* Models: (evaluate the models)
  + LDA
  + k-means
  + H-LDA?
* Cluster metrics
  + -

Focused vs. general knowledge graphs

* Vectorizations
  + Christine’s vectorization
  + Christine’s vectorization + TFIDF transformation
  + \_\_
  + \_\_ + TFIDF transformation
* Run a clustering
* Evaluate metrics

Project final report

Your final project report can be at most four pages long (include all text, appendices, figures, and anything else), with one additional page that can contain nothing but references, and must be written in LATEX.

If you did this work in collaboration with someone else, or if someone else (such as another professor) had advised you on this work, your report must fully acknowledge their contributions. At a minimum, your final report must describe the **problem/application and motivation, survey related work, discuss your approach, and explain your results/conclusions/impact of your project**. It should include enough detail such that someone else can reproduce

your method and results. You are also required to provide a link to a GitHub repository where your code is stored.

You may look at previous projects or papers from the list in section 2 to get an idea of what should be included in your project report. You will likely end up with a better report if you start by writing a 6-7 page report and then edit it down to 4 pages of well-written and concise prose. Keep in mind if you have an exciting and novel idea for this project, you can extend your work and submit it to an appropriate machine learning conference.

**Introduction/Motivation**

* Accessible mapping is important, but features to map (and understanding of accessibility-first GIS in general?) are still unclear.
* Understanding how to cluster documents offers the following benefits:
  + More organized view of the knowledge (literature)
    - Can determine what categories exist
  + Can train systems for distributing work more efficiently (ex. by specialization)
  + Can evaluate against external labels to see how categories overlap (and how content aligns with or differs from labels)
* In our work, we evaluate different methods of vectorization for their efficacy in forming well-defined clusters. We compare clusters obtained via unsupervised learning to the external label of disability category (often used to organize accessibility work).

**Related work**

* Organizing knowledge
  + Word embedding techniques
  + Topic modeling
  + Cluster analysis
  + Types of documents used (see health tweet paper)
* Organizing knowledge in accessibility in GIS (the state of accessibility in GIS)
  + (lead into our specific case study)

Questions to answer:

1. Definitions: What is topic modeling? What is cluster analysis? How do topic modeling and cluster analysis differ?
   1. This is good for broad definitions on topic modeling: <https://arxiv.org/abs/1309.6874>

* Good place to start: [Mining Text Data | SpringerLink](https://link.springer.com/chapter/10.1007/978-3-319-14142-8_13) (cited in the health tweets paper) (login via UNC libraries) (what about post-2012?)
  + Section 13.4.2 might also describe use of LDA in refining clusters

1. More definitions: What is document embedding? How can word embedding be useful for generating vectors? What techniques are often used?
   1. <https://arxiv.org/abs/1507.07998> this one is really good and talks about LDA and document embedding and bag of words with vector embedding in it

Topic Modeling: A topic model is a generative model for documents; it specifies a simple probabilistic procedure by which documents can be generated.It is a type of unsupervised learning technique which analyzes the words of the original texts to discover the themes that run through them, without needing any prior annotations or labeling. [<https://arxiv.org/abs/1309.6874> ,<https://scholar.google.com/citations?view_op=view_citation&hl=en&user=szUb_isAAAAJ&citation_for_view=szUb_isAAAAJ:qjMakFHDy7sC> ]

Cluster Analysis: is a statistical technique used in data analysis and machine learning 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 in-group and out-group classification. In topic modeling, this type of cluster analysis can be used to group/extract document topics, sentiment, and themes. [<https://www.semanticscholar.org/paper/Some-methods-for-classification-and-analysis-of-MacQueen/ac8ab51a86f1a9ae74dd0e4576d1a019f5e654ed>,<https://typeset.io/papers/data-clustering-a-review-4c452f5cab>]

Document Embedding: creates embeddings that transform 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 similarity searches, 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. [<https://paperswithcode.com/paper/an-empirical-evaluation-of-doc2vec-with>, <https://typeset.io/papers/distributed-representations-of-sentences-and-documents-3g8596uykm>]

Document Embedding:

1. What is the current state of the research on the comparison of document embedding techniques for clustering documents? Are there gaps in the research? (Ex., the health tweet paper mentions that less research is done on short documents.) Is clustering dependent on article content?
2. Our specific case study: Has clustering been applied to papers on accessibility in GIS? What do people often use as labels for organizing/clustering such data?

* not accessibility, but health: [Identifying built environmental patterns using cluster analysis and GIS: Relationships with walking, cycling and body mass index in French adults](https://ijbnpa.biomedcentral.com/articles/10.1186/1479-5868-9-59) albeit this is geospatial clustering. We’re looking for clustering of literature documents: ex. SemNet is used for traversing the knowledge in literature
* KGs v FKGs
* entity recognition as a step in KG formation
* Word embedding techniques (need to find a research article) <https://towardsdatascience.com/word-embedding-techniques-word2vec-and-tf-idf-explained-c5d02e34d08> - used to generate vectors
  + ~~One-hot encoding~~ more like binary encoding of multiple features (what we currently have for the selected words (can also do for entire corpus))
    - Did we want to do one-hot encoding and TF-IDF from LDA/HLDA results?
  + TF-IDF (weights words based on frequency)
    - *should* we weight words based on frequency? what are the implications of doing that at a document level?
  + Doc2vec
    - [Gensim: What is difference between word2vec and doc2vec? - Stack Overflow](https://stackoverflow.com/questions/42827175/gensim-what-is-difference-between-word2vec-and-doc2vec)
    - Vectorizes documents
    - What are we using as tags?
  + (Possibilities)
    - [LDA2vec: Word Embeddings in Topic Models | by Lars Hulstaert | Towards Data Science](https://towardsdatascience.com/lda2vec-word-embeddings-in-topic-models-4ee3fc4b2843)
    - HLDA2vec?
* why vectorize? why compare the two approaches to tokenization? why cluster?
  + vectorization - a way of representing documents; needed for clustering algorithms
  + What are the two approaches of “tokenization”? (general v focused) - is this really tokenization or a pre-processing step before vectorization?
  + clustering - grouping documents
    - What do we gain from grouping documents? (with the different vectorization techniques?) predictive analysis?
    - Techniques we are using (maybe just use one or two very common techniques?)
      * K-means
    - Obtain metrics (what do they mean?)
      * see 3.3.2 and 3.3.3
      * Internal indices
      * External indices (uses labels)
        + “**Adjusted Rand Index**: the rand index (RI) computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. RI gives a value between 0 and 1, where 1 indicates that the data in clusterings are the same. This measure can be seen as the percentage of correct decisions made by the algorithm”
        + “**Homogeneity**: a cluster has perfect homogeneity if all members of that cluster have the same external label. That is, the class distribution within each cluster contains only one class or equivalently has zero entropy.”

We could use this or other external indices to compare how well things are actually subdivided via disability label versus other things. What label is most representative of the clusters?

* what is the gap in literature we are trying to address? the q we want to answer?
  + how to best create representations of the documents? (what about for specific purposes?)
  + which vectorizations are best for clustering documents?
    - What is a “good clustering?”
      * distinctness?
  + how to organize documents for attending to by people?
    - Scientific documents v. other documents?
    - Are there any assumptions on how it would be organized? Categories?
    - Organizing content? (Note: judging importance of features via frequency over the *entire corpora* is flawed due to bias in populations studied)
* Do you mind finding a paper (maybe the one sent) describing the difference between topic modeling and clustering and putting some notes here about that?
  + <https://www.sciencedirect.com/science/article/abs/pii/S0933365721000890>
  + This is what I caught you reading the other day? Yup!
  + “existing topic modeling and clustering approaches have suboptimal performance on these short texts. Moreover, research over health-related short texts using these methods has become difficult to reproduce and benchmark, partially due to the absence of a detailed comparison of state-of-the-art topic modeling and clustering methods on these short texts”
  + <https://link.springer.com/chapter/10.1007/978-3-319-51133-7_91>
  + This is about “ The proposed method adopts K-means and Latent Dirichlet Allocation (LDA) algorithms for document clustering and topic modeling, respectively.”
  + It uses text mining and all that, just for manufacturing purposes?
  + <https://ieeexplore.ieee.org/abstract/document/6964656>
  + This one is about TFIDF models, which is a keyword model you mentioned earlier but I can’t get access to it all

**Methods (goes after related work)**

* We did a literature review in summer 2023 to get 86 articles (screened for relevance, etc.)
* We reviewed the articles for stated and actual target population, setting, country of origin, N, and GIS features
* Dimension 1: focused v. general
  + GIS features were manually constructed into focused knowledge graphs
    - should we still introduce FKGs given that we’re not making general KGs?
  + Topic modeling was used for obtaining word embeddings from general content analysis
* Dimension 2: weights
  + multi-hot encoding (T/F) v. word frequency
  + We judged word frequency *within* articles to be representative of perceived importance. (Vs. the entire corpora, where importance is limited by study populations selected)
* Cluster creation (k-means?)
* Cluster analysis
  + List metrics used (what do we think makes a “good” cluster? Distinctness?)

**Results**

Much tables

**Discussion**

* Which methods were best?
* Challenges encountered, problems learned

**Conclusion**

* include future work

Outline:

Contributions (Christine)

Intro-

Methods (how we got the dataset) (Aurorah, but monitored by Christine)

Problem/Application (Aurorah, but monitored by Christine)

Motivation (Anyone)

Survey Related Work (Christine probably has documents)

Discuss your approach (Please not Aurorah)

Explain results/conclusion/impact of work (Please not Aurorah)

Abstract:

Abstract - This is very abstract! We are looking at ways to cluster similar documents and comparing their results, I guess. \*\*\* replace with the real information please!!!

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.

Topic Modeling: A topic model is a generative model for documents; it specifies a simple probabilistic procedure by which documents can be generated.It is a type of unsupervised learning technique which analyzes the words of the original texts to discover the themes that run through them, without needing any prior annotations or labeling. [<https://arxiv.org/abs/1309.6874> ,<https://scholar.google.com/citations?view_op=view_citation&hl=en&user=szUb_isAAAAJ&citation_for_view=szUb_isAAAAJ:qjMakFHDy7sC> ]

Cluster Analysis: is a statistical technique used in data analysis and machine learning 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 in-group and out-group classification. In topic modeling, this type of cluster analysis can be used to group/extract document topics, sentiment, and themes. [<https://www.semanticscholar.org/paper/Some-methods-for-classification-and-analysis-of-MacQueen/ac8ab51a86f1a9ae74dd0e4576d1a019f5e654ed>,<https://typeset.io/papers/data-clustering-a-review-4c452f5cab>]

Document Embedding: creates embeddings that transform 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 similarity searches, 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. [<https://paperswithcode.com/paper/an-empirical-evaluation-of-doc2vec-with>, <https://typeset.io/papers/distributed-representations-of-sentences-and-documents-3g8596uykm>]