



# Internet Archive PDF

Huilin Chang, Yihnew Eshetu, Celeste Lemrow

Faculty Advisor : Professor Alvarado
Client : Internet Archive
Sponsor : Bryan Newbold

# Presentation Outline

- Problem Statement
- Data Overview
- Data Pipeline
- Data Feature Engineering
- Exploratory Data Analysis
- Initial Models
- Conclusion

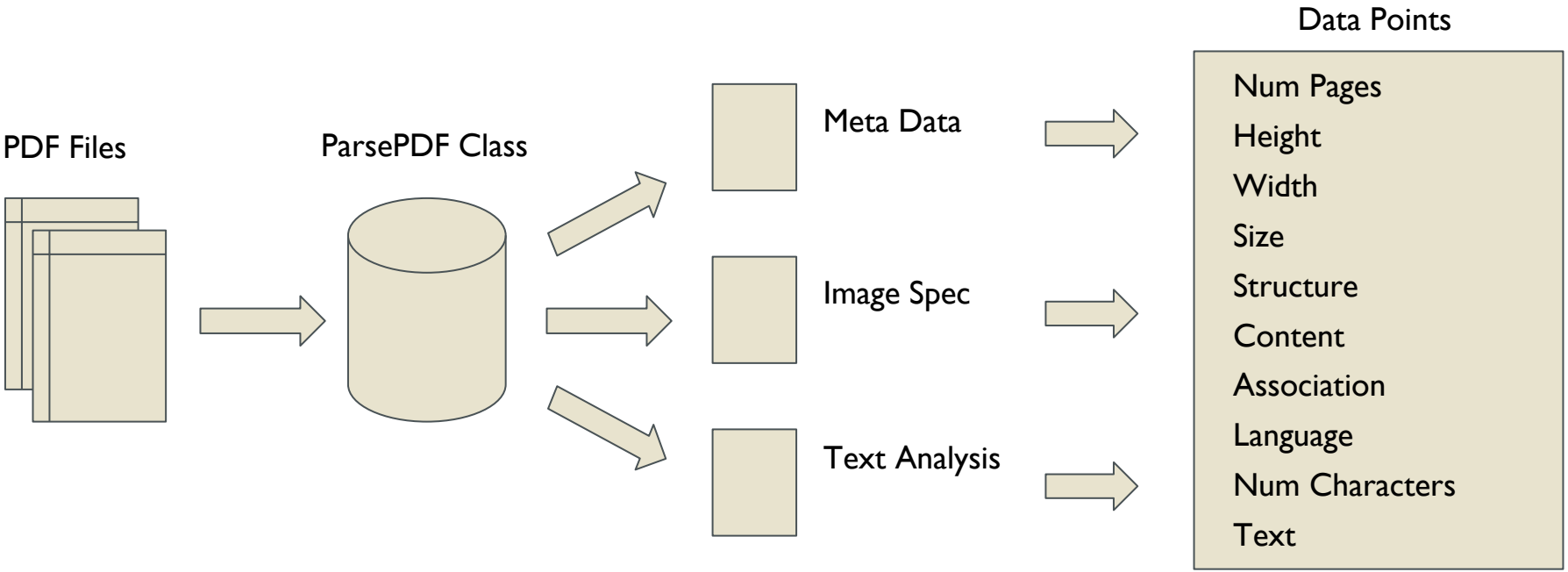
# Problem Statement

- One of the Internet Archive's mission areas is "Universal Access to All Knowledge", which includes collecting and providing access to the "scholarly web" -- research publications and datasets.
- Curation to accurately identify legitimate research publications is needed to help users find scholarly content
- An inclusive approach that accounts for diverse content, particularly from underrepresented geographic areas, groups, and content domains, is important to avoid excluding relevant content due to implicit bias and narrow criteria
- Our project aims to help this mission by implementing a fast PDF identification tool, which will score files on their likelihood of being a research publication

# Data Overview

- 4 IA Training Datasets, consisting of 40k PDFs
  - Global Wayback Random - Random sampled PDFs from the Wayback Machine
  - Fatcat - A set of PDFs from the existing ‘Fatcat’ catalog of research papers
  - Fatcat Longtail Language - Papers from less-represented languages
  - Longtail - A set of PDFs created using heuristics (GROBID)
- Minor issues with data
  - Password Protected
  - Corrupted/Unparsable
- Plan to branch out further into IA content archives as well as other known sources of PDF scholarly documents

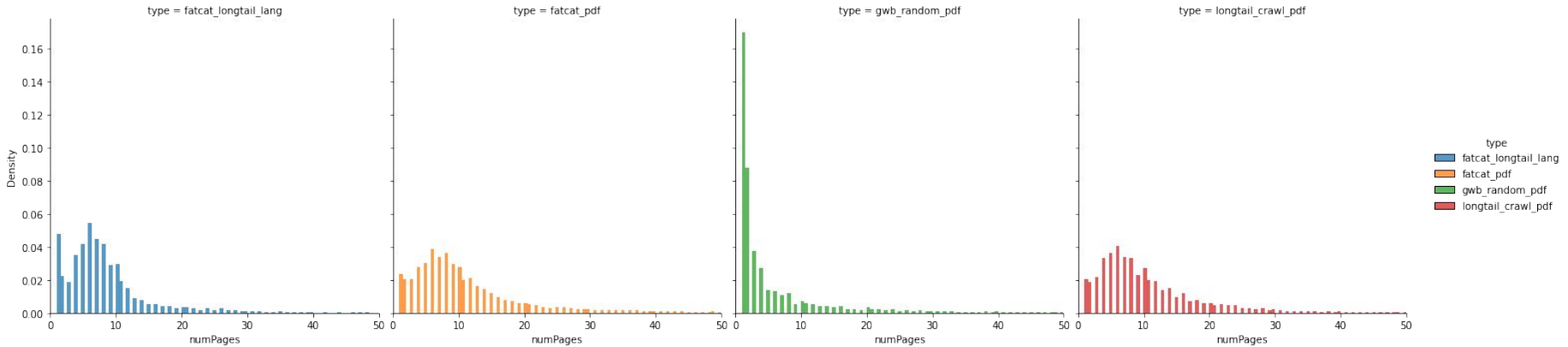
# Data Pipeline



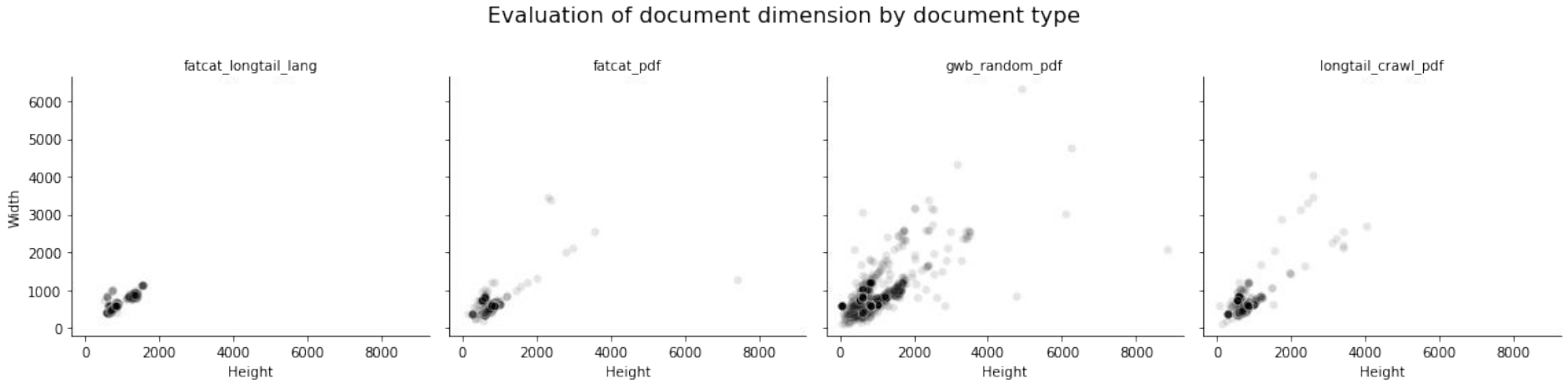
# Data Feature Engineering

- **Structural Features**:
  - Number of Pages
  - Height
  - Width
  - Size
  - Number of Characters
- **Content Features**:
  - Language (English, Romance, Other)
  - Structure - words that signify structure of a research paper
    - Abstract, introduction, conclusion, reference, table of contents
  - Content - words that represent the content of a paper
    - Research, analyze, result, table, investigation, explain, theory, study, paper, data, perform
  - Association - words that represent association
    - Journal, association, organization, doi, university, school, board

# Exploratory Data Analysis



# Exploratory Data Analysis



# Exploratory Data Analysis

# Initial Models

- Bayes Model Averaging
- Topic Modeling (LDA)
- Decision Trees
- K-Means Clustering

# Bayesian Model Averaging Results

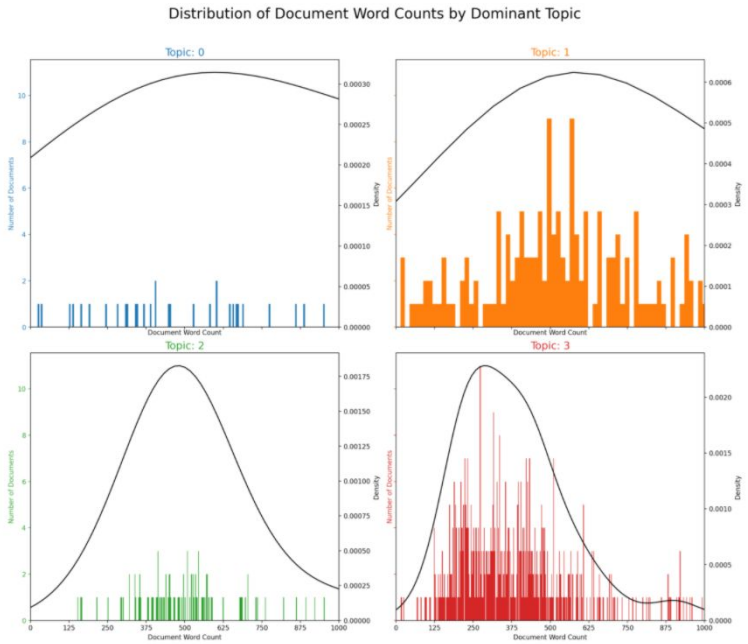
| Model | Features | Likelihood |
|---|---|---|
| 1 | width, structure | 6.20e-23 |
| 2 | number of pages, width, structure, number of characters | 1.03e-23 |
| 3 | number of pages, structure, number of characters | 5.71e-24 |

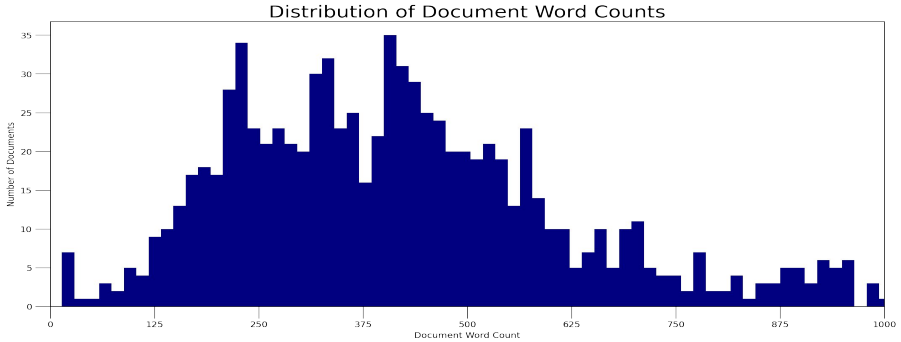
# Model Prediction Comparison

| Model | Features | Accuracy |
|---|---|---|
| 1 - Top model from BMA | width, structure | 75% |
| 2 - 2nd best model from BMA | number of pages, width, structure, number of characters | 78.3% |
| 3 - Selected for comparison to assess content-focused features | structure, content, association | 80% |
| 4 - Full model | number of pages, height, width, dimension, structure, content, association, language, number of characters | 77% |
| 5 - Simple model (for baseline comparison) | structure | 76.7% |

# Topic Modeling

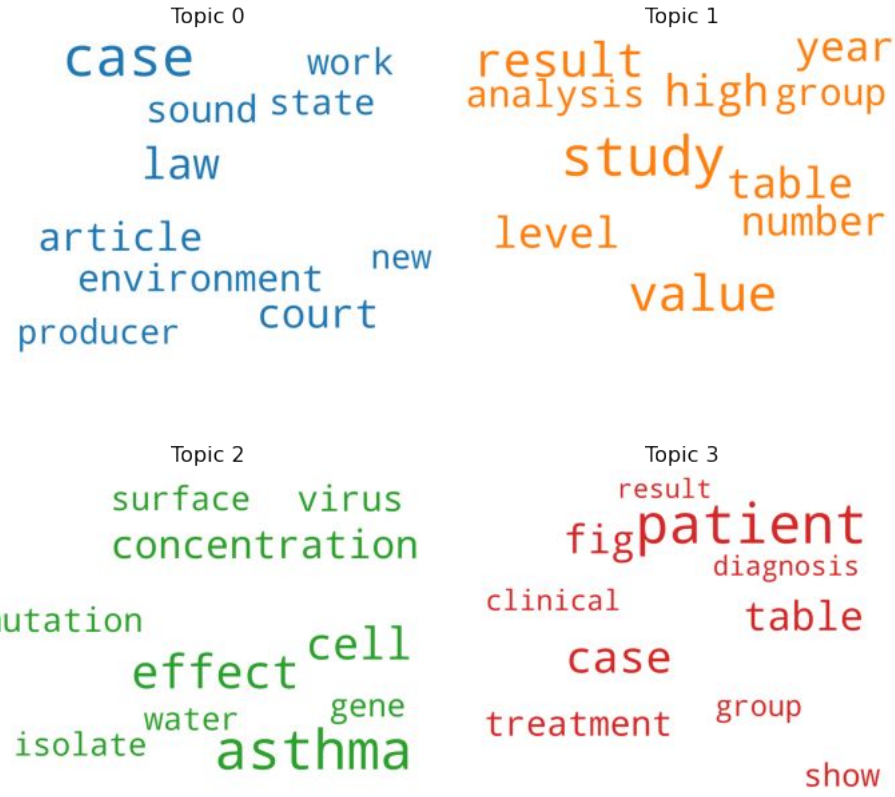
- Topic model based on LDA
  - Tokenize Sentences and Clean
  - Build the Bigram, Trigram Models and Lemmatize
  - Build the Topic Model
  - Dominant topic and percentage contribution in each document
  - Frequency distribution of word counts in documents



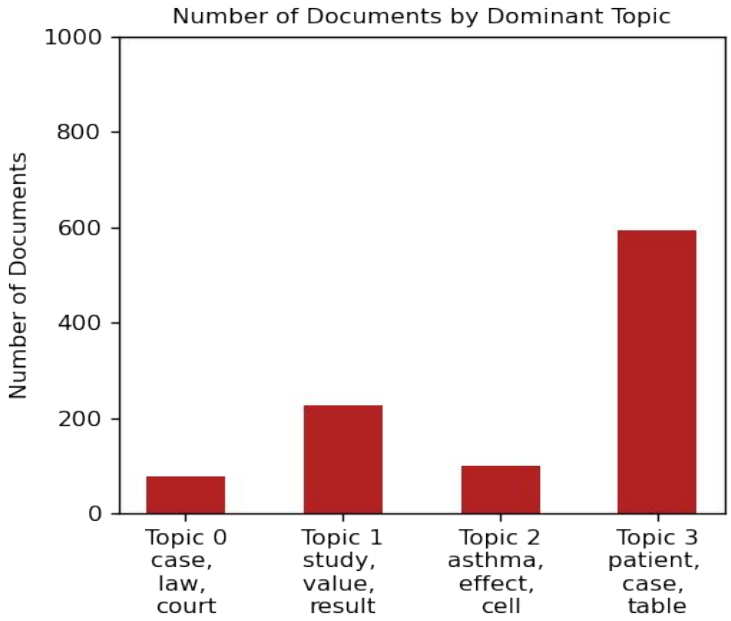



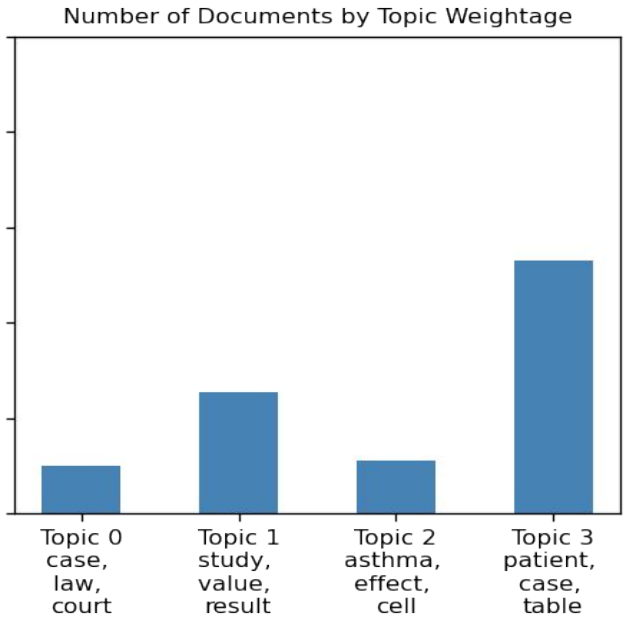
# Topic Word Clouds

Top 10 Keywords of each Topic



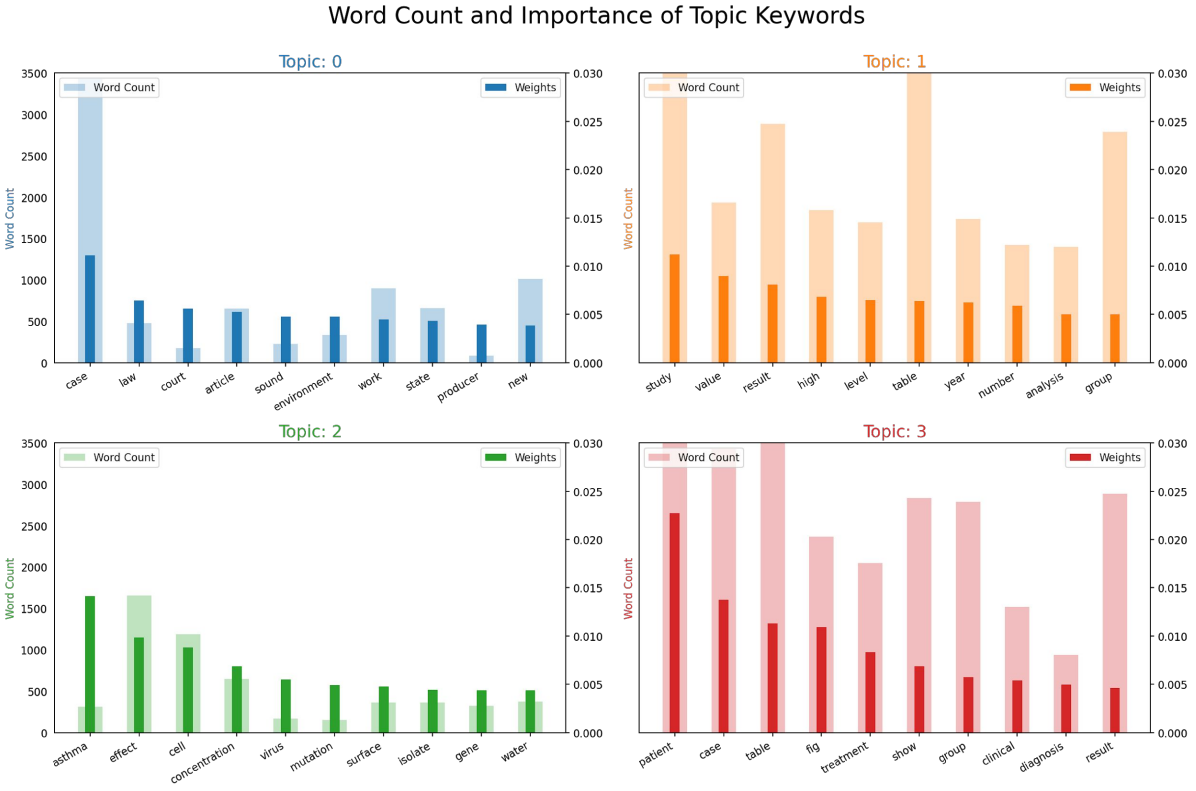
# Top Topics


Number of Documents by Dominant Topic
Number of Documents
1000
800
600
400
200
0
Topic 0
case,
law,
court
Topic 1
study,
value,
result
Topic 2
asthma,
effect,
cell
Topic 3
patient,
case,
table


Number of Documents by Topic Weightage
Topic 0
case,
law,
court
Topic 1
study,
value,
result
Topic 2
asthma,
effect,
cell
Topic 3
patient,
case,
table

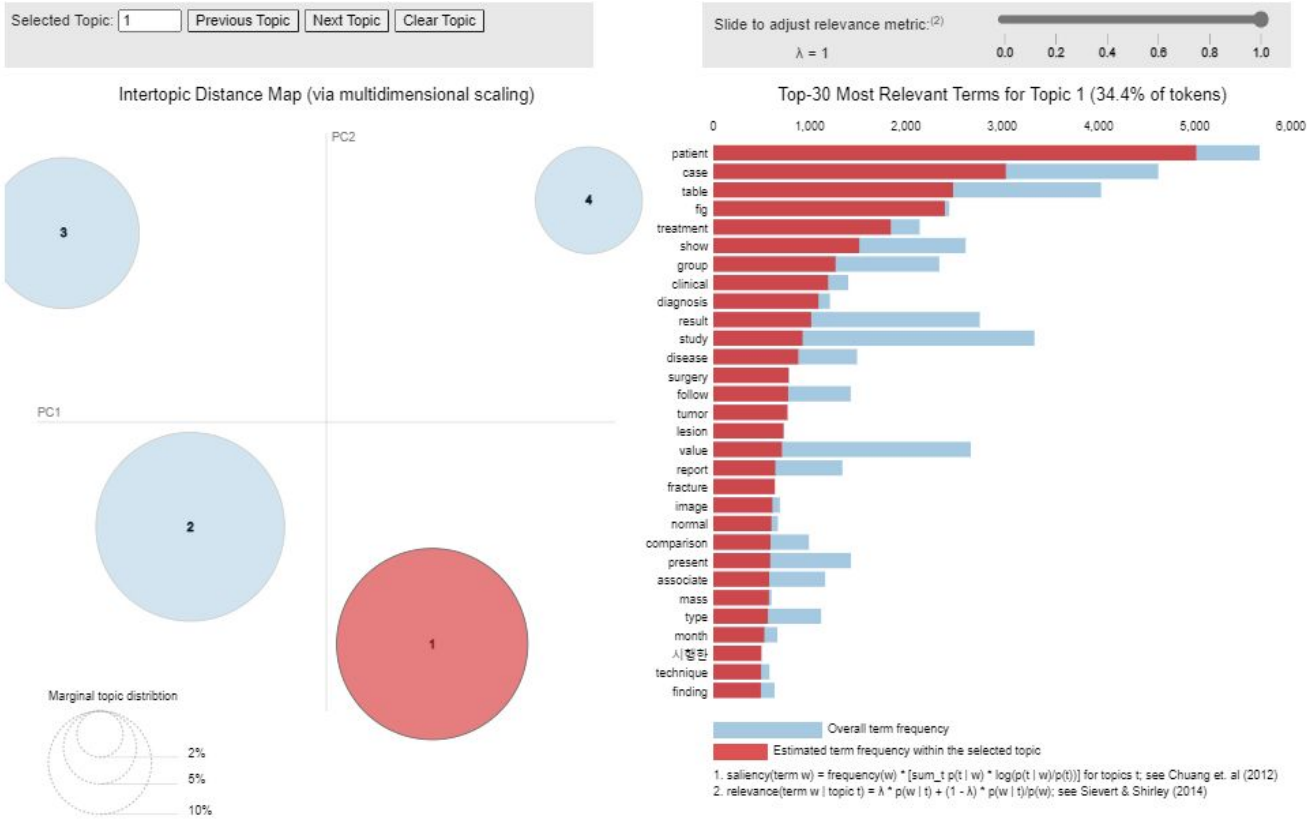
# Word Count and Weighted Keywords

Frequency of Keywords in each Topic



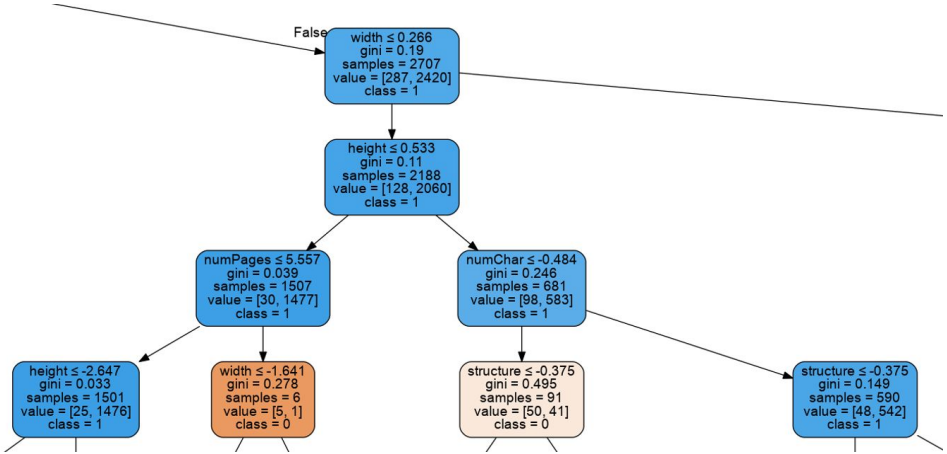
# pyLDAis

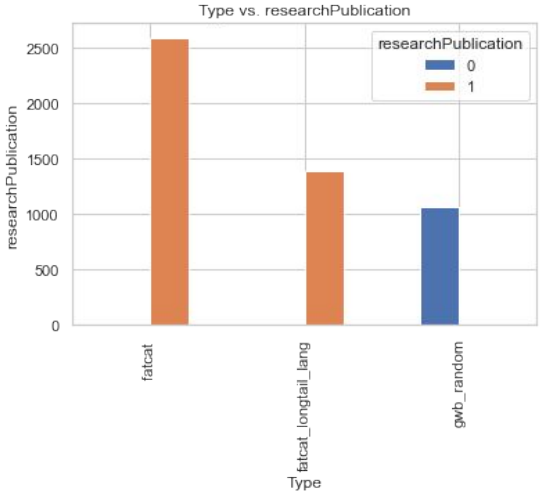
pyLDAVis Visualization of Topic Modeling



# Decision Tree

- Nodes : Test for value of a certain attribute
- Gini index: how often a randomly chosen would be incorrectly identified.
- Blue : 1 Research Publication
- Orange : 0 Not Research Publication



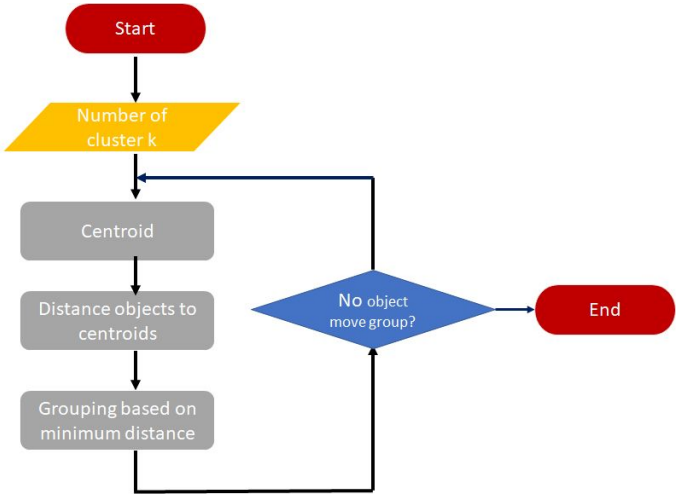
# K-Means Clustering

**Strengths**

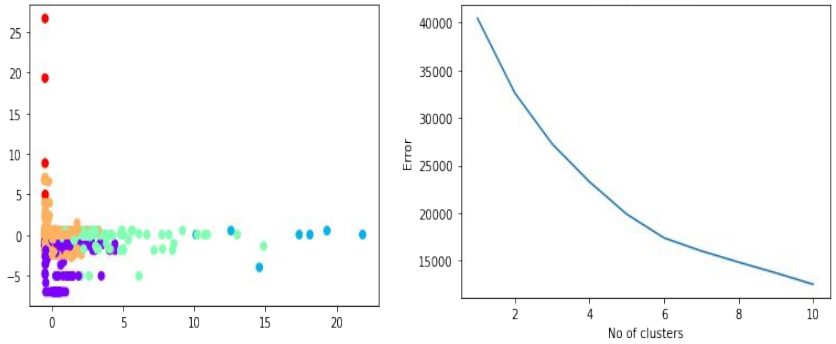
- Simple iterative method
- User provides “K”

**Weaknesses**

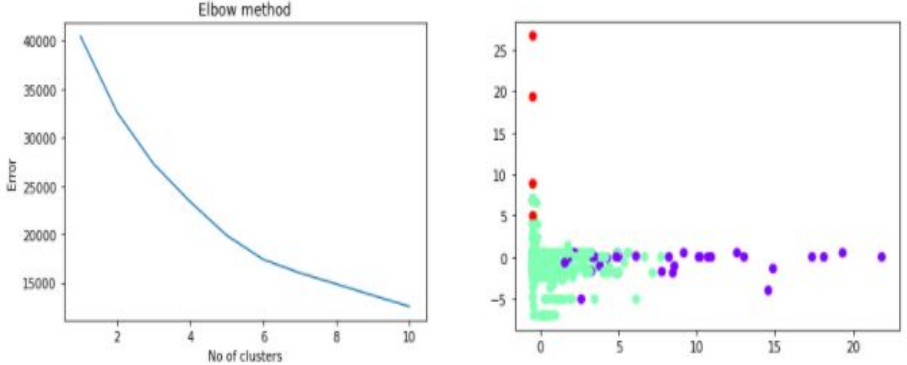
- Difficult to guess the correct “K”




K=5




K=3



# Conclusions

- Document structural parameters may be powerful predictors of a research paper
- The “look” of a document can be leveraged for accurate prediction
- Language model content markers also can predict research content at a substantive level
- Multiple language pose a challenge -- require further consideration

# Next Steps

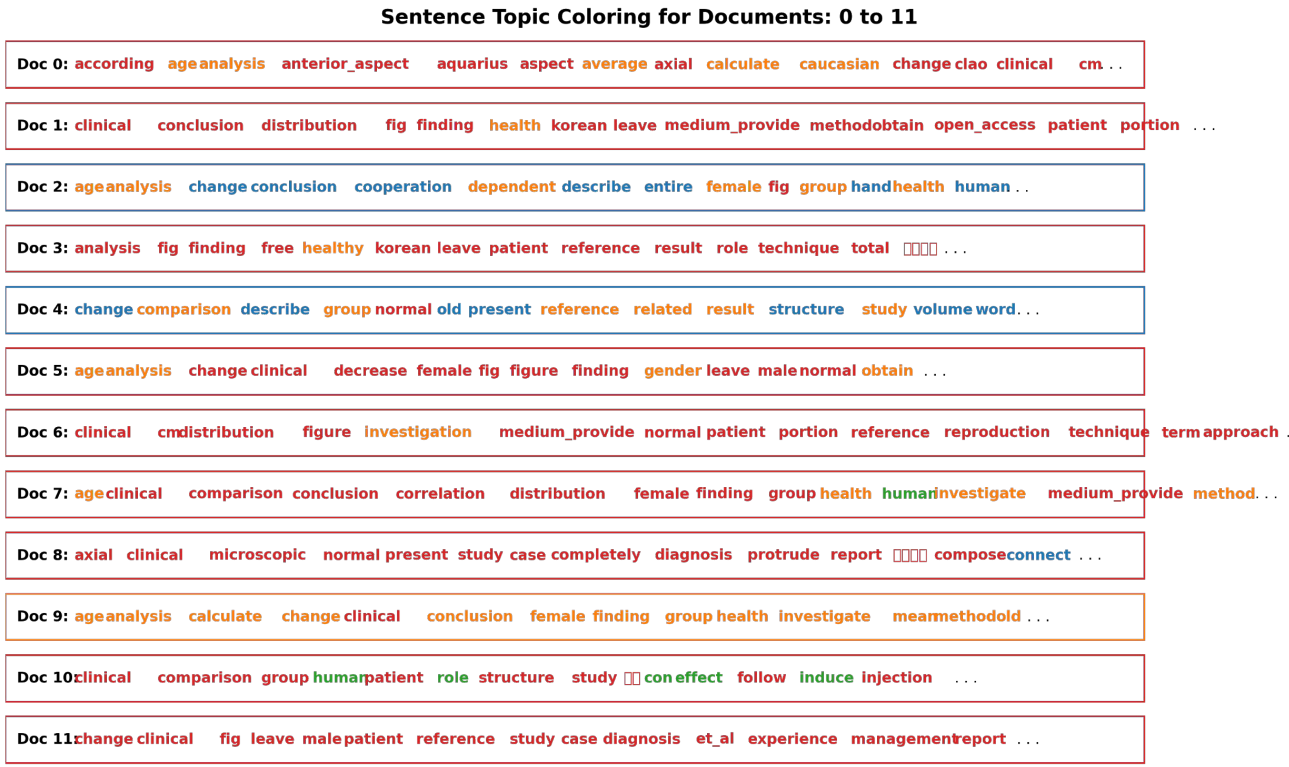
- Expand structural and language model analysis to a larger set of documents and refine further
- Further integrate initial topic modeling results into the language model
- Additional parsing of individual content language model elements to determine comparative predictive power (for simpler, more efficient modeling without sacrificing accuracy)
- Experiment with language model analysis in different languages
- Build on document structure findings with a computer vision/ image model
- Develop methods to address “edge cases”
- Assess inclusivity to align with “long-tail” objective

Thank You!

# Extra Slides

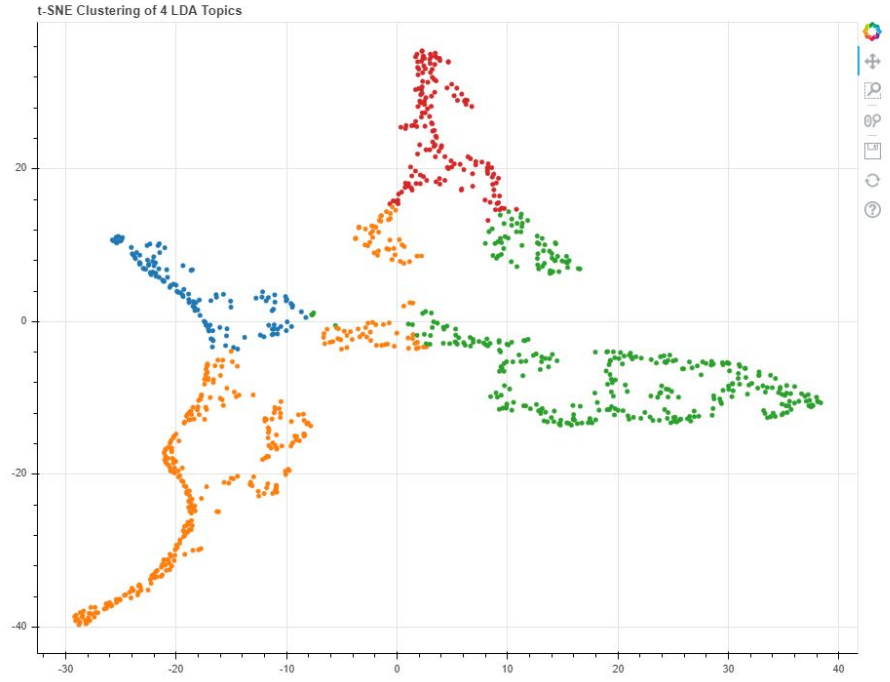
# Sentence Topic

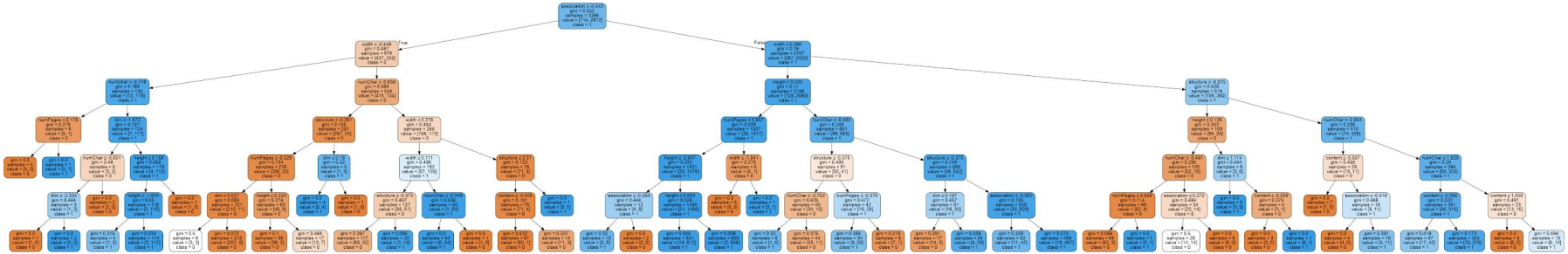
Words in documents are assigned color codes according to four topics



# t-SNE Clustering Chart

t-SNE visualize the cluster of documents in a 2D space (t-distributed stochastic neighbor embedding)




association ≤ 0.443
gini = 0.333
samples = 3288
value = [714, 2872]
class = 1
True
False
width ≤ 0.0448
gini = 0.497
samples = 575
value = [427, 252]
class = 0
width ≤ 0.296
gini = 0.29
samples = 2707
value = [287, 2620]
class = 1
numChar ≤ 0.278
gini = 0.5
samples = 192
value = [132, 118]
class = 0
numChar ≤ -0.606
gini = 0.469
samples = 383
value = [295, 134]
class = 0
height ≤ 0.033
gini = 0.25
samples = 2199
value = [128, 2085]
class = 1
structure ≤ 0.375
gini = 0.5
samples = 508
value = [159, 366]
class = 1
numPages ≤ 0.176
gini = 0.215
samples = 91
value = [16, 1]
class = 0
dim 1 ≤ 1.827
gini = 0.107
samples = 101
value = [7, 114]
class = 1
structure ≤ -0.249
gini = 0.198
samples = 192
value = [257, 34]
class = 0
width ≤ 0.779
gini = 1.484
samples = 191
value = [158, 110]
class = 0
numPages ≤ 0.157
gini = 0.5
samples = 1021
value = [1477, 1476]
class = 0
numChar ≤ 4.464
gini = 0.185
samples = 1178
value = [94, 963]
class = 1
height ≤ 0.136
gini = 0.5
samples = 119
value = [93, 24]
class = 0
content ≤ 0.527
gini = 0.469
samples = 26
value = [9, 17]
class = 1
numChar ≤ -0.854
gini = 0.01
samples = 361
value = [14, 539]
class = 1
gini = 0.0
samples = 5
value = [0, 7]
class = 1
num!Char ≤ 0.051
gini = 0.49
samples = 6
value = [4, 3]
class = 0
height ≤ 0.136
gini = 0.005
samples = 185
value = [3, 111]
class = 1
numPages ≤ 0.329
gini = 0.154
samples = 236
value = [257, 20]
class = 0
dim 0 ≤ 0.19
gini = 0.32
samples = 5
value = [1, 5]
class = 1
width ≤ 0.111
gini = 0.496
samples = 182
value = [157, 105]
class = 0
structure ≤ 2.31
gini = 0.278
samples = 76
value = [1, 6]
class = 1
height ≤ 0.033
gini = 0.5
samples = 1021
value = [1477, 1476]
class = 0
width ≤ 1.661
gini = 0.278
samples = 6
value = [1, 5]
class = 1
structure ≤ -0.375
gini = 0.495
samples = 91
value = [81, 69]
class = 0
structure ≤ 0.375
gini = 0.166
samples = 550
value = [13, 717]
class = 1
numChar ≤ 0.693
gini = 0.295
samples = 100
value = [93, 19]
class = 0
dim 0 ≤ 0.114
gini = 0.444
samples = 19
value = [1, 3]
class = 1
structure ≤ 0.375
gini = 0.5
samples = 508
value = [159, 366]
class = 1
numChar ≤ -0.854
gini = 0.01
samples = 361
value = [14, 539]
class = 1
dim 2 ≤ -2.329
gini = 0.444
samples = 3
value = [1, 2]
class = 1
gini = 0.0
samples = 2
value = [0, 2]
class = 1
gini = 0.0
samples = 1
value = [1, 0]
class = 0
height ≤ 0.033
gini = 0.496
samples = 118
value = [3, 106]
class = 1
gini = 0.0
samples = 67
value = [0, 67]
class = 1
gini = 0.0
samples = 4
value = [3, 0]
class = 0
gini = 0.5
samples = 4
value = [3, 3]
class = 0
gini = 0.375
samples = 8
value = [2, 6]
class = 1
gini = 0.0
samples = 110
value = [0, 110]
class = 1
gini = 0.0
samples = 5
value = [5, 0]
class = 0
gini = 0.0
samples = 5
value = [0, 5]
class = 1
gini = 0.0
samples = 5
value = [5, 0]
class = 0
gini = 0.0
samples = 1
value = [0, 1]
class = 1
gini = 0.0
samples = 1
value = [1

# Requirements for Project

Your presentation must have a minimum of 4 slides cover the material below.

(a) Title with the names of your capstone team, faculty advisor, client, and sponsor.

(b) Problem statement with a description of your problem and why it is important.

(c) The sources and types of data you are using for your project.

(d) Your data, ***exploratory***, and ***feature engineering*** tasks to help you address your problem statement.

Data :

- 40k pdf files from 4 datasets

Exploratory :

- Graphs and Dashboard (All of us can explore)
- Bayes Model Averaging (model performance and parameters)

Feature Engineering :

- Meta data (csv)
- Text data (create csv containing file, text, and text language) Yihnew
- Images of first page of pdf
- Multiple languages (translate text to english)
- Top Modeling (Huilin)