Unpacking Digital Grinnell

Exploring our college's values through institutional archives

Abstract

Our study investigates metadata of the objects in Digital Grinnell, an online archival platform hosting the collections of Grinnell College. The Digital Grinnell repository showcases scholarship, documents and specimens of historical significance to the college and surrounding community; we examine this source to gauge which values were most prevalent in the crafting of our institutional narrative. To scrape data for analysis, we run parallelized loops through the website's item listings to gather metadata for comparison across items and collections. We then derive descriptive statistics, and, to identify themes of importance to the college, prepare word frequency visualizations on the texts associated with items. To assess underlying trends in the language the College uses to describe its collections, we model our data via Latent Dirichlet Allocation to find that underlying "topics" of local spaces, students at the college, and the town's commemorative institutions emerge as clusters in Digital Grinnell's descriptive text.

Background and Research Objective: Digital Grinnell aims to document scholarly works by students and faculty alongside academic resources and materials of historical significance to the college and the town of Grinnell. As such it is an example of an archive, a construction fundamental to the creation and preservation of Grinnell's institutional memory. Examining the various collections housed within Digital Grinnell, we aim to identify themes in their objects' explanatory and descriptive texts to gain a better understanding of Grinnell's institutional values and goals as curators of a broader history. With this in mind, our research question is framed as follows: "What are the salient characteristics, themes, and values evidenced in Grinnell's archives of institutional memory?"

Methods

Data Scraping and Cleaning: Digital Grinnell stores each of its objects and its metadata as its own webpage, accessed by navigating through collections pages or through a sitewide search. Using a reference list exported from Digital Grinnell's search function, we identified a pattern in the URLs of individual web pages to construct a dataframe for html scraping from object IDs. Excluding cases we knew to be problematic (oral histories and art objects, which require different URL strings and document different metadata; webpages under maintenance; pages which require user login; and pages representing whole collections, which would compromise the integrity of our unit of analysis and duplicate information), we used packages `httr` and `rvest` within looped functions to visit each of over 9000 object pages and extract those which contained valid metadata. The filtering process took several hours, so we exported our cleaned data as a dataframe with each case representing an item in the collections[1].

Data Description and Variable Selection: Once we had compiled a listing of item webpages which were eligible for metadata extraction, we determined a list of the metadata variables on the site which would be consistently available for all items, from ancient Roman coins, to senior theses, to historical photography. Each item in Digital Grinnell has a unique table of metadata values, with no standard order or designated length--one item might have eight separate "keyword" fields on display, whereas another might not have any keywords at all, or might have the same information listed as "topic." To eliminate much of the inconsistency, and to find relevant fields for objects of all varieties in order to perform a holistic and unifying analysis on the collections, we selected 7 string variables: Title, Description, item, key, resource, topic, and geo; and one numeric variable, Date,[2] then ran a parallelized set of loops for extracting metadata[3] from the html and xml nodes on each object page, which we translated into additional columns in our dataframe for use in an overarching analysis.

Visualizations and Model: We address the question of Digital Grinnell's themes and emphases on two levels, looking at word frequencies and textual clustering. Our data is almost entirely categorical and textual, so our visualizations assessing different facets of the collections incorporate word-clouds and barplots. For our model, we selected Latent Dirichlet Allocation (an example of probablistic topic modelling, similar to k-means clustering but better suited to our questions, with overlap allowed between clusters and the probabilities of words to a topic constituting their proportion of cluster membership) , using Gibbs sampling for our iterations and the package topicmodels for our implementation. After experimenting with k-values between 2 and 10 in order to achieve the highest stability and interpretability.

Results

Visual Trends in Word Frequency: A comparison of word-clouds generated from subdivisions of the collections created during different time periods reveals an interesting finding: the older materials in Digital Grinnell emphasize "Iowa" most, but the descriptions of newer objects more frequently include "Grinnell"[4]; the institution's record-keeping seems to reflect an underlying strengthening of institutional identity, considering the college and town themselves to be the locus of culture and heritage rather than simply carried in the waves of statewide trends. Grinnell college was, after all, once called Iowa College, and recent projects such as Imagine Grinnell have intentionally mobilized the town's name to promote a sense

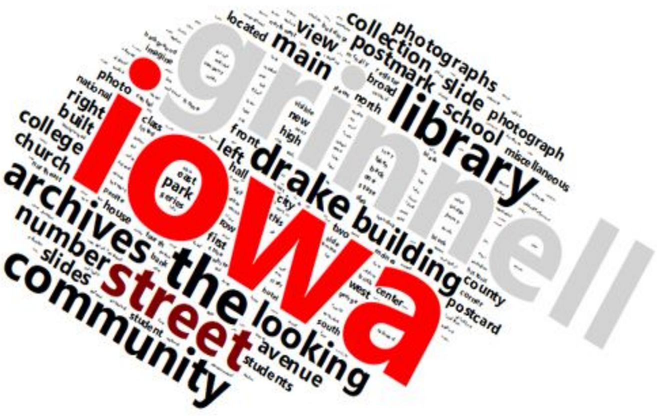
[1] Due to the nature of the website's upkeep, *exported data represented might look different if the script were run again today.*

[2] See appendix B for variable descriptions. Our analysis focused on Title, Description, item, and Date.

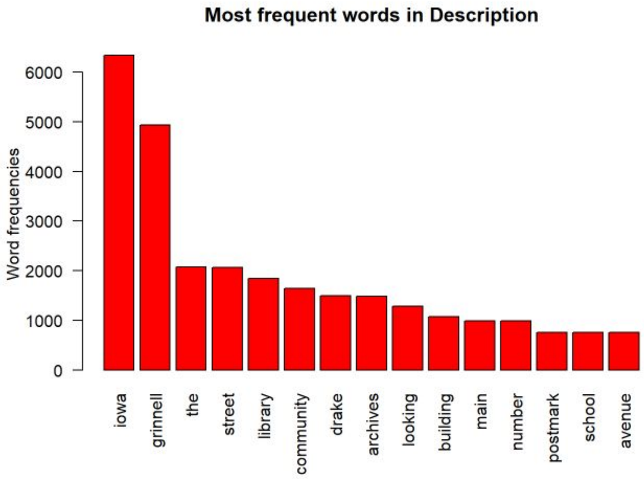
[3] *We had to decide before we ran our loop which metadata was useful to us; future users of this code to might have different priorities, especially if they choose to examine Faulconer Art*

[4] See appendix A for examples of static wordclouds examining separate time periods and specific collections

of local pride. Other interesting findings lie in the word-clouds which examine specific collections and observe what emphases seem to exist within their object preferences.[4] Digital Grinnell in its entirety unifies collections dealing with student life and the broader community, but through breakdowns like these we can begin to see how the sources of those themes are often divided between student-oriented and community-oriented collections[4].



“Topics” from LDA: The interpretation of the three clusters output from our LDA model has not been completely transparent, but a naive analysis could summarize the first as referring to a theme of local spaces (spatial terms, buildings and addresses in Grinnell), students at the college (yearbook photos, campus locations), and the town’s commemorative institutions (archives, libraries, chamber of commerce) emerge as clusters in Digital Grinnell’s descriptive text[5]



Discussion & Future Scholarship: Grinnell College has historically been engaged with its surrounding community, but its definition of that community’s scope has shifted through time towards more local pride. Local community institutions have clearly been integral in the formation of the archive; thus we frequently see references to examples in the community (“drake,” “church,” “bank”), the college campus (“burling,” “college”), and between them (“library,” “archives”). The curators of Digital Grinnell have taken care to describe their collections and community in concrete spatial terms, referencing specific streets and directions; they have chosen to represent their own identity and history largely in terms of buildings and urban development, rather than through an emphasis more intangible ideas. The college’s influence is clear even when looking only at the community-centered collections. We can see Digital Grinnell’s intentionality in combining community and academic sources in forming this website, which could easily have preoccupied itself only with the college but clearly brought in just as much from the surrounding community. It will be interesting to see how that pans out as the collections continue to be digitized, and if that balance will shift at all.

There were, of course, limitations to our study--some arising from the data source, with its inconsistent metadata, only partially accessible pages, and overall incompleteness; some arising from our methods, which focus on categorical data and are largely descriptive rather than predictive; and some arising from the limited scope of this short-term project, which meant we made decisions about where to focus which could be altered in future work. In a continuation of this project, we might consider weighting our comparisons across collections to account for their relative sizes. It might not be appropriate to make blanket statements about the entirety of the collections at this stage in Digital Grinnell’s development, as the born-digital collections were obviously overrepresented in the early uploads, and objects which need other kinds of processing simply require more effort to be entered into the system. The rest of the physical collections, once processed, will provide even more variety and only make a search for unifying concepts and themes all the more intriguing, providing us a more holistic picture of Grinnellian values.

[5] See appendix C for output from our LDA model, which shows the words with the highest-probability words included in each of the three clusters. Note that the words have been stemmed in order to better identify clusters.

References

Avila, Luis M. “A web scraper tutorial using R packages htts and rvest.” 30 October 2017: https://lmavila.github.io/markdown_files/PeruCurrencyScraper.html

Grinnell College Libraries. *Digital Grinnell:* https://digital.grinnell.edu

Jones, Matt. “Quick Intro to Parallel Computing in R.” 25 July 2017: https://nceas.github.io/oss-lessons/parallel-computing-in-r.html

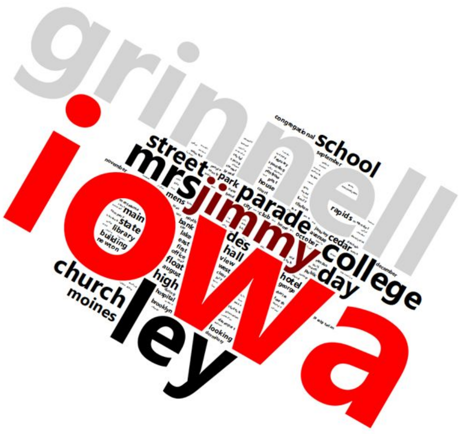
Lettier. *Your Easy Guide to Latent Dirichlet Allocation*. 23 Feb 2018: https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

Miller, Ryan. “Clustering (Part 1)” & “Clustering (Part 2).” *Sta-230 Introduction to Data Science:* https://remiller1450.github.io/sta230s19.html

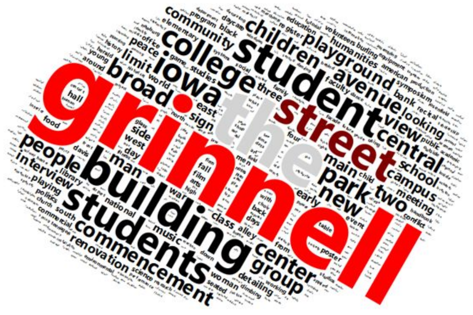
Appendix

A. Static Word Clouds

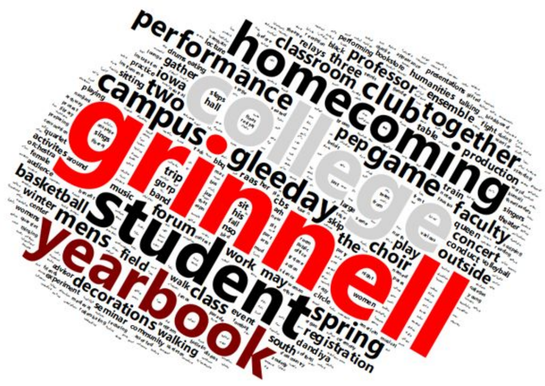
Descriptions of Items Created 1914-1949



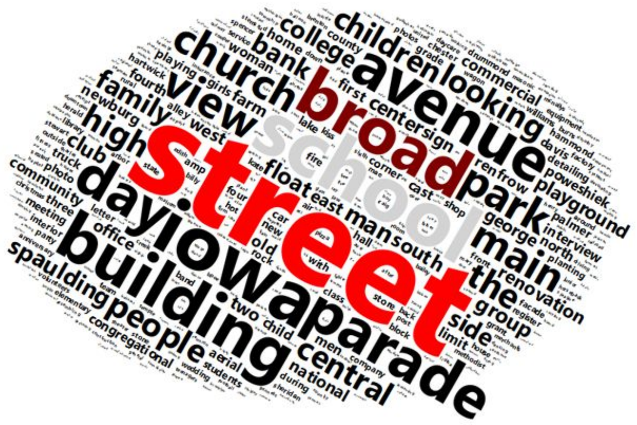
Descriptions of Items Created 1988-Present



Descriptions in “Life at Grinnell College”



Descriptions in “Poweshiek History Project”

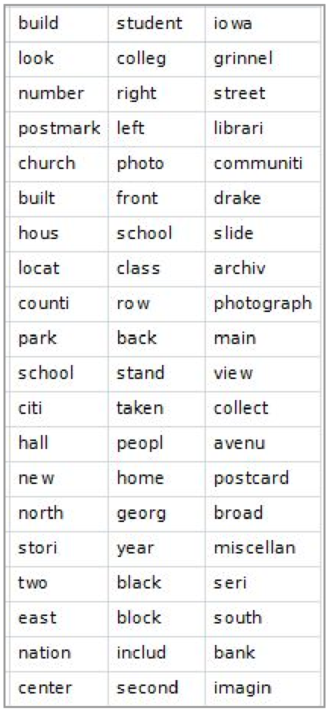


B. Descriptions of Variables in Metadata dataframe:

Variable	Meaning	Type	Example
Title	Name of the object and its webpage on Digital Grinnell.	string	“Axis of colony of Archimedes”
PID	Unique identifier used to construct URL string for an item’s webpage.	string of form “grinnell:####”	“grinnell:23150”
Description	Short Abstract describing the item and its relevance in the collections; displayed in results of search function and on item’s webpage.	string	“A pamphlet by L.F. Parker regarding Hester Hillis and her missionary experiences at home and abroad, written on the occasion of her passing in 1887. Includes letters written by Hillis from India about her experiences.”
date	Year the item was created. From “Index Date” or “Import Index” on	numeric	1885

	item’s webpage.		
item	From “Related Items” on Digital Grinnell website; indicates collections to which the object belongs.	string, containing substrings separated by commas	“Campus Events, Life at Grinnell College, Social Justice at Grinnell”
key	From item’s webpage: associated keywords. Excluded from analysis because it is inconsistently thorough based on collection, more useful for exploration within a collection.	string	“Latin, Quadrigatus, Janus, Jupiter, Quadriga, Victory, Roma”
resource	“Resource Type” from the item’s webpage, indicates form the item takes in its storage for Digital Grinnell; typically “still image.”	string	“Still image”
topic	From item’s webpage: associated topics. Unlike keywords, these are not standardized, but specific to the item. Excluded from analysis because it is inconsistently thorough based on collection, more useful for exploration within a collection.	string	“Internet and women, African Americans, Blacks, Internet -- Safety measures”
geo	Geographic location of the item; inconsistently formatted and inconsistently defined; overwhelming within Iowa and largely within Grinnell. Excluded from analysis.	string	“Grinnell (Iowa)”

C. LDA Output: first 20 high-probability words for each of the 3 underlying “topics”:



build	student	iowa
look	colleg	grinnel
number	right	street
postmark	left	librari
church	photo	communiti
built	front	drake
hous	school	slide
locat	class	archiv
counti	row	photograph
park	back	main
school	stand	view
citi	taken	collect
hall	peopl	avenu
new	home	postcard
north	georg	broad
stori	year	miscellan
two	black	seri
east	block	south
nation	includ	bank
center	second	imagin