Optimization of document vectorization for unsupervised clustering of literature on accessibility in the built environment

Christine Mendoza, Steven J. von Dohlen, Harrison Truscott, and Aurorah Arndt
Computer Science
University of North Carolina at Chapel Hill

May 2, 2024

Abstract

Wayfinding is a universal experience. Unfortunately for individuals with disabilities, navigating unfamiliar spaces can be particularly difficult and sometimes impossible. We performed a literature review on 86 papers discussing accessibility features in wayfinding to identify key features and manually created a bag of words for each article. However, this is an extremely inefficient process that doesn't scale well. While machine learning methods like clustering, topic modeling, and document embedding exist, they have not been widely applied to better understand how to make wayfinding more accessible.

1 Introduction

> The fact is, most of us will move in and out of disability in our lifetimes, whether we do so through illness, an injury or merely the process of aging. - Rosmarie Garland-Thomson [7]

This means that obtaining accessibility information may become important to each one of us for a period of time ranging from a few days to the rest of our lives. Yet while mainstream navigation apps include route outlines for the built environment, they often lack route details relevant to users of varying abilities. [11]

Text mining has long been used in multiple domains to extract key concepts and structured knowledge from the scientific literature. [5] However, limited information exists regarding applications of text mining to literature on the accessibility in the built environment, especially in relation to clustering documents. Understanding how best to cluster documents offers the following benefits: (i) a more organized view of the knowledge, and (ii) enhanced training methods for systems distributing review and analysis work, and (iii) opportunity for evaluation against external labels, such as disability category, to see how clusters within the knowledge compare to such labels. In our work, we evaluate different methods of document vectorization for their efficacy in forming well-defined clusters. Future work will compare clusters obtained via unsupervised learning to the external label of disability category (often used to organize accessibility work).

2 Related Work

Clustering of documents generally requires a representation of documents in numerical form. This is achieved via vectorization and word embedding.

2.1 Definitions

Topic Modeling: A type of unsupervised learning technique which analyzes the words of original texts to discover the themes that run through them, without needing any prior annotations or labeling. [4, 14]

Cluster Analysis: A statistical technique used to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. This is typically done through unsupervised learning techniques, like K-means clustering, where some multidimensional distance metric is used to quantify distance in some higher dimensional space. That distance is then used to define thresholds for classification. In topic modeling, this type of cluster analysis can be used to group/extract document topics, sentiment, and themes. [2, 10]

Document Embedding. Document embedding transforms textual information into numerical vectors in a high-dimensional space, where each document is represented by a single vector. This technique captures the semantic meaning of the document, enabling various machine learning and natural language processing tasks such as clustering, and classification. One popular document embedding approach is Doc2Vec, also known as Paragraph Vector, which provides a framework for learning fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. [3, 8]

TF-IDF. The term frequency–inverse document frequency (TF-IDF) measures the relative importance of a word to the document or corpus that it is in adjusted for how frequently the word appears. [12]

3 Methods

A case study was conducted on accessibility in the built environment.

3.1 Literature Review

Query A keyword/title/abstract search ((”mapping standards” OR ”pathfinding” OR ”navigation” OR ”built environment”) AND (”impairment” OR ”disability” OR ”handicap”)) was conducted on Scopus on June 20, 2023, with results restricted to articles from 2021 to 2022 in the subject areas of computer science, engineering, math, psychology, sociology, decision sciences, arts, material science, economics, and multidisciplinary fields. This returned an initial pool of 549 results.

Screening Papers were screened by title, abstract, and/or paper content to remove articles without built environment factors (N = 421) and articles for which the full text could not be reached through the authors' institutional libraries (N = 42). This narrowed the pool of eligible papers to 86 results.

Feature identification Eligible articles were analyzed manually for built environment terms and context on the study purpose, target population, actual sample population, and sample size. This information was then inputted into a graph database for analysis, with relationships between nodes determined manually based on the literature review.

3.2 Analysis

We used two clustering techniques, Latent Dirichlet Allocation (LDA) and K-means, to analyze our collected articles. LDA categorizes documents into topics based on their highest-probability topic, while K-means groups input vectors by minimizing the distance to each cluster's center.

LDA requires minimal preprocessing, accepting tokenized bag-of-words, tailored for natural language data. In contrast, K-means can take various vector inputs but lacks awareness of data semantics.

Our raw input consisted of tokenized, stemmed, and filtered text data, alongside a manually crafted bag-of-words representation by Christine Mendoza. We applied preprocessing steps leading to experimental pipelines for LDA (bag-of-words output) or K-means (vector output).

Bag-of-Words

- Manual Document BoW
- Stemmed Document BoW
- TFidf transformed Manual BoW

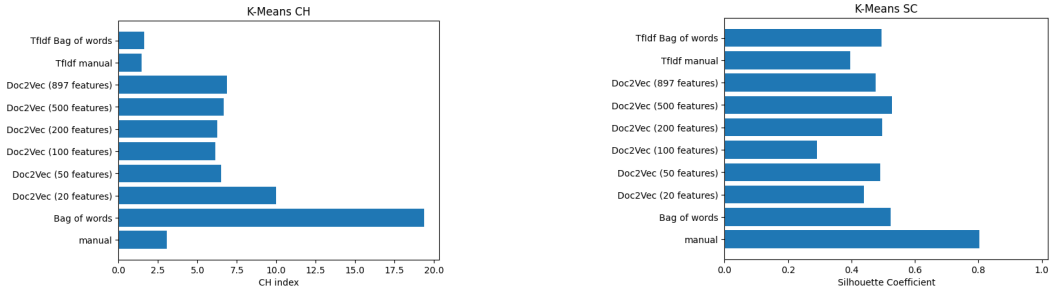
Vector

- All bag of words representations from above, encoded as count vectors
- Doc2Vec of stemmed documents. Feature counts [20,50,100,200,500,897] were chosen; the last was chosen because it matches the length of the manual bag-of-words representation. [1] [9]

LDA and KMeans models were then trained on the relevant vectors/bag-of-words. To evaluate the resulting clusters, two clustering metrics were used: the Calinski-Harabasz index (CH) [6] and the Silhouette Coefficient (SC) [13]. CH, which is sometimes also called "cohesion", is a measure of how related articles within each cluster are to each other; SC, which is calculated as a ratio of distances between clusters and distances within a cluster (and normalized to be between [-1 and 1], with 1 being maximal cluster separation), is a measure of how much farther apart clusters are than the clusters are spread out. We calculated these values for each of the clusters generated.

4 Results

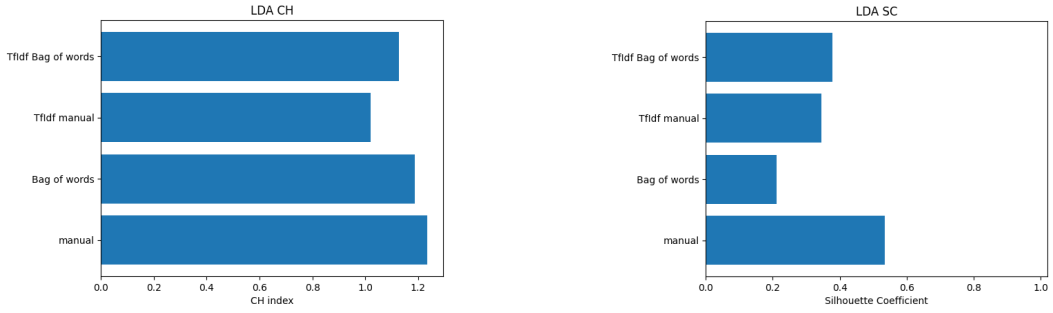
Using several preprocessers, clusterers, and parameters, we created several clusterings of our selected articles and compared them using two cluster metrics: CH ("cohesion") and SC (a measure of distance between clusters); See methods section. Figures 1 and 2 show SC and CH for K-means clustered data and LDA clustered data.



(a) K-Means CH

(b) K-Means SC

Figure 1: Cluster metrics for k-means methods



(a) LDA CH

(b) LDA SC

Figure 2: Cluster metrics for LDA methods

Since the best performing K-means algorithm in terms CH was above-and-beyond the others, we decided to graph the clusters (in the axes of the dataset's first two principal components). The resulting cluster plot can be seen below; triangles represent cluster means. As seen in Figure 3, the largest cluster has very tight groups; recall that this vectorization uses a bag-of-words generated from the entire article, implying that the majority of these articles all use very similar language. However, the notable exceptions are all exceptionally different, as can be seen by the large number of small and even single-member clusters farther from the primary clump.

5 Discussion

As seen in Figures 1 and 2, no one pipeline was the ultimate winner. However, surprisingly, K-means methods performed generally better across the board than LDA methods. The cohesions for LDA were all around 0.4-0.6; the highest cohesion for K-means was the raw bag-of-words with a value of 19. Clearly, K-means was producing much tighter clusters (in the vector space of the bag-of-words) than LDA. However, the same cannot be said for the relative spacing between clusters, measured by SC; fascinatingly, the manually constructed bag-of-words, which was one of the least cohesive clusterings by K-means, was by far the most separated, and the raw bag-of-words was around middle of the pack. The best K-means SC value (0.8) was still better than the best LDA SC value (0.55), but not by as much. Finally, of the bag-of-words representations used for LDA, the manual bagging was most effective for both SC and CH. This means that while the manual representation had an abysmal K-means CH score, it was the best pipeline for all three figures, having in both methods the best SC value and for both indices the best performance under LDA.

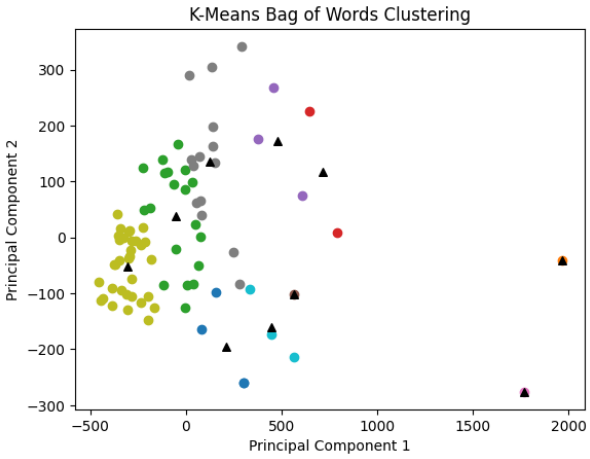


Figure 3: K-Means bag of words clustering

6 Conclusion

Our approaches had several tradeoffs. One big one is that we restricted document membership to a single cluster so that we could more easily compare results with clustering indices; however, the topic models used provide rich membership data per document, potentially allowing application of several labels per document. Additionally, the validity of applying vector-based cluster indices to the more categorical bag-of-words data is questionable; it is possible that the clusters that LDA produces are quite effective in ways that are hard to quantitatively measure. It is quite likely that the overall "improved performance" of K-means over LDA is due largely to the fact that K-means operates directly on the vectors analyzed by the cluster metrics, whereas LDA is not necessarily operating on or optimizing for the vector distances between clusters. Finally, the most direct applications of this work is to classifying documents based on cluster assignment; it could be interesting to do a more practical analysis of how well this works in situ for associating clusters with features of the data like disability type.

7 Acknowledgments

We gratefully acknowledge Anat Caspi of the Taskar Center for Accessible Technology for her mentorship of Christine Mendoza in creating the dataset and conducting the initial research.

This work is supported in part by the Distributed Research Experiences for Undergraduates (DREU) program, a joint project of the CRA Committee on the Status of Women in Computing Research (CRA-W) and the Coalition to Diversify Computing (CDC), which is funded in part by the NSF Broadening Participation in Computing program (NSF BPC-A #1246649).

We would also like to thank Hunter, a yellow Labrador that provided emotional support while we performed our work.

References

[1] [n. d.]. *Gensim Doc2vec paragraph embeddings*. https://radimrehurek.com/gensim/models/doc2vec.html Accessed: 2024-4-29.

[2] 1999. Data clustering: a review. , 264-323 pages. https://doi.org/10.1145/331499.331504

[3] 2014. Distributed Representations of Sentences and Documents. 4 (2014), 1188–1196.

[4] 2017. Probabilistic topic models. , 424-440 pages. Issue 7.

[5] Charu C. Aggarwal. [n. d.]. Mining Text Data. In *Data Mining: The Textbook*, Charu C. Aggarwal (Ed.). Springer International Publishing, 429–455. https://doi.org/10.1007/978-3-319-14142-8$_1$3

[6] T. Caliński and J Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3, 1 (1974), 1–27. https://doi.org/10.1080/03610927408827101 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101

[7] R. Garland-Thomson. 2016. Becoming Disabled. https://www.nytimes.com/2016/08/21/opinion/sunday/becoming-disabled.html. *The New York Times* (Aug. 2016). Accessed: 2024-4-29.

[8] Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. arXiv:1607.05368 [cs.CL]

[9] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. arXiv:1405.4053 [cs.CL]

[10] J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. https://api.semanticscholar.org/CorpusID:6278891

[11] Catia Prandi, Barbara Rita Barricelli, Silvia Mirri, and Daniela Fogli. [n. d.]. Accessible wayfinding and navigation: a systematic mapping study. 22, 1 ([n. d.]), 185–212. https://doi.org/10.1007/s10209-021-00843-x

[12] Juan Ramos. [n. d.]. Using TF-IDF to Determine Word Relevance in Document Queries. ([n. d.]).

[13] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[14] Pengtao Xie and Eric P. Xing. 2013. Integrating Document Clustering and Topic Modeling. arXiv:1309.6874 [cs.LG]