Evolution of Trending Topics in Mechanical Engineering Research Theses

by

Sofia Eva Leon

Submitted to the
Department of Mechanical Engineering
in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Mechanical Engineering

at the

Massachusetts Institute of Technology

May 2022

© 2022 Massachusetts Institute of Technology. All rights reserved.

Signature of Author: __________________________________________

Department of Mechanical Engineering

May 6, 2022

Certified by: ________________________________________________

Franz S. Hover
Senior Lecturer of Mechanical Engineering
Thesis Supervisor

Accepted by: ________________________________________________

Kenneth Kamrin
Associate Professor of Mechanical Engineering
Undergraduate Officer
Evolution of Trending Topics in Mechanical Engineering Research Theses

by

Sofia Eva Leon

Submitted to the Department of Mechanical Engineering
on May 6, 2022 in Partial Fulfillment of the
Requirements for the Degree of
Bachelor of Science in Mechanical Engineering

ABSTRACT

As new concepts and technologies emerge, researchers in mechanical engineering have focused on various areas of study. This thesis seeks to understand the evolution of research topics over time and identify which subjects have been favored at different points.

To accomplish this goal, the titles of MIT Mechanical Engineering theses were analyzed to measure the frequency with which certain keywords appear each year. Text data from theses published between 1975 and 2021 (inclusive) was collected from two MIT websites using an automated script. Using R, the text data was broken up and processed to count the number of occurrences of a particular word in each year. To account for differences in the amount of available data each year, the annual number of occurrences of a word was normalized by the total number of words that appeared that year.

Through this analysis, several interesting trends are revealed. Key words and phrases tend to have small time windows (about 5 to 10 years) where they experience heightened popularity and then see decreased usage. This likely represents a rise in research interest when a technology is novel, followed by a decrease in interest once the technology becomes commoditized or obsolete. Furthermore, we will take a deep dive into trends in manufacturing technologies and explore a case study comparing various manufacturing techniques. We take particular interest in additive manufacturing as the research interest in this topic has grown in recent years.

Thesis Supervisor: Franz S. Hover
Title: Senior Lecturer of Mechanical Engineering
ACKNOWLEDGEMENTS

I would like to thank Dr. Franz Hover for his assistance and insights in developing this thesis. I would also like to thank the MIT Mechanical Engineering Department for their patience and guidance during the course of this work. Finally, I would like to thank my peers Mason Massie and Gregory Xie for helping me think through the ideas in this thesis.
Table of Contents

Abstract
Acknowledgements
Table of Contents
List of Figures
List of Tables
1. Introduction
2. Methods
   1. Data Acquisition
   2. Data Processing
3. Broad Trends
   1. Common Words
   2. Bursts of Interests
   3. Commoditization of Computers
4. Manufacturing Trends
   1. Trendy Phrases
   2. Traditional Manufacturing Methods
5. Additive Manufacturing Case Study
   1. Part Design
   2. Simulating Part Performance
   3. Implications
6. Conclusion
Bibliography
Appendix
List of Figures

Figure 2.1: Number of theses available each year from digital MIT repositories. 10
Figure 3.1: Word cloud of the 100 most common words. 11
Figure 3.2: Plot depicting usage of the phrase “spectral element”. 12
Figure 3.3: Plot depicting usage of the word “computer”. 13
Figure 3.4: Plot depicting usage of word which contain “comput”. 13
Figure 4.1: Plot depicting usage of the phrase “computerized manufacturing”. 14
Figure 4.2: Plot depicting usage of the phrase “lean manufacturing”. 15
Figure 4.3: Plot depicting usage of the phrase “additive manufacturing”. 15
Figure 4.4: Plot depicting usage of the word “printing”. 16
Figure 4.5: Plot depicting usage of words containing “print”. 16
Figure 4.6: Plot depicting usage of the word “mill” 17
Figure 4.7: Plot depicting usage of the phrase “milling machine” 18
Figure 4.8: Plot depicting usage of the word “waterjet” 19
Figure 4.9: Plot depicting usage of the phrase “water jet” 19
Figure 5.1: Design of waterjet part. 20
Figure 5.2: Design of milled part. 21
Figure 5.3: Design of 3D printed part. 21
Figure 5.4: Diagram of FEA setup. 22
Figure 5.5: Simulated performance of three parts. 23
Figure 5.6: Zone of high stress in waterjet part. 24
Figure 5.7: Zone of high stress in milled part. 25
Figure 5.8: Zone of high stress in 3D printed part. 25
List of Tables

Table 2.1: Sample of thesis data. 9
1. Introduction

Every year, students in the MIT Department of Mechanical Engineering produce theses as part of their studies. A historical record of these theses is available online through several MIT digital repositories, making abundance of text data available for collection and processing. To better understand how mechanical engineering research may develop in the future, we must first look to the past to understand historical trends. By analyzing changes in word usage over time, we can identify shifting interests in research topics and learn how novel technologies evolve.

When a technology is first introduced, it can be difficult to imagine it becoming a part of everyday life. However, cutting edge research often becomes commoditized after many years of development. This thesis seeks to identify and understand how these development cycles appear in academic research. With greater historical context, we can change our perspective on novel technologies and realize that they too might become commoditized in the near future. Furthermore, by understanding which research topics were popular in the past, we can identify how they have shaped the current state of mechanical engineering research.

To accomplish these goals, this study will acquire data on the titles and publication dates of MIT Mechanical Engineering theses available on DSpace@MIT and the MIT Libraries “Search Our Collections” website. By counting how often a specific word or phrase was used in a particular year, we can learn about the research interest in that topic over time. Once trends are identified, a case study will be examined to understand how a topic which currently has high research interest (additive manufacturing) may evolve in the future.
2. Methods

The title and publication date of MIT Mechanical Engineering theses were collected from DSpace@MIT as well as the MIT Libraries “Search Our Collections” website. This data was then processed to understand trends in word usage over time.

2.1 Data Acquisition

Python scripts (shown in the Appendix) were written and run to automate acquisition of data from DSpace@MIT and the MIT Libraries website. The open-source tool Selenium enabled the automatic collection of text data which appeared on the websites’ search result pages. On DSpace, the search results of interest were all MIT theses connected to the Department of Mechanical Engineering. Similarly, the search results of interest on the MIT libraries website were all MIT theses in which the MIT Department of Mechanical Engineering was an author. Theses from all degree types were collected - including doctoral, graduate, and undergraduate. The automated scripts were able to extract the title and publication date of each thesis in the search results, producing a table as shown in Table 2.1 below.

<table>
<thead>
<tr>
<th>Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dispersion of water sprays in a transverse air jet and the aging of spray nozzles</td>
<td>1999</td>
</tr>
<tr>
<td>Scalable manufacturing of hierarchical nanostructures for thermal management</td>
<td>2012</td>
</tr>
<tr>
<td>Development of an inertial generator for embedded applications in rotating environments</td>
<td>2007</td>
</tr>
<tr>
<td>A study of system-induced instabilities in forced-convection flows with subcooled boiling</td>
<td>1965</td>
</tr>
<tr>
<td>A programming system for hybrid computer computation</td>
<td>1975</td>
</tr>
<tr>
<td>Design and construction of an electrically commutated A-C motor with a fixed torque angle</td>
<td>1974</td>
</tr>
<tr>
<td>Force feedback hydraulic servos for advanced assembly machines</td>
<td>1975</td>
</tr>
<tr>
<td>Incompressible journal bearings with combined hydrostatic-hydrodynamic action</td>
<td>1965</td>
</tr>
<tr>
<td>Thermal aspects of metal cutting</td>
<td>1952</td>
</tr>
<tr>
<td>Acoustic wave propagation and non-intrusive velocity measurements in highly concentrated suspensions</td>
<td>1991</td>
</tr>
</tbody>
</table>

Table 2.1: Sample of thesis data collected from MIT digital repositories using automated Python script.

Data from DSpace and the MIT Libraries website were combined and duplicate titles were removed. As a result, it was possible to attain 10,304 theses from 1873 to 2021. Figure 2.1 displays the number of theses that were able to be collected for this study for each year. Unfortunately, there were fewer theses available prior to 1975, so most of this study’s analysis will focus on the period from 1975 -2021 (inclusive). Note that a slight gap exists
from 1991-1994 due to MIT Department of Mechanical Engineering not appearing as an author for theses available on the MIT Libraries website. Accessing data via multiple websites helped to fill gaps in either repository and create a more complete dataset.

Figure 2.1: Number of theses that were available each year from DSpace@MIT and the MIT Libraries “Search Our Collections” website.

2.2 Data Processing

Once the data was assembled, the titles were tokenized using the R package tidytext. This broke up the titles to create pairs of words and the year in which the word appeared in a title. The tokenizing process was repeated to also extract pairs of two-word phrases and their publication year. Additionally, stop words (such as “the”, “is”, “a”) were removed to eliminate unnecessary and noisy information. With the data in this form, it was possible to count the number of times a specific word or phrase appeared in a particular year. Furthermore, the total number of words appearing in a certain year were counted. This made it possible to normalize the number occurrences of a word in a given year by the total number of words in that year, controlling for the differences in availability of data in different years. The absolute value of this metric is unimportant as further analysis will focus on its relative value compared to other years for the same word or phrase.
3. Broad Trends
By analyzing changes in the use of a particular word or phrase over time, it is possible to identify trends in mechanical engineering research topics. Notably, it is possible to identify periods of high interest in a specific topic during which the word or phrase of interest is used more than in other years.

3.1 Common Words
Looking over all years of available MIT Mechanical Engineering thesis data, we can get a first glance at areas of interest by counting the number of occurrences of a word over all time (from 1873-2021 inclusive). Figure 3.1 depicts a word cloud of the 100 words used the greatest number of times throughout the entire dataset. “Design” was the most used word with 1,998 occurrences. The second and third most used words were “analysis” and “system” with 797 and 767 occurrences, respectively.

Figure 3.1: Word cloud of the 100 most commonly occurring words in MIT Mechanical Engineering theses looked at by this study. Larger size indicates greater use.
3.2 Bursts of Interest

We can recognize bursts of interests in a topic by plotting the usage of a specific word or phrase over various years. To control for differences in availability of data in different years, the relevant metric to plot is the fraction of words (or phrases) in a particular year that match the target word (or phrase). Additionally, due to limited data prior to 1975, analyses of trends for specific words and phrases will focus on the period of 1975 through 2021. An example of this burst phenomena is the use of the phrase “spectral element” as depicted in Figure 3.2. The phrase experiences a surge of research interest after 1984 when it is introduced and popularized by Anthony Patera’s paper on fluid dynamics [1]. However, the phrase then experiences diminished research activity once the topic is no longer novel, and it does not see any use after 1996.

![Figure 3.2: Plot depicting usage of the phrase “spectral element” over time.](image)

3.3 Comoditization of Computers

Analyzing the research interest in a technology as it becomes commoditized reveals a counter-intuitive trend. One might think that the more a technology is used, the more it appears in MIT theses. However, the opposite trend appears in the data. For example, the word “computer” experiences the most usage in the late 1970s and 1980s and tends to drop off in subsequent years (Figure 3.3). Additionally, Figure 3.4 shows that the trend is similar for usage of words
that contain the string “comput” (for example “computational”). These patterns indicate that high research interest in a topic is proportional to the novelty and growth rate of the technology, not its prevalence in day-to-day scenarios.

**Figure 3.3**: Plot depicting usage of the word “computer” over time.

**Figure 3.4**: Plot depicting usage of word which contain “comput” over time.
4. Manufacturing Trends

Looking at the content of theses titles over time has revealed interesting patterns in manufacturing methodologies. Specific approaches toward manufacturing tend to experience bursts of research interest during 5 to 10 year periods and afterwards appear less frequently.

4.1 Trendy Phrases

The first trend in manufacturing is the phrase “computerized manufacturing” which appears from 1978 to 1981 as shown in Figure 4.1. Similar to the word “computer”, this phrase likely experienced heavy usage while the technology was novel and then decreased usage once computerized manufacturing becomes standardized.

![Figure 4.1: Plot depicting usage of the phrase “computerized manufacturing” over time.](image)

The next notable trend in manufacturing is the phrase “lean manufacturing” which saw a burst of interest from 1996 to 2007 (Figure 3.4). Lean methodologies were initially developed by Toyota as a method to reduce the cost of automobile production by keeping less inventory and using fewer resources. As reflected by the phrase usage, lean manufacturing techniques became popularized in the early 21st century as manufacturers sought to lower costs and improve their ability to adapt to consumer demand [2].
Finally, the most recent phrase of interest in manufacturing has been “additive manufacturing”. As shown in Figure 3.5, it began appearing in 2015 and continues to show strong usage through 2021. Since we are currently in the middle of the burst of interest in additive manufacturing, it is interesting to think about how the technology will develop in the future. Based upon previous trends, it’s not unreasonable to think that additive manufacturing might also become standardized similar to “computerized manufacturing”.

Figure 4.2: Plot depicting usage of the phrase “lean manufacturing” over time.

Figure 4.3: Plot depicting usage of the phrase “additive manufacturing” over time.
Furthermore, because additive manufacturing is relatively novel, there is MIT thesis data available to show how it developed. As early as 1989, there were MIT Mechanical Engineering theses about “three dimensional printing”. In fact, the word “printing” and words containing “print” saw considerable use in the 1990s (Figures 3.6 and 3.7). The shift from research about “3D printing” to “additive manufacturing” demonstrates a change in the attitude toward the technology from being a lab-scale innovation to a technology with broader manufacturing capabilities.

**Figure 4.4:** Plot depicting usage of the word “printing” over time.

**Figure 4.5:** Plot depicting usage of words containing “print” over time.
4.2 Traditional Manufacturing Methods

As previously established, the amount of research interest in a topic tends to correlate to its novelty. This trend continues to be observed when more traditional methods of manufacturing are analyzed. Looking back at history, milling machines developed from lathe-style gear cutters in the late 1700s and early 1800s. The first true milling machines were created by Eli Whitney and a group of English gunsmiths between 1818 and 1820 [3]. Although there continue to be advancements in milling technologies and milling machines are still frequently utilized, there have been no recent periods of heightened interest in this topic. Looking at usage of the word “mill” and the phrase “milling machine” between 1975 and 2021 revealed no strong trend or burst of interest (Figures 4.6 and 4.7). Instead, research interest is sparse and spread out.

![Graph depicting usage of the word “mill” over time.](image)

**Figure 4.6:** Plot depicting usage of the word “mill” over time.
Figure 4.7: Plot depicting usage of the phrase “milling machine” over time.

More recently developed than milling, waterjet cutting is another established manufacturing technique that is widely used today. This technology was first created to cut paper in 1933, improved to cut plastic in 1956, and ultimately able to cut metals in 1958 [4–6]. As shown in Figures 4.8 and 4.9, uses of the word “waterjet” and the phrase “water jet” do not exhibit any notable trends or periods of heightened interest throughout 1975 to 2021, similar to the previous example on milling. This analysis further demonstrates the tendency for research interest in a topic to peak when it is novel and rapidly evolving, prior to commoditization.
**Figure 4.8:** Plot depicting usage of the word “waterjet” over time.

**Figure 4.9:** Plot depicting usage of the phrase “water jet” over time.
5. Additive Manufacturing Case Study

Analysis of trends in manufacturing has thus far revealed that significant research interest has been placed on additive manufacturing in recent years. It’s natural to wonder whether 3D printing might also become commoditized in the same fashion that computers did. To begin answering this question, it is necessary to compare additive manufacturing with more traditional manufacturing techniques. This case study will compare an example part produced through 3D printing to similar parts created with a waterjet and milling machine.

5.1 Part Design

A simplified L-bracket part was designed in three different ways for manufacture via waterjet, mill, and 3D printing. These manufacturing methods were chosen for their common availability in modern machine shops. Part characteristics were kept similar to simplify comparison of part performance. All parts weigh 12.74 ounces, have length and width of 4 inches, and are made of 4142 steel. This material was chosen for its ability to be 3D printed with Desktop Metal’s Studio extrusion printer. Additionally, the three parts feature a similar level of complexity in design and manufacture, with few features and manufacturing operations. To accommodate fasteners, all parts needed material on the top and vertical faces. However, specific fasteners were not modeled in this study to avoid stress concentrations, so part performance resulting from the overall geometry could be more clearly compared. Figures 5.1, 5.2, and 5.3 depict the various designs for the three parts modeled with Autodesk Fusion CAD software.

![Figure 5.1](image)

**Figure 5.1:** Design of waterjet part. Made of 4142 steel plate with 0.5 inch thickness.
Figure 5.2: Design of milled part. Made of 4142 steel with 0.875 inch thickness. Minimum wall thickness is 0.125 inches.

Figure 5.3: Design of 3D printed part. Made of 4142 steel with total depth of 1 inch. Minimum wall thickness is 0.102 inches.

5.2 Simulating Part Performance

Autodesk Fusion was used to simulate a load of 10,000 N on the top face of all three parts. A fixed constraint was applied to the vertical face of each part to constrain movement. Additionally, the load was applied through a secondary steel block to constrain the displacement of the top face and prevent stress concentrations near hollow regions. Figure 5.4 depicts this setup.
The load simulation allows us to see the stresses in each part under the same load. The results are shown in Figure 5.5 with a constant color scale ranging from 0 to 100 MPa for all parts. Due to the geometric constraints of the simulation, a stress concentration appears at the corner where the top and vertical faces meet. The highest stress in each of the parts is experienced at this stress concentration, with values of 99.45 MPa, 67.61 MPa, and 69.46 MPa for the waterjet, milled, and 3D printed parts, respectively. The waterjet part experienced the highest maximum stress while the milled and 3D print parts performed similarly with the milled part performing slightly better. However, these high stresses from the stress concentration are highly affected by the particular geometric constraints used in this simulation and may not be representative of more realistic connections.
Figure 5.5: Simulated performance of waterjet, milled, and 3D printed parts, respectively, using constant color scale ranging from 0 to 100 MPa.
5.3 Implications

Another key area in which all three parts experienced high stress was the diagonal face. This region is far from the contact point between the block and the part, so the stresses were less affected by the specific geometric constraints used for the simulation. The waterjet part experienced the highest stress in this region of all three parts, experiencing 52 MPa (Figure 5.6). The milled part was in the middle with a highest stress of 41.4 MPa on the diagonal face (Figure 5.7). Finally, the 3D printed part had the lowest maximum stress on this face with a value of 35.6 MPa (Figure 5.8). This low stress on the diagonal face of the 3D printed part was made possible by having material spread out to the regions that were more directly in the load path. These results demonstrate that additive manufacturing can enable novel geometries that improve load distribution and can decrease stresses in a part.

![Image](image.png)

**Figure 5.6:** Zone of high stress far from boundary conditions in waterjet part. Stress reaches a peak of about 52 MPa on the outer diagonal face.
**Figure 5.7:** Zone of high stress far from boundary conditions in milled part. Stress reaches a peak of about 41.4 MPa on the outer diagonal face.

**Figure 5.8:** Zone of high stress far from boundary conditions in 3D printed part. Stress reaches a peak of about 35.6 MPa on the outer diagonal face.
6. Conclusion

From 1975 through 2021, this study has tracked the usage of words in MIT Mechanical Engineering thesis titles to identify trends in research interest. Specifically, it was possible to identify 5 to 10 year bursts of interest in a particular topic when the underlying methods and technologies were novel. Trends in manufacturing techniques were particularly discernable, as there were clear shifts in interest from “computerized manufacturing” to “lean manufacturing” to “additive manufacturing”. Additionally, a counter-intuitive pattern emerged in which the research interest in a topic decreased as a technology became additionally commoditized. For example, the word “computer” experienced the highest use in the late 1970s to early 1980s and then decreased in popularity. This trend demonstrates that the prevalence of a technology in mechanical engineering research is proportional to its novelty and rate of development, not its presence in everyday environments. As this study focused exclusively on MIT theses, this trend may represent the tendency for MIT researchers to focus on highly novel innovations.

Once the historical trends in mechanical engineering research were established, it became possible to look critically at the current topics of interest. Notably, the phrase “additive manufacturing” has seen a rise in interest over the past several years (from 2015-2021). To understand whether this technology might become commoditized and see a decline in research interest, a case study was examined to compare additive manufacturing with more traditional techniques that use a waterjet or milling machine. Design of sample parts and simulation of a load using Autodesk Fusion demonstrated that 3D printing can enable novel geometries with lower internal stresses. As a result of this analysis, it is possible that additive manufacturing may experience broader use and further commoditization in future years.
Bibliography


Appendix

Python Script Written to Collect Data from DSpace@MIT

```python
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
import numpy as np
import pandas as pd
import time

driver = webdriver.Chrome()
website = "https://dspace.mit.edu/discover?field=department&scope=1721.1/7582&filtertype=department&filter_relational_operator=authority&filter=5f65f0e89567129019da97cf3541bb09"
driver.get(website)

df = pd.DataFrame(columns = ['Title', 'Date'])
for i in range(761):
    time.sleep(1)
    for i in range(1,11,1):
        title = driver.find_element_by_xpath("(//div[@class = 'row ds-artifact-item'])["+ str(i)+ "]//h4[@class = 'artifact-title']").text
        print(title)
        raw_date = driver.find_element_by_xpath("(//div[@class = 'row ds-artifact-item'])["+ str(i) + "]//span[@class = 'date']").text
        print(raw_date)
        date = int(''.join(filter(str.isdigit, raw_date)))
        df = df.append({'Title' : title, 'Date' : date}, ignore_index = True)

    df.to_csv("DSpaceProgress.csv")
    next_button = driver.find_element_by_class_name('next-page-link')
    next_button.click()

time.sleep(1)
for i in range(1,8,1):
    title = driver.find_element_by_xpath("(//div[@class = 'row ds-artifact-item'])["+ str(i)+ "]//h4[@class = 'artifact-title']").text
    print(title)
    raw_date = driver.find_element_by_xpath("(//div[@class = 'row ds-artifact-item'])["+ str(i) + "]//span[@class = 'date']").text
    print(raw_date)
    date = int(''.join(filter(str.isdigit, raw_date)))
    df = df.append({'Title' : title, 'Date' : date}, ignore_index = True)

    df.to_csv("DSpaceComplete.csv")
```

Python Script Written to Collect Data from MIT Libraries “Search Our Collections” Website

```python
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
import numpy as np
import pandas as pd
import time

driver = webdriver.Chrome()
website = "https://mit.primo.exlibrisgroup.com/discovery/search?query=creator,contains,mechanical%20engineering,AND&tab=all&search_scope=MIT_theses&sortby=date_d&vid=01MIT_INST:MIT&facet=rtype,exclude,books&facet=lang,include,eng&facet=creator,include,Massachusetts%20Institute%20Of%20Technology%20Department%20Of%20Mechanical%20Engineering&lang=en&mode=advanced&offset=0&came_from=sort"
driver.get(website)

df = pd.DataFrame(columns = ['Title', 'Date'])
for i in range(510):
    time.sleep(5)
    date = 0
    title = driver.find_element_by_xpath("//div[@class='list-item-wrapper first-in-page']//h3[@class='item-title']").text
    type = driver.find_element_by_xpath("//div[@class='list-item-wrapper first-in-page']//div[@class='media-content-type align-self-start']").text
    if type=="THESIS/DISSERTATION":
        raw_date = driver.find_element_by_xpath("//div[@class='list-item-wrapper first-in-page']//span[@data-field-selector='creationdate']").text
        date = (''.join(filter(str.isdigit, raw_date)))
        date = int(date[ : 4])
    df = df.append({'Title' : title, 'Date' : date}, ignore_index = True)
for i in range(1,9,1):
    date = 0
    title = driver.find_element_by_xpath("//div[@class='list-item-wrapper']"+str(i)+")//h3[@class='item-title']").text
    type = driver.find_element_by_xpath("//div[@class='list-item-wrapper']"+str(i)+")//div[@class='media-content-type align-self-start']").text
    if type=="THESIS/DISSERTATION":
        raw_date = driver.find_element_by_xpath("//div[@class='list-item-wrapper']"+str(i)+")//span[@data-field-selector='creationdate']").text
        date = (''.join(filter(str.isdigit, raw_date)))
        date = int(date[ : 4])
```
df = df.append({'Title': title, 'Date': date}, ignore_index = True)

date = 0
title = driver.find_element_by_xpath("//div[@class='list-item-wrapper last-item']/h3[@class='item-title']").text
type = driver.find_element_by_xpath("//div[@class='list-item-wrapper last-item']//div[@class='media-content-type align-self-start']").text

if type="THESIS/DISSERTATION":
    raw_date = driver.find_element_by_xpath("//div[@class='list-item-wrapper last-item']//span[@data-field-selector='creationdate']").text
    date = (''.join(filter(str.isdigit, raw_date))
    date = int(date[ : 4])

df = df.append({'Title': title, 'Date': date}, ignore_index = True)
df.to_csv("TestLibThesesProgress.csv")

next_button = driver.find_element_by_class_name("counter-nav.counter-next")
next_button.click()

time.sleep(5)
date = 0
title = driver.find_element_by_xpath("//div[@class='list-item-wrapper first-in-page']/h3[@class='item-title']").text
type = driver.find_element_by_xpath("//div[@class='list-item-wrapper first-in-page']//div[@class='media-content-type align-self-start']").text

if type="THESIS/DISSERTATION":
    raw_date = driver.find_element_by_xpath("//div[@class='list-item-wrapper first-in-page']//span[@data-field-selector='creationdate']").text
    date = (''.join(filter(str.isdigit, raw_date))
    date = int(date[ : 4])

df = df.append({'Title': title, 'Date': date}, ignore_index = True)

for i in range(1,9,1):
    date = 0
    title = driver.find_element_by_xpath("//div[@class='list-item-wrapper']"+%str(i)+"//h3[@class='item-title']").text
    type = driver.find_element_by_xpath("//div[@class='list-item-wrapper']"+%str(i)+"//div[@class='media-content-type align-self-start']").text

    if type="THESIS/DISSERTATION":
        raw_date = driver.find_element_by_xpath("//div[@class='list-item-wrapper']"+%str(i)+"//span[@data-field-selector='creationdate']").text
        date = (''.join(filter(str.isdigit, raw_date))
        date = int(date[ : 4])

        df = df.append({'Title': title, 'Date': date}, ignore_index = True)
```python
date = 0
title = driver.find_element_by_xpath("//div[@class='list-item-wrapper last-item']/h3[@class='item-title']/text")
type = driver.find_element_by_xpath("//div[@class='list-item-wrapper last-item']/div[@class='media-content-type align-self-start']/text")

if type=="THESIS/DISSERTATION":
    raw_date = driver.find_element_by_xpath("//div[@class='list-item-wrapper last-item']/span[@data-field-selector='creationdate']/text")
    date = (''.join(filter(str.isdigit, raw_date)))
    date = int(date[ : 4])
df = df.append({'Title' : title, 'Date' : date}, ignore_index = True)

df.to_csv("TestLibThesesComplete.csv")
print(df)
```

---

**R Code Written to Process Text Data**

```r
# Text mining packages
library(tm)
library(tidyverse)
library(tidytext)
library(wordcloud)

# Combining Data

old_theses <- read.csv("/Users/sleon/Documents/MITStuff/Thesis/ThesisCode/AllLibThesesProgress.csv")
dspace_theses$Date <- substr(dspace_theses$Date, 1, 4) %>% strtoi()

all_theses <- rbind(old_theses, new_theses,dspace_theses)
all_theses$Title <- str_trim(all_theses$Title)
all_theses$Title <- str_replace_all(all_theses$Title, "([.])", "")
all_theses$Title <- tolower(all_theses$Title)

all_theses <- all_theses[!duplicated(all_theses$Title), ]
all_theses_words <- all_theses %>% unnest_tokens(word,Title) %>% anti_join(stop_words)
```
```
```r Counting
all_theses_words %>% count(word,sort=TRUE)
all_theses_words %>% count(word,sort=TRUE) %>% with(wordcloud(words=word, n, max.words = 100))
all_theses_words %>% count(word,Date,sort=TRUE)
thesis_count <- all_theses %>% count(Date)
thesis_count <- thesis_count[thesis_count$Date > 1900, ]
ggplot(thesis_count, aes(x=Date, y=n)) + geom_bar(stat="identity") + xlab("Year") + ylab("Number of Available Theses") + theme(text = element_text(size = 18))
```

### Tokenizing Functions
```
```{r}
getwords1 <- function() {
  thesis <- all_theses
  tidytext <- thesis %>% unnest_tokens(word,Title) %>% anti_join(stop_words)
  return(tidytext)
}

getwords2 <- function() {
  thesis <- all_theses
  tidytext <- thesis %>% unnest_tokens(ngram,Title,token="ngrams",n=2) %>% separate(ngram,c("word1","word2"),sep=" ") %>%
  filter(!(word1 %in% stop_words$word) & !(word2 %in% stop_words$word))%>%
  unite(ngram, word1, word2, sep=" ")
  return(tidytext)
}
```

### Grouping
```
```{r}
thesis1 <- getwords1()
word_count <- thesis1 %>% count(word,sort=TRUE)
word_year_count <- thesis1 %>% count(Date,sort=TRUE)
colnames(word_year_count) <- c('Date','NumWords')
raw_grouped_words <- thesis1 %>% count(Date,word,sort=TRUE)
grouped_words <- merge(word_year_count, raw_grouped_words , by = "Date")
```
grouped_words$Fraction <- grouped_words$n / grouped_words$NumWords

thesis2 <- getwords2()
phrase_count <- thesis2 %>% count(ngram, sort=TRUE)
year_count <- thesis2 %>% count(Date, sort=TRUE)
colnames(year_count) <- c('Date', 'NumPhrases')
raw_grouped_phrases <- thesis2 %>% count(Date, ngram, sort=TRUE)
grouped_phrases <- merge(year_count, raw_grouped_phrases, by = 'Date')
grouped_phrases$Fraction <- grouped_phrases$n / grouped_phrases$NumPhrases
...

## Plotting

```r
plotWordFraction <- function(targWord) {
  target_word <- grouped_words[grouped_words$word == targWord,]
  target_word <- target_word[target_word$Date >= 1975, ]
  target_word <- target_word[target_word$Date <= 2021, ]
  ggplot(target_word, aes(x=Date, y=Fraction)) + geom_bar(stat="identity") + xlab("Year") + ylab(paste("Fraction of Words which Match", "n\n", targWord,"\n")) + theme(text = element_text(size = 18))
}

plotPhraseFraction <- function(phrase) {
  target_phrase <- grouped_phrases[grouped_phrases$ngram == phrase,]
  target_phrase <- target_phrase[target_phrase$Date >= 1975, ]
  target_phrase <- target_phrase[target_phrase$Date <= 2021, ]
  ggplot(target_phrase, aes(x=Date, y=Fraction)) + geom_bar(stat="identity") + xlab("Year") + ylab(paste("Fraction of Phrases which Match", "n\n", phrase,"\n")) + theme(text = element_text(size = 18))
}

plotWordStemFraction <- function(targStem) {
  target_word <- grouped_words[grepl(targStem, grouped_words$word, fixed = TRUE),]
  target_word <- target_word[target_word$Date >= 1975, ]
  target_word <- target_word[target_word$Date <= 2021, ]
  ag_target_word <- aggregate(target_word$Fraction, list(target_word$Date), FUN=sum)
}
ggplot(ag_target_word, aes(x=Group.1, y=x)) + geom_bar(stat="identity") + xlab("Year") + ylab(paste("Fraction of Words which Contain", "n", targStem, "n")) + theme(text = element_text(size = 18))

plotPhraseStemFraction <- function(targStem1, targStem2) {
  target_phrase <- grouped_phrases[grepl(targStem1, grouped_phrases$ngram, fixed = TRUE) & grepl(targStem2, grouped_phrases$ngram, fixed = TRUE),]
  target_phrase <- target_phrase[target_phrase$Date >= 1975,]
  target_phrase <- target_phrase[target_phrase$Date <= 2021,]
  ag_target_phrase<- aggregate(target_phrase$Fraction, list(target_phrase$Date), FUN=sum)
  ggplot(ag_target_phrase, aes(x=Group.1, y=x)) + geom_bar(stat="identity") + xlab("Year") + ylab(paste("Fraction of Phrases with", "n", targStem1, "n" and "n", targStem2, "n")) + theme(text = element_text(size = 18))
}

##Test Plot
```{r}
plotWordFraction('computer')

plotWordStemFraction('comput')

plotPhraseFraction('additive manufacturing')```