# ViziTex: Interactive Visual Sense-Making of Text Corpora

**Natraj Raman**[1], **Sameena Shah**[2], **Tucker Balch**[2], **Manuela Veloso**[2]
J.P.Morgan AI Research
[1]London, UK.
[2]New York, USA.
first.last@jpmorgan.com

## Abstract

Information visualization is critical to analytical reasoning and knowledge discovery. We present an interactive studio that integrates perceptive visualization techniques with powerful text analytics algorithms to assist humans in sense-making of large complex text corpora. The novel visual representations introduced here encode the features delivered by modern text mining models using advanced metaphors such as hypergraphs, nested topologies and tessellated planes. They enhance human-computer interaction experience for various tasks such as summarization, exploration, organization and labeling of documents. We demonstrate the ability of the visuals to surface the structure, relations and concepts from documents across different domains.

## 1 Introduction

Despite admirable progress in machine learning, human participation in data analysis and decision making is a reality. Human efforts are often required for bootstrapping labels, interpreting decisions and verifying outcomes. It is important to design intuitive visualizations that can exploit the pattern recognition and spatial reasoning capabilities of humans in order to transform the human-computer interaction experience. While traditional bar charts and heat map displays hold value, complex interactive graphical representations (Yuan et al., 2020) are often required to effectively slice and dice high-dimensional data. Furthermore, it is essential for these visuals to encode all the features delivered by machine learning models.

Interactive information processing in large complex text corpora pose a significant challenge due to the sheer volume, lack of structure and multi-faceted nature of text material. Existing efforts around visual interfaces for sense-making of text documents do not characterize the true potential of the text analytics algorithms (Liu et al., 2012), are often tied to a particular model (Vig, 2019) or remain fragmented with task specific solutions (Wang et al., 2016).

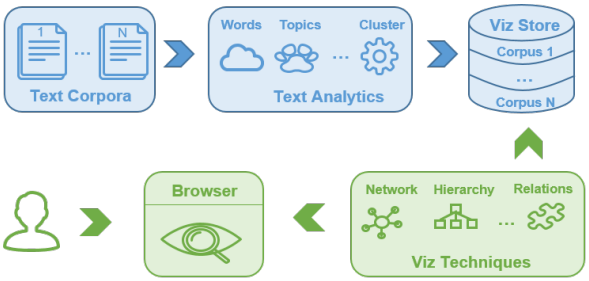


Figure 1: Architecture Overview.

There is a compelling opportunity for perceptive visualization techniques that fully leverage the capabilities of text mining models and cater to analysis at various levels of task granularity and human expertise. Towards this effort, we propose an interactive studio that delivers novel visual representations for common text oriented tasks such as theme discovery, document organization and label exploration. Visualizations presented here include a hypergraph that encodes distributional similarity between words, a multi-level radial layout to capture distinguishing terms, a clutter-free parallel coordinate plot of topic relations, a nested topology for document hierarchies and a tessellated plane to capture boundary points. These visualizations highlight interesting linguistic patterns in the corpus, surface complex relations between documents and reduce the burden of annotations for labeling exercises.

The structure of the framework, which follows a loosely coupled architecture pattern, is outlined in Fig. 1. There are three main components that drive the system: a text corpora, a suite of text analytics algorithms and a set of visualization techniques. We particularly focus on metadata rich corpus with multiple facets or data dimensions along which a corpus can be subdivided. The visualizations are



*Proceedings of the 2nd Workshop on Data Science with Human in the Loop: Language Advances*, pages 16–23

independent of the analytical models and newer algorithms can be flexibly plugged-in. All the generated graphical elements are interactive, with the end user being able to zoom, pan, hover and click for receiving contextual information. The users merely require a web-browser to access the visuals.

We demonstrate the domain agnostic nature of the visuals by providing illustrations from publicly available datasets that span across informal, legal and scientific language formats. In the following sections, we review related efforts and present eight different visualizations.

## 2 Related Work

Research in visual text analytics has gained prominence and surveys such as (Liu et al., 2018) provide an overview of recent progress. Differently, Kucher and Kerren (2015) present a visual survey by collating the images generated by the various visualization methods and offer an interactive filter for exploration.

There has been several efforts towards the development of software tools for analyzing text data. For example, the Leximancer (Angus et al., 2013) application plots word frequency statistics to help an analyst examine concepts in text. Tiara (Liu et al., 2012) is a visual text analysis tool that uses topic models to summarize documents. The popular pyLDAViz package (Sievert and Shirley, 2014) offers interactive visualization for topic models. Our work differs from these by introducing new metaphors and integrating a variety of text mining tasks.

Designing interactive graphics for the creation of interesting visualization techniques is popular. StoryPrint (Watson et al., 2019) is a visualization method for script-based media that presents prominence and emotion of characters in a scene. The visual analytic system in Verifi (Karduni et al., 2019) enables investigation of misinformation on social media. Vig (2019) introduced a multi-scale visualization tool to illustrate the inner workings of attention patterns generated by neural Transformers. Unlike these application and model specific efforts, our work is intended to be agnostic both to the data domain and underlying algorithm.

## 3 Visualizations

We present several techniques for visually analyzing a corpus at word, topic and document levels below. Samples from three different datasets namely Amazon Reviews (McAuley and Leskovec, 2013), Arxiv Abstracts[1] and Code of Federal Regulations (CFR)[2] are used to illustrate the visualizations.

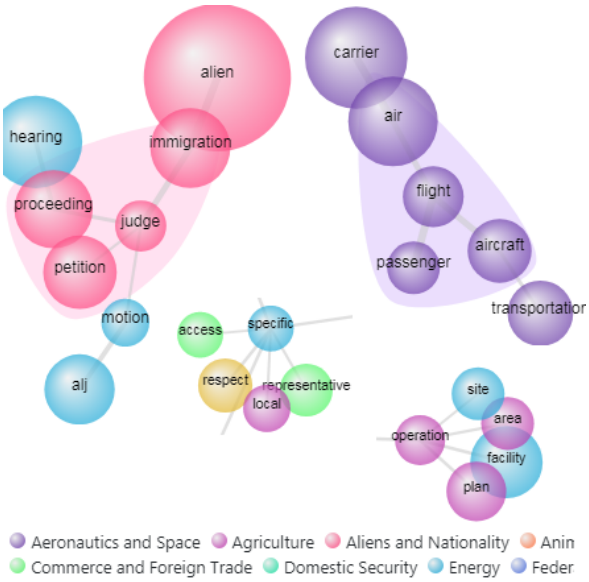


Figure 2: Hypergraph depicting word co-occurrences.

### 3.1 Word Hypergraph

A usual first step in text analytics is to plot the frequency of words in the corpus with a word cloud. However, the absence of context limits the ability of a word cloud visual to provide any insights beyond a basic overview. Following the principle of "characterizing a word by the company it keeps", we depict the co-occurrences of words (Weeds and Weir, 2005) to indicate semantic proximity. Rather than the structure-less cloud visual, a graph format with word nodes inter-connected by weighted edges is used. The words are scaled by a measure of how often they appear and colored by their dominant facet. The co-occurrence strength between words is encoded in the edge thickness. Furthermore, hyper-edges are used to connect the linked nodes that share similar attributes.

Formally, we are given a corpus with $D$ documents comprising of $N$ terms and a discrete attribute associated with each document. Let $G = (V, E, H)$ be a hypergraph with term nodes $\{v_n\}_{n=1}^{N}$, a set of edges $E \subset V \times V$ and hyper-edges $H \subset \mathcal{P}(V)$. We set the dyadic connections between terms $i$ and $j$ based on weighted mutual

[1] https://doi.org/10.6078/D1708G
[2] https://www.ecfr.gov/

information as

$$e_{ij} = \frac{\mid t_i \cap t_j \mid}{D} \log \frac{N^2 \mid t_i \cap t_j \mid}{D \mid t_i \mid\mid t_j \mid}, \quad (1)$$

where $|t|$ is an occurrence measure and $e_{ij} \in E$. Let $c_i \in \{C_1 ... C_P\}$ be the dominant attribute of term $i$. An hyper-edge $h \in H$ connecting potentially arbitrary number of nodes is defined as

$$h_i = \{v_i\} \cup \{v_k : c_i = c_k \wedge e_{ik} > \tau, \forall k \in V_{\setminus i}\} \quad (2)$$

where $\tau$ is a threshold to control visual clutter.

This hypergraph visualization allows the user to identify words that are central to characterizing a particular subset of the corpus. For example, by paying attention to the hyper-edges connecting nodes *air, flight, passenger* and *aircraft* in Fig. 2, the user can conclude that *flight* is a key-word in the *Aeronautics* subset of the CFR corpus while words such as *operation* or *access* is more ambiguous in describing the corpus.

### 3.2 Word Relations

Domain experts are often interested in understanding how subsets of a text collection differ. The identification of terms that are distinct to particular subsets will aid in this effort. To achieve this, we construct a radial layout of the top relevant terms that are shared across the various subsets. Each subset occupies a non-uniform slice based on its bandwidth in an inner concentric circle while its corresponding terms appear along the outer circle. The prominence of a term to a particular subset is reflected in its font-size. The relations between the terms are modeled as a B-Spline curve (Holten, 2006) in order to reduce visual clutter. The curve is drawn with a linear interpolation of the term colors and its width depends on the relationship strength.

In detail, let $\eta_p = \{w_i\}_{i=1}^{N_p}, \exists p^{'} : w_i \in \eta_{p^{'}}$ be the set of relevant terms in subset $p$ of the corpus. The arc length for $p$ is set to $N_p / \sum_{p^{'}} N_{p^{'}}$ and the curve width between $p$ and $p^{'}$ for term $i$ is computed as

$$\gamma^i_{pp^{'}} = \big[1/Z\big] f(w^p_i) + f(w^{p^{'}}_i), \quad (3)$$

where $f$ is a measure of term occurrence and $Z$ is a normalization constant. Fig. 3 shows the relation between words across different facets of the Amazon corpus. When inspecting the word *music*, the user can visually infer that this word is common to *CDs, Android Apps, Movies/TV* and *Video Games* subsets unlike a word such as *great* that is prevalent across all subsets, thereby unearthing distinguishing terms.

### 3.3 Topic Graph

Summarizing a corpus using a small set of underlying topics is a popular text mining technique to discover semantic structures. We improve over existing topic model visualizations by introducing three new features: the ability to capture correlations between topics, rank a topic by the significance of its semantic content, and associate meaningful labels with a topic. Consequently, the topics are now represented as a graph with the links between topic nodes denoting the extent of their correlation (Blei and Lafferty, 2006). While a node is colored by the dominant facet of its topic, its opacity is controlled by the topic's significance (Röder et al., 2015). Thus topics that are less coherent are de-emphasized, blending into the background. Both the automatically extracted topic label (Mei et al., 2007) and the top ranked terms of a topic are displayed, with the latter decorated in an elliptical arc around a node.

In order to extract the topic label, we first construct a set of candidate phrases $1...L$ and score the semantic relevance of a phrase $l$ to topic $k$ as

$$score(l, k) = \sum_m \log \frac{p(w_{mk})}{p(w_m)}, \quad (4)$$

where $p(w_{mk})$ denotes the probability of the $m^{th}$ term in the phrase for topic $k$ and $p(w_m)$ is the probability of the term across all topics. The topic label is then selected from the top ranked scores.

The utility of this visualization is evident in Fig. 4. The presence of labels such as *convex optimization problem* and *multivariate asset return* makes the topic theme of Arxiv corpus more interpretable than merely viewing a generic term such as *problem* or *model*.

### 3.4 Topic Relations

It is useful to discover differentiating topics across corpus subsets, similar to the identification of distinguishing words. However, it is difficult to accommodate the additional topic level grouping in the radial layout discussed above without disrupting legibility. Hence we employ a parallel co-ordinate (Siirtola et al., 2009; Collins et al., 2009) representation to portray this high-dimensional data. Specifically, the subsets are visualized along parallel columns and the top ranked topics for a subset


Figure 3: Relations between words across different subsets are displayed in a radial layout.

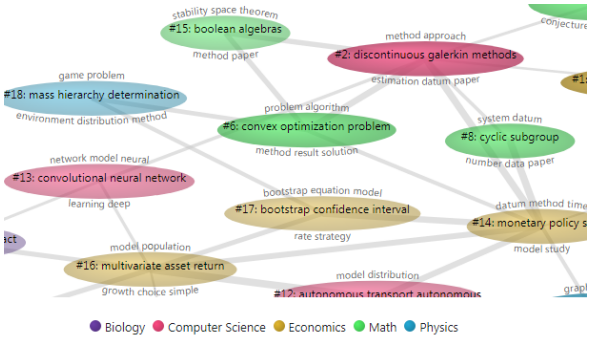


Figure 4: Topic Graph encoded with correlations, coherence and labels.

are scaled by their corresponding topic distribution. The topic terms themselves are relatively sized.

Naively showing all the links between related topics will clutter the visual. Hence the edges appear clipped by default, and are expanded only when the user hovers over a topic of interest. Fig. 5 demonstrates this concept, with the full-links shown only for *Topic 3* and rest of the elements are de-emphasized. The user can judge whether a topic is distinctive or not from the presence or absence of the clipped edges. For example, unlike *Topic 1*, *Topic 2* does not contain any edge implying that it captures *Aliens* subset specific terms.

### 3.5 Document Clusters

Visualizing the documents in the corpus in a manner that reflects the similarity and differences between them is essential for efficient organization and navigation. The spatial relations between the documents can be determined by comparing their embedding representations, which may range from a simple bag-of-words model to a modern pre-trained contextual text encoder (Devlin et al., 2018). Instead of simply plotting a 2D projection of these document embeddings, we cluster the documents using their original high-dimensional representation and visualize their relative positions in the clustered space.

Formally, we convert a document $d$ to a fixed length continuous vector of size $m$ through a function $\phi : d \rightarrow \mathbb{R}^m$. The pdf of the document is modeled based on this vector as a mixture of $K$ multi-variate Gaussian densities as follows:

$$p(d) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k). \quad (5)$$

Here $\mu_k \in \mathbb{R}^m$, $\Sigma_k \in \mathbb{R}^{m \times m}$ and $\pi_k \in \mathbb{R}$ denotes a mixture proportion. The above model partitions the corpus into $K$ different clusters and the cluster index of a document sampled from this density function is used to determine its position in the cluster network.

Fig. 6 illustrates such a cluster network of documents for the CFR corpus. The documents are centered around their corresponding cluster and colored by their facet. Nearby clusters are linked together denoting their similarity and a cluster can be collapsed interactively to simplify the view. The visual enables the user to reason say "Why is an *Aeuronatics* document grouped in a cluster with predominantly *Federal Elections* documents?" and provide feedback, thereby improving the tagging process in an active learning setting.

### 3.6 Document Hierarchy

Examining the relationship between documents within a same cluster is critical to gaining granular insights about the corpus structure. Instead

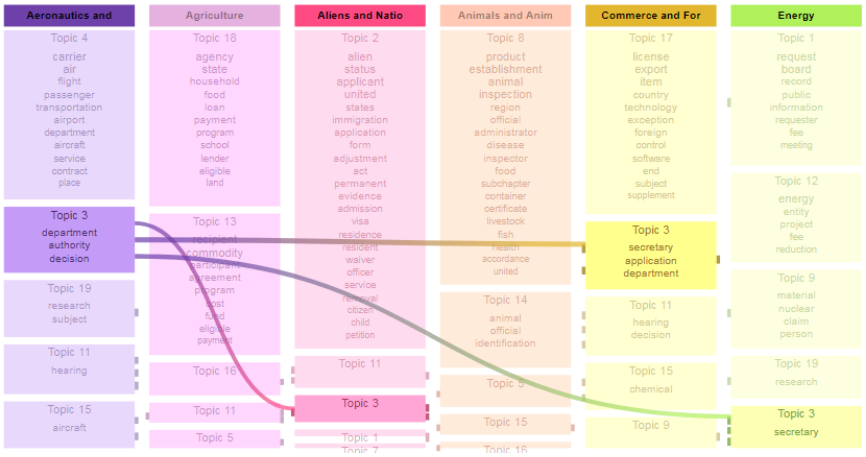


Figure 5: Topic Relations rendered using a parallel coordinate plot and clipped edges. Topic boxes and links are highlighted on hover to reduce clutter.

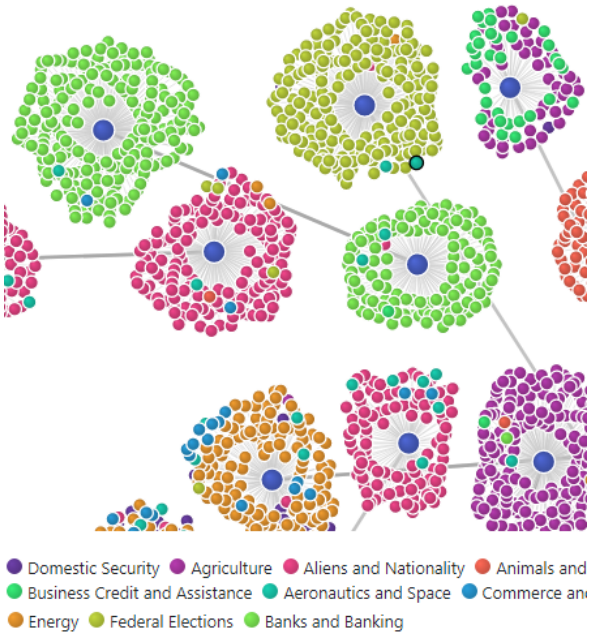


Figure 6: Network representation of document clusters.

of partitioning the documents exclusively, a better alternative is to organize them in a hierarchical fashion, from generic to specific (Ibrahim et al., 2019). This would empower the users to decide the level of detail, as dictated by their target task. We recursively partition the corpus to create hierarchical clusters and visualize them using a nested structure enclosure diagram.

Each circle in Fig.7 denotes an hierarchical level, with the circles contained inside the same parent being more similar. Leaf level circles denote the documents and are colored by their facet. The user can zoom-in to each circle and access the document content to explore anomalous patterns. For example, we investigated the reasons for the placement of an orange point (*Clothing/Shoes*) in the midst of violet points (*Android Apps*) and observed that it was a data quality issue.

### 3.7 Document Boundaries

Selecting the right data to label is important for annotation exercises and in active learning tasks. An effective strategy when sampling the data points is to identify points that are near decision boundaries (Monarch, 2021). The idea being that such uncertain points may have subjective interpretation and hence are worthy of human attention. We focus on presenting such boundary documents to the user in conjunction with documents that the machine is confident about.

In detail, the documents are first partitioned using a flat clustering algorithm based on their embedding representations using (5). Let $\mathcal{D}_k \subset \mathbb{R}^m$ denote the set of documents in cluster $k$. A convex hull encompassing the points in this cluster is defined from their convex combinations of $\mathcal{D}_k$ as

$$\Big\{ \sum_j \lambda_j \mathcal{D}_{kj} : \sum_j \lambda_j = 1 \wedge \lambda_j \geq 0 \wedge \mathcal{D}_{kj} \subset \mathcal{D}_k \Big\}. \tag{6}$$

The vertices of the hull are treated as the boundary points of a cluster. For visualization, a Voronoi diagram (Phillips, 2021) is constructed by using the cluster centroids as seed points. Thus each cluster is now visualized as a Voronoi cell bounded by a polygon with the polygon segments overlapping for nearby cells. The boundary points of a cluster are placed adjacent to the polygon sides while the interior points are arranged in a radial fashion at the center. Fig. 8 depicts this structure. The user

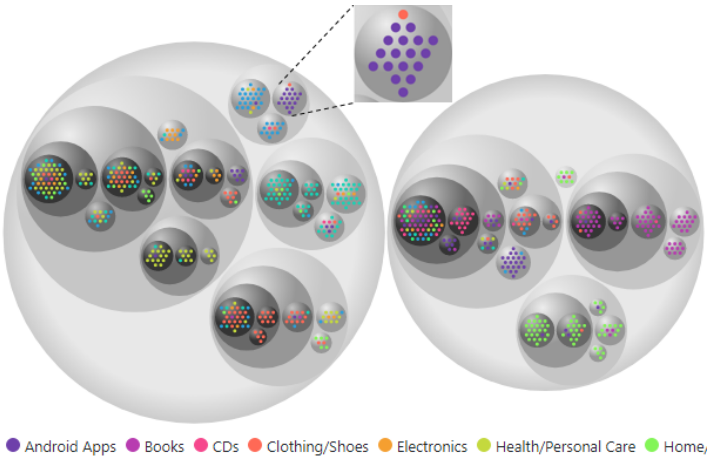


Figure 7: Hierarchical relations between the documents portrayed using a nested topology. Documents inside the same circle are more similar than the documents in sibling and parent circles.

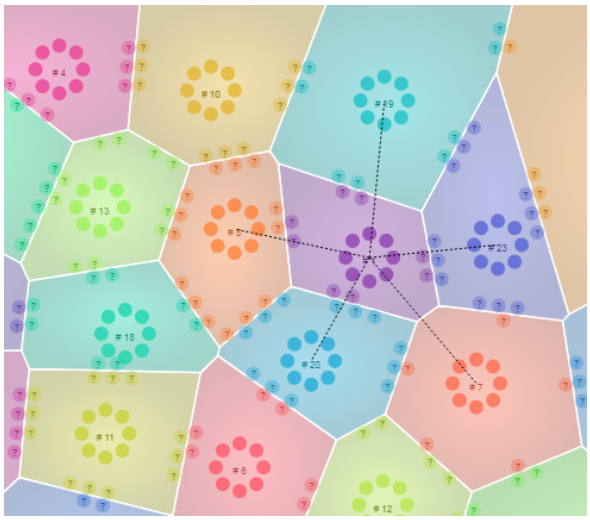

Figure 8: Voronoi tessellation of the cluster space showing both boundary and interior points.

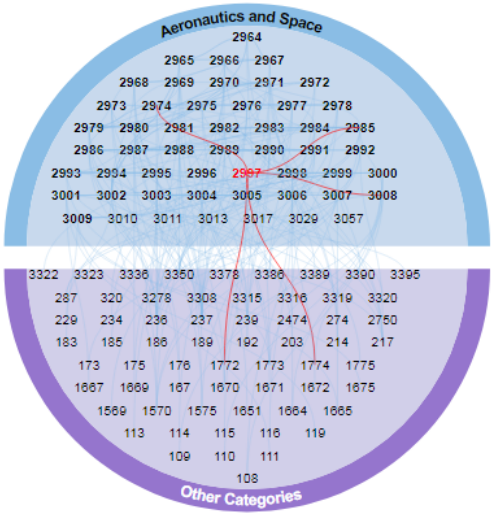


Figure 9: Relations (intra vs inter) between documents.

can drill down to see details about the boundary points and the sampled interior points. The Voronoi cells adjacent to *Cluster 1* is highlighted, signifying that the boundary points for this cluster may be assigned to its neighbors such as *Cluster 20* or *23*.

### 3.8 Document Relations

All the document specific visualizations outlined above consider the corpus holistically. Sometimes it is required to anchor the analysis to a particular subset of the corpus and compare with the rest of the subsets in a one-vs-all setting. Such intra and inter subset relations is explored in Fig. 9. The top hemisphere contains the ids of documents only from *Aeuronatics and Space* subset of CFR corpus. The documents from other subsets that are close to these documents in the embedding space are listed in the bottom hemisphere, with related documents being linked. The documents with strong intra-segment links are highlighted in bold font. The user can analyze the similarity and differences between a select set of documents based on the link cues.

## 4 Conclusion

Large and complexly related text collections require perceptive information visualization techniques to assist human understanding and reasoning. The interactive visualizations proposed here facilitates discovering concepts, themes, clusters, outliers and structure in a corpus by integrating text analytics models with novel visual representations. In future, we wish to extend the suite of statistical models for selection and incorporate new visuals for temporal analysis that exploits animations.

## Acknowledgments

This paper was prepared for information purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. J.P. Morgan makes no representation and warranty whatsoever and disclaims all liability for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful. © 2021 JP Morgan Chase & Co. All rights reserved.

## References

Daniel Angus, Sean Rintel, and Janet Wiles. 2013. Making sense of big text: a visual-first approach for analysing text data using leximancer and discursis. *International Journal of Social Research Methodology*, 16(3):261–267.

David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.

Christopher Collins, Fernanda B Viegas, and Martin Wattenberg. 2009. Parallel tag clouds to explore and analyze faceted text corpora. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 91–98. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Danny Holten. 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics*, 12(5):741–748.

R Ibrahim, S Zeebaree, and K Jacksi. 2019. Survey on semantic similarity based on document clustering. *Adv. Sci. Technol. Eng. Syst. J*, 4(5):115–122.

Alireza Karduni, Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin L Arendt, Samira Shaikh, and Wenwen Dou. 2019. Vulnerable to misinformation? verifi! In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 312–323.

Kostiantyn Kucher and Andreas Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pages 117–121. IEEE.

Shixia Liu, Xiting Wang, Christopher Collins, Wenwen Dou, Fangxin Ouyang, Mennatallah El-Assady, Liu Jiang, and Daniel A Keim. 2018. Bridging text visualization and mining: A task-driven survey. *IEEE transactions on visualization and computer graphics*, 25(7):2482–2504.

Shixia Liu, Michelle X Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. 2012. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):1–28.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499.

Robert Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*, volume 1. Manning, Shelter Island, NY.

M Jeff Phillips. 2021. *Mathematical Foundations for Data Analysis*, volume 1. Springer.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Carson Sievert and Kenneth Shirley. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Harri Siirtola, Tuuli Laivo, Tomi Heimonen, and Kari-Jouko Räihä. 2009. Visual perception of parallel coordinate visualizations. In *2009 13th International Conference Information Visualisation*, pages 3–9. IEEE.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Xiting Wang, Shixia Liu, Junlin Liu, Jianfei Chen, Jun Zhu, and Baining Guo. 2016. Topicpanorama: A full picture of relevant topics. *IEEE transactions on visualization and computer graphics*, 22(12):2508–2521.

Katie Watson, Samuel S Sohn, Sasha Schriber, Markus Gross, Carlos Manuel Muniz, and Mubbasir Kapadia. 2019. Storyprint: an interactive visualization of stories. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 303–311.

Julie Weeds and David Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475.

Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2020. A survey of visual analytics techniques for machine learning. *Computational Visual Media*, pages 1–34.