16 Exploratory Swedish text analysis using notebooks – a *smörgåsbord* of basic corpus linguistic insights

Dimitrios Kokkinakis
University of Gothenburg

Gerlof Bouma
University of Gothenburg

The computational notebook has established itself as a significant tool for conducting exploratory data analysis, which aims at investigating characteristics of a dataset without preformulated expectations. Computational notebooks are a type of interactive document, that supports mixing prose, executable code and its output, such as a calculated result, a table, or a graphic. Data, process, and narrative are effectively integrated into one environment, which makes notebooks ideal for documenting exploratory research. Notebooks also facilitate sharing research in a reproducible way for teaching, collaboration or dissemination.

This chapter demonstrates basic exploratory techniques for Swedish text analysis implemented as *Jupyter notebooks*, a popular computational notebook implementation. Using a selection of documents from a Swedish corpus of COVID-19-related materials, we show some of the kinds of text analysis that can easily be performed using readily available software libraries. The examples in this chapter rely only on automatic annotation, requiring minimal manual processing.

1 *Notebooks for exploratory data analysis in the humanities*

As the field of Digital Humanities continues to evolve, researchers are increasingly exploring new ways to integrate computational tools into traditional humanistic study. The significant expansion in the volume of accessible materials has fuelled the development of methodologies that leverage data-driven analysis, visualization techniques, and machine learning to uncover new insights into historical, literary, and cultural phenomena. A critical

HUM INFRA Dimitrios Kokkinakis & Gerlof Bouma. 2025. Exploratory Swedish text analysis using notebooks – a *smörgåsbord* of basic corpus linguistic insights. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 541–563. University of Tartu Library. DOI: 10.58009/aere-perennius0185

area of growth is also the use of open-source platforms and collaborative tools, which allow scholars to perform complex data analysis while sharing methodologies transparently.

Large-scale analysis of textual sources is one of the core methodologies of Digital Humanities. Computational linguistics and corpus linguistics have contributed greatly to this methodology and will continue to do so. This chapter showcases some basic techniques from these linguistic fields and their use in exploratory data analysis in the context of text-based Digital Humanities (York 2017). Exploratory data analysis is "a collection of techniques for identifying the main characteristics of a novel dataset, about which one may initially know nothing" (York 2017: 462) and involves looking at the materials that form a study's empirical basis from many different angles and at different levels of abstraction, qualitatively and especially, quantitatively. Documenting this iterative process is very important. It supports the researchers themselves, but it also provides a means to communicate to others how the researchers came to draw conclusions from the dataset or formulate hypotheses based on it. *Computational notebooks* have proven to be an effective way of creating this documentation as part the analysis process.

1.1 *Introduction to notebooks*

A computational notebook is an interactive document that lets the notebook author combine formatted text, source code and program output into one flow. The researcher can thus perform exploratory data analysis inside the notebook environment, and add explanatory narrative, be it private or public, along the way.

Because it places code and explanatory narrative on equal footing, the computational notebook concept aligns within the paradigm of *literate programming* (Knuth 1984). In this paradigm documented computer programs are written in the form of coherent texts focussed on the human reader, which include the executable source code of the program that is being documented. Notebooks add to this paradigm the ability to directly execute code and display its results as part of such documents. A computational notebook thus consists of three types of content, or "cells" in notebook terminology: formatted text, program code and program output. The latter may be in the form of text, tables, visualizations, etc. During creation of a notebook, the text and code cells are directly editable. Notebooks can be converted to formats suitable for publishing, for instance PDFs or webpages. These are typically not dynamic in the same way, although some notebook implementations may allow for interactive content in published notebooks. Also, when the underlying notebook files themselves are made accessible

– an increasingly common dissemination strategy – other researchers may directly use them as dynamic documents to edit and build upon.[1]

Notebooks are actively deployed in a wide range of scientific contexts, including education, economics, engineering, data science, (corpus) linguistics and digital humanities (Alderson 2021, Barba et al. 2019, Granger & Pérez 2021, Hardebolle 2023, Dombrowski et al. 2019, Bednarek et al. 2024). Interestingly, although notebooks can be and are used for dissemination, their primary use is, as their name suggests, as personal documentation and support in data exploration (Rule et al. 2018). Notebooks are also valued as pedagogical tools (Blanke et al. 2023), amongst other things because they allow instructive text to be interspersed with editable and executable code. The adoption of notebooks in Digital Humanities research represents a clear method advancement in the field. By combining the power of computer-supported data analysis with the flexibility of narrative and collaboration, these tools reshape how scholars may engage with large and complex datasets. Computational notebooks provide a versatile platform that enhances both the analytical process and the communication of research findings. Through their support for iterative analysis and dynamic visualization, these tools lower the threshold for addressing complex research questions in the humanities using data-intensive methods. They are increasingly central to interdisciplinary research, bridging computational and humanistic approaches (Ahnert et al. 2023). In the quickly evolving practices of digital humanists, notebooks can supplement "fixed" graphical user interfaces to give more technical users flexible access to archives and datasets (Melgar-Estrada et al. 2019).

1.2 *Implementations and deployment of notebooks*

There are many modern implementations of the notebook idea; open-source ones such as Jupyter Notebook[2] (Granger & Pérez 2021), currently the de-facto standard, and Spyral (Land et al. 2021) as well as proprietary ones such as MATLAB Live Editor[3] and the seminal notebooks of Wolfram Mathematica.[4] Depending on the implementation, notebooks can be created and viewed locally or online. For the widely used Jupyter Notebook, commercial online services include Google Colaboratory[5] (Colab), Kaggle[6], and Mi-

1 As just a tiny example of this, https://github.com/quinnanya/dh-jupyter list 100 notebooks / notebook-based resources with Digital Humanities relevance.
2 https://jupyter.org
3 https://www.mathworks.com/products/matlab/live-editor.html
4 https://www.wolfram.com/notebooks/
5 https://colab.research.google.com/
6 https://www.kaggle.com

crosoft's GitHub Codespaces and Azure Machine Learning.[7] One advantage of using an online service is that there is no need to install any software: the functionality to run the notebooks is supplied by the service and typically there is a rich, pre-installed ecosystem of auxiliary packages that add ready-made capabilities to the notebook. Moreover, online services may provide access to powerful computing servers that allow the researcher to work on a scale that is not easily achievable on their local hardware. Some services also offer tools to facilitate collaboration, for instance through joint notebook editing and publication. On the other hand, the advantage of running notebooks locally is that the researcher keeps full control over code and data, imperative when working with sensitive or otherwise restricted materials, and does not need a user account, crucial when researcher anonymity is important. Also, no internet connectivity is needed for local execution, even when Jupyter Notebook uses the browser as their user interface.

In this chapter, we hope to demonstrate notebooks in exploratory analysis of (large-scale) text corpora in a way that is both efficient for the researcher and capable of supporting insightful interpretations. We will use Jupyter Notebook as our notebook platform and Python as our programming language. Figure 1 contains an overview of the Jupyter Notebook interface. This chapter will not include detailed instructions on how to install or use Jupyter Notebook, neither will it introduce the Python language. For this we refer the reader to the many online tutorials on these topics, to Dombrowski et al. (2019) for an article length instructive introduction to Jupyter Notebook for researchers in the humanities, and to Karsdorp et al. (2021) for a book length introduction to humanities data analysis using Python. The notebooks will be available from spraakbanken.gu.se/en/projects/huminfra/exploratory-text-analysis.

2 *Research scenario*

The following sections present an outline of the various text analysis components that are implemented in our notebooks. These components not only serve as a guide to the implemented methodologies but also offer a dynamic showcase of the diverse analyses conducted using the platform. Furthermore, they stimulate hands-on exploration and experimentation by individuals with minimal programming skills, thereby democratizing access to advanced data analysis tools and enabling broader participation in computational research. This transparency ensures the reproducibility of the research, allowing others to access and verify the data sources used. All

7 https://visualstudio.microsoft.com/vs/features/notebooks-at-microsoft/

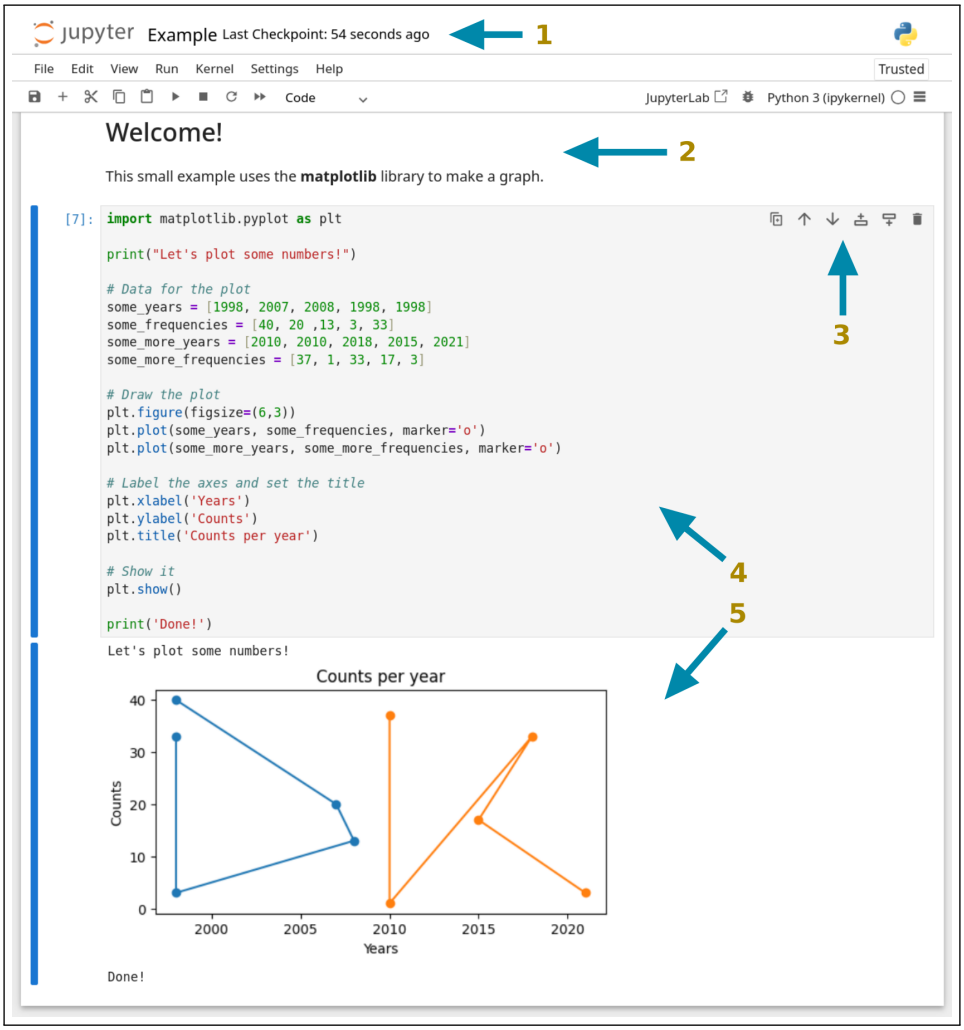


Figure 1: The user interface in Jupyter Notebook: (1) The name of the notebook and information about the last manual save. (2) A *markdown cell* containing formatted prose text. (3) Controls to move the active cell or add new cells around it. (4) A *code cell* containing Python code. The coloured code formatting is provided by the notebook software. (5) The output of the code cell above it, containing both textual and graphical elements.

resources used in the demonstrations, the notebooks as well as the used text corpus and lexical resources, are available online and are linked from the notebooks.

Although interactive notebooks are available for many different programming languages, we use Python as our language becuase of its accessability

and the rich ecosystem of libraries for data analysis, NLP and graphing. In our notebooks, we use for instance NLTK[8], spaCy[9] and BERTopic[10] for NLP, Pandas[11] or Polars[12] for data handling, NumPy[13] and SciPy[14] for numerical computing, and Matplotlib[15] for graphing.

After a brief introduction of the dataset used in all of our notebooks in Section 2.1, Sections 2.2–2.6 discuss the different case studies on this data.

2.1 *Dataset overview*

The text corpus used in the experiments below consist of approximately 2 700 documents published online in 2021, during the COVID-19 pandemic. This dataset is part of the sv-COVID-19 corpus (Kokkinakis 2021), which contains articles published in Swedish on the internet and covers a range of topics related to the COVID-19 pandemic from different sources. The corpus has been systematically classified by the first author into eight distinct stylistic genres, determined by the original publication medium and contextual factors surrounding each text.[16] The figure between parenthesis is the number of documents for each genre:

- Authorities (64): e.g., https://skr.se/
- Blog (189): http://www.islandsbloggen.com/
- Medical (460): e.g., https://www.lif.se/
- News (850): e.g., https://www.gp.se/
- Public media (759): e.g., https://sverigesradio.se/
- Periodicals (338): e.g., https://borsvarlden.com/
- Research (44): e.g., https://www.forskning.se/
- Social media (8): e.g., https://www.familjeliv.se/

These genres provide different perspectives on the pandemic, highlighting the varied approaches to the topic across different platforms. The ability to perform genre-specific searches facilitates the exploration of how different

8 https://www.nltk.org/
9 https://spacy.io/
10 https://maartengr.github.io/BERTopic/
11 https://pandas.pydata.org/
12 https://pola.rs/
13 https://numpy.org/
14 https://scipy.org/
15 https://matplotlib.org/
16 This categorization enables targeted analysis and querying within SpråkBanken Text's word research platform *Korp*: https://spraakbanken.gu.se/korp/#?corpus=sv-covid-19.

types of media and publication contexts influence the portrayal and discussion of COVID-19. Researchers can also use the data to examine shifts in language use, thematic focus, and, to some extent, public sentiment over time. Analysis of data like these may offer insights into pandemic-related communication and its broader impact, and hopefully deepen our understanding of how the crisis was discussed and perceived.

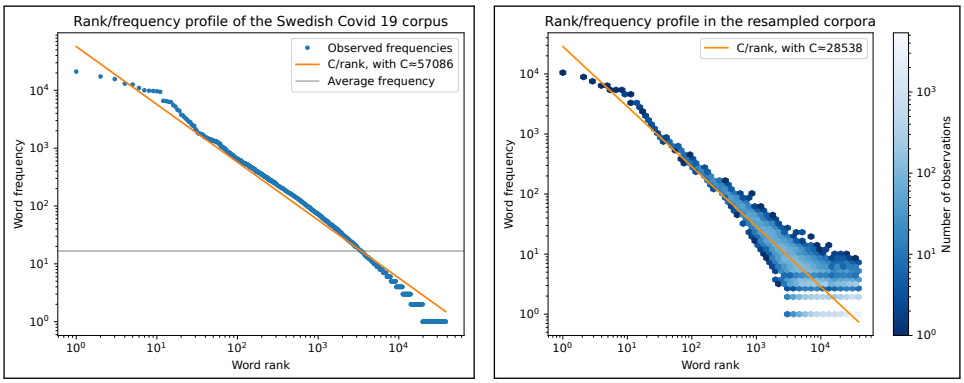
2.2 *Word distributions and word frequencies: Zipf and beyond*

In linguistics, the study of word distributions and word frequencies, particularly in large corpora, has been significantly shaped by what has become to be known as Zipf's Law (Zipf 1949). This model posits that the frequency of a word is inversely proportional to its rank: A small number of high-frequency words dominate, while the vast majority of words occur only very infrequently. Easy access to text materials at scale for many languages has allowed for precise word frequency analysis, revealing both adherence to and deviations from Zipf's model (Piantadosi 2014).

Figure 2a shows a classical plot of the *Zipfian distribution* of word frequencies (rank–frequency plot) in the comparatively small, examined dataset. The distribution is plotted on logarithmic scales for both axes; the Zipfian model says that the graph should follow a straight line in this case. The model's predicted relation between rank and frequency of is included in the graph in orange. A few function words and closed class words, such as *att* 'to/that' (infinitive marker/subordinator), *i* 'in', *och* 'and', *det* 'it', and *som* 'that/as' (subordinator), are used extremely often, while an abundance of words are rarely used and thus appear in the right tail of the diagram.

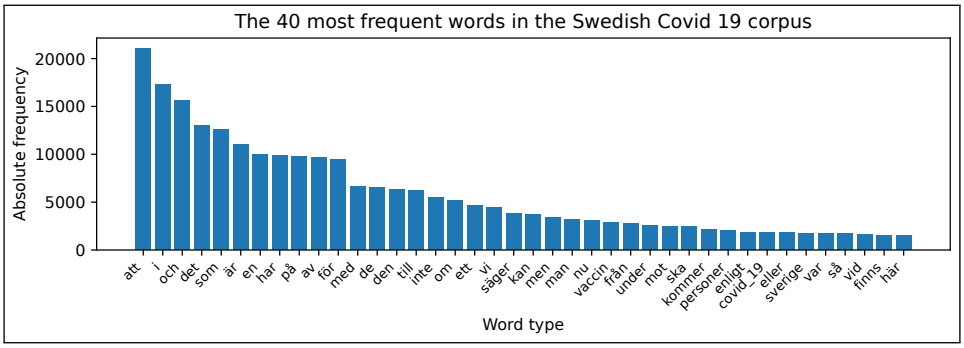
Piantadosi (2014) argues that estimating both rank and frequency from exactly the same data is a mistake, and proposes to create two random splits of a dataset: one for the estimation of a word's frequency and one for the estimation of its rank. The resulting graph is in Figure 2b. The overall shape is very similar, with a strong linear trend on the log–log scale. But the right tail now fans out, because at the low frequency end, small differences in frequency may lead to relatively large differences in rank. This uncertainty in the relation between rank and frequency is not shown at all in the classic way of constructing the Zipf plot, but it shows (correctly) in this alternative method.

The next two graphs connect the previous Zipfian discussion with an alternative visualisation of word frequencies. Figure 3a shows the frequencies of the top 40 most frequent words as a bar chart. What is a linear trend in the Zipf plots, can be seen here as exponential decay. The sharp drop after *för* 'for' (at 11th place) in the bar chart can be recognized as a bend in the

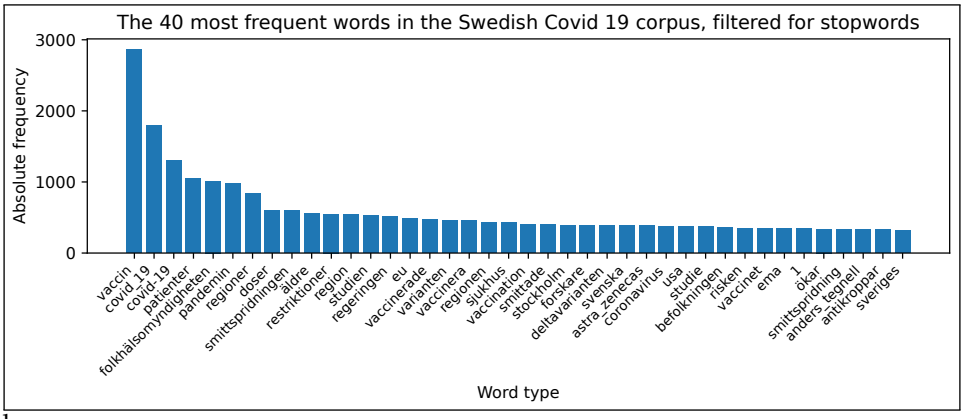


a. b.

Figure 2: The Zipfian distribution of the words in the dataset, in (a) the classic style and (b) the style of Piantadosi (2014) with rank and frequency estimated from different random subsets of the corpus.



a.



b.

Figure 3: The 40 most frequent unigrams in the dataset, (a) unfiltered, (b) without stopwords.

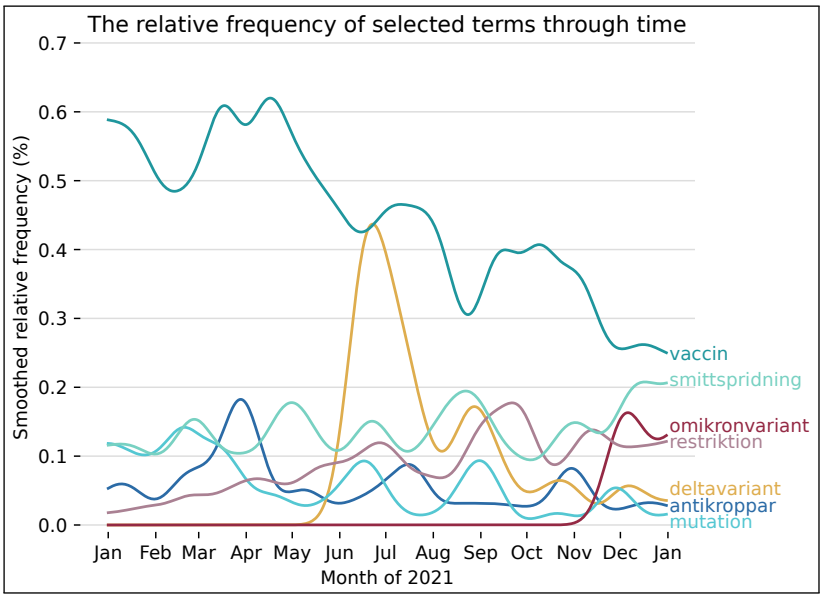


Figure 4: Smoothed word frequency trends for a small number of selected words.

Zipf plot. Figure 3b shows the same, but after so-called "stop words" have been filtered out. Stop words typically provide minimal semantic value in isolation, making them less relevant for tasks like text classification, sentiment analysis, and information retrieval, and are often filtered out during preprocessing stages, as their high frequency across texts can overshadow more meaningful content. We see that the overall shape of the distribution is similar (quickly dropping in frequency initially, more slowly towards the right-hand side), but from a completely different starting point in terms of the frequency of the most common words. Looking at the words themselves, they are not just general domain content words, but many of them are highly specific to the corpus' common denominator of the COVID-19 pandemic, such as *vaccin* 'vaccine', *covid-19*, and *folkhälsomyndigheten* 'Public health agency of Sweden'.

The corpus contains materials collected over a whole year, and it is therefore possible to study the changes in word frequencies over this period. Figure 4 shows the word frequency distributions over the examined year for a small number of selected words, namely *omikronvariant* 'omicron variant', *deltavariant* 'delta variant', *vaccin* 'vaccine', *restriktion* 'restriction', *smittspridning* 'spread of infection', *antikroppar* 'antibodies', and *mutation*, including inflected forms of these words. In this diagram one can clearly see fluctu-

ations and areas where word frequencies spike, like the brief heightened frequency of mention of the *delta variant* in the middle of the year and the rise of the *omicron variant* at the end of 2021. These observations can be linked to real-world events, such as the dominance of these variants in Sweden during this year.

2.3 *Significant word associations and collocations*

Collocations are words that tend to occur in each other's company. At a macro-level, the study of collocations can tell us things about the nature of the text material we are studying and at a micro-level, we can learn about the meaning of a word by studying its collocates.[17] Looking at collocations may also provide the empirical foundation for lexical studies of multiword expressions such as idioms, phrasal verbs, proper names, etc.

The notion of collocation can be operationalized in many different ways, for instance by the choice of *association measure*, that is, a score to summarize how often words co-occur and how surprising this level of co-occurrence is. These measures typically rely not just on the observed frequency of co-occurence, but also on an expected frequency, based for instance on what we would expect to see if the distribution of two words were not related to eachother.

Table 1 shows the 20 bigrams, directly adjacent words, from the COVID-19 corpus that score highest on the association measure of *log-likelihood ratio* (LLR),[18] after some preprocessing like removing case distinctions and filtering out combinations containing stop words.[19] This LLR top 20 contains personal (*Sara Byfors*) or geographic (*Västra Götaland*) proper names, but we can furthermore see a recurrent pattern descriptive noun–proper name (*statsepidemiolog Anders Tegnell* 'chief epidemiologist Ander Tegnell', *socialminister Lena Hallengren* 'minister of social affairs Lena Hallengren', *läkemedelsmyndigheten EMA* 'medicinal product authority EMA', amongst others).

The table also gives scores for four other association measures: frequency, point-wise mutual information (PMI), t-score, and chi-squared (χ^2). To show that these measures may disagree wildly about which bigrams form strong collocations, we have included the rank of each bigram for the re-

17 Or, as the famous and ubiquitously cited adage from Firth (1957: 11) has it: "You shall know a word by the company it keeps!"

18 We refer the reader to Evert (2009) for definitions of LLR and the other association measures mentioned here, and for a good general discussion of the collocation concept.

19 The used corpus contains some preprocessing that combines first name–last name combinations. For simplicity, we have treated these underscore-separated combinations as one "word". This means that the list of bigrams may contain combinations of more than two words, such as in *astra_zenecas vaccin* 'Astra Zeneca's vaccine'.

Table 1: The 20 strongest bigram collocations in the dataset according to the log-likelihood ratio (LLR), for pairs that do not contain stopwords and occur at least 5 times. Scores (and ranks) for four other association measures are also given.

Bigram	LLR	Freq	PMI	t-score	χ^2
astra_zenecas vaccin	3043.7	315 (1)	7.4 (965)	17.6 (1)	52513.6 (366)
statsepidemiolog anders_tegnell	2068.1	140 (2)	10.7 (558)	11.8 (2)	230143.3 (166)
tegmark wisell	1829.0	95 (6)	12.5 (355)	9.8 (6)	557066.6 (1)
socialminister lena_hallengren	1698.6	99 (5)	11.9 (413)	10.0 (5)	374281.3 (108)
europeiska läkemedelsmyndigheten	1544.5	102 (4)	11.1 (509)	10.1 (4)	217260.4 (176)
sara byfors	1444.5	78 (10)	12.5 (352)	8.8 (10)	468851.2 (49)
karin tegmark	1386.6	83 (8)	11.8 (418)	9.1 (8)	305002.5 (137)
västra götalandsregionen	1203.8	73 (12)	11.6 (452)	8.6 (12)	220265.8 (172)
västra götaland	1203.8	73 (12)	11.6 (452)	8.6 (12)	220265.8 (172)
johan carlson	1097.9	74 (11)	10.8 (539)	8.6 (11)	136540.0 (223)
region stockholm	1092.7	108 (3)	8.3 (888)	10.4 (3)	34492.6 (439)
emma spak	1027.0	56 (20)	12.7 (338)	7.5 (20)	370954.6 (109)
sveriges kommuner	1018.7	83 (8)	9.7 (687)	9.1 (9)	67857.1 (333)
matti sällberg	991.3	50 (22)	13.3 (279)	7.1 (22)	492057.8 (44)
kristine rygge	954.7	47 (24)	13.4 (264)	6.9 (24)	526653.4 (20)
läkemedelsmyndigheten ema	947.6	68 (14)	10.6 (572)	8.3 (14)	102524.6 (273)
låga nivåer	936.8	65 (16)	10.9 (526)	8.1 (16)	127783.3 (235)
karolinska universitetssjukhuset	914.4	57 (18)	11.8 (422)	7.6 (18)	207728.8 (179)
brittiska varianten	902.2	84 (7)	8.8 (836)	9.2 (7)	36223.9 (424)
johan bratt	875.0	57 (18)	11.0 (519)	7.6 (19)	115870.4 (252)

maining association measures. We see for instance that LLR, frequency and t-score agree to a high extent, since the latter two also give prominent ranks to the cases that LLR brings up. PMI and χ^2 on the other hand would place the shown combinations much further down on the list. An important part of studying collocations is choosing the association measure that gives the relevant insights into the data.

Collocations need not just concern directly adjacent words. Figure 5 is a network visualisation of the most common words to occur within a window of four words before or after the focus terms of *positiv*, *positivt*, *positiva* 'positive(ly)' and *negativ*, *negativt*, *negativa* 'negative(ly)'. The placement of nodes and the thickness of connecting edges reflect how common a collocation is. The lack of a connection between nodes indicates that the frequency of co-occurrence falls below a pre-set threshold (and may be zero). Only

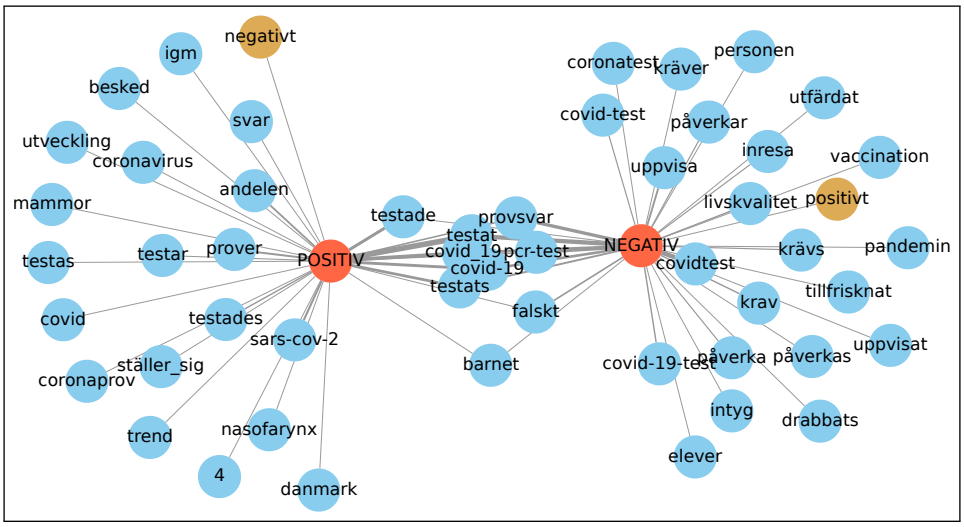


Figure 5: Terms strongly collocating with "positive" and "negative" as a network graph.

a small set of words are collocations of both *positive* and *negative*. A clear real-world effect on the collocation pattern can be seen in the collocates of *negative* that have to do with restrictions on free movement and the demands on having proof of recent negative test results, such as: *intyg* 'certificate', *krav* 'demand', *krävs* 'is demanded', *inresa* 'entry', *uppvisa* 'show'.

Association measures can also be used to study the association between words and a category. The heatmap in Figure 6 visualises the association between singular personal pronouns and genre. The map uses a normalized variant of pointwise mutual information, NPMI, which falls between -1 (never cooccur) and 1 (always cooccur). The map shows for instance that the first and second person pronouns (*jag* 'I', *mig* 'me', *du* 'you.sub.sg', *dig* 'you.obj.sg') are associated with the Social media and Blog genres, and that the gender neutral *hen* '(singular) they' is associated with the Medical genre. The second person pronouns are also associated with the Research and Authorities data. This high-level overview can be the starting point for a fine-grained, qualitative investigation for the mechanisms behind these associations.

2.4 *Grammatical analysis and Universal Dependencies*

SpaCy is an open-source, industrial-strength NLP pipeline with models for many languages. It offers for instance facilities for part-of-speech tagging,

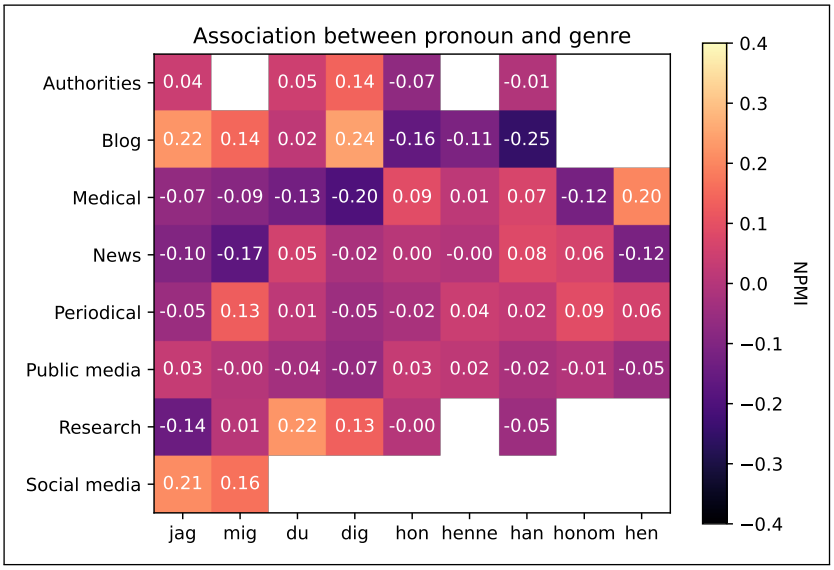


Figure 6: Association between singular person pronouns and genre in the dataset. Cells without observations (by definition NPMI $= -1$) are left uncoloured to avoid clutter.

dependency parsing, named-entity recognition and text classification. Here we have used SpaCy's Swedish model to part-of-speech tag and parse the COVID-19 data.[20] Table 2 shows the distribution of part-of-speech tags in the data. Nouns are most common part of speech with just over 22%, whereas verbs make up about 16% (main and auxiliary verbs) of the data.

The syntactic structures assigned by SpaCy are dependency structures following the Universal Dependencies framework (de Marneffe et al. 2021). An example parse can be seen in Figure 7. The automatic parser manages to correctly recognize many grammatical relations in this sentence, such as the PP postmodifying structures in *exemplaren av nytrycket av boken* 'copies of the reprint of the book', but it also makes mistakes (in red in the figure), for instance when it treats *Corona-PLANdemin* 'Corona PLANdemic' as an adverbial, rather than part of the NP *boken Corona-PLANdemin* 'the Corona PLANdemic book', or when it treats *har* 'has' as the root instead of as a temporal auxiliary to the main verb *kommit* 'come'.

An example use of parses is to construct a measure of syntactic complexity of the material. One such measure is the average number of finite subordinate structures per sentence: having more such subordinate clauses is associated

20 The Swedish pipeline in SpaCy can be found here: https://spacy.io/models/sv (version: 3.7.0). The Swedish model applied is called `sv_core_news_sm`.

Table 2: Distribution of part-of-speech labels in the corpus according to the spaCy analysis.

Part of speech		Count	%
NOUN	Noun	157512	22.3
ADP	Preposition	83873	11.8
VERB	Main verb	81221	11.5
PUNCT	Punctuation	71037	10.0
ADJ	Adjective	57588	8.1
PRON	Pronoun	48894	6.9
ADV	Adverb	46357	6.5
AUX	Auxiliary verb	34772	4.9
DET	Determiner	30295	4.2
PROPN	Proper noun	27915	3.9
CCONJ	Coordinator	22755	3.2
SCONJ	Subordinator	19233	2.7
PART	Particle	15646	2.2
NUM	Numeral	7269	1.0
SYM	Other symbol	521	0.0
INTJ	Interjection	211	0.0

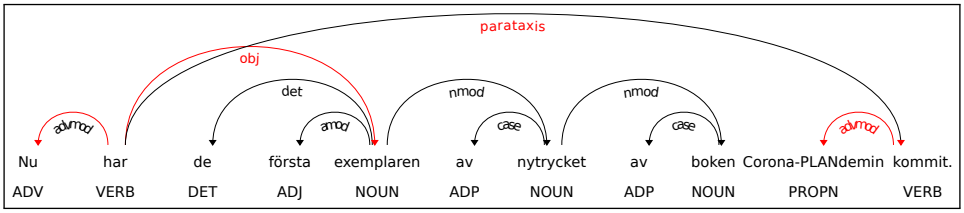


Figure 7: An example of a dependency tree from the dataset using the Swedish SpaCy model.

with increased complexity of the text. Subordinate clauses can be recognized by looking at the grammatical functions assigned by the parser. Figure 8 plots the relation between the number of subordinate clauses and sentence length for the Blog data and the News data. Perhaps somewhat surprisingly, the Blog data is more complex on several accounts: the average number of subordinate clauses is higher and, additionally, the subordinate clauses themselves are on average longer.

2.5 *Sentiment analysis*

Sentiment analysis is a computational technique for evaluating the polarity of opinions expressed in textual data. It enables scholars to track shifts in

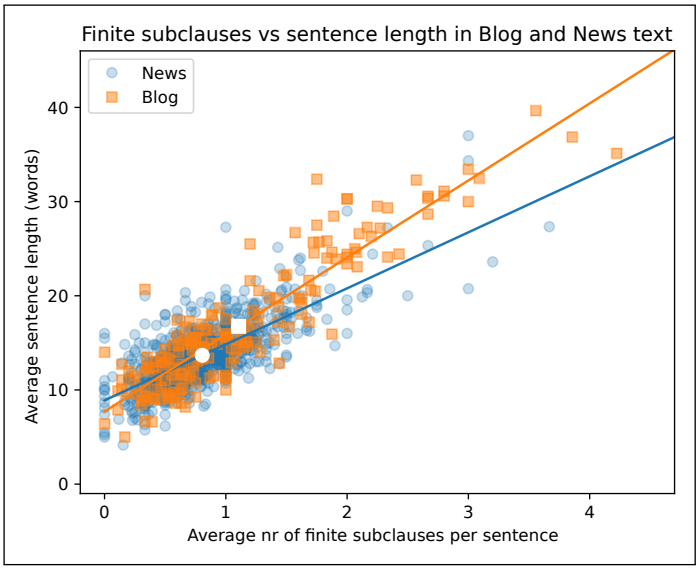
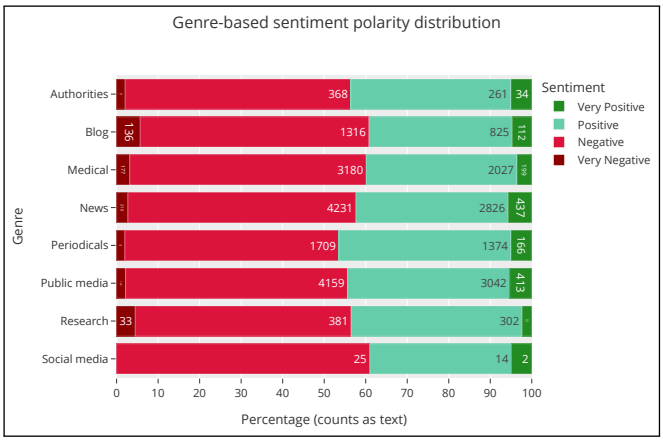


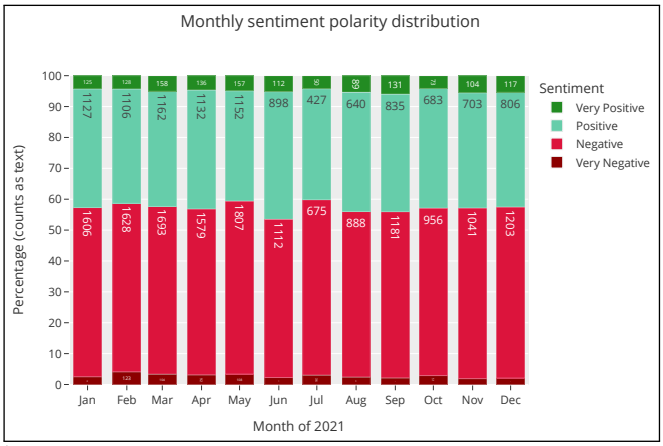
Figure 8: Texts from the Blogs genre are more complex than those from the News genre, both in terms of average sentence length and average number of finite subordinate clauses (the white points). Moreover, each additional finite subordinate clause is longer in Blog text than in News texts by on average 2.3 words (difference in regression line slopes, 95% confidence interval: 1.6–2.9 words/clause).

public opinion over time, assess attitudes toward significant historical events, and analyze the portrayal of individual characters in literature and other narratives, on a large scale. This quantitative approach offers a complementary lens through which to view qualitative sources, enabling researchers to uncover patterns and trends that may not be immediately apparent through traditional analysis alone.

A common method of sentiment analysis looks for the use of emotionally charged words from a so-called *sentiment lexicon*. However, scholars must carefully apply these computational tools to avoid overlooking the inherent subtleties and complexities in the texts under analysis. Literary works may contain complex meanings that require a deep understanding of their contexts. Texts from various historical periods or genres may use language and rhetorical strategies that complicate such direct sentiment analysis. Nevertheless, the "distribution of emotion-laden words" (Öhman 2021) can still provide valuable insights into the emotional tone of a text, serving as a foundation for more nuanced analysis. More advanced approaches to sentiment analysis involve machine learning, NLP and hybrid lexical-based



a.



b.

Figure 9: The sentiment distribution per genre (a) and per month (b) for 2021 in the dataset.

techniques, which enhance the ability to interpret sentiment by considering context, tone, and even the cultural background of the text. These advanced methods do come at the price of added complexity and development cost.

We apply a lexicon-driven approach to our material, using a list of about 2000 words marked up with polarity and strength (Nusko et al. 2016, Språkbanken Text 2017). We distinguish four categories of emotion-laden words, ranging from very negative to very positives. Results are visualised in Figure 9. The most striking thing about both pictures is how even the distribution is, throughout genres as well as over time. The (very) negative words have a fairly constant slight overhand (between 50-60%). The Blog

and Social media genres have the highest share of (very) negative words amongst emotion-laden words, followed by the Medical genre. If we were to pursue the topic further, we would probably look into what makes the Medical genre so negative. In addition, the Blog data contains the highest proportion of very negative terms.

2.6 *Topic modeling*

Topic modeling aims to uncover themes or topics in a collection of texts and is a hugely popular method in Digital Humanities.[21] Texts in a collection are stochastically assigned to a number of undefined topics on the basis of their linguistic content. These topics can then be interpreted by the researcher, by looking at the words associated with them, and their distribution throughout the material can be studied. Topic modelling can be contrasted with the way we have used genres in our dataset: whereas our "genres" are a classification imposed by us (by design) on the data, the categories from topic modelling come from the distributional properties of the data itself.

There are many implementations of the topic modeling idea, the most established technique being *Latent Dirichlet Allocation* (LDA, see Jelodar et al. 2019 for an overview). We, however, use the topic modelling library BERTopic, Grootendorst (2022), which presuposes a BERT-style language model that can create numerical meaning representations of the documents in the collection. For our Swedish data, we have access to the BERT models developed at KBLab at the National Library of Sweden (Rekathati 2021).[22] BERTopic requires minimal data preprocessing and its availability as a Python library means it is easy to integrate in our setup.[23]

The topics found using BERTopic in our data are summarized in Figure 10. We can interpret these topics by looking at the words associated with them. For instance, Topic 1, the most commonly present topic, appears to deal with pressure on the hospital care system (*patienter* 'patients', *krislägesavtalet* 'emergency labour agreement [in hospitals]', *IVA* 'ICU'), Topic 4 with speculations about the origin of the virus (*Kina* 'China', *Wuhan*, *ursprung* 'origin'), and Topic 7 with tracking the spread of the virus in the population (*avlopsvattnet* 'sewage', *PCR*, *smittspårning* 'contact tracing')

Looking at the distribution of these topics in the dataset now gives us new perspectives on the data. For instance, Figure 11 shows the frequency of

21 Already more then 10 years ago, Schmidt (2012: 49) wrote: "So many scholars in humanities departments are turning to the tool [of topic modeling] in their research that it is sometimes described as part of the digital humanities in itself."

22 Available from https://huggingface.co/KBLab/sentence-bert-swedish-cased.

23 See Haffenden & Sikora (2025; Chapter 2 in this handbook) for more information about the use of KBLab's BERT models.

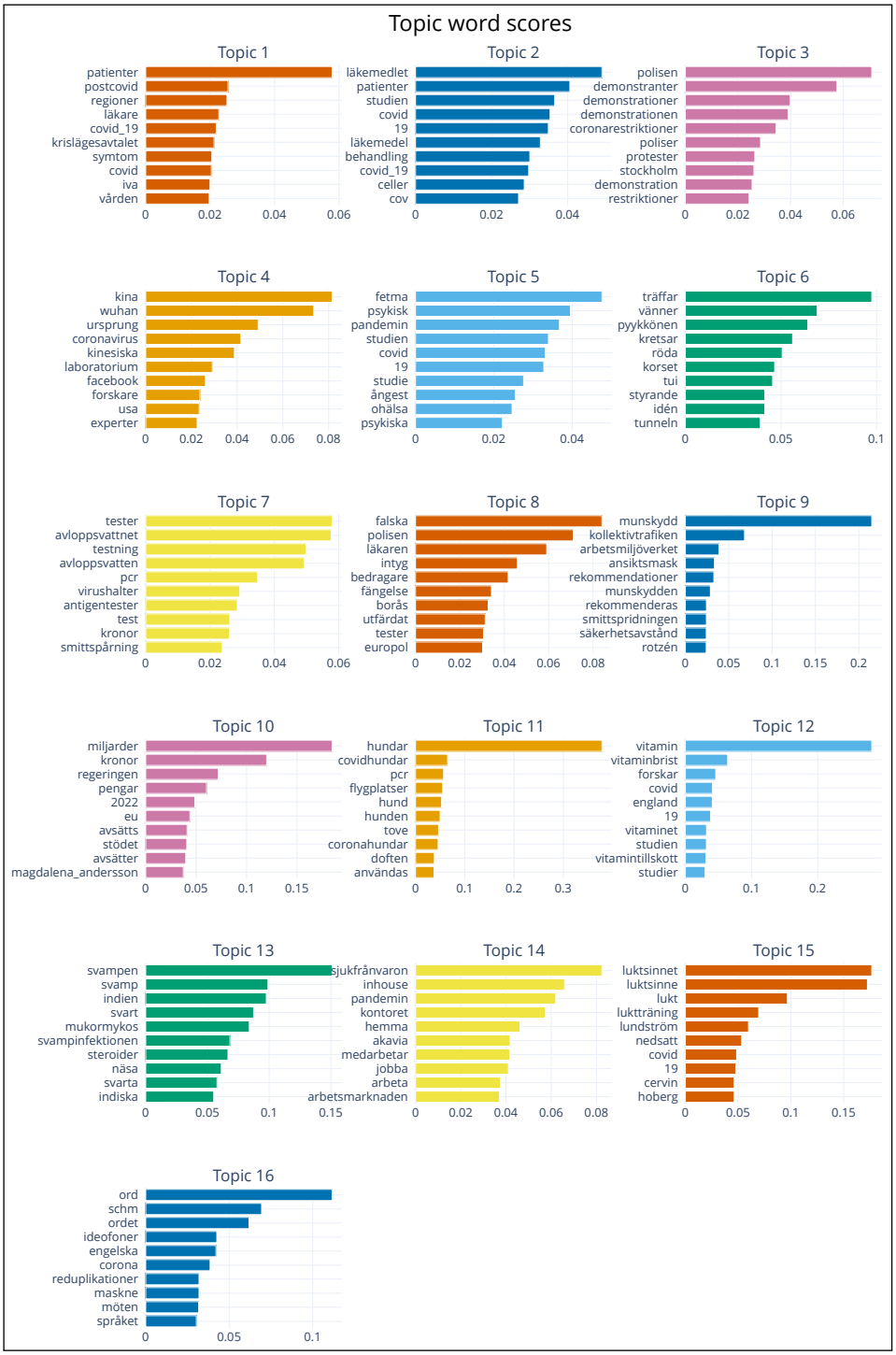


Figure 10: Topic word scores for the generated topics.

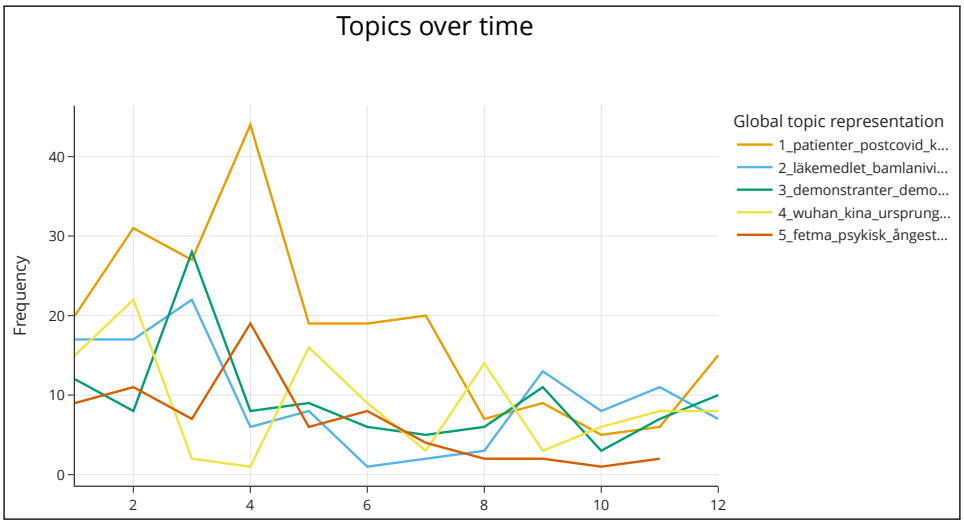


Figure 11: The distribution of the top 5 most common topics over time.

the five most common topics throughout the year. All of these "big topics" decrease in frequency, suggesting that the discourse becomes less dominated by them towards the end of the year, and more diverse.

3 *Final remarks*

In this chapter, we have given several basic examples of using interactive notebooks to perform text exploration tasks. Notebooks can be a powerful tool for researchers in Digital Humanities as pedagogical, analytical, and scholarly tools (De Keulenaar et al. in press; Talboom & Bell 2022, Chun & Elkins 2023), offering a flexible and efficient environment with rich textual documentation alongside the code, ease of collaboration and code interactivity and data visualization capabilities to help convey findings and insights. The ease with which interactive notebooks can be shared and published can also improve reproducibility, allowing other researchers to reproduce results, build upon existing work, and contribute to a collective understanding. This transparency is important in ensuring the robustness and credibility of humanities research.

Many tasks covered in this chapter, such as trend analysis of selected words or part-of-speech tagging, can be performed using dedicated tools like

AntConc[24], LancsBox[25], or the Korp and Mink tools from Språkbanken Text[26]. The notebooks presented in this chapter belong instead to a more general coding approach, suitable for those who wish to customize and automate their analysis. These tools are valuable for quick, hands-on exploration, but coding provides more flexibility for complex or large-scale analyses.

Acknowledgments

The research presented here was supported by *Huminfra*, a Swedish national infrastructure for the Humanities, funded by the Swedish Research Council and the consortium nodes (grant numbers 2021-00176 and 2023-00171), as well as by the Swedish national research infrastructure *Språkbanken* – jointly funded by its 10 partner institutions and the Swedish Research Council (grant numbers 2017-00626 and 2023-00161).

References

Ahnert, Ruth, Emma Griffin, Mia Ridge & Giorgia Tolfo. 2023. *Collaborative historical research in the age of big data: Lessons from an interdisciplinary project* (Elements in Historical Theory and Practice). Cambridge University Press. DOI: 10.1017/9781009175548.

Alderson, David. 2021. Interactive computing for accelerated learning in computation and data science. *INFORMS Transactions on Education* 22(2). 130–145. DOI: 10.1287/ited.2021.0261.

Barba, Lorena, Lecia Barker, Douglas Blank, Jed Brown, Allen Downey, Timothy George, Lindsey Heagy, Kyle Mandli, Jason Moore, David Lippert, Kyle Niemeyer, Ryan Watkins, Richard West, Elizabeth Wickes, Carol Willing & Michael Zingale. 2019. *Teaching and learning with Jupyter*. https://jupyter4edu.github.io/jupyter-edu-book/index.html.

Bednarek, Monika, Martin Schweinberger & Kelvin Lee. 2024. Corpus-based discourse analysis: From meta-reflection to accountability. *Corpus Linguistics and Linguistic Theory* 20(3). 539–566. DOI: 10.1515/cllt-2023-0104.

Blanke, Tobias, Giovanni Colavizza & Zarah van Hout. 2023. An open educational resource to introduce data analysis in python for the humanities. *Education for Information* 39(2). 105–119. DOI: 10.3233/EFI-230020.

24 https://www.laurenceanthony.net/software/antconc/
25 https://lancsbox.lancs.ac.uk/
26 https://spraakbanken.gu.se/en/tools

Chun, Jon & Katherine Elkins. 2023. The crisis of Artificial Intelligence: A new Digital Humanities curriculum for human-centred AI. *International Journal of Humanities and Arts Computing* 17(2). 147–167. DOI: 10.3366/ijhac.2023.0310.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402.

Dombrowski, Quinn, Tassie Gniady & David Kloster. 2019. Introduction to Jupyter Notebooks. *Programming Historian* 8. DOI: 10.46430/phen0087.

Evert, Stefan. 2009. Corpora and collocations. In *An international handbook*. Anke Lüdeling & Merja Kytö (eds.). Berlin: De Gruyter Mouton. Chap. 58, 1212–1248. DOI: 10.1515/9783110213881.2.1212. https://doi.org/10.1515/9783110213881.2.1212.

Firth, John. 1957. A synopsis of linguistic theory, 1930–55. In *Studies in linguistic analysis. Special volume of the Philological Society*, 1–32. Oxford: Blackwell.

Granger, Brian & Fernando Pérez. 2021. Jupyter: Thinking and storytelling with code and data. *Computing in Science & Engineering* 23(2). 7–14. DOI: 10.1109/MCSE.2021.3059263.

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Haffenden, Chris & Justyna Sikora. 2025. Doing digital research at KBLab: A practical introduction to using the National Library of Sweden's data lab. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 17–56. University of Tartu Library. DOI: 10.58009/aere-perennius0171.

Hardebolle, Cécile. 2023. *Online interactive textbooks with Jupyter Notebooks*. https://www.epfl.ch/education/educational-initiatives/cede/teaching-interactively/jupyter-notebooks-for-education/teaching-and-learning-with-jupyter-notebooks/online-interactive-textbooks-with-jupyter-notebooks/.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li & Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications* 78(11). 15169–15211. DOI: 10.1007/s11042-018-6894-4.

Karsdorp, Folgert, Mike Kestemont & Allen Riddell. 2021. *Humanities data analysis: Case studies with Python*. See https://www.humanitiesdataanalysis.org/ for the book's contents in notebook form. Princeton University Press.

de Keulenaar, Emillie, Thomas Poell, Anne Helmond, Bernhard Rieder & Jasmijn Van Gorp. In press. Computational cross-media research: Trac-

ing divergences between normative Dutch television and social media discourses on the 'refugee crisis' (2013-2018). *Convergence*. DOI: 10.1177/13548565241258956.

Knuth, D. E. 1984. Literate programming. *The Computer Journal* 27(2). 97–111. DOI: 10.1093/comjnl/27.2.97.

Kokkinakis, Dimitrios. 2021. Insights on a Swedish covid-19 corpus. In Monica Monachini & Maria Eskevich (eds.), *CLARIN Annual Conference proceedings*, 31–34. https://office.clarin.eu/v/CE-2021-1923-CLARIN2021_ConferenceProceedings.pdf.

Land, Kaylin, Geoffrey Rockwell & Andrew MacDonal. 2021. Spyral notebooks as a supplement to voyant tools. In *CSDH-SCHN 2021: Making the net work, virtual*. Canadian Society for Digital Humanities. DOI: 10.17613/2bsr-xp53.

Melgar-Estrada, Liliana, Marijn Koolen, Kaspar Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martinez-Ortiz, Jaap Blom & Roeland Ordelman. 2019. The CLARIAH Media Suite: A hybrid approach to system design in the humanities. In *Proceedings of the 2019 conference on human information interaction and retrieval*, 373–377. Glasgow, Scotland UK: Association for Computing Machinery. DOI: 10.1145/3295750.3298918.

Nusko, Bianka, Nina Tahmasebi & Olof Mogren. 2016. Building a sentiment lexicon for Swedish. In *Digital humanities 2016. from digitization to knowledge 2016: Resources and methods for semantic processing of digital works/texts*, vol. 126 (Linköping Electronic Conference Proceedings 6), 32–37. LiU Electronic Press. https://ep.liu.se/ecp/126/006/ecp16126006.pdf.

Öhman, Emily. 2021. The validity of lexicon-based sentiment analysis in interdisciplinary research. In Mika Hämäläinen, Khalid Alnajjar, Niko Partanen & Jack Rueter (eds.), *Proceedings of the workshop on natural language processing for digital humanities*, 7–12. NIT Silchar, India: NLP Association of India (NLPAI). https://aclanthology.org/2021.nlp4dh-1.2/.

Piantadosi, Steven. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* 21(5). 1112–1130. DOI: 10.3758/s13423-014-0585-6.

Rekathati, Faton. 2021. *Introducing a Swedish sentence transformer*. The KBLab Blog. https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/.

Rule, Adam, Aurélien Tabard & James Hollan. 2018. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3173574.3173606.

Schmidt, Benjamin. 2012. Words alone: Dismantling topic models in the Humanities. *Journal of Digital Humanities* 2(1). 49–66.

Språkbanken Text. 2017. *Sentimentlexikon*. Dataset. DOI: 10.23695/1yyf-6409.
Talboom, Leontien & Mark Bell. 2022. Keeping it under lock and keywords: Exploring new ways to open up the web archives with notebooks. *Archival Science* 22(3). 393–415. DOI: 10.1007/s10502-022-09391-6.
York, Christopher. 2017. Exploratory data analysis for the Digital Humanities: The Comédie-Française registers project analytics tool. *English Studies* 98(5). 459–482. DOI: 10.1080/0013838X.2017.1332024.
Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Addison-Wesley.

List of abbrevations

BERT	Bidirectional Encoder Representations from Transformers
Colab	Google Colaboratory
LLR	Log-likelihood ratio
NLP	Natural Language Processing
NPMI	Normalized pointwise mutual information
PMI	Pointwise Mutual Information

Corresponding author

Dimitrios Kokkinakis
Språkbanken Text
Department of Swedish,
Multilingualism, Language
Technology
University of Gothenburg
dimitrios.kokkinakis
@svenska.gu.se