6 Multimodal analysis and data mining

Large datasets and complex computational problems call for more processing power than single laptops or workstations can provide. These approaches also require thinking about the features that can be processed in datasets and how to specify the procedures for identifying, counting, and comparing them. Much progress has been made in recent decades so that solutions are available in what is off-the-shelf software and platforms though more sophisticated research requires customization of the algorithms and programming as well as the methods of analysis and display.

6a Data mining and high-performance computing for humanities

Bonus content: high-performance computing in humanities

High-performance computing (HPC) refers to the use of powerful computers and parallel processing techniques to perform complex calculations at high speed. HPC systems are often comprised of clusters of interconnected computers, known as supercomputers, that work together to process large volumes of data and execute intensive computational tasks. These systems can process data much faster than standard computing systems, enabling researchers to analyze, model, and simulate complex phenomena that would be infeasible with ordinary computers. In other words, this is not the type of computing that your laptop can do on its own. The infrastructure required for this type of supercomputing encompasses a variety of systems and resources designed to support complex and large-scale computational tasks.

One such system is a computational cluster—a collection of interconnected computers, often called nodes, that work together as a single, powerful system. Each node is a separate computer, but when combined in a cluster, they can perform parallel processing tasks, dividing the workload to complete complex computations more quickly. Clusters are commonly used in scientific research, and humanities scholars can utilize them for large-scale text mining, image analysis, or data-intensive historical research. Clusters typically require some familiarity with the command line to operate, and access is often provided through research institutions or dedicated HPC facilities.

Computational clusters are often housed within a data center. Data centers provide the physical infrastructure for HPC resources. These facilities are designed to manage, process, and store vast amounts of data securely and efficiently. Data centers have climate-controlled environments, backup power, and high-security protocols to ensure reliable and continuous operation. For humanities scholars, data centers can offer access to storage and processing power that would be difficult or costly to maintain independently. Universities and research organizations often provide access to data centers as part of their HPC resources. There are also national resources for humanities HPC available through a partnership between the National Science Foundation and the National Endowment for the Humanities.

Cloud computing platforms, such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure, offer on-demand access to HPC resources through the internet. Cloud services allow humanities scholars to scale their computational needs based on the project's requirements, without investing in physical infrastructure. They provide storage, computing, and software tools that can be accessed and managed remotely, making them highly accessible. Additionally, many cloud providers offer pricing models suited for occasional, project-based usage, making HPC more affordable and accessible to scholars who need to analyze large datasets but lack dedicated resources.

Traditionally, HPC infrastructure was more common in fields like physics or bioinformatics, but it is increasingly accessible for humanities researchers. Many academic institutions now provide training programs, support services, and partnerships to help scholars learn how to use HPC resources effectively. Additionally, initiatives in digital humanities, along with grant funding and collaborative research opportunities, have made HPC infrastructure more available to humanities scholars who seek to explore large-scale questions in literature, history, and cultural studies.

Bonus content: techniques for big data and HPC integration

One task the computational clusters are used for is parallel processing. **Parallel processing** is a computing technique that divides a large task into smaller, independent subtasks that can be processed simultaneously across multiple processors or computing nodes. By distributing these subtasks, parallel processing significantly speeds up data processing and analysis, making it ideal for handling extensive datasets, complex simulations, or high-volume computations that would take too long on a single processor.

In parallel processing, tasks are often broken down using data, task, or pipeline parallelism. **Data parallelism** involves dividing data into chunks that can be processed independently. For example, in a large text analysis project, each node might be assigned a subset of texts from a corpus, performing the same analytical tasks (e.g., word frequency analysis or sentiment analysis) on different segments. Once complete, results from each node are aggregated to produce the final analysis. When using **task parallelism**, different tasks or functions within a program are assigned to different processors. For example, in image analysis, one processor might perform edge detection, while another handles color analysis. Each task runs concurrently, and their outputs are combined as needed. **Pipeline parallelism** breaks tasks into a sequence of stages, similar to an assembly line. Each stage performs part of the overall task and passes results to the next stage. This approach is often used in real-time data processing where data needs to be processed in a sequence of operations, such as in streaming data analysis (Czech 2017).

How do we apply these methods to humanistic data? Suppose a humanities researcher is analyzing a massive corpus of digitized books for themes and sentiment over time. Using parallel processing, the corpus can be divided into smaller subsets, with each subset sent to a different computing node. Each node processes its subset independently, performing sentiment analysis and extracting key themes. The results from each node are then aggregated to produce a comprehensive view of themes and sentiments across the entire corpus. This distributed approach allows the researcher to complete the analysis in a fraction of the time it would take on a single processor. Much more manipulation of data can be done using Python. **[See CW for resources.]**

To fully harness the potential of HPC techniques in humanities research, Python has emerged as an indispensable tool. Python's extensive ecosystem of libraries provides powerful solutions for parsing, manipulating, and processing large datasets. For example, pandas enables efficient management of tabular data, while text analysis libraries like NLTK and spaCy simplify tasks such as tokenization, named entity recognition, and sentiment analysis. When scaling workflows to accommodate parallel processing, tools like Dask and PySpark allow researchers to distribute Python tasks across clusters of computing nodes, seamlessly integrating HPC into their workflows. A researcher working with a vast corpus of historical texts could use Dask to partition the dataset into manageable chunks, perform analysis on each partition simultaneously, and aggregate the outputs to derive insights. By combining parallel processing techniques with Python's capabilities, humanities researchers can tackle complex questions at a scale that was previously unimaginable.

Exercise 6.1: Google Ngram Viewer

Open the Google Ngram Viewer (https://books.google.com/ngrams/) and select a date range. Enter several terms or names and see what has changed over time. How much can you trust your results? What are they based on?

How-to example

If you enter the following terms "digital humanities, data science, media studies" and limit your search to 1960–2022, you will see that looks something like (Figure 6.1):

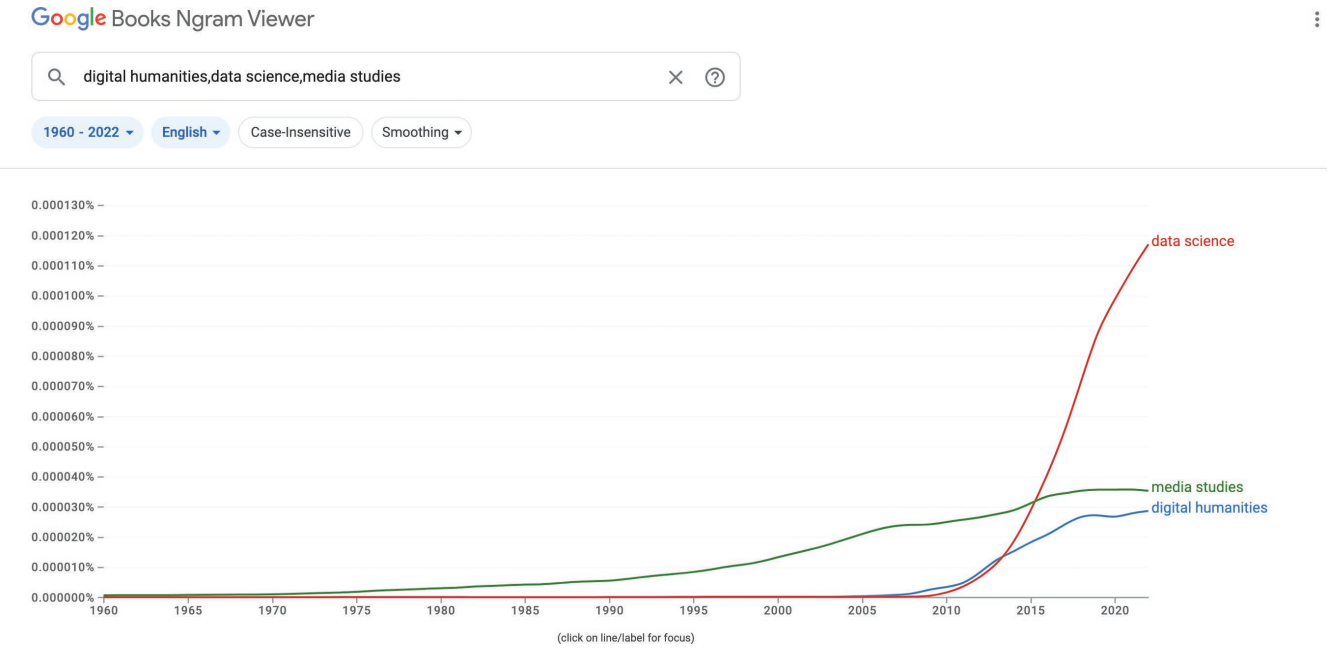

Figure 6.1 Google Ngram Viewer result for digital humanities, data science, media studies (1960–2022).

Google Ngram Viewer is a valuable tool for exploring language trends over time, but its results should be interpreted with caution due to several limitations. The data comes from a large corpus of digitized books, which skews toward Western, academic, and English-language texts, potentially introducing sampling bias. It measures the frequency of n-grams normalized by total word count per year but lacks context, leading to issues like ambiguous meanings or metadata errors. Additionally, it excludes other media forms like newspapers and oral traditions, which can omit significant cultural insights. For accurate interpretation, cross-check results, consider historical context, and use advanced filters to refine searches. Treat the findings as indicative rather than definitive of linguistic or cultural trends.

Exercise 6.2: Reading an API

Look at the Australian National Library Trove API console (https://troveconsole.herokuapp.com/#otherzones). What can you learn from reading the documentation? Construct a query and assess the results. What ideas does this give you for designing an API? Compare the results of a search in the library's catalog and the format of its results, with the XML output generated by Trove which is useful for data analysis. (Also useful: Getting Data for Digital Humanities with APIs: A Gentle Introduction (https://studentwork.prattsi.org/dh/2019/05/13/getting-data-for-digital-humanities-with-apis/) by Abigail Walker from 2019)

How-to example

The Trove API allows users to query the National Library of Australia's digital collections in various categories, such as newspapers, images, and archives. The API can be queried in several ways, offering rich filtering options for refining search results. Some key parameters include:

- Search query (q): This is the core parameter for defining the search terms, such as a name or topic of interest.
- Categories: You can specify the type of content you want to search, like newspaper, image, book, or archive.
- Filters: Filters allow you to limit results by factors such as publication date, place, or the presence of illustrations. For instance, using l-state=Victoria will restrict results to those published in Victoria, while one-year = 1924 narrows the search to that year.

For example, querying for the term "wragge" in the newspaper category with results in the state of Victoria for the decade of 1920–29 can be achieved through this URL structure: https://api.trove.nla.gov.au/v3/result?q=wragge&category=newspaper&encoding=json&l-state=Victoria&l-decade=192.

The Trove API illustrates a well-structured approach to handling large datasets, with its clear separation of query terms and result constraints (e.g., limiting by decade or place). Its response format, typically JSON or XML, is useful for both end-users and data analysis. The JSON format provides structured metadata for easy parsing and integration into data workflows. For data analysis in the digital humanities, the JSON format enables easy manipulation of the data. Each article's metadata (like publication date, article content, and associated media) is returned in a consistent, machine-readable format, making it efficient for building datasets, visualizations, or conducting large-scale text mining.

A typical library catalog would also allow for you browsing, searching, and filtering. The display of the information would prioritize legibility for the user. The Trove API results are not as immediately legible to the average user; however, the results are machine readable, allowing users to make use of the data for their own purposes.

Recommended readings

Bajorek, Joan Palmiter. 2019. "Voice Recognition Still Has Considerable Race and Gender Biases." *Harvard Business Review*. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases.

Hartquist, John. 2018. "Audio Classification Using FastAI and On-the-Fly Frequency Transforms." *Towards Data Science*. https://towardsdatascience.com/audio-classification-using-fastai-and-on-the-fly-frequency-transforms-4dbe1b540f89.

Harwell, Drew. 2019. "Federal Study Confirms Racial Bias of Many Facial Recognition Systems." *Washington Post*. https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/.

Manovich, Lev. 2011. "How to Compare One Million Images." *Cultural Analytics*. https://manovich.net/index.php/projects/how-to-compare

6b Multimodal analysis

Multimodal analysis refers to the study of multiple types of data or "modes" (e.g., text, images, audio, video, gestures) to create a more comprehensive understanding of communication, cultural artifacts, or social phenomena. Unlike

single-mode analysis, which might focus on text alone, multimodal analysis integrates various forms of information, enabling researchers to analyze how different modes interact to convey meaning.

Bonus content: beginnings of automated humanistic scholarship

A milestone frequently cited in early digital humanities projects is the work of Father Roberto Busa. He was engaged with text analysis in the form of a concordance—a list of all words in a work or body of work. This was a form of analysis with a long history within religious and classical scholarship, but Father Busa's project was ambitious intellectually as well as logistically. He was focused on the concept of "presence" in the Latin texts of the 13th-century scholar Thomas Aquinas. This was a metaphysical concept and thus had no simple literal meaning.

Busa had tracked the instances of the words *praesens* and *praesentia* to address their contexts in the 1940s (Busa 1980). He created thousands of index cards for individual instances and the phrases in which they were found. When he realized that the full corpus exceeded ten million words, he began to consider mechanical aids. This led him to a collaboration with IBM, thanks to the support of its CEO Thomas Watson. Many of the approaches Busa designed for his project, such as identifying text types (for example, the use of citations) have become part of standard markup and statistical analysis. Working with punch cards and a list of typological codes, Busa established a systematic approach to the analysis of natural language, including linking all forms and versions of a word to its root (Terras and Nyhan 2016). Busa was working in analog materials but developing formal methods compatible with automated processes.

A second area of early automated text analysis was in the area of stylometrics or stylometry (Hai-Jew 2015). Long-standing debates about whether or not William Shakespeare was the author of all of his plays, or whether some were actually composed by Christopher Marlowe, remain pressing matters for many scholars (Fox et al. 2012). The idea that statistical approaches could be brought to bear on the problem motivated formal analysis of style. Sentence length, grammatical structure, vocabulary choices, and other features of the texts were used to make comparisons. Many of the features on which style was formally addressed had a history in analog scholarship. The task of making formal parameters on which to analyze style is a useful intellectual exercise, as is any other attempt to make explicit parameters on which to formalize traditional humanistic approaches for computational purposes.

Methods for statistical processing are now far more complex than the counting and sorting of words into lists that were central to Busa's early project or the techniques developed for analysis of style (Sculley and Pasanek 2008). Current tools and platforms combine counting and sorting techniques with statistically driven capacities. The differences between these will be a recurring theme. The following sections will focus on different modalities (e.g., text, images, audio recordings, video, and film) of data and the types of data mining and analysis that can be utilized for digital humanities research. Every medium poses its own set of challenges for extracting information in a meaningful way. But each process has in common the same set of requirements—to translate analog or digital materials into a form in which a **feature set** can be identified, parameterized, tokenized, and processed computationally. This generally involves making discrete features from continuous phenomena. As in all such processes, the conceptual work and the technological developments have to coordinate. The ways we think about music or images will structure some of the ways digital representations are created—and the purposes to which they are put. In whose interest is it to do data mining of images or social media? Art historians, or police surveillance units? And music? Artists looking for inspiration, scholars studying historical materials, or industry sleuths tracking piracy—or those indulging in it (Kennedy and Moss 2015)?

Exercise 6.3: Voyant and Mallet

Identify a block of text you wish to analyze. (Consider making use of a resource like Project Gutenberg (https://www.gutenberg.org/) to find a text.) Go to https://voyant-tools.org/ and enter a block of text with which you are familiar. Look at the first display and figure out what each pane is showing. Play with the other tools and display modes. Do they match? Why is there a range of results in the displays? Then, see if you can understand the workings of Mallet by following this tutorial (https://programminghistorian.org/en/lessons/topic-modeling-and-mallet).

How-to example

From Project Gutenberg, use the *State of the Union Presidential Addresses from 1790–2006* and copy from the Plain Text UTF-8 file and paste it into Voyant (Figure 6.2).

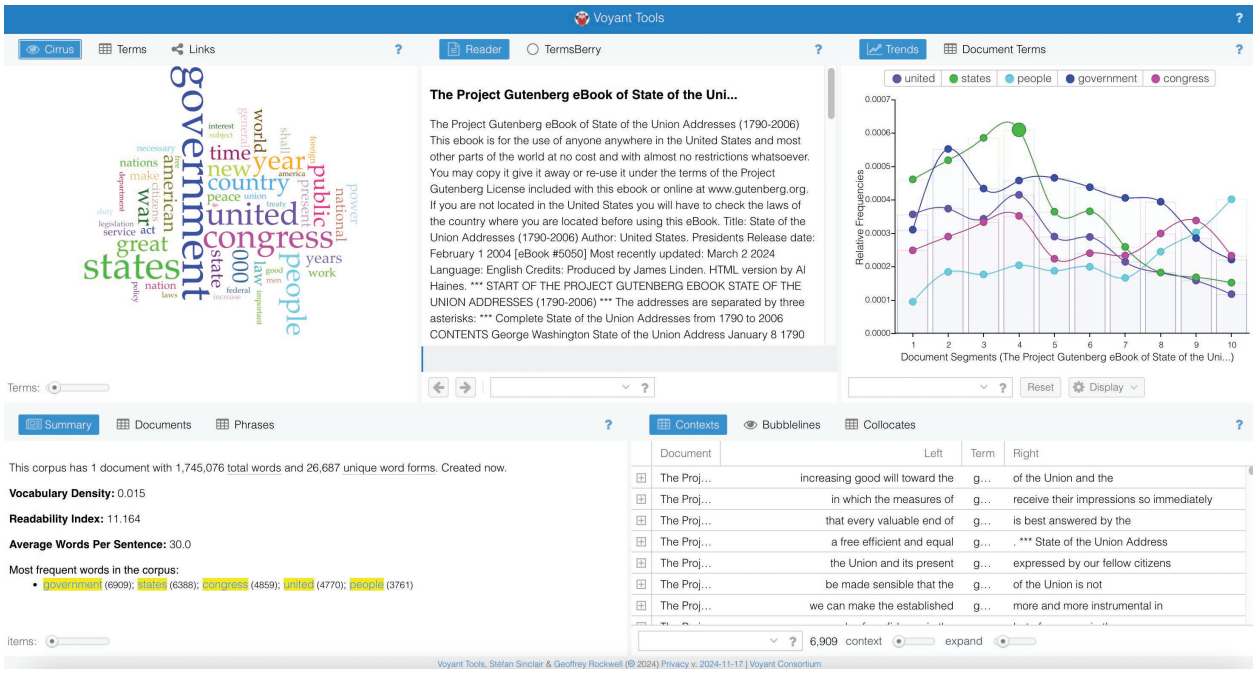

Figure 6.2 Voyant dashboard interface with the state of the union presidential addresses from 1790 to 2006.

Hovering over the words in the top left pane will show the number of occurrences for each word. The words are sized relative to the count. The upper center panel displays the text. The top right pane shows the most frequently used words and their relative frequency throughout the text. The bottom left provides a summary overview of the document, and the bottom right lists the text that comes right before and right after the most frequently used words within the text.

If you select a single word, you will notice that all the panes will update to focus on the particular term that has been chosen. For example, if you select "congress," the windows will adjust to look something like (Figure 6.3):

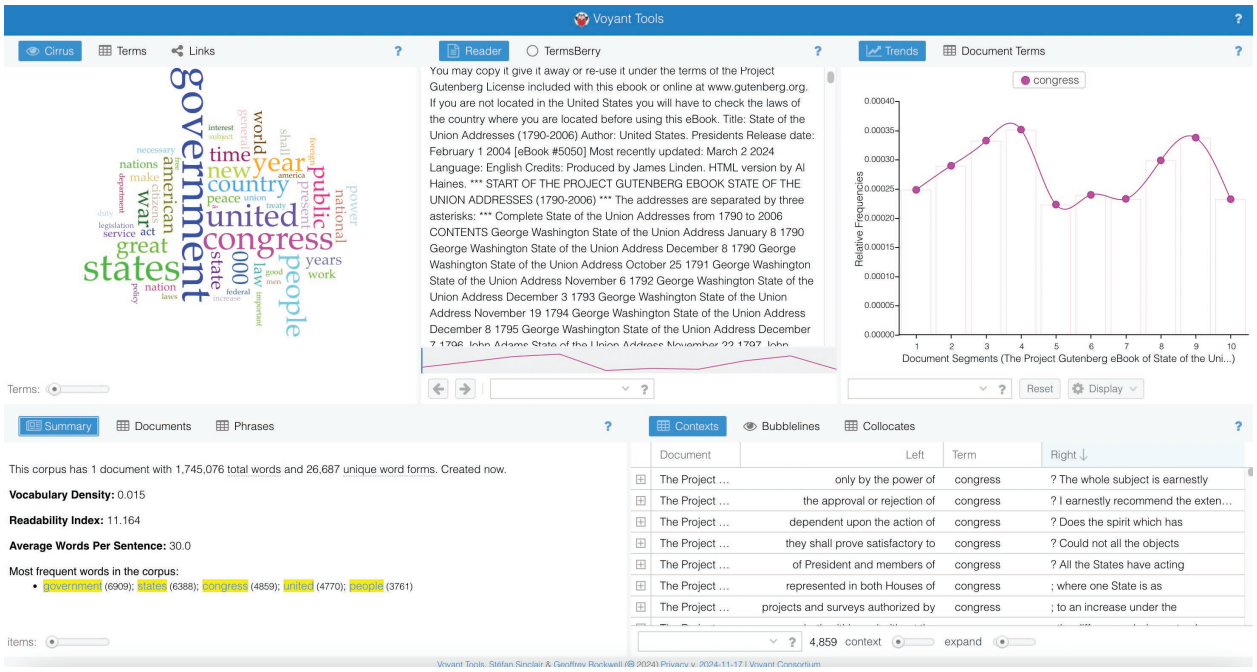

Figure 6.3 A Voyant tools dashboard presents a focused analysis of the word "congress" within the Project Gutenberg eBook of the state of the union addresses (1790–2006).

This will direct you to specific parts of the text where the word "congress" is used. You can select multiple words and consider the ways that these topics have shifted emphasis over the years by different US presidents.

Exercise 6.4: Cultural analytics

Look at Lev Manovich's Cultural Analytics Lab overview (https://blog.softwarestudieslab.com/p/overview-slides-and-video-articles-why.html) and projects (https://lab.culturalanalytics.info/p/projects.html). Design a project for which cultural analytics would be useful. Think in terms of the large scale of comparison and stay within humanities disciplines.

How-to example

A DH project could be designed around the covers of the Nancy Drew Mystery Series (https://en.wikipedia.org/wiki/Nancy_Drew_Mystery_Stories) to consider how the visual representation of the leading female character changed over the seven decades and 175 books within the series. Using ImagePlot (https://blog.softwarestudieslab.com/p/software-for-digital-humanities.html), this book covers could be visualized chronologically, but also by visual similarity. Similar visualizations could be done for the covers of Hardy Boys Mystery Stories (https://en.wikipedia.org/wiki/The_Hardy_Boys#Books), which were published during a similar time period and include 190 volumes within the series. The visual analysis would be fertile ground for a historical comparison of visualizing gender for youth adult Americana literature.

Exercise 6.5: Sound files

Examine the project by Tanya Clement, Hipstas "John A. Lomax and Folklore Data" (https://hipstas.org/2015/05/11/john-a-lomax-and-folklore-data/) What are the ways in which these folklore files from the early 20th century become more useful as a result of the digital interventions? What other kinds of materials do you think would benefit from such research?

How-to example

The HiPSTAS project (High Performance Sound Technologies for Access and Scholarship), led by Tanya Clement, significantly enhances the utility of John A. Lomax's early 20th-century folklore recordings through innovative digital interventions. By digitizing and annotating these archives, HiPSTAS makes them more accessible and easier to navigate, enabling researchers to locate specific pieces, such as sing, speech, and instrumental sections, within vast collections. Advanced computational tools like clustering, spectrographic visualization, and machine learning are applied to analyze audio files, uncovering patterns and features that might otherwise remain undetected. These technologies not only help to preserve these delicate recordings for future generations but also encourage critical re-evaluations of their content, including considerations of the sociopolitical contexts in which they were created and the agency of the contributors.

Beyond folklore, similar digital methodologies could benefit a variety of materials. For instance, indigenous language recordings could be preserved and analyzed to support revitalization efforts. Historical radio broadcasts, rich in cultural and historical insights, could be systematically studied using these tools. Music ethnography archives from diverse traditions could reveal patterns of cross-cultural influence, while oral history projects could become more accessible and analyzable through transcription and metadata enrichment. These efforts not only safeguard heritage but also open new avenues for interdisciplinary research.

Exercise 6.6: Media processing

If you were given the task of teaching an automated system to distinguish between news stories and advertisements in a television broadcast, what features would you identify for digital processing? Keep in mind that the task is to identify features that can be distinguished on the basis of their formal properties.

How-to example

To teach an automated system to distinguish between news stories and advertisements in a television broadcast, it is essential to focus on measurable and distinguishable formal properties across visual, audio, textual, and temporal dimensions. Visually, advertisements often feature prominent logos, vibrant colors, and rapid scene transitions, whereas news broadcasts rely on subdued tones, consistent studio settings, and longer visual sequences. Text overlays in news stories tend to follow a standardized format, such as ticker tapes or static headlines, while advertisements employ dynamic, eye-catching fonts and designs.

From an audio perspective, news segments typically emphasize neutral speech with minimal background music, whereas advertisements frequently include enthusiastic, sales-pitch-like tones, jingles, and sound effects. The language

used in news content is formal and fact-driven, while advertisements use persuasive, hyperbolic, and product-centered language, often highlighting keywords like "sale" or "free." Temporal properties also provide key distinctions: advertisements are generally short (15–60 seconds) and appear during predictable breaks, while news occupies longer, uninterrupted time blocks.

By combining these features, a machine learning model can be trained to classify content effectively. A convolutional neural network could analyze visual elements, while natural language processing and audio processing techniques would address textual and sound-based distinctions, ensuring comprehensive classification across media types.

Recommended readings

Hiippala, T. 2021. "Distant Viewing and Multimodality Theory: Prospects and Challenges." *Multimodality & Society* 1 (2): 134–52. https://doi.org/10.1177/26349795211007094.

Janicke, S., G. Franzini, M. F. Cheema, and G. Scheuermann. 2015. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." In *EuroGraphics Conference on Visualization*. https://pdfs.semanticscholar.org/20cd/40f3f17dc7d8f49d368c2efbc2e27b0f2b33.pdf.

Jewitt, Carey, Josephus Johannes Bezemer, and Kay L. O'Halloran. 2016. *Introducing Multimodality*. London: Routledge, Taylor & Francis Group. https://site.ebrary.com/id/11176915.

Kress, Gunther R., and Theo Van Leeuwen. 1996. *Reading Images: The Grammar of Visual Design*. London: Routledge.

Piper, Andrew, and Richard Jean So. 2015. "Quantifying the Weepy Bestseller." *The New Republic*. https://newrepublic.com/article/126123/quantifying-weepy-bestseller.

Schulz, Kathryn. 2011. "What Is Distant Reading?" *New York Times Book Review.* https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html.

Smits, Thomas, and Melvin Wevers. September 2023. "A Multimodal Turn in Digital Humanities. Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections." *Digital Scholarship in the Humanities* 38 (3): 1267–80, https://doi.org/10.1093/llc/fqad008.

Bibliography

Acerbi, Alberto, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. "The Expression of Emotions in 20th Century Books." *PLoS One* 8 (3). www.ncbi.nlm.nih.gov/pmc/articles/PMC3604170/.

Anderson, Steve. 2017. *Technologies of Vision*. Cambridge: MIT Press.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication, & Society* 15 (5): 662–79.

Busa, Robert. 1980. "The Annals of Humanities Computing: The Index Thomisticus." www.alice.id.tue.nl/references/busa-1980.pdf.

Clement, Tanya. 2012. "Announcing High Performance Sound Technologies for Access." https://tanyaclement.org/2012/08/09/hipstas/ and https://hipstas.org/2015/05/11/john-a-lomax-and-folklore-data/.

Czech, Zbigniew J. 2017. *Introduction to Parallel Computing*. Cambridge: Cambridge University Press. https://login.proxy.bib.uottawa.ca/login?url=https://doi.org/10.1017/9781316795835.

Delice, Ali. 2010. "The Sampling Issues in Quantitative Research." *Educational Sciences: Theory and Practice* 10 (4). https://files.eric.ed.gov/fulltext/EJ919871.pdf.

Dunne, Carey. 2015. "Microsoft's New Emotion-Detecting App Deems the Mona Lisa 43% Happy." *Hyperallergic*. https://hyperallergic.com/261508/microsofts-new-emotion-detecting-app-deems-the-mona-lisa-43-happy/.

Fox, Neal, Omran Ehmodea, and Eugene Charniak. 2012. "Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate." https://cs.brown.edu/research/pubs/theses/masters/2012/ehmoda.pdf.

Giannakopoulos, Theodoros, and Aggelos Pikrakis. 2014. *Introduction to Audio Analysis*. Academic Press. www.sciencedirect.com/book/9780080993881/introduction-to-audio-analysis.

Guldi, Jo. 2018. "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora." *Journal of Cultural Analytics*. https://culturalanalytics.org/article/11028.

Hai-Jew, Shalin. 2015. "A Light Stroll through Computational Stylometry and Its Early Potential." *C2C Digital Magazine*. https://scalar.usc.edu/works/c2c-digital-magazine-fall-winter-2016/a-light-stroll-through-computational-stylometry-and-its-early-potential.

Jofre, Ana, Josh Cole, Vincent Berardi, Carl Bennett, and Michael Reale. 2020. "What in a Face? Gender Representations of Faces in Time, 1940s–1990s." *Journal of Cultural Analytics*. https://doi.org/10.22148/001c.12266.

Karsdorp, Folgert, Kestemont, Mike, and Riddell, Allen. 2021. *Humanities Data Analysis: Case Studies with Python*. Princeton University Press.

Kennedy, Helen, and Giles Moss. 2015. "Known or Knowing Publics? Social Media Data Mining and the Question of Public Agency." *Big Data & Society* 2 (2), SAGE Publications Ltd. https://doi.org/10.1177/2053951715611145.

Kuhn, Virginia. 2018. "Images on the Move." In *The Routledge Companion to Media Studies and Digital Humanities Routledge*. New York: Routledge.

Lamarche, Stephen. 2012. *LARB*. Los Angeles, October. https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/.

Lee, Changsoo. 2019. "How Are 'Immigrant Workers' Represented in Korean News Reporting?—A Text Mining Approach to Critical Discourse Analysis." *Digital Scholarship in the Humanities* 34 (1): 82–99. https://doi.org/10.1093/llc/fqy017.

Mandell, Laura. 2019. "Gender and Cultural Analytics: Finding or Making Stereotypes?" In *Debates in Digital Humanities*. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2019.

Ogihara, Mitsunori, and George Tzanetakis. 2014. "Special Section on Music Data Mining." *IEEE Transactions on Multimedia* 16 (5): 1185–87. https://ieeexplore.ieee.org/document/6856270.
Patchet, François, Gert Westermann, and Damien Laigre. n.d. "Musical Data Mining for Electronic Music Distribution." www.music.mcgill.ca/~ich/classes/mumt621_09/Query%20Retrieval/Pachetwedelmusic.pdf.
Sandelowski, Margarete. 1995. "Sample Size in Qualitative Research." *Research in Nursing and Health*. https://onlinelibrary.wiley.com/doi/abs/10.1002/nur.4770180211.
Schmidt, Benjamin M. 2013. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities*. https://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.
Sculley, D., and B. M. Pasanek. 2008. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities." *Literary and Linguistic Computing* 23 (4): 409–24. https://doi.org/10.1093/llc/fqn019.
Serlen, Rachel. 2010. "The Distant Future? Reading Franco Moretti." *Literature Compass* 7. https://warwick.ac.uk/fac/arts/english/currentstudents/undergraduate/modules/fulllist/special/en264/serlen_reading_franco_moretti.pdf.
So, Richard Jean, and Edwin Roland. 2020. "Race and Distant Reading." *PMLA* 135 (1): 59–73. www.mlajournals.org/doi/abs/10.1632/pmla.2020.135.1.59?journalCode=pmla.
Terras, Melissa, and Julianne Nyhan. 2016. "Father Busa's Female Punch Card Operatives." In *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/debates/text/57.
Tilton, Lauren, Taylor Arnold, Thomas Smits, Mark Williams, Lorenzo Torresani, Maksim Bolonkin, John Bell, and Dimitrios Latsis. 2018. "Computer Vision in DH." In *Digital Humanities*. Mexico City. https://dh2018.adho.org/computer-vision-in-dh/.
Underwood, Ted. 2017. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11 (2). https://digitalhumanities.org:8081/dhq/vol/11/2/000317/000317.html.

Resources

- Introduction to Populating a Website with API Data (https://programminghistorian.org/en/lessons/introduction-to-populating-a-website-with-api-data)
- Creating Web APIs with Python and Flask (https://programminghistorian.org/en/lessons/creating-apis-with-python-and-flask)
- Computer Vision (Heidelberg University) (https://hci.iwr.uni-heidelberg.de/compvis/projects/digihum)
- Cultural Analytics (http://lab.culturalanalytics.info/p/projects.html)
- Emulation (https://libguides.bodleian.ox.ac.uk/digitalpreservation/emulation)
- Image-Net (http://www.image-net.org)
- Inscriptifact (http://www.inscriptifact.com/aboutus/index.shtml)
- Natural Language Processing (https://nlp.stanford.edu/software/)
- Python (an introduction) (https://wiki.python.org/moin/SimplePrograms)
- Quantitative history (http://historymatters.gmu.edu/mse/numbers/what.html)
- R (an introduction) (http://www.r-project.org/about.html)
- Voyant (https://voyant-tools.org/)
- Zotero (http://www.zotero.org/)