DET KGL. BIBLIOTEK AUL ARTS Nobelparken

Introduction to Digital Methods for Literary Studies

(some) resources, tools and approaches

Karoline Liv Vildlyng, Information Specialist

Hi!

- I am an information specialist at the department for Library Services and Partnerships at AU Library // The Royal Library
- I am specialized in communication and digital culture, digital methods, library resources and digital collections
- I mainly teach and do course creation, but I also work as a librarian at Kasernen, Nobelparken and Moesgaard
- Masters degree in Culture and Communication from the University of Copenhagen
- *This is a brand new workshop!*

#1 discussion excercise - 5 minutes

1. Turn to the person sitting next to you
2. Discuss your expectations for the workskop
3. What do you expect to know after today's workshop that you do not already know?
4. What is (if any) your experience with doing digital methods?

What will you know after today's workshop?

You will know what Voyant is and how to use the main affordances of the app

You will also know:

- What a wordcloud is and what a network graph looks like
- (more about) how a computer reads
- How to use stopwords to clean your data
- How to make and analyze a corpus
- How to analyze, contrast and compare large bodies of text through computational text analysis aka *text mining*

Workshop aims for students

That you:

- Learn basic concepts of computational text analysis and why you might want to incorporate the methodologies into your academic research practice

- Develop critical awareness of Voyant’s interface, tools, functions and affordances, as well as its value as a potential research tool

- Feel confident exploring computational text analysis or text mining further on your own

The program for the workshop

First block - talk the talk (45 minutes)

- An intro to natural language processing, computational text analysis or text mining
- Key awarenesses when doing digital methods for literary studies and literary history
- Why mixhed methods? Fairy tale edition

Second block - walk the walk (45 minutes)

- Github
- Demonstration of Voyant
- Group excercise: hands-on text mining with Jane Austen

Third block - fairy tales, critique, and discussions (45 minutes)

- Group excercise: contrast and compare two corpora of fairy tales
- Group discussion: what can these methods show, what are the limitations and how are they relevant to you?

An intro to computational text analysis

Computational text analysis is:

- (arguably) a different term for *text mining* or *natural language processing*
- A process of deriving information from (large bodies of) texts, such as novels, monographs, articles, web pages etc. aka a *corpus*
- Generally involves detecting patterns, such as identifying word frequency or associative links between words, co-ocurrence and text density
- A mixed methods approach, it combines a qualitative and quantitative approach to research in the humanities
- Not as new as we think

An intro to computational text analysis

Distant reading was formalized as such and introduced to the field by Franco Moretti around 2005.
But **quantitative text analysis** as a set of practices have been practiced for decades. Using paper as the medium, it is extremely labor intensive and time consuming.

A relevant example for us is: Ruth Bottigheimer in 1986

- She analyzed *speaking* versus *silence* across different fairy tales, especially the Grimm brothers' tales
- She counted the frequency of words related to speech and what patterns they take with regard to the character's gender or role

1 game exercise - Key word bingo

1. Write down what you think each term means/is
2. Together we will go through them one by one
3. Count how many you got right along the way

Natural language processing

Corpus

OCR

Token

Stopwords

Vocabulary density

Some key words and terms

Corpus: Pl. Corpora:

A large and structured set of texts. Can be in multiple or just one language. It is not equivalent to a database or an anthology or archive. Can be a great body of work on a specific subject by one author. Great for hypothesis texting or statistical analysis of texts like studying linguistic structures, frequencies, etc.

Natural language processing:

A natural language is English. Or Spanish or Danish. A natural language is not Javascript or Python or a spreadsheet which a computer can read or decode.

Some key words and terms

Token:

Is what we call a 'word' in the computational reading field. It is anything that isn’t a space.

It is a unit.

For computers to read, they have to **tokenize** to be able to transform the corpus into numbers. It is akin to cutting a text into little pieces consisting of symbols, words and numbers.

It is common in natural language processing.

Some key words and terms

OCR (optical character recognition):

The use of computer technologies to convert scanned images of typewritten, printed, or handwritten text into machine-readable text. OCR misrecognition is also the reason why you might find errors in you datasets.

Vocabulary density:

A measurement of vocabulary usage in comparison to the length of a text. The ratio of the number of words in the document to the number of unique words in the document.
A lower vocabulary density indicates complex text with lots of unique words, and a higher ratio indicates simpler text with words reused.

Key words and terms

Stopwords:

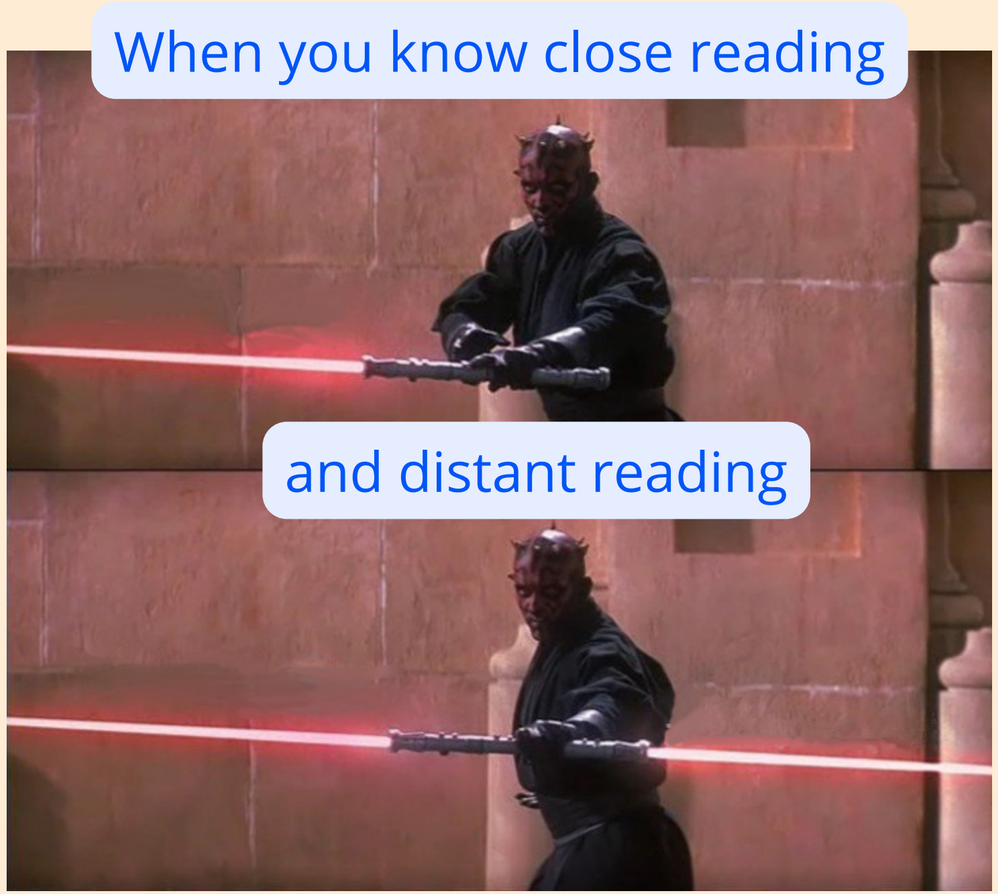
Words that are repeated at high frequency without conveying more meaning simply because they are repeated. *And*, *or*, *of*, *that*, *if*, *then*, *the*. They take up space, but do not necessarily provide analytical insight.

When analyzing a corpus, we need to make sure we have the most optimal **stop word list**.

Optimal, in this case, means the optimal one for your research.

Why mixed methods? Fairy tale edition

- The corpora for today is fairy tales or folk tales by H. C Andersen and the Grimm brothers
- A common characteristic of fairy tales is how they are presented. The words themselves state the narrative with little inner psychology of the characters
- This can make fairy tales very appropriate for quantitative textual analysis. The narrative structure is simple and the reader is already supplying the subtext



Why mixed methods? Fairy tale edition

- Many research projects involving quantitative analysis on fairy tales has been in the field of gender studies

A relevant example is Jeana Jorgensen and Scott Weingart (2013):

- Through digital quantitative tools they analyzed age and gender in fairy tale characters in relation to how their physical appearences were described
- They found that females were described more than males, and that the old were described in physical terms more than the young. ->The assumed narrative viewpoint is male and not-old
- Jorgensen (2014) found that descriptions for women prioritize beauty, morality, blood, hair, and skin, whereas for men the majority involve violence, transformation, size, and age

Why mixed methods? Fairy tale edition

- Fairy tales are partly defined by being stories that are told and retold countless time and exist in many different versions throughout time. This also makes them suited for computational text analysis

- A possible research design is to use Voyant to analyze different versions of the same fairy tale. How does the word frequency change throughout four different versions of Little Red Riding Hood?

- That *could* also be done with a version of the original fairy tale vs a transcript of a new Disney-retelling of that story. Like *Rapunzel* and *Tangled* or *The Snow Queen* and *Frozen*. How would the vocabulary density change or the word frequency shift?

Why we're not analyzing *Frozen* today...

- As a rule, works in Denmark are protected by copyright law spanning a period from the time the work is created until **70 years** after the author's death
- During that period, the author or her heirs, if she is deceased, must give permission (consent) if others want to use the work. It also means that the author or heirs may claim a fee when others use their work
- So: We (at the Royal Library) are using the Gutenberg Project to access non-copyrighted works

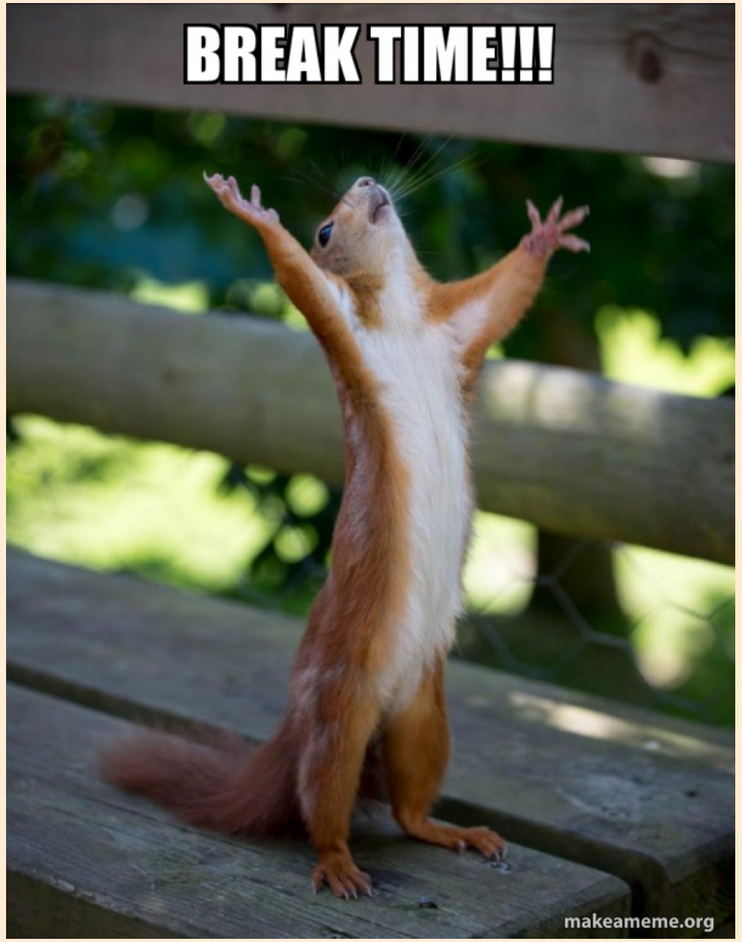
Break -15 minutes

Voyant walk-through and hands-on text mining

First step:

Go to https://github.com/karolinevildlyng/Introduction-to-Voyant

Break -15 minutes



Criticisms and critiques

- ICT's fundamentally change how we learn - and read?

- Moretti (2005) describes distant reading as *'a little pact with the devil: we know how to read texts, now let's learn how not to read them*'.

2 Discussion exercise - 5 minutes

1. Turn to the person sitting next to you
2. Discuss: Is uploading a corpus to Voyant and analyzing the numbers, graphs and charts, *reading*?
3. If yes, then how? If no, why not?

Criticisms and critiques

"Literary criticism is, in many ways, the art of telling stories about stories. Using textual evidence, convincing critics weave alternative narratives around texts that re-contextualize them forever, making it impossible to re-read a work in the same way again. Thus, criticism is a performance of a new narrative and a deformation of the original, now-reshaped text. Literary criticism routinely undertakes, then, what Lisa Samuels and Jerome McGann call 'deformance'."

Martin Paul Eve (2022) *The Digital Humanities and Literary Studies*

Criticisms and critique

- Methodology within the humanities is often critiqued for being 'anecdotal'

"The literary scholar of the twenty-first century can no longer be content with anecdotal evidence, with random “things” gathered from a few, even “representative,” texts. We must strive to understand these things we find interesting in the context of everything else, including a mass of possibly “uninteresting” texts."

Matthew L. Jockers (2013) *Macroanalysis*

- But quantitative data is not necessarily neutral either...

Criticisms and critique

- No software is perfect and a computer does not read exactly like a human being (yet)

Example: The way Voyant calculates the length of sentences should be considered very approximate, because it is complicated to distinguish between the end of an abbreviation and that of a sentence or other uses of punctuation

- Stanley Fish (2012) argues that the main issue with distant reading is "*first you run the numbers, and then you see if they prompt an interpretive hypothesis. The method, if it can be called that, is dictated by the capability of the tool*".

- It becomes a question of the tail wagging the dog

Criticisms and critiques - and answers?

Digital methods do not replace traditional methodologies in the humanities

- They are a **supplement** to existing scientific research practices
- Through them, we can gain insight we would otherwise have trouble reaching
- They can automate otherwise manual research processes
- They can spark new **debates and critical reflection**
- They are flawed and carry blind spots just like traditional methodologies

Thank you for listening and participating!

The library is here to help - you are always welcome to reach out to me with questions at kavk@kb.dk