Confusing the Modern Breakthrough: Naïve Bayes Classification of Authors and Works

Peter M. Broadwell and Timothy R. Tangherlini, UCLA

The Modern Breakthrough marks an important turn towards realism in Scandinavian literature, and is broadly recognized as one of the most important periods in modern Nordic literary history. Georg Brandes's lectures on main currents in nineteenth century European literature at the University of Copenhagen and his later work Det moderne gennembruds mænd *(1883) provided the foundations for understanding this important movement. While his lectures grounded his appeals for naturalism in developments in European literature, his portraits of the male authors he considered to be at the core of the Modern Breakthrough offered a touchstone for a deeper understanding of this movement. One hundred years after the publication of Brandes's work, Pil Dahlerup published an important corrective to it, with her* Det moderne gennembruds kvinder*, a series of portraits and analyses of late nineteenth century female authors largely overlooked by the deeply biased literary establishment of the time.*

A great deal of scholarship on the Modern Breakthrough considers the rich network of literary cross-influence that characterized the period. Influence, however, is a complex phenomenon and one that is hard to formalize. In the following work, we propose to explore the related phenomenon of similarity, predicated on the notion that the most sincere form of flattery is imitation. To what extent do writers from this period share aspects of language? Can we capture this sharing in a useful manner computationally?

Broadwell, Peter M. & Tangherlini, Timothy R. "Confusing the Modern Breakthrough: Naïve Bayes Classification of Authors and Works." *HUMAN IT* 14.2 (2018): 19-42.
Published by University of Borås

Keywords: Classification, Modern Breakthrough, Periodicity, Naïve Bayes, Visualization

The Modern Breakthrough is widely considered to be one of the most important turning points in late nineteenth century Nordic literature, ushering in a period of literary experimentation predicated on a pivot toward naturalism. Georg Brandes's iconic work, *Det moderne gennembruds mænd* (1883), provides a literary historical framework for the consideration of the movement, outlining in broad strokes the contours of this shift in literature and, through the portraits of a series of featured male authors, presenting a touchstone for broader understanding of this movement. In 1983, Pil Dahlerup offered a corrective to Brandes's work with *Det moderne gennembruds kvinder*. Here, Dahlerup surfaced the numerous female authors who were writing groundbreaking work in the shadows of the male-dominated literary world. These women writers often considered many of the same themes as their male counterparts, albeit from markedly different perspectives. In her critique of Brandes, Dahlerup noted that he provided no justification of his numerous exclusions of other classes of authors active at the time: "...han forklarer ikke, hvorfor han ikke medtager en eneste kvindelig forfatter, en eneste bondeforfatter eller en eneste arbejderforfatter" [he does not explain why he does not include a single female author, a single peasant author or a single worker author] (Dahlerup, 1983, 62). Brandes's silence – and Dahlerup's strong rejoinder – affords an opportunity to explore the contours of this movement from a broader, computational perspective, and to explore the degree to which other authors of the period were inspired by each other.

A great deal of scholarship on the Modern Breakthrough considers the rich network of literary cross influence that characterized the period. Influence, however, is a complex phenomenon and one that is hard to formalize. In the following work, we propose to explore the related phenomenon of similarity, predicated on the notion that the most

sincere form of flattery is imitation. Here, similarity is based on aspects of language that we can measure, and thus models the extent to which writers from this period shared aspects of language. We had two main interrelated research questions: First, given a large corpus of Nordic literary works spanning several centuries, could we identify periodicity, and specifically the "Modern Breakthrough", as defined narrowly by Brandes and more broadly by Dahlerup? Second, given a series of authors identified as being part of the Modern Breakthrough, could we detect overlap in language usage that would help determine the boundaries of the movement and the authors who most likely influenced each other? A final goal of our work was to devise user interfaces that could present these experiments in a visually engaging and meaningful manner to support research into literary movements and the broader question of periodicity.

In earlier work, Tangherlini and Leonard (2013) showed how probabilistic topic modeling could be deployed to help discover similarities across the works of male and female authors of the period. Working at the level of the passage (Algee-Hewitt, Heuser & Moretti, 2015), they used a model of male modern breakthrough authors to identify passages that shared topic similarity drawn from a large, poorly labeled corpus, in their case all of the works in Google books written in Danish until 1923. By modeling passages from the works of male authors from the movement, they were able to identify, among other things, passages from contemporaneous female authors, thereby confirming Dahlerup's identification of numerous female "modern breakthrough" authors.

In this work, we focus on a smaller corpus of works curated by *Det Danske Sprog- og Litteraturselskab* and the Danish Royal Library under the rubric of the *Arkiv for Dansk litteratur* (ADl; Archive of Danish Literature). To model the corpus, we use straightforward computational methods that treat entire authorships or entire works each as a "whole", inspired by earlier work on folklore classification (Broadwell, Mimno &

Tangherlini, 2017). To make the system useful for literary scholars, we present the analysis macroscopically, deploying three different scales (Tangherlini, 2013). On the broadest macro scale, works in the entire corpus are compared, and their similarities are represented through a two-dimensional heat map. Regular square patterns in the map identify areas of significant similarity, highlighting authorships and potentially helping to identify periods. These patterns can be explored in greater detail through a zoom function. On the intermediate, meso scale, we apply two methods of exploration. First, we present a work similarity map as a two-dimensional cluster map, which provides a simple method for n-way comparisons between authorships. Second, we aggregate all of the works of a particular author into a single grouping and represent authorship similarities through a simplified confusion matrix, facilitating 2-way comparisons across authors. Areas of authorial overlap are easily determined by finding the intersection of the authors listed on the *x* and *y* axes. On the most focused micro scale, each work (*e.g.*, a novel) is considered individually. A similar confusion matrix to the one generated for authors is used to represent works that share similar features. On drill-down, the interface provides access to the underlying works and a visual representation of the linguistic features driving the similarity.

Given these three scales of representation, from corpus (macro) to authors (meso) to works (micro), users can explore the various overlaps between authors and works. This approach can be extended to complicate binary classes of authors, such as the male-female divide that Dahlerup considers, and can also allow for various other groupings of authors, including the rural, proletariat, school teacher, and bourgeois authors who were active during this same period. As a test case for a different "class" of author, beyond the male and female Modern Breakthrough authors, we include the late-Biedermeier author Sophus Bauditz, who otherwise would not be considered in the context of this movement, to see where his works and authorship would appear in these representations. Our motivation for considering Bauditz and other

classes of authors is the belief that the confusion of otherwise accepted categories can be a productive contribution to literary history, stimulating new ideas concerning periods and movements.

Resources

In this work, we augment the 498 volumes available through the *Arkiv for Dansk litteratur* (ADl), a collaborative project between the Danish Royal Library and *Det Danske Sprog- og Litteraturselskab*, with a small group of additional texts covering authors of interest not included in the ADl. We selected the ADl texts because they were all in excellent shape and free from the common OCR errors that are found in less curated collections. These texts also provided a substantial temporal spread, from Saxo Grammaticus (12th century) to Gustaf Munch-Petersen (1912-1938), and covered a broad range of literary genres. Despite these benefits, this reliance on a highly curated corpus greatly limits the number of works available for consideration, particularly of non-canonical works. Because of the limitations of the ADl list, particularly in regards to female authors or non-canonical authors, we augmented the corpus with a series of texts from the female authors Amalie Skram, Erna Juel-Hansen, and the pre-Breakthrough writer Mathilde Fibiger, as well as Sophus Bauditz, to create our experimental Modern Breakthrough-plus (MB+) corpus. With these additions, two authors identified by Dahlerup (1983) as modern breakthrough authors are included in this corpus, with Skram's works being selected from Danish translations of her oeuvre. As noted, we add Bauditz to the corpus as a test case for other contemporaneous authors who would not be considered a part of the target movement, and we added Fibiger to test how divergent her earlier writing was from the core Modern Breakthrough. Although the source texts that we used to augment the corpus are of fairly high quality, they are not as free from errors as those from ADl, as potential infelicities may have been introduced through OCR faults and inconsistent orthographic normalization. Since the aim of the project is experimental,

we believe this relatively constrained group of texts provides a reasonable corpus for testing our multiscale macroscopic approach.

Methodology and Interface

Calculating similarity is an ongoing challenge in the study of literature. There are numerous approaches to classification, each with their own advantages. One of our goals is to use the most straightforward approaches possible so as to reduce computational complexity and to make interpretation of our results more accessible to non-specialists. Extending earlier work by Broadwell, Mimno and Tangherlini (2017) on the classification of folk legends, we develop a "hold one out" Naïve Bayes (NB) classifier trained on the machine actionable works of the authors in our corpus, and also apply software modules to run standard text-similarity calculations including cosine similarity based on TF-IDF scores for unigrams, bigrams, and trigrams, and LDA topic inference (Blei, Ng & Jordan, 2003; Salton, 1991; Salton & Buckley, 1988). This work therefore occupies something of a methodological middle ground between comparisons that focus on stylometric features (Eder, *et al.*, 2016) and those that quantify content overlap based on "fuzzy" string matching (Smith, *et al.*, 2013).

As a precursor to our analysis, each machine-actionable work is chunked into 500-word passages to be fed to the classifier after applying basic orthographic normalization. We then run the groupings described above through the NB classifier and text similarity calculations. Instances of classification "confusion" – where the NB classifier "fails" in assigning all passages to their original grouping – suggest significant overlaps in style and content within or between authors' oeuvres. We compare these to the output from the text similarity computations. Such comparisons enact a fundamental principle of the "macroscope" as introduced by Börner (2011) and extended to the humanities by Tangherlini (2013), namely the greater degree of insight made available when one can switch rapidly between multiple analytical and perspectival scales on complex

cultural phenomena. In this literary history macroscope, we begin with a corpus-level view of Danish literature, represented as a heat map. Once we identify a period for closer evaluation – here the Modern Breakthrough – we move to visualizations on the meso scale for authors and the micro scale for individual works.

At the corpus-wide macro scale, the results of the text cosine and LDA topic similarity comparisons are visualized via two separate similarity matrices, analogous in format to a confusion matrix, with the degree of shading in each cell *x*,*y* indicating the similarity of the full texts associated with column *x* and row *y* (Figure 1a and Figure 1b).

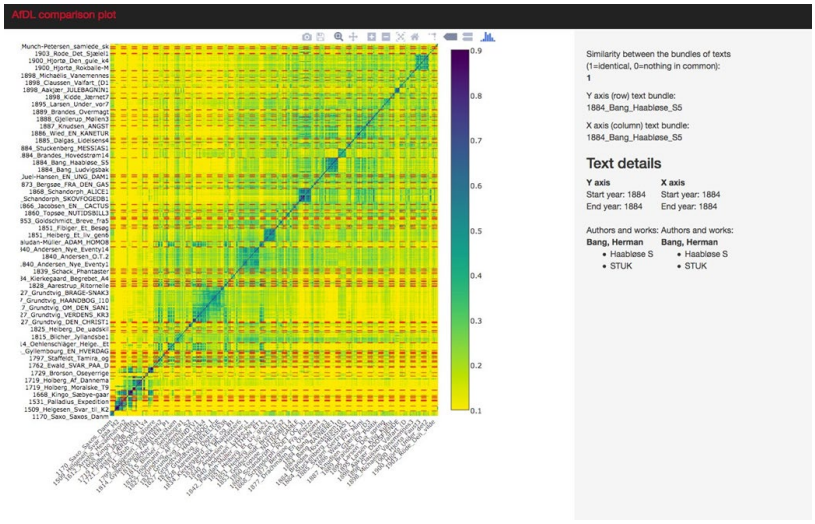


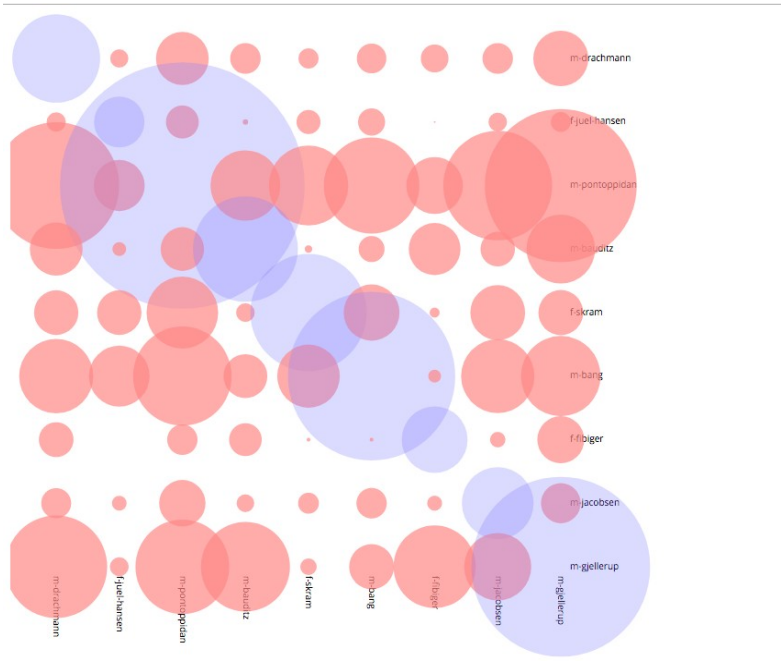
Figure 1a. A text similarity matrix of all the works in the corpus, based on the cosine similarity of the TF-IDF weights of the unigrams, bigrams, and trigrams of each work. Darker shading at the intersection of each pair of texts from the two axes indicates higher similarity. This interface can be accessed at http://babylon.library.ucla.edu/~broadwell/adl_sim/simmap.html

Such matrices can also be converted to distance plots wherein points representing texts are placed closer together when they are more similar. In our interface, on mouse over, the right-hand side of the user interface

displays the intersecting works, or bundles of works, a similarity measure for the two bundles, and a listing of the texts in those bundles. Clicking on an intersection brings up a list of the terms in the text bundles, ordered by frequency, with shared words highlighted.

Figure 1b. A text similarity matrix of all the works in the corpus, based on the cosine similarity of the topic weights for each work compared to every other work, as calculated via LDA topic modeling. This interface can be accessed at http://babylon.library.ucla.edu/~broadwell/adl_sim/ldamap.html

The ability to zoom into these matrices is a necessary feature that facilitates moving from macro to meso and micro perspectives. As a means for exploring the comparison in greater detail, the user can choose areas of interest, and zoom in on those either through interface controls or by drawing an arbitrarily sized bounding box on any part of the visualization. The resulting visualization on the left includes the closeup view of the heat map and allows for finer grained exploration of similarities.



Original label	Stories w/orig. label	Stories w/label from NB	Matches	Precision	Recall	F-score
m-drachmann	533	530	523	0.9868	0.9812	0.984
f-juel-hansen	182	171	170	0.9942	0.9341	0.9632
m-pontoppidan	4101	4135	4099	0.9913	0.9995	0.9954
m-bauditz	754	751	751	1	0.996	0.998
f-skram	935	925	923	0.9978	0.9872	0.9925
m-bang	1936	1905	1904	0.9995	0.9835	0.9914
f-fibiger	300	297	294	0.9899	0.98	0.9849
m-jacobsen	356	353	349	0.9887	0.9803	0.9845
m-gjellerup	2187	2217	2186	0.986	0.9995	0.9927

Figure 2a. A meso scale confusion matrix visualization of the Modern Breakthrough corpus. Here, the actual and computationally inferred authorships are compared to each other. The table beneath the visualization reveals the very high accuracy of the NB classifier for authorship. This interface can be accessed at http://etkspace.scandinavian.ucla.edu/~broadwell/mg_confusion/mg_authors.html

Figure 2b. "Drill-down" on the meso scale, where authorships are comprised of the aggregated works for each author.

On the intermediate, meso scale, we visualize an interactive confusion matrix for authorship (Figure 2a). A table beneath the visualization provides statistics related to the accuracy of the NB classifier underlying the matrix, including precision, recall, and F-score (the harmonic mean of the precision and recall). On this scale and on the micro (work-oriented) scale, the interface includes a drill-down interface that visualizes how the classifier has assigned a label to a work or authorship (Figure 2b). Clicking on a blue dot brings up a view of the words that are most highly predictive of the label, along with the individual passages and their potential additional labels. Clicking on a red dot, which indicates a disagreement between the original label and the NB classifier assigned label, presents a list of words printed along a color gradient from red (predictive of the original label) to blue (predictive of the label proposed by the NB classifier).

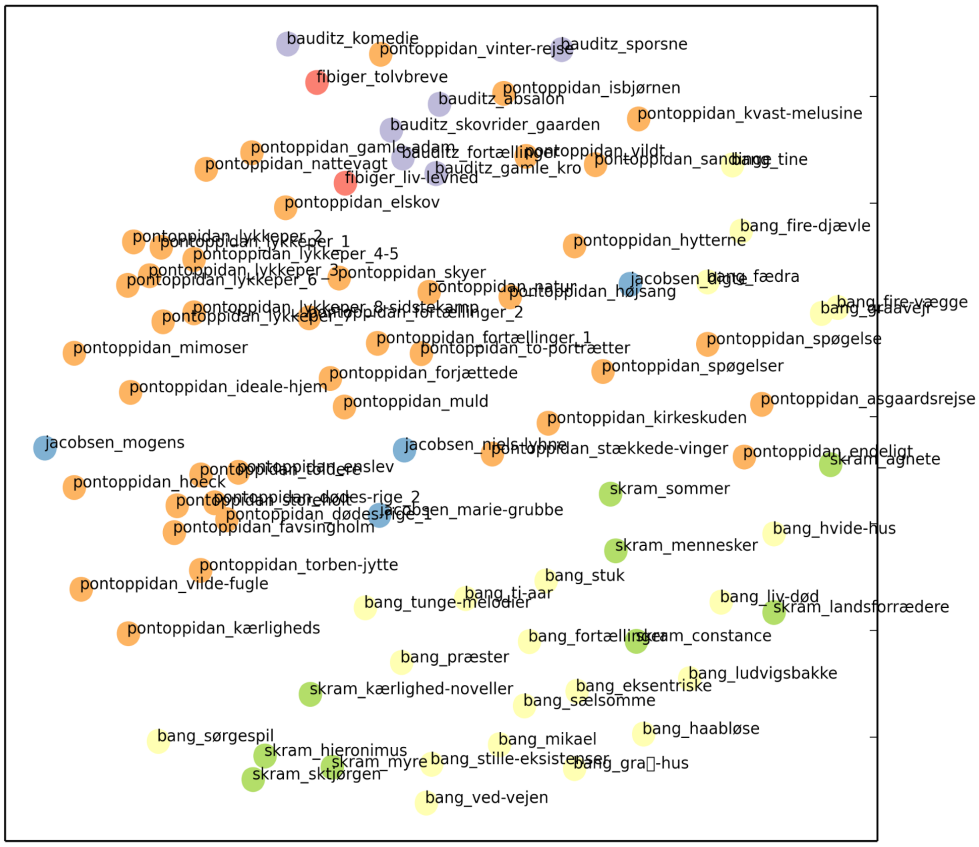


Figure 3. A micro scale text clustering plot of the works in the Modern Breakthrough test corpus. The distance between the points (works) is indicative of their textual similarity as calculated for the similarity matrix (See Figure 1a).

On the micro scale, we present two visualizations. A simple cluster point plot (Figure 3) shows the relative similarity of all the works in the Modern Breakthrough test corpus in a two-dimensional space. The distance between works is based on their cosine similarity measures. Also, the confusion matrix approach is repeated, with the level of comparison

now the work (as opposed to the authorship). Here, the sizes of the dots drawn on the cells of the matrix indicate the number of passages with the actual "label" (author+work) in the same row on the vertical axis that were assigned by the classifier to the proposed label at the corresponding column on the horizontal axis. For instance, a passage from Herman Bang's *Ved Vejen* may be properly classified by the NB classifier as a Bang passage, or it may be assigned to another author in the corpus. The strong diagonal of blue circles that emerges in these visualizations represents those passages that the NB classifier has placed into the expected category. The presence of red dots off the main diagonal indicates where passages have "confused" the classifier (Figure 4).

The drill-down interface operates in much the same way as the drill-down at the meso scale. Given the generally high accuracy of the NB classifier when predicting both author and work-level labels for substantial text excerpts, we add to each cell the lower ranked choices for the classified document, weighted in inverse proportion to their rank, thereby increasing the degree of confusion. Once again, the most predictive words from the work (as opposed to authorship as on the meso scale) are printed along a color gradient from those most highly associated with the original label (red) to those most highly related to the proposed label (blue). Below this list of words, one finds a ranked list of possible passage labels along with the negative log-likelihood of each label according to the NB classifier (Figure 5).

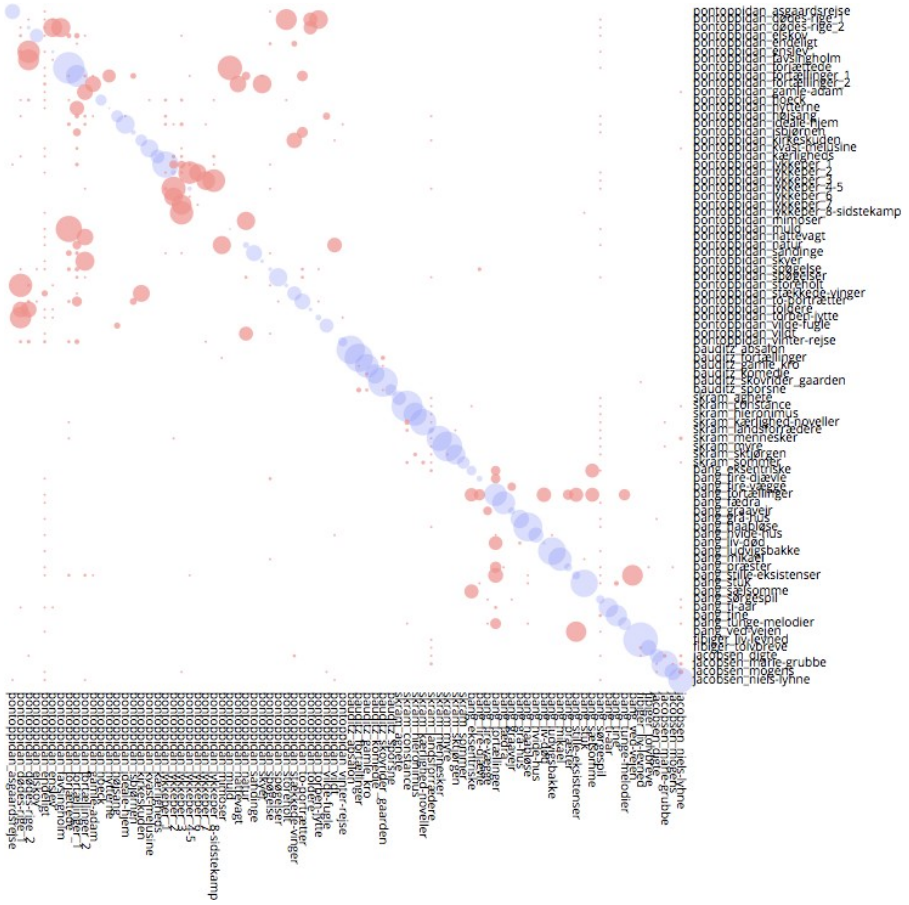


Figure 4. A micro scale confusion matrix of a subset of Danish language works, with the size of the dots indicating the number of passages from each author and work on the horizontal axis that were categorized by a Naïve Bayes classifier as belonging to the author and work on the corresponding row of the vertical axis. This interface can be accessed at http://etkspace.scandinavian.ucla.edu/~broadwell/mg_confusion/mg_books.html

Figure 5. "Drill-down" detailed view of the text passages from a single work (Bang's novella "Stille eksistenser" from 1886) identified by the Naïve Bayes classifier as most likely belonging to Bang's novel Ved vejen *(also written in 1886). The color-coding of the words indicates that the classifier considered reddish words to be more closely associated with "Stille Eksistenser," while the blue-tinted words are more closely related to* Ved Vejen.

Results and Discussion

In our consideration of the entirety of the ADl corpus, we find an interesting series of structures in the cosine similarity heat map that are strongly indicative of authorships, but do not clearly represent movements or periods. The strong self-similarity in authorships visually represented by the heat map is confirmed by the similarity metrics. It is worth noting that several of the authorships stand out in clear relief, notably those of the theologian N.F.S. Grundtvig, the mid-19th century

fairy tale write Hans Christian Andersen, and the novelist and poet Sophus Schandorf.

Shifting to the topic heat map, the concept of period becomes more pronounced, although authorships are still predominant. This difference between the maps makes sense, as the cosine similarity heatmap represents similarities in language use across n-grams while the topic models cut across texts. While we might expect a high degree of language use consistency within authorships, we might expect a high degree of topic consistency within periods or movements.

Of particular interest is the box-like pattern bounded by the works of J.P. Jacobsen (1847-1885) and Bang (1857-1912) (Figure 6), which corresponds with the Modern Breakthrough authors included in the ADl corpus and suggests that there may be topic similarity in this group of texts. Importantly, the graph identifies the works of Vilhelm Bergsøe (1835-1911) as an isolate within the Modern Breakthrough. Even though some of Bergsøe's works include realistic descriptions of social relationships, he is more closely aligned with Romanticism. Immediately above the Modern Breakthrough group, the graph identifies Brandes's lectures on the main currents in European literature, which are a touchstone in Scandinavian literary history. There is a striking dissimilarity in the graph between the upper right quadrant, corresponding to the Modern Breakthrough and subsequent literary movements, and the lower left quadrant, corresponding to pre-Modern Breakthrough literature, where individual self-similar authorships predominate.

The meso and micro scales of analysis rely on the "hold one out" NB classifier, which classifies authorships and texts with extremely high accuracy. We believe that this accuracy may be due in part to the relatively small number of labels. Training on the limited set of Modern Breakthrough authors and their works available in the ADl results in a classifier that is overly attuned to the various label classes. Consequently, to make the system more useful for research purposes, we include the

ranked list of labels that the classifier proposed for any given work, as opposed to the customary procedure of only taking the top ranked label.

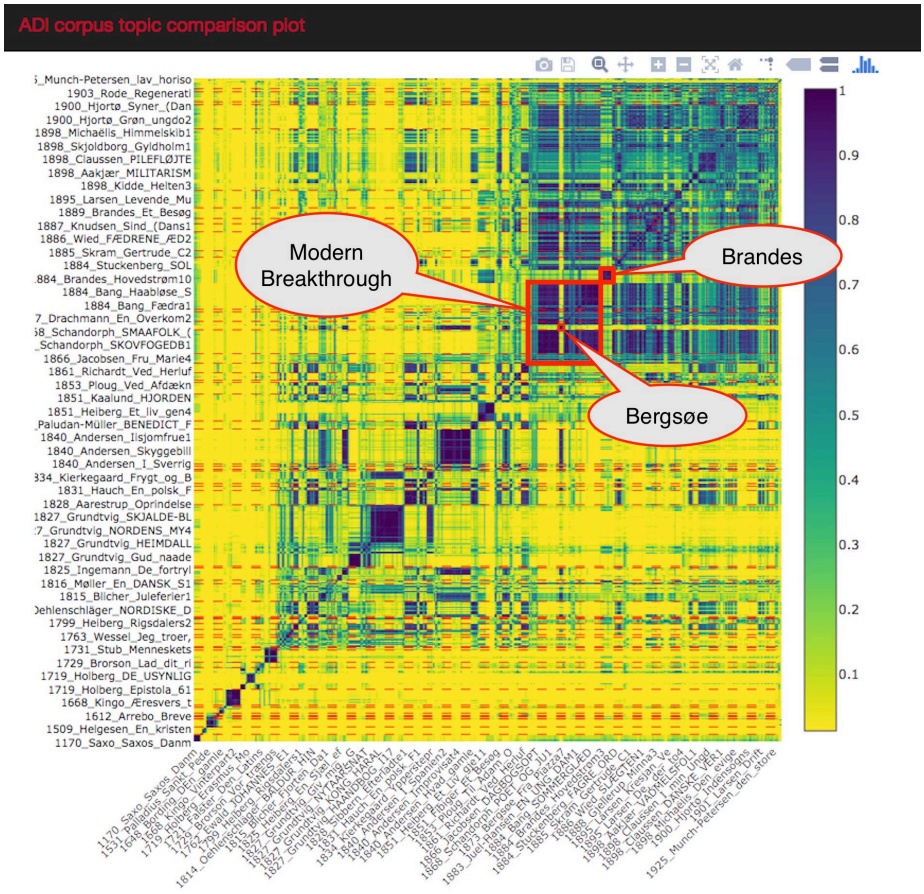


Figure 6. Possible identification of the Modern Breakthrough in a corpus-wide macro view on the LDA topic-based similarity heatmap.

Looking at the ranked list of labels – in effect "detuning" the classifier to the task for which it was designed – leads to a more nuanced view of similarity across works and authorships. This modification to how we

present results of the classification consequently opens up the corpus considerably, in effect offering the "near misses" of classification as possible areas of overlap and potentially fruitful areas for exploration. These "liminal" areas in the classification landscape have proven themselves to be productive in other Humanistic inquiries (Broadwell, Mimno, Tangherlini, 2017), and thus might be broadly applicable in other realms such as literary history.

On the meso scale, at which we aggregate authors' works, there is only a very small degree of classifier confusion. Our initial experiments with this classifier indicate a low degree of inter-author similarity and confusion, with the F-score of the NB classifier averaging over 95% for each author "label" when only the passages' authors are considered. By considering lower ranked labels, however, we are able to uncover areas of overlap between authorships. Given the limited size of the corpus, there are not many such overlaps, although we do find a considerable number of overlaps between Pontoppidan and Bang.

Results related to the second-ranked labels are worth considering. It is on this second order of classification that interesting aspects of influence are most likely to be found. Indeed, it would be surprising if authors were not most like themselves. It is more interesting to see whom they are also like once those top-level labels have been disregarded. For example, one discovers 61 Pontoppidan text passages that could, if the second-order label is used, be classified as texts from Bauditz; or 126 Skram excerpts that could have been classified as written by J.P. Jacobsen. We expect that, as more texts are added to the corpus, these second-order confusions will increase, and could serve as a fertile area for understanding overlap and influence among these authors.

On the micro scale, we identify similarities between works, with the simple similarity cluster graph highlighting some intriguing aspects of this corpus (Figure 3). The cluster graph immediately makes apparent the high representation of Pontoppidan in the corpus, while providing a clear indication of the linguistic separation of many of the authors.

Although the overall placement of works on the graph is arbitrary, similar works are placed closer to each other. On this graph, Bang and Skram appear to occupy one portion of the graph, while Bauditz, the one non-Modern Breakthrough author, and Fibiger, the pre-MB female author, occupy the opposite corner. Here, the comparison has identified two authors who are at best marginal to the movement, although for different reasons. Pontoppidan and Jacobsen intermingle through the middle of the graph. Jacobsen's spread across the entire horizontal axis seems to confirm his range as an author and language stylist.

Our second micro-scale representation, the confusion matrix interface, reveals a great deal of intra-author confusion, although surprisingly little cross-author confusion. Indeed, this result speaks to the strong consistency in style of individual authors, at least on the language features that we used for the classifier. One provocative overlap, however, is that between Jacobsen's "Mogens" and Pontoppidan's *Lykke-Per*, a curious juxtaposition not least because of the considerable difference in scale of the two works. Another interesting overlap is that between Skram's *Constance Ring*, and Juel-Hansen's *En ung dames historie*, since the works offer two distinct perspectives on women's lives.

On drill-down, the interface offers clues to how the works are similar, particularly in word use. Continuing with the *Constance Ring/En ung dames historie* comparison noted above, for example, we discover that Skram's work has a series of discriminatory words that include introspection, shouting and love, while Juel-Hansen's work includes words such as angst, blood, and youth (Figure 7a).

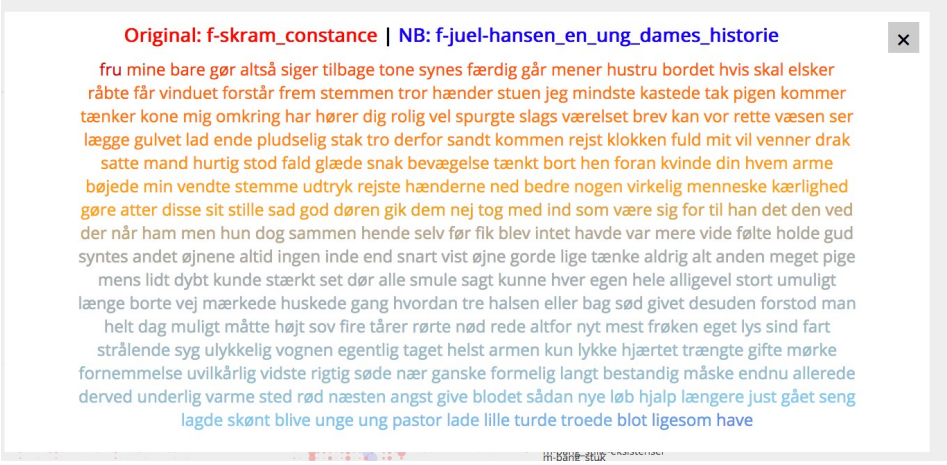


Figure 7a. Drill-down showing the ranked words that drive the similarity between Amalie Skram's Constance Ring *and Erna Juel-Hansen's* En ung dames historie.

Returning to the "Mogens"/*Lykke-Per* comparison, we discover a series of discriminatory words that deal with air, the heavens and fatigue for "Mogens" and relationships for *Lykke-Per* (Figure 7b). While neither of these word lists could in and of themselves form the basis for a discussion of literary influence between these pairs of authors, they do offer an opportunity to not only discover similarities across works by different authors, but also to reveal how word use influences similarity. The confusion matrix itself therefore does little to answer the question of how these works are similar – rather it proposes these similarities for consideration and helps to focus thought.

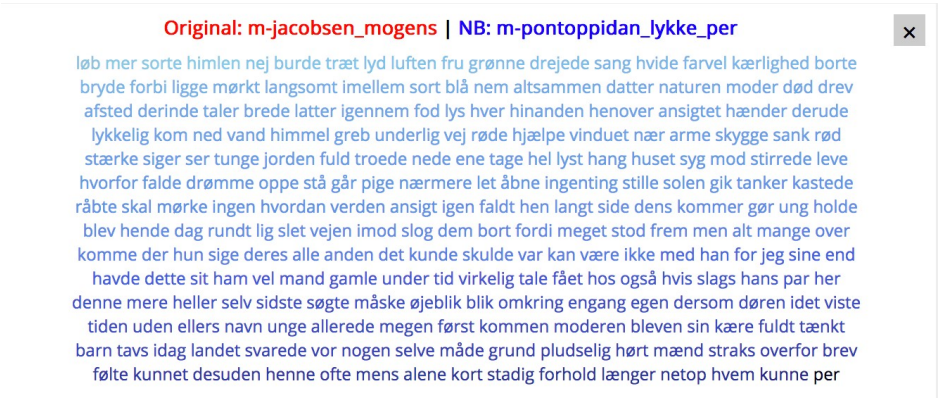


Figure 7b. Drill-down showing the ranked words that drive the similarity between Jacobsen's "Mogens" and Pontoppidan's Lykke-Per. *The lack of a large color gradient suggests a close alignment of word choice.*

Conclusion

Computational methods for discovering the boundaries of movements, and the interdependence of authors within a movement, show great promise for supporting literary historical scholarship. In these very limited experiments, we showed how several deliberately simple approaches to authorships and literary works can help identify literary periods and potentially challenge the pre-existing boundaries of those periods. Methods such as confusion matrix visualizations of classifications based on simple features can help identify passages or works that might otherwise be ignored. Our inclusion of lower-ranked prediction labels that would, in more traditional classification work, be discarded, allows for a gradual increase in recall over otherwise canonical (high precision) groupings of authors and works.

We recognize the need for caution in drawing conclusions from these experiments. The considerable constraints on the corpus size represent a

significant source of bias, since it is almost certain that we are modeling a fairly conservative view of the Danish literary canon. That said, there is little doubt that these works are of considerable importance in Nordic literary history. Indeed, we find these experiments to be encouraging. Adding texts and authorships to the corpus will lead to an increasing understanding of the intersections in language use across periods and across authorships. The strong performance of the classifiers lends support to established scholarship, while the moments of misclassification – particularly those based on the lower-ranked labels – offer an opportunity to understand the fluidity of periods, movements, and genre.

In future work, we plan to use these moments of "misclassification" and overlap between authors and within the works of a single author to develop a further understanding of stylistic and topical similarity and possible influence among authors. In particular, incorporating a temporal dimension into these analyses may help to estimate authorial influence by determining whether the classificatory "confusion" of a given text favors the authors that are considered to have influenced it. Alternately, such an analysis can suggest instances of text similarity and potential influence that extend or even contradict accepted narratives of Nordic literary history.

Peter M. Broadwell is Academic Projects Developer, Digital Library Program, UCLA. He works with faculty and library specialists to use emerging technologies in computational analysis, digital archiving, and multimedia presentation to support collaborative scholarship in fields from the humanities to the sciences. Recent projects in which he has participated include work in literary history, computational folkloristics, linked open data, online news archives, and audiovisual cultural analytics.

Contact: broadwell@library.ucla.edu

Timothy R. Tangherlini is Professor, Scandinavian Section, UCLA. His work focuses on folklore, Nordic literature and literary history, computational approaches to problems in the Humanities, and the study of culture. He has received support from the National Endowment for the Humanities, the National Science Foundation, the National Institutes of Health, the Fulbright Foundation, the Nordic Council of Ministers, the Mellon Foundation, the John Simon Guggenheim Memorial Foundation, the American Council of Learned Societies, Apple, and Google.

Contact: tango@humnet.ucla.edu

References

ALGEE-HEWITT, MARK, HEUSER, RYAN & MORETTI, FRANCO (2015). "On Paragraphs: Scale, Themes and Narrative Form." *Literary Lab*, 2015.

BLEI, DAVID, M., NG, ANDREW Y. & JORDAN, MICHAEL I. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022.

BRANDES, GEORG (1883). *Det moderne gennembruds mænd*. København: Gyldendal.

BROADWELL, PETER, MIMNO, DAVID & TANGHERLINI, TIMOTHY R. (2017). "The Telltale Hat: Surfacing the Uncertainty in Folklore Classification." *Journal of Cultural Analytics* 1. 2.

BÖRNER, KATY (2011). "Plug and Play Macroscopes." *Communications of the ACM* 54. 3: 60-69.

DAHLERUP, PIL (1983). *Det moderne gennembruds kvinder*. København: Gyldendal.

EDER, MACIEJ, RYBICKI, JAN & KESTEMONT, MIKE (2016). "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 8. 1: 107-121.

SALTON, GERARD (1991). "Developments in automatic text retrieval." *Science* 1991: 974-980.

SALTON, GERARD & BUCKLEY, CHRISTOPHER (1988). "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24. 5: 513-523.

SMITH, DAVID A., CORDELL, RYAN & MADDOCK DILLON, ELIZABETH (2013). "Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers."

Proceedings of the Workshop on Big Humanities Data (IEEE Computer Society Press, 2013).

TANGHERLINI, TIMOTHY R. (2013). "The Folklore Macroscope: Challenges for a Computational Folkloristics." The 34th Archer Taylor Memorial Lecture. *Western Folklore* 72. 1: 7-27.

TANGHERLINI, TIMOTHY R. & LEONARD, PETER (2013). "Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research." *Poetics* 41. 6: 725-749.