Millican (Oxford University) and Patrick Juola (Duquesne University) with The Sunday (UK) Times

JK Rowling is the true author, not a former military policeman
CARLO ALLEGRI/REUTERS
LARGEST VOCABULARY IN HIP-HOP

Polygraph

# OF UNIQUE WORDS USED WITHIN ARTIST'S FIRST 35,000 LYRICS

2,900 Words
3,600
4,300
5,000
5,700
6,400

All Artists
View by Region
Just

Notes/sources:
(1) I used the first 5,000 words from 7 of Shakespeare's works: Hamlet, Romeo and Juliet, Othello, Macbeth, As You Like It, Winter's Tale, and Troilus and Cressida. For Melville, I used the first 35,000 words of Moby Dick.
All lyrics are provided by Rap Genius, but are only current to 2012. My lack of recent data prevented me from using quite a few current artists.
This data viz uses code by Amelia Bellamy-Royds in this juliadilla.
SMALL DETAILS & CHANGE OVER TIME

Bankspeak: The Language of World Bank Reports, 1946-2012 by Franco Moretti and Dominique Pestre (Stanford Literary Lab)
frequency of the word "love" in the 574 Browning love letters

January 1845 to September 1846
DIGITALLY AVAILABLE

Is it machine readable?

What format are the files?

Visit library.georgetown.edu/digital-scholarship for help
PRINT ONLY

Is it a printed text? Modern fonts?

How will you make it machine readable?
-OCR (Optical Character Recognition) -
-Transcription-

Can you scan it yourself?

Visit library.georgetown.edu/digital-scholarship for help
Finding the Right Tool

Consider

- Do I want a browser based program or downloadable software?
- What kind of questions do I want to ask about the corpus?
- Am I comfortable working in the command line or do I prefer a graphical user interface?
- Do I know (or want to learn) a programming language like R or Python?
- What do I want the outputs of my work to look like?

It's ok not to have an answer to every question!!
SameDiff

Answers:

Do I want a browser based program or downloadable software? Browser

What kind of questions do I want to ask about the corpus? I want to compare 2 texts to each other

Am I comfortable working in the command line or do I prefer a graphical user interface? GUI

Do I know (or want to learn) a programming language like R or Python? Nah

What do I want the outputs of my work to look like? I don’t need the tool to create visualizations for me
USING SAME DIFF
Voyant

Answers:

Do I want a browser based program or downloadable software? Browser

What kind of questions do I want to ask about the corpus? I want to know about word usage in different works in one author’s entire body of work.

Am I comfortable working in the command line or do I prefer a graphical user interface? GUI

Do I know (or want to learn) a programming language like R or Python? Nope

What do I want the outputs of my work to look like? I’d like some visualizations created for me.
USING VOYANT