

國立臺灣大學管理學院資訊管理研究所

碩士論文

Graduate Institute of Information Management

College of Management

National Taiwan University

Master Thesis

MAD EDGAR: 10-K 報表變動比對與檢視系統

Making A Difference for SEC EDGAR (MAD EDGAR):
Highlighting 10-K Year-over-Year Differences for Better
Awareness

余俊廷

Chun-Ting Yu

指導教授: 盧信銘 博士

Advisor: Hsin-Min Lu Ph.D.

中華民國 111 年 8 月

August, 2022






Acknowledgements

即將完成的碩士論文也代表著二十多年來的學生生涯總算要告一段落，回顧在臺大資管的兩年當中，撇除疫情之外算是非常幸運的不論是在課程、研究或是實習的過程中都受到了很多人的幫助。

首先，我要感謝我的指導教授—盧信銘博士，感謝盧老師在兩年前願意收留我進入到老師的實驗室，也感謝老師給了我一個如此適合我的研究主題，從最一開始的系統發想、實作到後續的實驗設計與分析，整個過程提供了我很多資源以及自由度，同時也會在我遇到瓶頸時給予我協助，讓我能夠順利的完成系統以及此篇論文。也感謝洪茂蔚口委以及畢南怡口委，有了口委們提供的許多寶貴建議讓此篇論文能更臻完善。

感謝前來支援實驗的朋友—岱妤、旻桓、彥翔、君儒、貴晴、哲豪、漢威、維辰、彥銘、芷伶、立勤、延東、雨澤、宏穎、雅汝、柏辰、煥勳、亭遠、富纖以及正宇，還有實驗室的夥伴—雅汝、煥勳、聖典、君儒以及富纖學姊，沒有你們我同樣也無法完成此篇論文，知道有你們跟我一樣在為同樣的目標努力使我能夠保持研究的動力，也因為有你們讓我的碩士生活過得比想像中還要滋潤跟充實。



感謝一直在身邊的家人還有岱好，從三年前準備考研到現在即將要畢業，一路陪我走過所有的跌宕起伏，謝謝你們。最後，也要感謝自己算是沒有辜負入學前對自己的期許，在臺大的兩年的確增廣見聞了不少，希望未來在開始工作後能夠不要忘記現在寫程式帶給我的這份快樂跟成就感。謹以此篇論文獻給我的家人、師長以及朋友們，並在此獻上我最高的謝意。

余俊廷 謹誌

于國立臺灣大學資訊管理研究所

中華民國 111 年 8 月



摘要

在經歷 1930 年代經濟大蕭條後，美國國會為了確保市場的公平、透明及穩定性而制定與通過了 1934 年的證券交易法，該法案的出現賦予了美國證券交易委員會 (SEC) 管理監督與標準制定的權限。依據法條中的申報規定，資產在 1000 萬美元以上且股東人數超過 500 人的公司需提供年度報告及其他定期財務報告 (主要為 Form 10-K、10-Q 及 8-K) 給 SEC，而這些報告皆可於 SEC 提供的資料庫 EDGAR 中查詢。然而，受限於 SEC EDGAR 本身提供的搜尋功能只有簡單的財報搜尋，對於想要取得更深入資訊的使用者來說沒辦法輕易的達成，過去也有研究提到許多重要的資訊往往是分散在多個財報之中，例如說是希望了解一份財務報告相較去年時期的改變。在這個研究之中，我們首先基於 SEC EDGAR 的系統重新設計與實作了一套新的資訊系統 MAD EDGAR 輔助使用者閱讀 10-K 年報，我們的系統將年報中兩年間的變化以紅色及綠色的醒目提示呈現在報告之中以輔助使用者閱讀。其次，為了系統成效的評估我們也設計了兩套實驗與任務並招募了 20 名受測者實際操作系統。我們的研究結果顯示，當任務涉及到跨年報的資訊時，MAD EDGAR 在回答問題的效率上確實表現的比 SEC EDGAR 更加良好。

關鍵字：美國證券交易委員會、EDGAR、財務報表、10-K 年度報告、變化偵測



Abstract

This paper addresses the problem of SEC EDGAR's limitations in identifying year-over-year differences among financial reports. We presented Making A Difference for SEC EDGAR (MAD EDGAR), a new web-based information system that facilitates the analysis of year-over-year modifications in 10-K reports. MAD EDGAR highlights the differences between 10-K reports to help investors efficiently comprehend the modifications in documents. The year-over-year differences in a 10-K report are presented in a colorized format, where the color green stands for new statements added in the current year of the report, while the color red stands for deletion from the previous year on the contrary. Twenty graduate students from four universities were recruited and observed while completing two types of tasks specifically designed to evaluate MAD EDGAR's usefulness during the experiment. The results indicated that our system out-performed SEC EDGAR in terms of identifying year-over-year changes.

Keywords: SEC, EDGAR, Financial Report, 10-K Report, Change Detection



Contents

	Page
Acknowledgements	i
摘要	iii
Abstract	iv
Contents	v
List of Figures	viii
List of Tables	x
Chapter 1 Introduction	1
1.1 Background	1
1.2 Motivation	4
1.3 Objectives	7
Chapter 2 Literature Review	8
2.1 SEC EDGAR	8
2.1.1 XBRL	10
2.1.2 EDGAR System Usage Data	11
2.2 Text Analysis in Financial Reports	12



2.2.1	Tone	13
2.2.2	Readability	13
2.2.3	MD&A Modifications	14
2.3	Information System for Financial Information	15
Chapter 3	Research Gaps and Research Questions	17
3.1	Research Gaps	17
3.2	Research Questions	17
Chapter 4	System Design and Implementation	18
4.1	System Requirements	18
4.2	User Interface Design of Functionality	21
4.3	System Architecture	26
4.3.1	Data Tier	27
4.3.1.1	Master Server	28
4.3.1.2	Task Queue Server	28
4.3.1.3	Worker Servers	29
4.3.1.4	The Differencing Algorithm	29
4.3.2	Application Tier	32
4.3.3	Presentation Tier	33
Chapter 5	Experimental Design	35
5.1	Setup	35
5.1.1	Participants	35
5.1.2	Systems	36

5.1.3	Technical Environment	36
5.1.4	Tasks	36
5.1.5	Hypothesis Settings and Expected Experimental Results	39
5.2	Experimental Procedure	40
5.2.1	Data Collection	40
5.2.2	Coding and Scoring	41
5.2.3	Data Analysis	41
Chapter 6	Result and Discussion	43
6.1	Outliers	43
6.2	Summary Statistics	46
6.3	Analysis of Variance (ANOVA)	48
6.3.1	Result	48
6.3.1.1	First Task	48
6.3.1.2	Second Task	49
6.3.2	Tests of Normality	51
6.3.3	Homogeneity of Variances	52
6.4	Conclusions	53
	References	55
	Appendix A — Data Collection Form	61

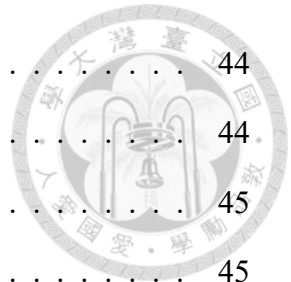




List of Figures

1.1	Screenshot of SEC EDGAR	2
1.2	10-K Report Sections	5
4.1	Sample Diff Output in Git	19
4.2	Sample Diff Output in Microsoft Word	19
4.3	Sample Diff-Highlighted Report in MAD EDGAR	20
4.4	Screenshot of the Main Page	22
4.5	Screenshot of the Search Tool	23
4.6	Screenshot of the Report Viewer	24
4.7	Screenshot of a Sample Report	25
4.8	Overall Architecture in MAD EDGAR	26
4.9	Data Tier in MAD EDGAR	27
4.10	Differencing Algorithm in Action	30
4.11	Application Tier in MAD EDGAR	32
4.12	Endpoints in the Application Tier	33
4.13	Presentation Tier in MAD EDGAR	33
5.1	Example Question in the First Task	37
5.2	Example Question in the Second Task	37
5.3	Counterbalanced Design	38
5.4	Screenshot of Data Collection Form Prepared using Google Form	39
5.5	Screenshot of Experiment Result	41

6.1	Box-Plot from the First Task using SEC EDGAR	44
6.2	Box-Plot from the First Task using MAD EDGAR	44
6.3	Box-Plot from the Second Task using SEC EDGAR	45
6.4	Box-Plot from the Second Task using MAD EDGAR	45
6.5	Profile Plot of the Mean from the First Task	46
6.6	Profile Plot of the Mean from the Second Task	48
A.7	Data Collection Form (P.1)	62
A.8	Data Collection Form (P.2)	63
A.9	Data Collection Form (P.3)	64
A.10	Data Collection Form (P.4)	65
A.11	Data Collection Form (P.5)	66
A.12	Data Collection Form (P.6)	67
A.13	Data Collection Form (P.7)	67
A.14	Data Collection Form (P.8)	68
A.15	Data Collection Form (P.9)	69
A.16	Data Collection Form (P.10)	70
A.17	Data Collection Form (P.11)	71
A.18	Data Collection Form (P.12)	72
A.19	Data Collection Form (P.13)	73
A.20	Data Collection Form (P.14)	74
A.21	Data Collection Form (P.15)	75





List of Tables

1.1	Top 20 SEC EDGAR Filing Type between January 1993 to March 2022	3
4.1	System Function Comparison	21
4.2	Differences Table between Both Paragraphs	31
6.1	Descriptive Statistics	47
6.2	ANOVA Table from the First Task	49
6.3	ANOVA Table from the Second Task	50
6.4	Estimate Table from the Second Task	50
6.5	Test of Normality from the First Task	51
6.6	Test of Normality from the Second Task	52
6.7	Levene's Test for Equality of Variances	52
6.8	Levene's Test from the First Task	53



Chapter 1 Introduction

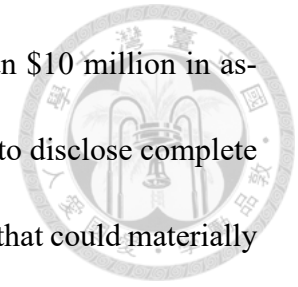
1.1 Background

In the effort to resolve the Great Depression back in the mid-1930s, U.S. President Roosevelt made it clear that more securities regulation was needed and promised to deliver economic reform. Roosevelt sought to bring back public confidence in the securities markets and was convinced that truthful and full disclosure was essential to this goal.

To protect investors, the United States Congress crafted a mandatory disclosure process designed to force companies to disclose information that investors would find pertinent when making investment decisions. The Securities and Exchange Act of 1934 (Exchange Act) was created to ensure an environment of fairness, financial transparency, and accuracy.

The Securities and Exchange Commission (SEC) was then established as a federal administrative agency tasked with monitoring markets, enforcing securities laws, and developing new regulations. Since its inception in the mid-1930s, the primary mission of the SEC has been to protect investors and maintain the integrity of securities markets. As

part of this effort, domestic, publicly held companies with more than \$10 million in assets whose securities are held by more than 500 owners are required to disclose complete and accurate information about their operations, as well as any event that could materially impact them.



The SEC has developed a searchable online database known as SEC Electronic Data Gathering, Analysis and Retrieval (SEC EDGAR) system, which companies are required to use to file reports, forms, and other information required by the SEC. Users can search the database presented in Figure 1.1 similarly to an online search portal, using either the name of a company or individual. Search results can be further narrowed down by date, location of the company’s executive offices, or the type of file sought.

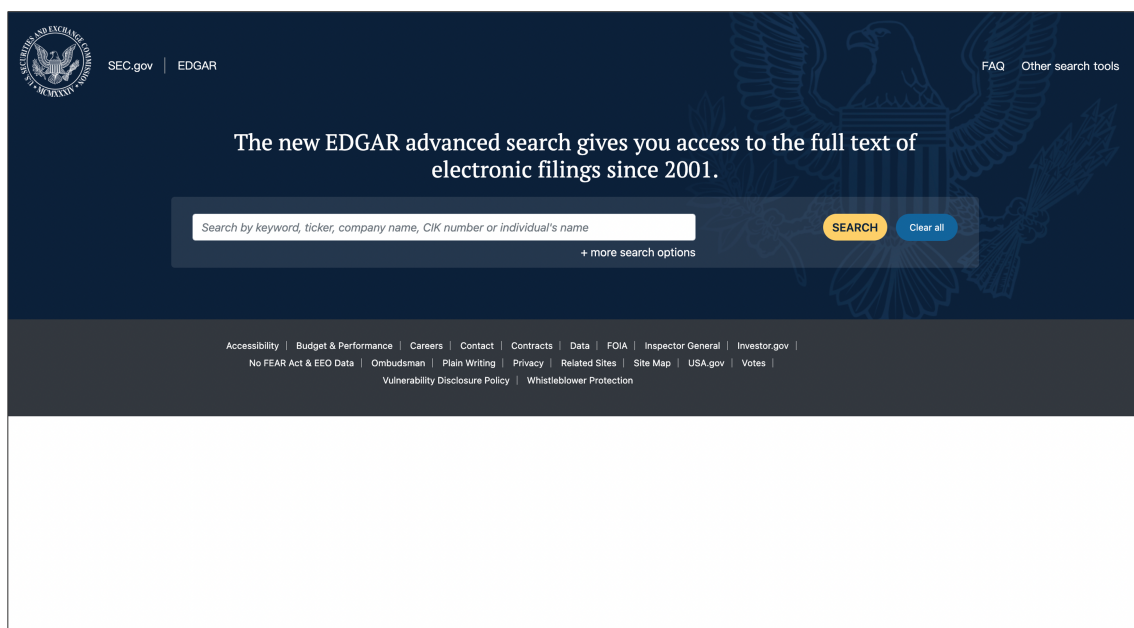


Figure 1.1: Screenshot of SEC EDGAR

SEC EDGAR is a valuable resource for recent research. It provides a web-based interface to all companies’ reports filed with the SEC since 1993. Reports can be downloaded by investors or researchers for free and in masses. As it can be seen from Table 1.1,

there were over 14,986,862 filings stored in the SEC EDGAR between January 1993 to March 2022.



Filing Type	Count
4	3,883,749
8-K	1,726,665
10-Q	613,076
6-K	448,612
SC 13G/A	392,377
497	382,755
3	346,059
424B2	342,194
D	318,706
13F-HR	312,002
497K	241,058
SC 13G	214,005
D/A	203,828
10-K	202,170
485BPOS	190,096
24F-2NT	187,042
FWP	181,410
CORRESP	180,720
DEF 14A	172,294
UPLOAD	171,079
Total	14,986,862

Table 1.1: Top 20 SEC EDGAR Filing Type between January 1993 to March 2022

The company's periodic documents such as annual (10-K) and quarterly (10-Q) reports receive the most attention from investors. Although a less timely than other reports such as 10-Q reports, 10-K reports still contain comprehensive information on past company achievements, therefore facilitating the confirmation of readers' investment decisions about a company.

1.2 Motivation



A 10-K report is a comprehensive document filed annually by a publicly-traded company about its financial performance and is required by the SEC. Some of the information about a company is required to disclose in the 10-K report such as its history, organizational structure, financial statements, earnings per share, subsidiaries, executive compensation, and any other relevant data. Figure 1.2 presents a typical 10-K report from Alphabet Inc., which along with other include at least the following four distinct sections:

1. Business: This section gives a general summary of the company's primary business, including its products and services.
2. Risk factors: This section lists all risks that the company is currently facing or may encounter in the future.
3. Management's discussion and analysis: Also known as MD&A, this section gives the company an opportunity to explain its business results from the previous fiscal year.
4. Financial statements and supplementary data: This section includes the company's audited financial statements including the income statement, balance sheets, and statement of cash flows.

Alphabet Inc. Form 10-K For the Fiscal Year Ended December 31, 2021		Alphabet Inc.
TABLE OF CONTENTS		Page
Table of Contents		3
Note About Forward-Looking Statements		3
PART I		
Item 1.	Business	4
Item 1A.	Risk Factors	10
Item 1B.	Unresolved Staff Comments	24
Item 2.	Properties	24
Item 3.	Legal Proceedings	24
Item 4.	Mine Safety Disclosures	24
PART II		
Item 5.	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities [Reserved]	25
Item 6.	Management's Discussion and Analysis of Financial Condition and Results of Operations	27
Item 7.	Quantitative and Qualitative Disclosures About Market Risk	28
Item 7A.	Financial Statements and Supplementary Data	42
Item 8.	Changes in and Disagreements With Accountants on Accounting and Financial Disclosure	45
Item 9A.	Controls and Procedures	86
Item 9B.	Other Information	86
Item 9C.	Disclosure Regarding Foreign Jurisdictions that Prevent Inspections	87
PART III		
Item 10.	Directors, Executive Officers and Corporate Governance	88
Item 11.	Executive Compensation	88
Item 12.	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters	88
Item 13.	Certain Relationships and Related Transactions, and Director Independence	88
Item 14.	Principal Accountant Fees and Services	88
PART IV		
Item 15.	Exhibits, Financial Statement Schedules	89
Item 16.	Form 10-K Summary	92
Signatures		92

Figure 1.2: 10-K Report Sections

10-K reports are fairly long and complicated due to the amount of the information they contain. *Ravula (2021)* found out that in order to reduce the effort while remaining compliant with regulatory requirements, companies have an incentive to use their own template when preparing their 10-K reports, which often makes 10-K reports even more lengthy and complex to analyze overall.

Furthermore, over the last twenty years, researchers around the world in many disciplines have leveraged SEC EDGAR as their primary data source getting financial reports to answer many important research questions (*Bommarito et al., 2018*). However, limited by the function of SEC EDGAR search tools, it is not easy for end-users to extract information from multiple reports using a single query. Knowledge in SEC EDGAR is usually distributed in different documents. It could be a report or a statement, but information such as management decision modifications or year-over-year changes is not straightforward

to get out without labor power (*Han et al.*, 2016).

Cohen et al. (2020) also constructed a panel dataset of 10-K report downloading activities from SEC EDGAR and correspondingly identified that a large percentage of investors not only downloaded the current year's 10-K report, but also the prior year's 10-K report in tandem. In order to analyze the lengthy 10-K reports and extract the distributed knowledge such as year-over-year changes in them, investors have to identify them by comparing two successive reports word-by-word.

Yet, 10-K reports are lengthy, which may restrict investors' ability to fully process and understand the information. Furthermore, the number of 10-K reports had also exceeded 202,170 as of March 2022. With over 6,000 new 10-K reports each year, investors have reached their limits in exploring this dataset without the assistance of computer systems.

Moreover, individual investors may not have the same advantage as a fund manager or analyst, who looks at hundreds of financial reports year after year. Individual investors need the ability to extract reliable information and knowledge from financial reports with minimum manual effort.



1.3 Objectives



To address the above issues and SEC EDGAR's limitations in extracting distributed knowledge among reports, we present Making A Difference for SEC EDGAR (MAD EDGAR), a system that facilitates the analysis of financial reports by highlighting the year-over-year differences in 10-K reports through an integrated web interface.



Chapter 2 Literature Review

While there are information sources other than SEC EDGAR to acquire financial reports, the SEC EDGAR still represents one of the primary data sources for individual investors doing fundamental research on security evaluation (*Ravula, 2021*). This review covers the previous work in:

Section 2.1 SEC EDGAR: How the SEC have attempted to enhance SEC EDGAR's information extractability and how do the analysts and researchers use SEC EDGAR?

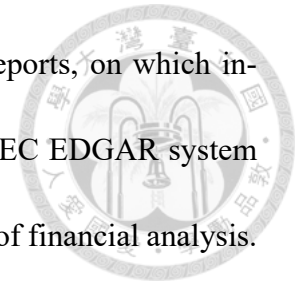
Section 2.2 Text Analysis in Financial Reports: How the researchers have attempted to extract information from financial reports using text analysis?

Section 2.3 Information System for Financial Information: How the researchers have proposed information systems specifically designed for financial data analysis?

2.1 SEC EDGAR

To enhance and facilitate market information flow, the SEC began an initiative to develop an electronic disclosure system in 1983, which was later made available to the

public in 1992, currently known as SEC EDGAR. SEC EDGAR reports, on which investors rely, have been shown to improve market efficiency. The SEC EDGAR system and the reports it keeps play a significant part in the present research of financial analysis.



Market reaction to SEC EDGAR filings was studied, among others, by *You and Zhang* (2007, 2011). *You and Zhang* (2007) established that investors use information from 10-K and earnings announcements differently altogether, and *You and Zhang* (2011) discovered that investors under-react to 10-K information.

The negative effect of the delay in SEC EDGAR filings was also another popular research question. *Duarte-Silva et al.* (2013) studied the market reaction to earnings delay announcements and concluded that these delays provide a signal of financial performance deterioration.

Brown et al. (2020) use a machine learning technique to assess whether the 10-K narratives from SEC EDGAR are incrementally informative in predicting intentional misreporting, while *McMullin et al.* (2018) use 8-K reports from SEC EDGAR to show that price formation can be enhanced by the increase of mandated disclosures.

Cannon et al. (2019) create a textual measure of corporate social responsibility (CSR) disclosure that aggregates CSR keywords found in 10-K reports and conclude that CSR measurement provides information about companies' competitive advantages. There are other numerous studies also based on the analysis of SEC reports (*Ege et al.*, 2019; *Hasan*, 2018; *Kim et al.*, 2019; *Lopatta et al.*, 2017; *Mitra et al.*, 2019).



2.1.1 XBRL

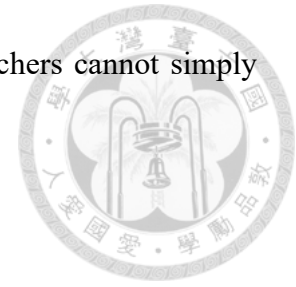
To enhance the information extractability and facilitate the analysis of investors and researchers, the SEC adopted the eXtensible Business Reporting Language (XBRL) in SEC EDGAR and mandated that all companies use XBRL by 2014.

XBRL is essentially a dictionary of tags that can be applied to each element in a financial report. This enables software to identify what the item represents and how it relates to others. This also enables the gathering of accounting numbers to be automatic, which provides a great opportunity for financial analysis (*Henselmann et al.*, 2014).

Numerous researchers had studied the effect of mandatory XBRL disclosure across various aspects of financial information, such as market efficiency and price discovery. *Kim et al.* (2012) suggested that mandatory XBRL filings may reduce information risk and improve information efficiency. *Efendi et al.* (2014) found a decline in post-earnings announcement drift for good news in the post-XBRL adoption period, while *Yen and Wang* (2015) discovered that the adoption of XBRL is positively related to market reactions in terms of revenue shocks and their research may alleviate companies' concerns regarding the benefits of adopting XBRL.

The XBRL intends to streamline the analysis of financial reports by providing a standard dictionary for collecting and analyzing financial information. However, considering that other than those quantitative information that can be categorized and tagged such as accounting numbers, there is significantly more qualitative information presented in the

form of unstructured text in financial reports. Analysts and researchers cannot simply extract qualitative information using the traditional XBRL way.



2.1.2 EDGAR System Usage Data

In order to have a better understanding of how investors use SEC EDGAR, the SEC had also started releasing the EDGAR Log File Dataset to the public, it captured investors' access to individual filings, alongside meta-data about the filing that is being accessed.

Currently, the EDGAR Log File Dataset covers the period from February 2003 through June 2017. Each record in the EDGAR Log File Dataset contains information about the user's partially anonymized unique Internet Protocol (IP) address, timestamp, company (identified by the Central Index Key (CIK)), and specific filing type (identified by the unique SEC accession number). Since the release of the EDGAR Log File Dataset, researchers have been studying investors' accessing patterns in financial reports.

Drake et al. (2012) found out that investors access mandatory financial filings during news release periods and the demand increases during times of negative news and increased uncertainty about the firm's business. *Lee et al.* (2015) identified economically related peer firms by analyzing the co-searches of companies by the same users on SEC EDGAR.

Gandhi et al. (2019) studied the consumption of financial information in filings by analyzing the distribution of daily filing requests. *Cohen et al.* (2020) also identified that a large percentage of investors not only downloaded the current year's 10-K, but also the

prior year's 10-K in tandem.

Previous studies indicated that investors and analysts access mandatory disclosures through SEC EDGAR and request millions of filings from SEC EDGAR each week. However, considering that historical filings are frequently accessed and that companies file two million disclosures every year, an automated extraction mechanism is needed for timely information discovery and knowledge dissemination.

2.2 Text Analysis in Financial Reports

Financial reports such as 10-K and 10-Q are rich with unstructured text information and provide a comprehensive insight into the future of the company. However, due to its enormous volume and unstructured nature, humans are not able to analyze it thoroughly without the assistance of the computer. In recent years, text analysis in financial reports has seen a dramatic increase in attention, *Ravula* (2021) covered the previous work in unstructured data analysis in Financial and Accounting and highlighted the limitations of the current focus. In this paper, we provide a short summary that classified this research area by the following three aspects:

1. Tone
2. Readability
3. MD&A Modifications



2.2.1 Tone

Investors' judgment is often influenced by forward-looking disclosures (*Allee and DeAngelis, 2014*), therefore, the manipulation of tone in text content has the potential to mislead the analysts and provide a distorted view of the company to investors. The tone analysis is used to extract the tone of the content in corporate disclosure and may reveal useful information about the managers' intentions.

Feldman et al. (2009) measured the changes in financial reports' tone and examined the information content of tone change. They found out that short window market reactions around the SEC filing are associated with the MD&A section's tone.

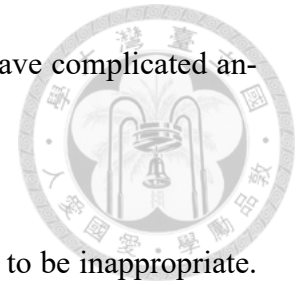
Loughran and McDonald (2010) introduced a sentiment word list for financial data using 10-K from 1994 to 2008. They discovered that their negative word lists are positively related to abnormal trading volume.

2.2.2 Readability

Another important construct in the emerging financial reports on text analysis is the notion of readability. Readability attempts to measure the ease of understanding a text content for reports' readers such as investors.

The first study to examine the relation between readability and the company's performance for a sufficient sample is *Li (2008)*. *Li (2008)* measured annual reports' readability using the Fog Index and researched the relation with the company's subsequent perfor-

mance. They concluded that poor performance companies tend to have complicated annual reports.



However, the use of the widely popularized Fog Index is likely to be inappropriate.

Loughran and McDonald (2010) reported that the file size of 10-K reports as a readability proxy that outperforms the Fog Index. They proposed using the file size as an easier and better calculation for financial report readability.

2.2.3 MD&A Modifications

Management Discussion and Analysis (MD&A) is a specific part of a company's overall financial disclosure. As MD&A section often contains new and useful information about the company, researchers and investors use it for financial analysis purposes. In recent years, more and more studies focus on the aspect of MD&A modifications, which are the text changes of MD&A section in 10-K reports between the current year and the previous year.

Brown et al. (2020) studied year-over-year changes in MD&A section and introduced a measure for disclosure to represent the modification score. They found that companies with larger financial performance changes (e.g., changes in Earning Per Share) modify the MD&A section more and that the modification scores have declined in the past declined even as MD&A sections have become longer.

Cohen et al. (2020) found out that when companies break their routines when phrasing and preparing their MD&A, this action contains rich, important information for future

companies' outcomes. They concluded that observing changes in MD&A section yields a powerful and robust indicator of future company performance.



2.3 Information System for Financial Information

As the volume of financial reports continuously increases, the development of new information systems that may help investors on accessing, searching, filtering, and understanding these documents has also begun to gain attention in recent years.

The FinIR 2020 workshop was the first to bring together a diverse set of researchers interested in exploring IR technology in finance (*Feng et al.*, 2020). The FinWeb 2021 workshop further explores the usefulness of information on the Web for financial technology. Other related works proposed several information search systems specifically designed for financial reports.

Liu et al. (2016) introduced a web-based information system, FIN10K, to retrieve relevant financial reports and visualize the analyzed results. It visualizes high and low-risk words learned via the ranking models, as well as the syntactic and contextual information among financial sentiment words.

Plachouras et al. (2016) introduced a system for users to search financial data using both keyword and natural language queries. After training model on the World Bank macro-economic indicator data, their system answers the queries with an automatically generated textual description using Natural Language Generation (NLG).

Liu et al. (2018) presented RiskFinder, a web-based information system that provides the highlight of the high-risk sentences in the MD&A section. They demonstrated the importance of sentence-level analysis and the integration of soft and hard information in finance.

Du et al. (2019) presented a financial risk information detecting and analyzing system integrating with multiple NLP models trained on financial reports. They provided two types of user interfaces: one for the assessment of a single report, and the other for an overview of a given company. Their system is aimed at helping users efficiently comprehend financial sentiment and risk delivered by financial reports.



Chapter 3 Research Gaps and Research Questions

3.1 Research Gaps

Previously proposed systems have tended to focus on enabling investors to efficiently comprehend financial reports by supporting visual assistance in a specific part of the documents such as MD&A section. Furthermore, to our best knowledge, few prior studies take year-over-year changes into account when designing their systems. Complementary to previous works, we focus on developing a novel system that facilitates the analysis of year-over-year differences among financial reports in a complete document aspect.

3.2 Research Questions

1. How can we design a system to facilitate the analysis of 10-K reports with regard to identifying year-over-year changes?
2. How can we evaluate the usefulness of our system compared to SEC EDGAR?



Chapter 4 System Design and Implementation

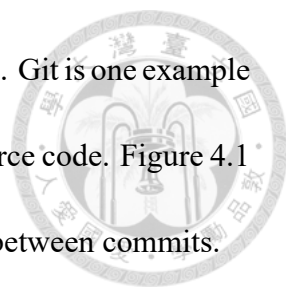
As mentioned in the previous chapters, the objective of this research is to develop a novel system, MAD EDGAR, to facilitate the analysis of the financial reports by highlighting the year-over-year differences in 10-K filings.

With an integrated web interface, investors can better identify what has been modified in the current year's financial report without needing to compare it against the previous one. In this chapter, we will discuss the system's requirements, design of functionality, and system's architecture detail.

4.1 System Requirements

To facilitate the analysis of year-over-year differences in 10-K reports, we explore the idea of providing visual support in the form of highlighting textual modifications between two successive years in this study.

A differencing algorithm outputs a set of differences between two inputs. These



algorithms are the basis of a number of commonly used developer tools. Git is one example where a developer can read, commit, pull, and merge diffs in their source code. Figure 4.1 presents a sample output in Git that shows changes in a source code between commits.

```
~/playground/mad-edgar-lab/master-server/web/edgar-web-app on master !5 ..... with kevinyu@baeir at 01:40:37 [3/31]
> git diff
diff --git a/src/components/FilingDisplay/FilingDisplay.js b/src/components/FilingDisplay/FilingDisplay.js
index 2d37bc0..9c74508 100644
--- a/src/components/FilingDisplay/FilingDisplay.js
+++ b/src/components/FilingDisplay/FilingDisplay.js
@@ -1,5 +1,5 @@
import React, { useState, useEffect } from 'react';
-import { Modal } from 'antd';
+import { Modal, PageHeader, Button } from 'antd';
import './FilingDisplay.css';

export const FilingDisplay = ({ data, setData }) => {
@@ -16,14 +16,22 @@ export const FilingDisplay = ({ data, setData }) => {
  return (
    <div className="FilingDisplay">
      <Modal
-        title={{(data === null) ? "" : "[" + data.company + "]" + data.filing_name}}
        centered
        visible={state.visible}
        onOk={() => setData(null)}
        onCancel={() => setData(null)}
        width={1400}>
+        <PageHeader
+          className="site-page-header"
+          title={{(data === null) ? "" : "[" + data.company + "]" 10-K differences}}

```

Figure 4.1: Sample Diff Output in Git

Microsoft also created a built-in feature that allows users to compare two Microsoft Word documents using such algorithms. Figure 4.2 presents a sample output that shows changes in two Microsoft Word documents.

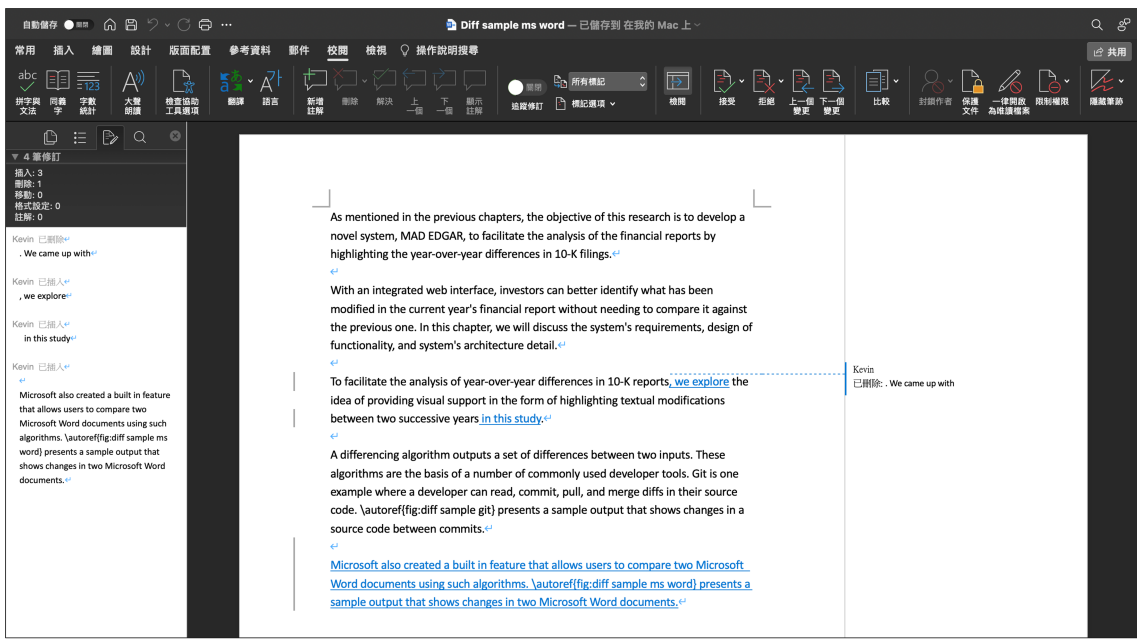
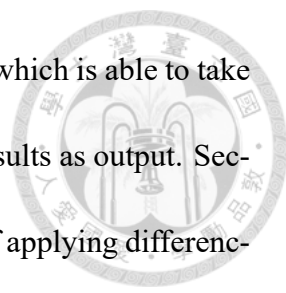


Figure 4.2: Sample Diff Output in Microsoft Word



Inspired by these insightful tools, we developed MAD EDGAR which is able to take a pair of 10-K reports as an input and generate the diff-highlighted results as output. Section 4.3 discusses about how we overcome the technical difficulties of applying differencing algorithm on HTML documents. Figure 4.3 presents a sample diff-highlighted 10-K report in MAD EDGAR.

<p>ITEM 1A. RISK FACTORS</p> <p>Our operations and financial results are subject to various risks and uncertainties, including but not limited to those described below, which could harm our business, reputation, financial condition, and operating results.</p> <p>Risks Specific to our Company</p> <p>We generate a significant portion of our revenues from advertising, and reduced spending by advertisers, a loss of partners, or new and existing technologies that block ads online and/or affect our ability to customize ads could harm our business.</p> <p>We generated over 80% of total revenues from the display of ads online in 2020. Many of our advertisers, companies that distribute our products and services, digital publishers, and content providers can terminate their contracts with us at any time. These partners may not continue to do business with us if we do not create more value (such as increased numbers of users or customers, new sales leads, increased brand awareness, or more effective monetization) than their available alternatives. Changes to our advertising policies and data privacy practices, as well as changes to other companies' advertising and data privacy practices may affect the advertising that we are able to provide, which could harm our business. In addition, technologies have been developed that make customized ads more difficult or that block the display of ads altogether and some providers of online services have integrated technologies that could potentially impair the availability and functionality of third-party digital advertising. Failing to provide superior value or deliver advertisements effectively and competitively could harm our reputation, financial condition, and operating results.</p> <p>In addition, expenditures by advertisers tend to be cyclical, reflecting overall economic conditions and budgeting and buying patterns. Adverse macroeconomic conditions including COVID-19 and its effects on the global economy (as discussed in greater detail in our COVID-19 risk factor under 'General Risks' below) also have impacted the demand for advertising and resulted in our advertisers reducing the amounts our advertisers spend on advertising, and could have an adverse impact on such demand and spend, which could harm our financial condition and operating results.</p> <p>We face intense competition. If we do not continue to innovate and provide products and services that are useful to users, we may not remain competitive, which could harm our business and operating results.</p> <p>Our business environment is rapidly evolving and intensely competitive. Our businesses face changing technologies, shifting user needs, and frequent introductions of rival products and services. To compete successfully, we must accurately anticipate technology developments and deliver innovative, relevant and useful products, services, and technologies in a timely manner. As our businesses evolve, the competitive pressure to innovate will encompass a wider range of products and services. We must continue to invest significant resources in research and development, including through acquisitions, in order to enhance our technology and new and existing products and services.</p> <p>We have many competitors in different industries. Our current and potential domestic and international competitors range from large and established companies to emerging start-ups. Some competitors have longer operating histories and well established relationships in various sectors. They can use their experience and resources in ways that could affect our competitive position, including by making acquisitions, continuing to invest heavily in research and development and in talent, aggressively initiating intellectual property claims (whether or not meritorious), and continuing to compete aggressively for users, advertisers, customers, and content providers. Further, discrepancies in enforcement of existing laws may enable our lesser known competitors to aggressively interpret those laws without commensurate scrutiny, thereby affording them competitive advantages. Our competitors may also be able to innovate and provide products and services faster than we can or may foresee the need for products and services before us. For example, we are investing significantly in subscription-based products and services such as YouTube, which face intense competition from large, experienced companies with well established relationships with users.</p> <p>Our operating results may also suffer if our products and services are not responsive to the needs of our users, advertisers, publishers, customers, and content providers. As technologies continue to develop, our competitors may be able to offer experiences that are, or that are seen to be, substantially similar to or better than ours. This</p>	<p>10</p> <p>Table of Contents Alphabet Inc.</p> <p>may force us to compete in different ways and expend significant resources in order to remain competitive. If our competitors Table of Contents Alphabet Inc. are more successful than we are in developing compelling products or in attracting and retaining users, advertisers, publishers, customers, and content providers, our operating results could be harmed.</p> <p>Our ongoing investment in new businesses, products, services, and technologies is inherently risky, and could disrupt our current operations and harm our financial condition and operating results.</p> <p>We have invested and expect to continue to invest in new businesses, products, services, and technologies. The investments that we are making across Google Services, Google Cloud and Other Bets reflect our ongoing efforts to innovate and provide products and services that are useful to users, advertisers, publishers, customers, and content providers. Our investments in Google Services, Google Cloud and Other Bets span a wide range of industries beyond online advertising. Such investments ultimately may not be commercially viable or may not result in an adequate return of capital and, in pursuing new strategies, we may incur unanticipated liabilities. These endeavors may involve significant risks and uncertainties, including diversion of management resources and, with respect to Other Bets, the use of alternative investment, governance, or compensation structures that may fail to adequately align incentives across the company or otherwise accomplish their objectives.</p> <p>Within Google Services, we continue to invest heavily in hardware, including our smartphones and home devices, which is a highly competitive market with frequent introduction of new products and services, rapid adoption of technological advancements by competitors, short product life cycles, evolving industry standards, continual improvement in product price and performance characteristics, and price and feature sensitivity on the part of consumers and businesses. There can be no assurance we will be able to compete hardware that competes effectively.</p> <p>Within Google Cloud, we devote significant resources to develop and deploy our enterprise-ready cloud services, including Google Cloud Platform and Google Workspace Suite. We are incurring costs to build and maintain infrastructure to support cloud computing services and hire talent, particularly to support and scale our cloud salesforce. At the same time, our competitors are rapidly developing and deploying cloud-based services. Pricing and delivery models are competitive and evolving, and we may not attain sufficient scale and profitability to achieve our business objectives.</p> <p>Within Other Bets, we are investing significantly in the areas of health, life sciences, and transportation, among others. These investment areas face intense competition from large experienced and well-funded competitors and our offerings may not be able to</p>
--	--

Figure 4.3: Sample Diff-Highlighted Report in MAD EDGAR

4.2 User Interface Design of Functionality

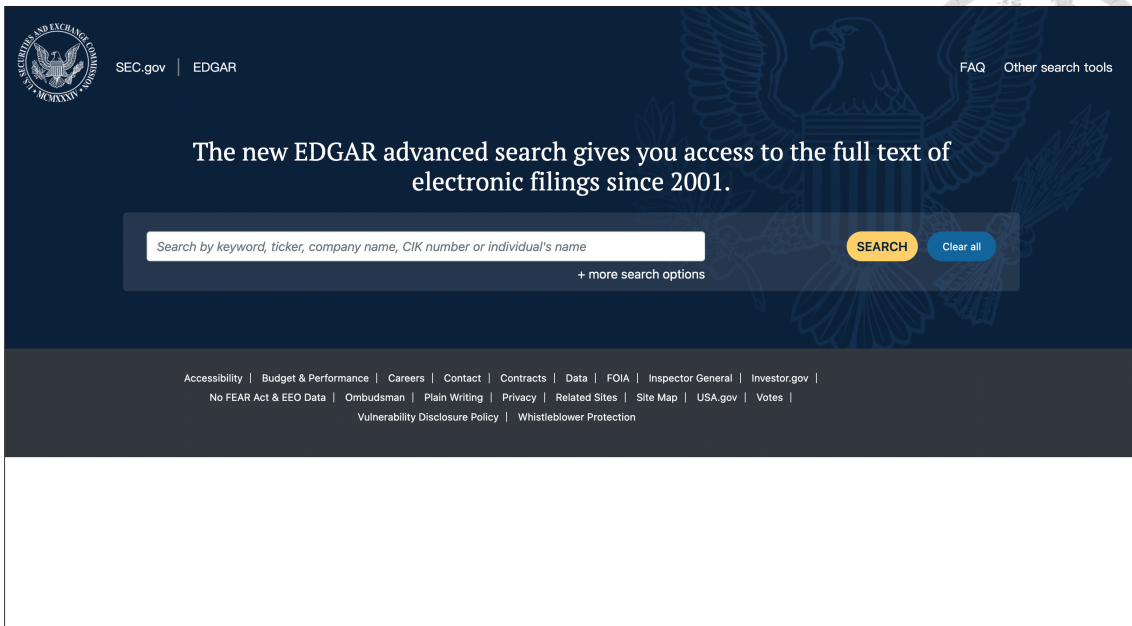


Our goal is to make changes to SEC EDGAR that help investors better identify year-over-year differences in 10-K reports. To introduce as minimal changes as possible to the original workflow and minimize the learning curve of the users, we designed the user interface of MAD EDGAR based on SEC EDGAR. Table 4.1 presents a brief comparison of functionality between both systems.

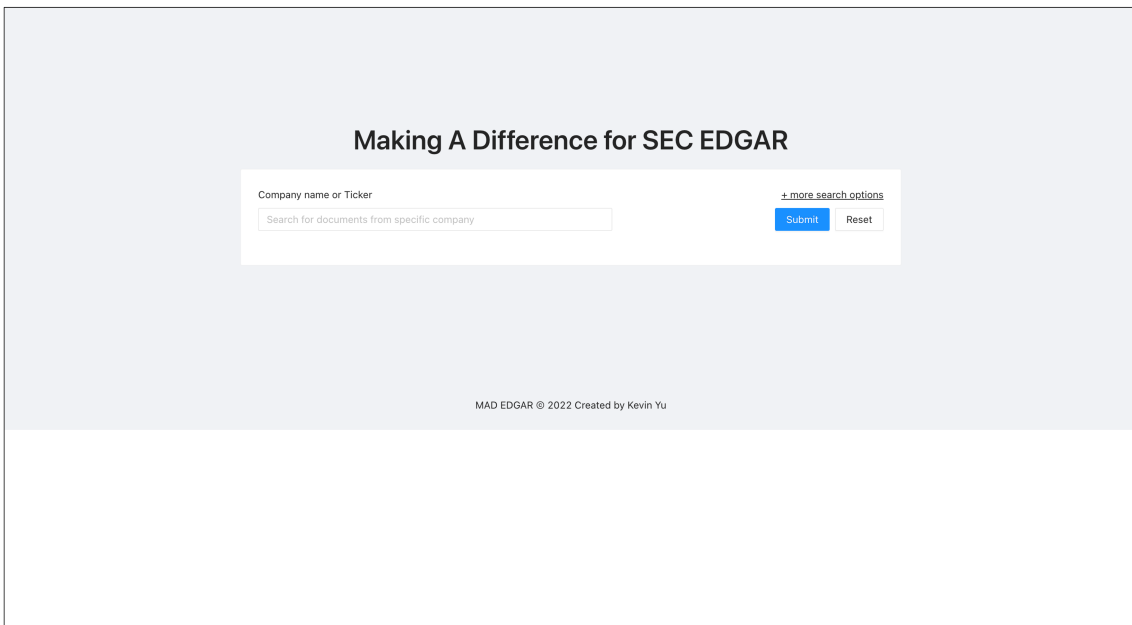
Function	SEC EDGAR	MAD EDGAR
Data coverage	✓ access to all SEC forms (10-K, 10-Q, 8-K, ...)	△ currently only support a subset of 10-K reports
Document search	✓ full-text search on company name or terms in documents	△ full-text search on company name
Advanced filter	✓ filtered by date range	✓ filtered by date range
Year-over-Year comparison	✗	✓ modifications are highlighted in the report

Table 4.1: System Function Comparison

Figure 4.4 presents the screenshots of the main page for both systems, an input search bar allows users to search for reports by a specific company.



(a) SEC EDGAR



(b) MAD EDGAR

Figure 4.4: Screenshot of the Main Page

Figure 4.5 presents the screenshots of the search tool for both systems. By extending the previous input search bar, search results can be further narrowed down by date.

(a) SEC EDGAR

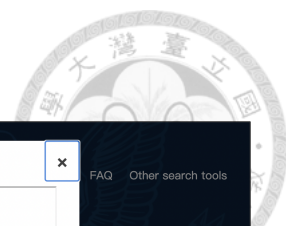
Filing Name	CIK	Company	Ticker	SIC Major Groups	Filing Type	Year	Filed Date	Similarity Ratio
0001018724-21-000004.htm	1018724	AMAZON COM INC	AMZN	Miscellaneous Retail	10-K	2021	2021-02-03	0.5243
0001018724-20-000004.htm	1018724	AMAZON COM INC	AMZN	Miscellaneous Retail	10-K	2020	2020-01-31	0.6744
0001018724-19-000004.htm	1018724	AMAZON COM INC	AMZN	Miscellaneous Retail	10-K	2019	2019-02-01	0.8182

(b) MAD EDGAR

Figure 4.5: Screenshot of the Search Tool

Figure 4.6 and Figure 4.7 show the screenshots of the report viewer and a sample report for both systems. In MAD EDGAR’s current form, the year-over-year differences in a 10-K report are presented in a colorized format, where the color green stands for new statements added in the current year of the report, while the color red stands for deletion

from the previous year on the contrary.



SEC.gov

Table of Contents

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

(Mark One)
 ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the fiscal year ended December 31, 2021
 or
 TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the transition period from to

Commission File No. 000-22513

AMAZON.COM, INC.
 (Exact name of registrant as specified in its charter)

Delaware
 (State or other jurisdiction of incorporation or organization)

91-1646860
 (I.R.S. Employer Identification No.)

Open document Open filing Close

Refine search results

Entity Form Principal executive offices located in

Show Columns

Filed Reporting for CIK Located Incorporated File number Film number

Form & File	Filed	Reporting for	Filing entity/person
10-K (Annual report)	2022-02-04	2021-12-31	AMAZON COM INC (AMZN)
10-K (Annual report)	2021-02-03	2020-12-31	AMAZON COM INC (AMZN)

Click headings to show top filters.
Document counts shown in 7

(a) SEC EDGAR

[AMAZON COM INC] 10-K differences

Open Diff-Highlighted Document Open Current Year

10-K 1 amzn-20201231.htm 20191231x10k.htm 10-K

Table of Contents

UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-K

(Mark One)
 ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the fiscal year ended December 31, 2021
 or
 TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the transition period from to

Commission File No. 000-22513

AMAZON.COM, INC.
 (Exact name of registrant as specified in its charter)

Cancel OK

0001018724-20-000004.htm	1018724	AMAZON COM INC	AMZN	Miscellaneous Retail	10-K	2020	2020-01-31	0.6744
0001018724-19-000004.htm	1018724	AMAZON COM INC	AMZN	Miscellaneous Retail	10-K	2019	2019-02-01	0.8182

(b) MAD EDGAR

Figure 4.6: Screenshot of the Report Viewer

**UNITED STATES
SECURITIES AND EXCHANGE COMMISSION**
Washington, D.C. 20549

FORM 10-K

(Mark One)
 ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the fiscal year ended December 31, 2021

or
 TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the transition period from _____ to _____

Commission File No. 000-22513

AMAZON.COM, INC.
 (Exact name of registrant as specified in its charter)

Delaware
 (State or other jurisdiction of incorporation or organization)

410 Terry Avenue North
 Seattle, Washington 98109-5210
 (206) 266-1000
 (Address and telephone number, including area code, of registrant's principal executive offices)

91-1646860
 (I.R.S. Employer Identification No.)

<p>Title of Each Class Common Stock, par value \$0.01 per share</p>	<p>Securities registered pursuant to Section 12(b) of the Act: Trading Symbol(s) AMZN</p> <p>Securities registered pursuant to Section 12(g) of the Act: None</p>	<p>Name of Each Exchange on Which Registered Nasdaq Global Select Market</p>
---	---	--

Indicate by check mark if the registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. Yes No
 Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Exchange Act. Yes No
 Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. Yes No
 Indicate by check mark whether the registrant has submitted electronically every Interactive Data File required to be submitted pursuant to Rule 405 of Regulation S-T during the preceding 12 months (or for such shorter period that the registrant was required to submit such files). Yes No
 Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, a smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer," "accelerated filer," "smaller reporting company," and "emerging growth company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer <input checked="" type="checkbox"/>	Accelerated filer <input type="checkbox"/>
Non-accelerated filer <input type="checkbox"/>	Smaller reporting company <input type="checkbox"/>
	Emerging growth company <input type="checkbox"/>

If an emerging growth company, indicate by check mark if the registrant has elected not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange Act.
 Indicate by check mark whether the registrant has filed a report on and attestation to its management's assessment of the effectiveness of its internal control over financial reporting under Section 404(b) of the Sarbanes-Oxley Act (15 U.S.C. 7262(b)) by the registered public accounting firm that prepared or issued its audit report.
 Indicate by check mark whether the registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act). Yes No
 Aggregate market value of voting stock held by non-affiliates of the registrant as of June 30, 2021 \$ 1,507,362,696,975
 Number of shares of common stock outstanding as of January 26, 2022 508,844,410

(a) SEC EDGAR

**UNITED STATES
SECURITIES AND EXCHANGE COMMISSION**
Washington, D.C. 20549

FORM 10-K

(Mark One)
 ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the fiscal year ended December 31, 2021

or
 TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
 For the transition period from _____ to _____

Commission File No. 000-22513

AMAZON.COM, INC.
 (Exact name of registrant as specified in its charter)

Delaware
 (State or other jurisdiction of incorporation or organization)

410 Terry Avenue North
 Seattle, Washington 98109-5210
 (206) 266-1000
 (Address and telephone number, including area code, of registrant's principal executive offices)

91-1646860
 (I.R.S. Employer Identification No.)

<p>Title of Each Class Common Stock, par value \$0.01 per share</p>	<p>Securities registered pursuant to Section 12(b) of the Act: Trading Symbol(s) AMZN</p> <p>Securities registered pursuant to Section 12(g) of the Act: None</p>	<p>Name of Each Exchange on Which Registered Nasdaq Global Select Market</p>
---	---	--

Indicate by check mark if the registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act. Yes No
 Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Exchange Act. Yes No
 Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. Yes No
 Indicate by check mark whether the registrant has submitted electronically every Interactive Data File required to be submitted pursuant to Rule 405 of Regulation S-T during the preceding 12 months (or for such shorter period that the registrant was required to submit such files). Yes No
 Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, a smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer," "accelerated filer," "smaller reporting company," and "emerging growth company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer <input checked="" type="checkbox"/>	Accelerated filer <input type="checkbox"/>
Non-accelerated filer <input type="checkbox"/>	Smaller reporting company <input type="checkbox"/>
	Emerging growth company <input type="checkbox"/>

If an emerging growth company, indicate by check mark if the registrant has elected not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange Act.
 Indicate by check mark whether the registrant has filed a report on and attestation to its management's assessment of the effectiveness of its internal control over financial reporting under Section 404(b) of the Sarbanes-Oxley Act (15 U.S.C. 7262(b)) by the registered public accounting firm that prepared or issued its audit report.
 Indicate by check mark whether the registrant is a shell company (as defined in Rule 12b-2 of the Exchange Act). Yes No
 Aggregate market value of voting stock held by non-affiliates of the registrant as of June 30, 2021 \$ 1,174,267,987,295
 Number of shares of common stock outstanding as of January 30, 2022 303,564,146

(b) MAD EDGAR

Figure 4.7: Screenshot of a Sample Report



4.3 System Architecture

The overall infrastructure is illustrated in Figure 4.8. We design MAD EDGAR based on the three-tier architecture model, where the system is organized into three logical layers:

1. The Presentation Tier, also known as the User Interface Tier;
2. The Application Tier, where the REST-APIs are deployed;
3. The Data Tier, where the data is stored and managed.

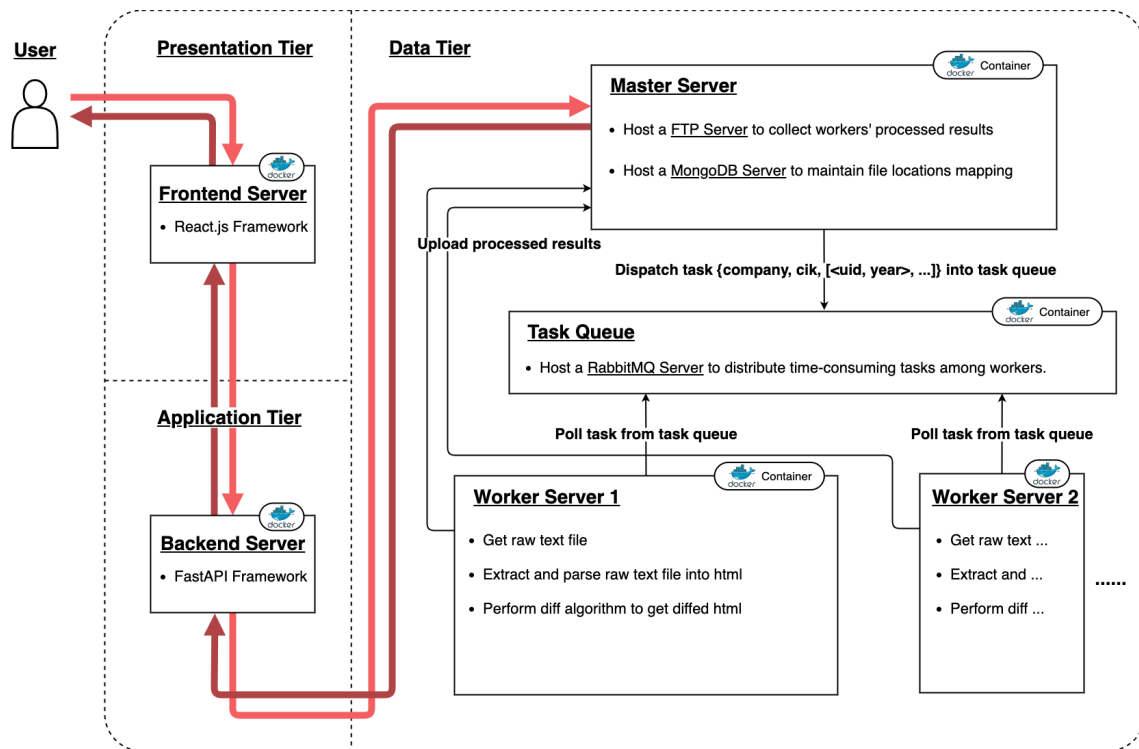


Figure 4.8: Overall Architecture in MAD EDGAR



4.3.1 Data Tier

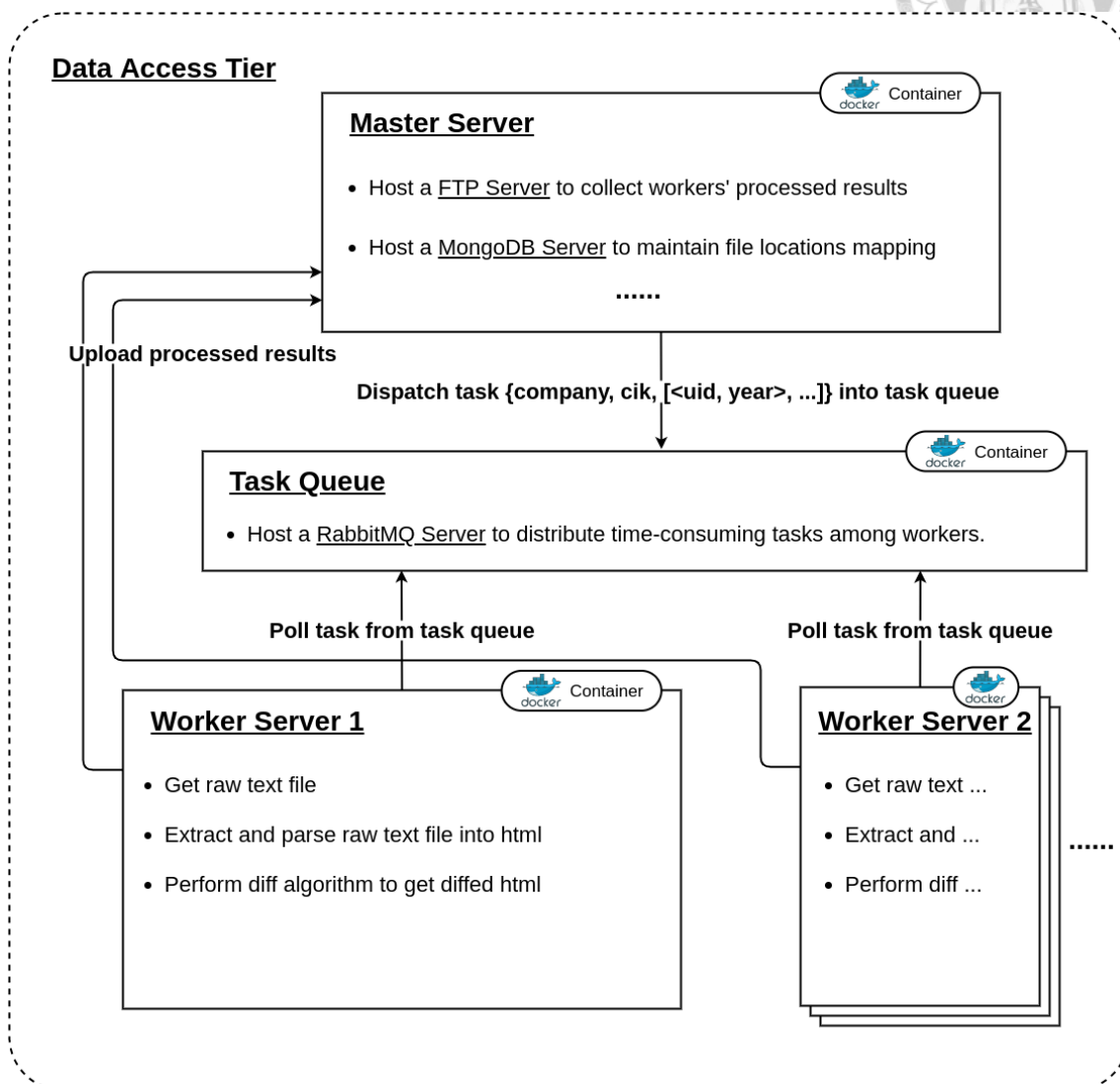
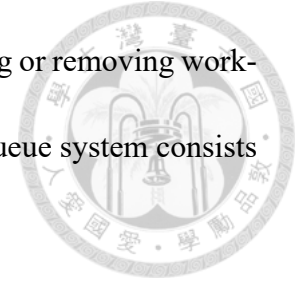


Figure 4.9: Data Tier in MAD EDGAR

The data tier mainly holds the system's data and responds to the application tier's data query. In order to populate MAD EDGAR with the diff-highlighted results, we created a distributed task queue system that handles the batch processing presented in Figure 4.9. Each task represents performing a differencing algorithm on a pair of two consecutive year 10-K reports from the same company and is independent of each other.

The task queue system can be scaled up or scaled down by adding or removing workers flexibly to make sure that our work can be processed. Our task queue system consists of the following three components:



1. A master server that is in charge of dispatching tasks and collecting diff-highlighted results.
2. A task queue server keeps the tasks until an available worker pops them off.
3. A pool of worker servers that perform the actual differencing algorithm on a pair of 10-K reports

4.3.1.1 Master Server

The master server in our task queue system acts as a central controlling center in charge of dispatching tasks to the workers. FTP server and MongoDB server are also maintained by the master server to collect workers' processed diff-highlighted results and to keep the mappings between files and their locations.

4.3.1.2 Task Queue Server

Communication between the master server and worker servers in our task queue system is done via RabbitMQ. A RabbitMQ server is hosted to maintain an asynchronous task queue and distribute time-consuming tasks among multiple workers.



4.3.1.3 Worker Servers

A dedicated pool of workers constantly monitors the task queue for any new task to process. After receiving a task from the task queue, workers perform the following series of processes to get the diff-highlighted results.

Step 1. Fetching the raw report in the form of .TXT extension

Step 2. Extract and parse the raw .TXT file into .HTML extension

Step 3. Perform the actual differencing algorithm on a pair of 10-K .HTML reports to get our final diff-highlighted .HTML result

Step 4. Upload the diff-highlighted .HTML result back to the master server

4.3.1.4 The Differencing Algorithm

We generate our diff-highlighted reports by extending a flexible class *SequenceMatcher* provided by Python 3.10.5 that can be used for comparing files and producing information about file differences in various formats. The basic algorithm is an algorithm published in the late 1980s by Ratcliff and Obershelp under the hyperbolic name “gestalt pattern matching” (*Ratcliff and Metzener, 1998*). It calculates string similarity based on the length of the longest common subsequence and recursive lengths of common characters in other parts of the string. The execution time of the algorithm is $O(n^2)$ in the worst case and $O(n)$ in the best case.

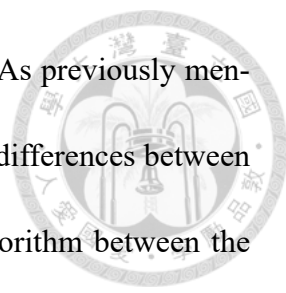


Figure 4.10 demonstrates our differencing algorithm in action. As previously mentioned, our system takes two paragraphs as inputs and calculates the differences between them. Table 4.2 presents the differences table generated by the algorithm between the two paragraphs. The differences table contains information about the operation (replace, delete, insert, equal) needed to perform on specific position to transform *para1* into *para2*. Our diff-highlighted result is then created by looking up the differences table and inserting corresponded labeled HTML span tags into the original paragraph.

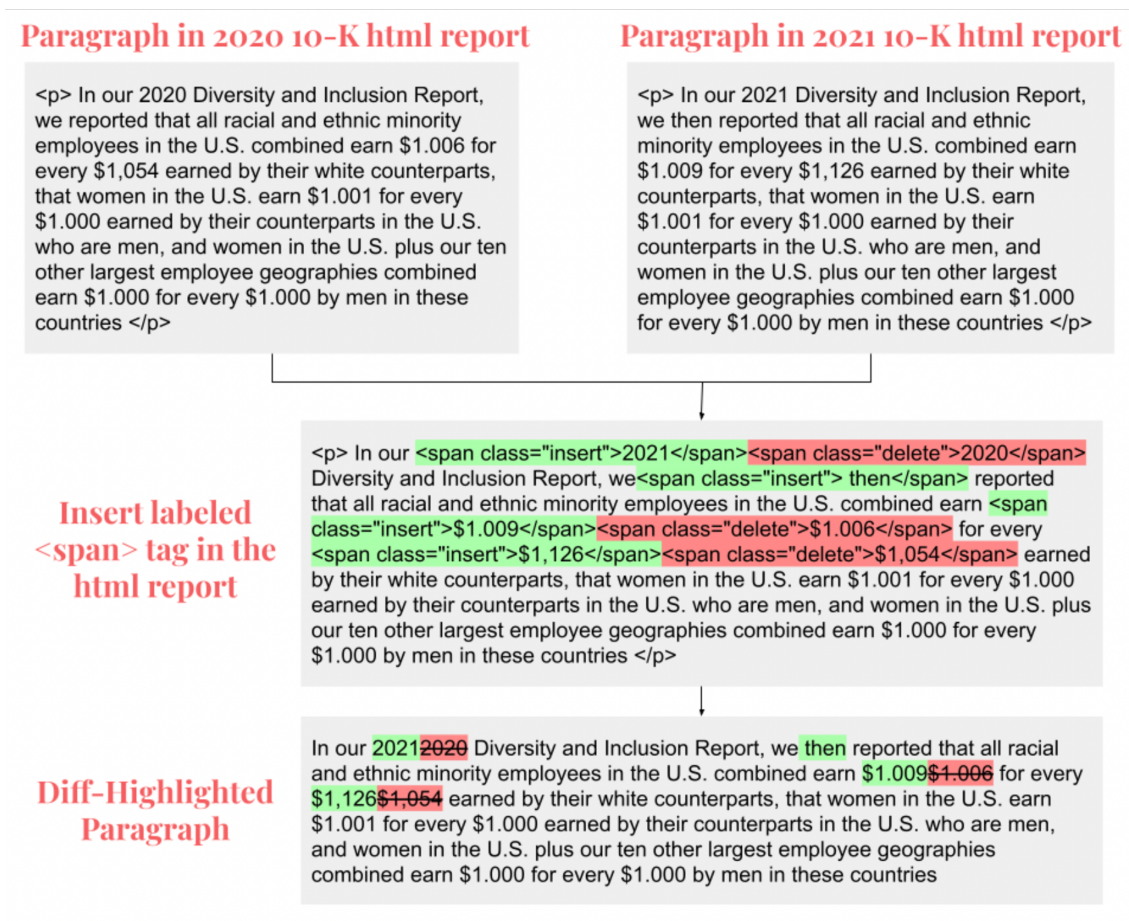


Figure 4.10: Differencing Algorithm in Action



Tag	Position	Description
equal	para1[0:6] → para2[0:6]	[‘<p>’, ‘’, ‘In’, ‘’, ‘our’, ‘’] → [‘<p>’, ‘’, ‘In’, ‘’, ‘our’, ‘’]
replace	para1[6:7] → para2[6:7]	[‘2020’] → [‘2021’]
equal	para1[7:17] → para2[7:17]	[‘’, ‘Diversity’, ‘’, ‘and’, ‘’, ‘Inclusion’, ‘’, ‘Report,’ ‘’, ‘we’] → [‘’, ‘Diversity’, ‘’, ‘and’, ‘’, ‘Inclusion’, ‘’, ‘Report,’ ‘’, ‘we’]
insert	para1[17:17] → para2[17:19]	[] → [‘’, ‘then’]
equal	para1[17:44] → para2[19:46]	[‘’, ‘reported’, ‘’, ‘that’, ‘’, ‘all’, ‘’, ‘racial’, ‘’, ‘and’, ‘’, ‘ethnic’, ‘’, ‘minority’, ‘’, ‘employees’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘combined’, ‘’, ‘earn’, ‘’] → [‘’, ‘reported’, ‘’, ‘that’, ‘’, ‘all’, ‘’, ‘racial’, ‘’, ‘and’, ‘’, ‘ethnic’, ‘’, ‘minority’, ‘’, ‘employees’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘combined’, ‘’, ‘earn’, ‘’]
replace	para1[44:45] → para2[46:47]	[‘\$1.006’] → [‘\$1.009’]
equal	para1[45:50] → para2[47:52]	[‘’, ‘for’, ‘’, ‘every’, ‘’] → [‘’, ‘for’, ‘’, ‘every’, ‘’]
replace	para1[50:51] → para2[52:53]	[‘\$1,054’] → [‘\$1,126’]
equal	para1[51:149] → para2[53:151]	[‘’, ‘earned’, ‘’, ‘by’, ‘’, ‘their’, ‘’, ‘white’, ‘’, ‘counterparts,’ ‘’, ‘that’, ‘’, ‘women’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘earn’, ‘’, ‘\$1.001’, ‘’, ‘for’, ‘’, ‘every’, ‘’, ‘\$1.000’, ‘’, ‘earned’, ‘’, ‘by’, ‘’, ‘their’, ‘’, ‘counterparts’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘who’, ‘’, ‘are’, ‘’, ‘men,’ ‘’, ‘and’, ‘’, ‘women’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘plus’, ‘’, ‘our’, ‘’, ‘ten’, ‘’, ‘other’, ‘’, ‘largest’, ‘’, ‘employee’, ‘’, ‘geographies’, ‘’, ‘combined’, ‘’, ‘earn’, ‘’, ‘\$1.000’, ‘’, ‘for’, ‘’, ‘every’, ‘’, ‘\$1.000’, ‘’, ‘by’, ‘’, ‘men’, ‘’, ‘in’, ‘’, ‘these’, ‘’, ‘countries’, ‘’, ‘<p>’] → [‘’, ‘earned’, ‘’, ‘by’, ‘’, ‘their’, ‘’, ‘white’, ‘’, ‘counterparts,’ ‘’, ‘that’, ‘’, ‘women’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘earn’, ‘’, ‘\$1.001’, ‘’, ‘for’, ‘’, ‘every’, ‘’, ‘\$1.000’, ‘’, ‘earned’, ‘’, ‘by’, ‘’, ‘their’, ‘’, ‘counterparts’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘who’, ‘’, ‘are’, ‘’, ‘men,’ ‘’, ‘and’, ‘’, ‘women’, ‘’, ‘in’, ‘’, ‘the’, ‘’, ‘U.S.’, ‘’, ‘plus’, ‘’, ‘our’, ‘’, ‘ten’, ‘’, ‘other’, ‘’, ‘largest’, ‘’, ‘employee’, ‘’, ‘geographies’, ‘’, ‘combined’, ‘’, ‘earn’, ‘’, ‘\$1.000’, ‘’, ‘for’, ‘’, ‘every’, ‘’, ‘\$1.000’, ‘’, ‘by’, ‘’, ‘men’, ‘’, ‘in’, ‘’, ‘these’, ‘’, ‘countries’, ‘’, ‘<p>’]

Tag	Meaning
‘replace’	para1[i1:i2] should be replaced by para2[j1:j2].
‘delete’	para1[i1:i2] should be deleted. Note that j1 == j2 in this case.
‘insert’	para2[j1:j2] should be inserted at para1[i1:i1]. Note that i1 == i2 in this case.
‘equal’	para1[i1:i2] == para2[j1:j2] (the sub-sequences are equal).

Table 4.2: Differences Table between Both Paragraphs



4.3.2 Application Tier

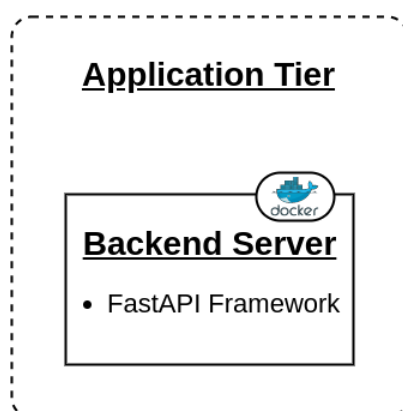


Figure 4.11: Application Tier in MAD EDGAR

The application tier lies between the presentation tier and the data tier and is also in charge of controlling the functionality of MAD EDGAR. Our application tier is structured as a standard REST-API, consisting of endpoints for interactions. We built our REST-APIs in a Python framework called FastAPI, which provides HTTP support for communicating with the application tier.

FastAPI is a modern, fast and robust framework that helps in building REST-APIs with Python 3.6+ versions. The framework is designed to optimize the development experience so that we can write simple code to build REST-APIs with efficiency. It offers a simple and easy-to-use dependency injection system, which enables us to declare relevant dependencies in the path operation functions assigned to the API endpoints.

FastAPI also provides a built-in web-based documentation user interface that interactively documents our APIs. We can test each endpoint through this interactive documentation. Different endpoints presented in Figure 4.12 are used by the presentation tier

to retrieve data from the database through HTTP methods.

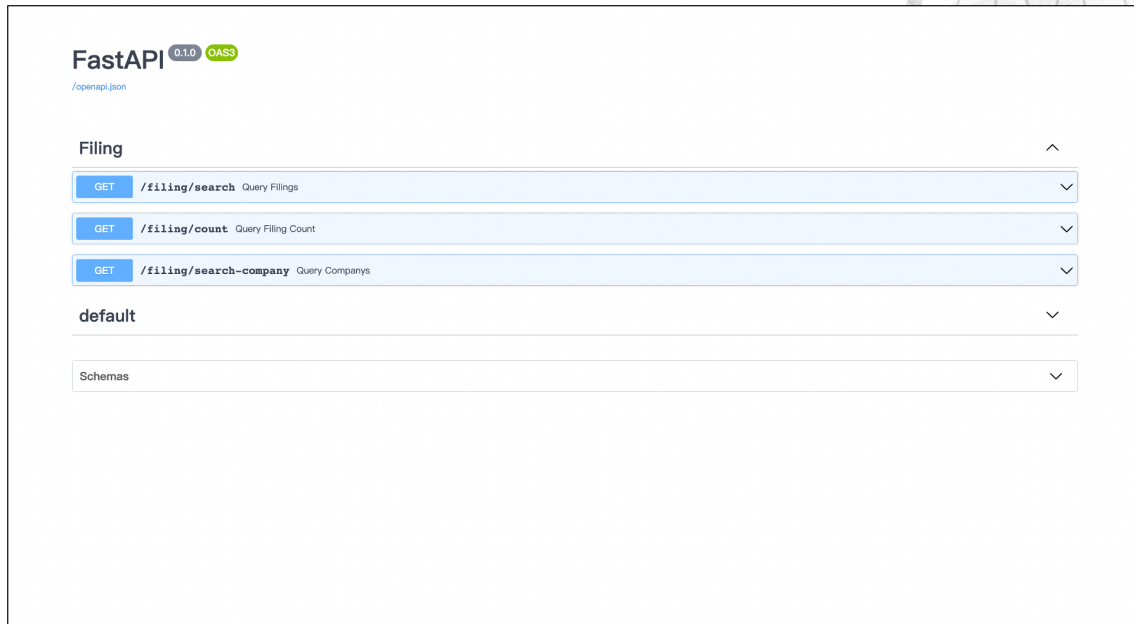


Figure 4.12: Endpoints in the Application Tier

4.3.3 Presentation Tier

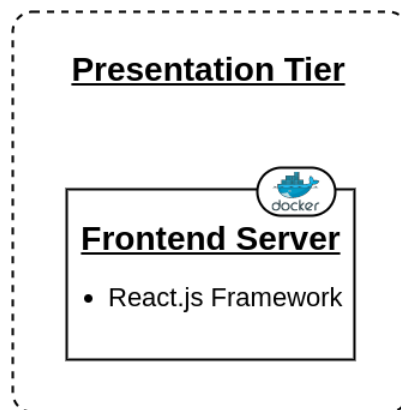


Figure 4.13: Presentation Tier in MAD EDGAR

Last but not least, the presentation tier represents the topmost level of the system and is also where the end-users interact with our system. We built our user interface in a JavaScript framework called React.js, which is currently one of the most popular libraries on the market developed by Meta (previously known as Facebook).

A React.js application consists of several components, each responsible for rendering a small piece of HTML that can be reused. To allow complex applications to be built out of simple building blocks, components can be nested within other components. We only need to design simple views for each state in our application, and React.js will efficiently update and render just the right components when our data changes. It's declarative nature and component-based structure make the task of building and maintaining our user interfaces much easier.





Chapter 5 Experimental Design

The purpose of this experiment is to evaluate MAD EDGAR's usefulness in facilitating the year-over-year comparison of 10-K reports. To this end, we designed two types of tasks with a total of sixteen questions for the participants to answer using MAD EDGAR or SEC EDGAR.

Our focus is to compare the *speed of financial report comprehensibility* between two systems, which is measured by the time required to complete each question. In order to discover whether the effect of MAD EDGAR on financial report comprehensibility is different from SEC EDGAR, a two-way ANOVA was conducted to assess the statistical significance.

5.1 Setup

5.1.1 Participants

Twenty graduate students from four universities participated in the experiment. There were 14 males and 6 females with a mean age of 25.5 ($SD = 1.9$).



5.1.2 Systems

SEC EDGAR and MAD EDGAR were used by the participants to complete the experiments in turn as instructed. For MAD EDGAR, our processed diff-highlighted reports covered a subset of S&P 500 companies' 10-K reports from 1993 to 2021 with a total of 8,377 distinct documents.

5.1.3 Technical Environment

A webcam, microphone, and internet connectivity were needed for the experiment to be carried out on the participant's laptop. We conducted the experiment on Google Meet and observed the video, voice, and participants' actions from the screen.

5.1.4 Tasks

During the experiment, the participants were required to complete two types of tasks using either SEC EDGAR or MAD EDGAR. Both types of tasks consist of eight questions, with a total of 16 questions in each experiment. Each question contains a target company and a target year for the participant's information.

The first task consists of multiple-choice questions focusing on finding several specific facts in a certain company's 10-K report. Figure 5.1 shows an example question that asked the participant to find out how many searches Google serves a year under a given context.

An example of the question in task 1 will be something like:

Company: Alphabet Inc.
Ticker: GOOGL, GOOG
CIK: 1652044
Filed year: 2021

How many searches does Google serve a year?

- (A) hundreds of millions
- (B) hundreds of thousands
- (C) trillions
- (D) billions

Figure 5.1: Example Question in the First Task

The second task consists of true-and-false questions focusing on identifying new risk factors in a certain company's 10-K report. Figure 5.2 shows an example question that asked the participant to determine whether these risk factor statements are recently disclosed in MCDONALDS CORP's 10-K report or not.

An example of the question in task 2 will be something like:

Company: MCDONALDS CORP
Ticker: MCD
CIK: 63908
Filed year: 2019

Whether or not the following risk factors are first included in 2019 as compared to the disclosure in 2018 (we=MCD)?

- (1) Our investments to enhance the customer experience, including through technology, may not generate the expected returns.
- (2) Supply chain interruptions may increase costs or reduce revenues.
- (3) Challenges with respect to labor availability and cost could impact our business and results of operations.
- (4) The global scope of our business subjects us to risks that could negatively affect our business.

Figure 5.2: Example Question in the Second Task

Both types of tasks differed with regard to the degree of complexity as the first task only requires the participants to open one report to answer the questions, while the second task requires the participants to identify year-over-year changes by comparing risk factors



in two successive reports.

During the experiment, to avoid order effects which could lead to practice effects such as the participant performing better on their second trial answering the same question using another system. The first task was handed out using a counterbalanced administration presented in Figure 5.3 with one group of participants answering the first four questions in MAD EDGAR and the last four questions in SEC EDGAR to preclude order effects and vice versa.

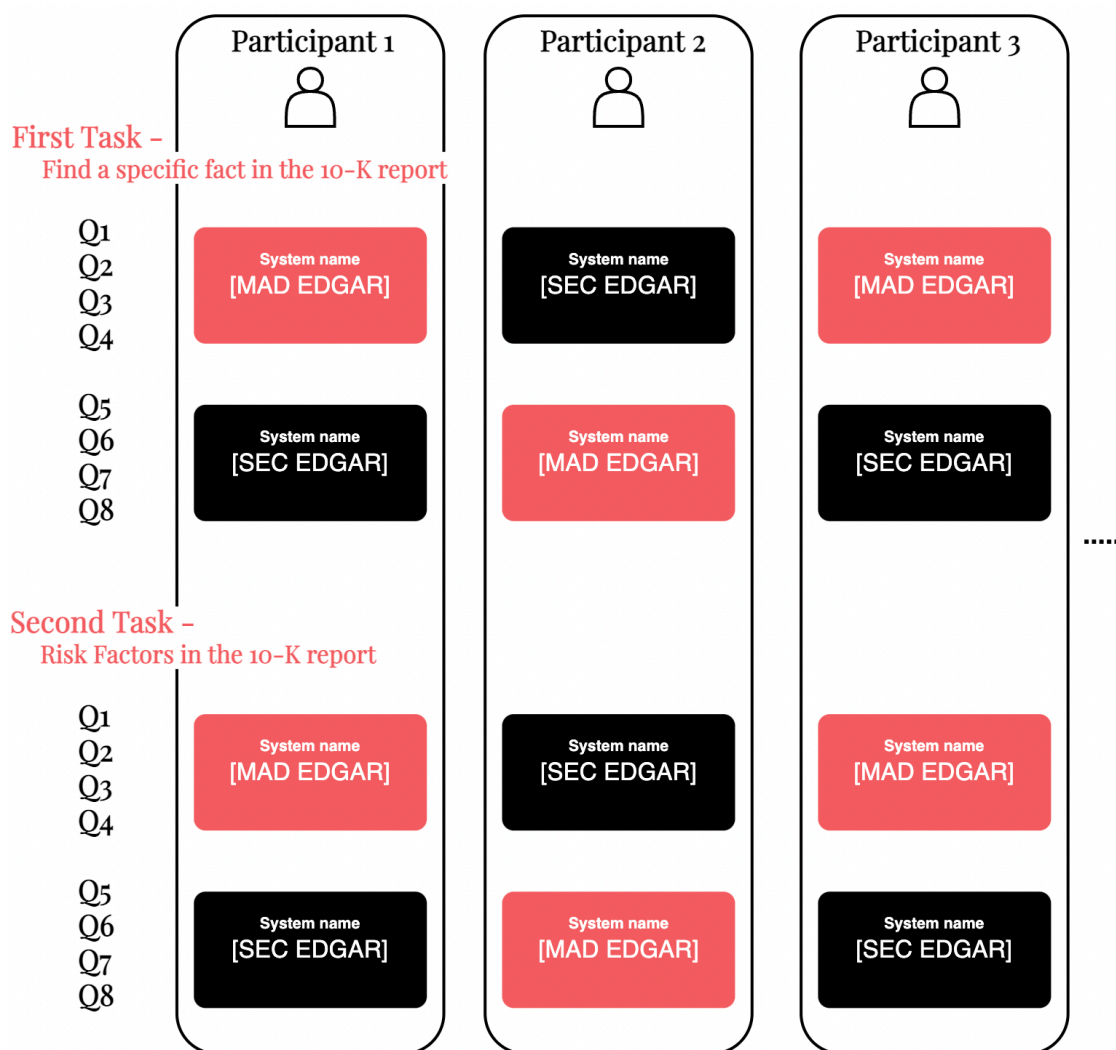


Figure 5.3: Counterbalanced Design

All of the questions were prepared in a Google Form and delivered to the participants at the beginning of each experiment. Figure 5.4 shows a screenshot of the Google Form.

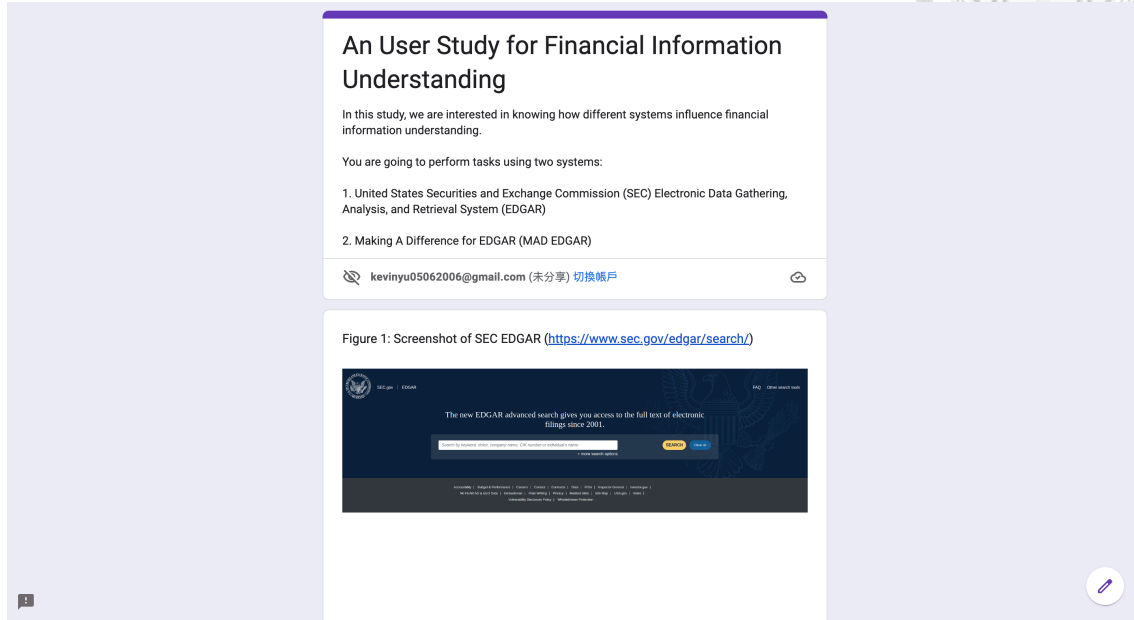


Figure 5.4: Screenshot of Data Collection Form Prepared using Google Form

5.1.5 Hypothesis Settings and Expected Experimental Results

For the first task, we are interested in testing the following hypothesis:

$H_{1,0}$: There is no significant difference in the *speed of financial report comprehensibility* when completing the first task using either SEC EDGAR or MAD EDGAR.

$H_{1,1}$: Users using MAD EDGAR can complete the first task faster compared to those using SEC EDGAR.

No performance differences were expected between the two systems on the first task because those questions only require participants to find several specific facts in a single financial report, without the need for a year-over-year comparison.

For the second task, we are interested in testing the following hypothesis:

$H_{2,0}$: There is no significant difference in the *speed of financial report comprehensibility* when completing the second task using either SEC EDGAR or MAD EDGAR.

$H_{2,1}$: Users using MAD EDGAR can complete the second task faster compared to those using SEC EDGAR.

The distinction in the *speed of financial report comprehensibility* between both systems was expected to reveal itself, especially on the second task. As MAD EDGAR was designed specifically to answer year-over-year comparison kind of questions.

5.2 Experimental Procedure

5.2.1 Data Collection

The experiment was conducted on Google Meet individually where the participant answers all sixteen questions using either SEC EDGAR or MAD EDGAR. Each session lasted for about 50 minutes.

At the beginning of the experiment, participants received a link to the Google Form that contains the question sets and were informed of the experiment's guide and goal. After a brief instruction on how to answer the questions using both systems, the participants started the hands-on part of the experiment.

All of the participants were asked to choose the correct answer for each question

until all questions in the first task were completed, and the second task was administered similarly.



5.2.2 Coding and Scoring

The *speed of financial report comprehensibility* was analyzed by recording the time the participants took to answer each question using either SEC EDGAR or MAD EDGAR.

Figure 5.5 shows a few data points we collected from the experiments, where there are two independent variables, *system* and *question*, and one dependent variable, *time_to_complete_task*.




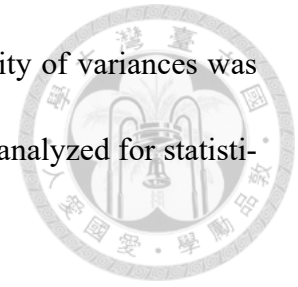
	 System	 Question	 time_to_complete_task
7	SEC EDGAR	Question 1	92
8	SEC EDGAR	Question 1	80
9	SEC EDGAR	Question 1	59
10	SEC EDGAR	Question 1	88
11	SEC EDGAR	Question 2	286
12	SEC EDGAR	Question 2	132
13	SEC EDGAR	Question 2	97

Figure 5.5: Screenshot of Experiment Result

5.2.3 Data Analysis

A two-way ANOVA was conducted to examine the effects of different *system* and *question* on *speed of financial report comprehensibility* for the first task and the second task. Residual analysis was performed to verify the assumption of the two-way ANOVA. Outliers were assessed by inspection of a box-plot, normality was tested using Shapiro-

Wilk's normality test for each cell of the design, and the homogeneity of variances was evaluated by Levene's test. Last but not least, the main effects were analyzed for statistically significant outcomes.





Chapter 6 Result and Discussion

In this chapter, we will first discuss how we handled the outliers in the data we previously collected from the experiment and provide a summary statistic about the time both SEC EDGAR and MAD EDGAR took to complete every question. Then, two-way ANOVA is conducted to examine the differences between both systems for the first task and the second task. Tests of normality and homogeneity of variances are also provided to make sure our data meet the assumptions of the two-way ANOVA. Finally, results and future works are presented and discussed in the last section.

6.1 Outliers

By inspecting box-plots generated from the first task in Figure 6.1 and Figure 6.2, we discovered that there were six outliers from both SEC EDGAR and MAD EDGAR, as assessed as being greater than 1.5 box-lengths from the edge of the box in the box-plots.

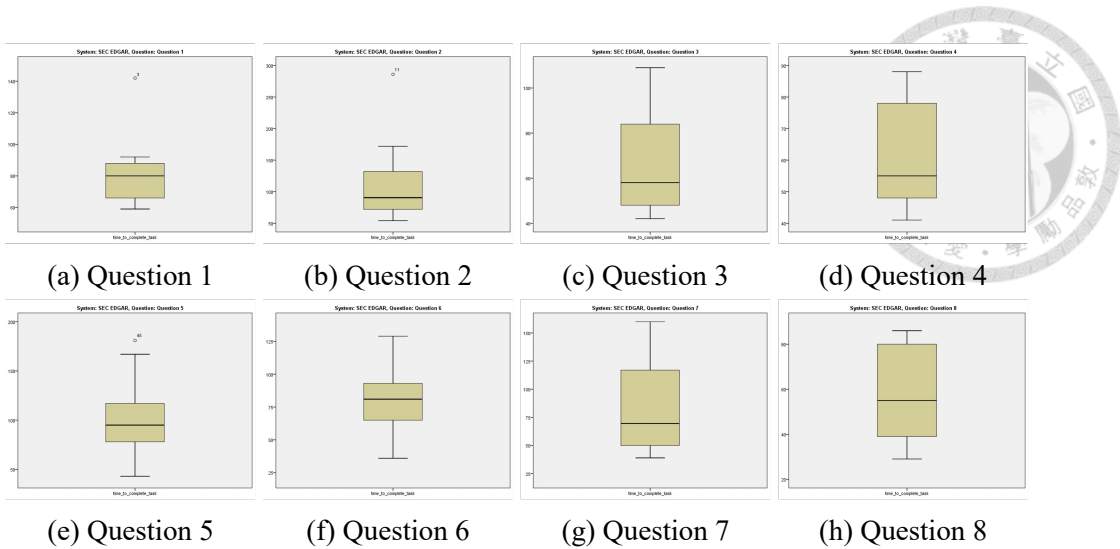


Figure 6.1: Box-Plot from the First Task using SEC EDGAR

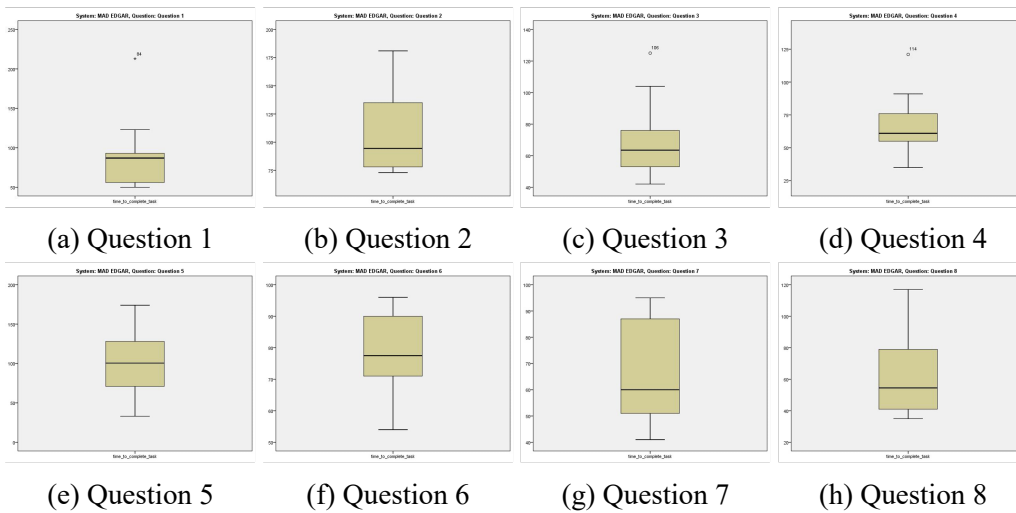


Figure 6.2: Box-Plot from the First Task using MAD EDGAR

Likewise, by inspecting box-plots generated from the second task in Figure 6.3 and Figure 6.4, we also discovered that there were seven outliers from both SEC EDGAR and MAD EDGAR, as assessed as being greater than 1.5 box-lengths from the edge of the box in the box-plots.

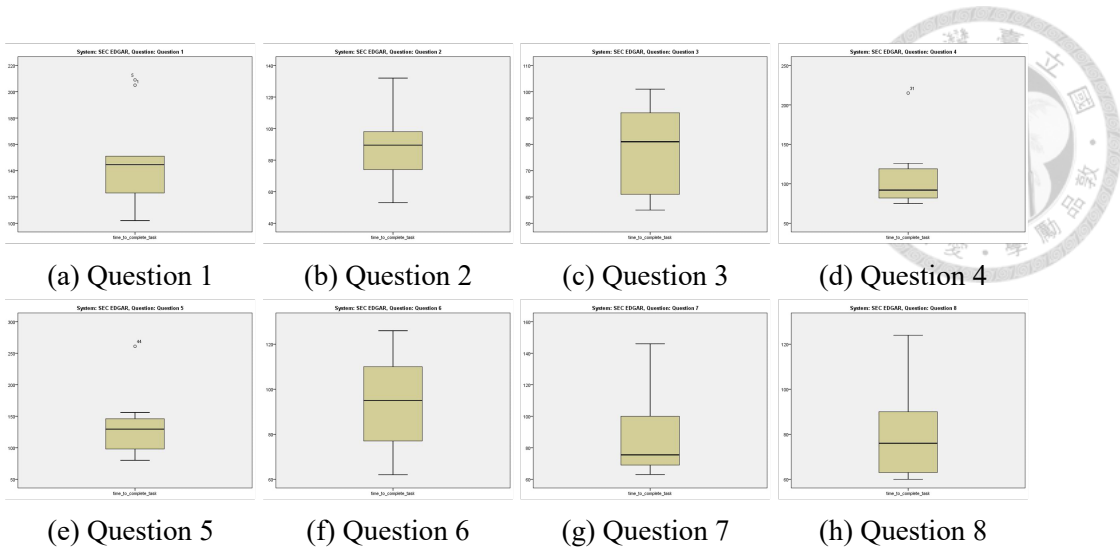


Figure 6.3: Box-Plot from the Second Task using SEC EDGAR

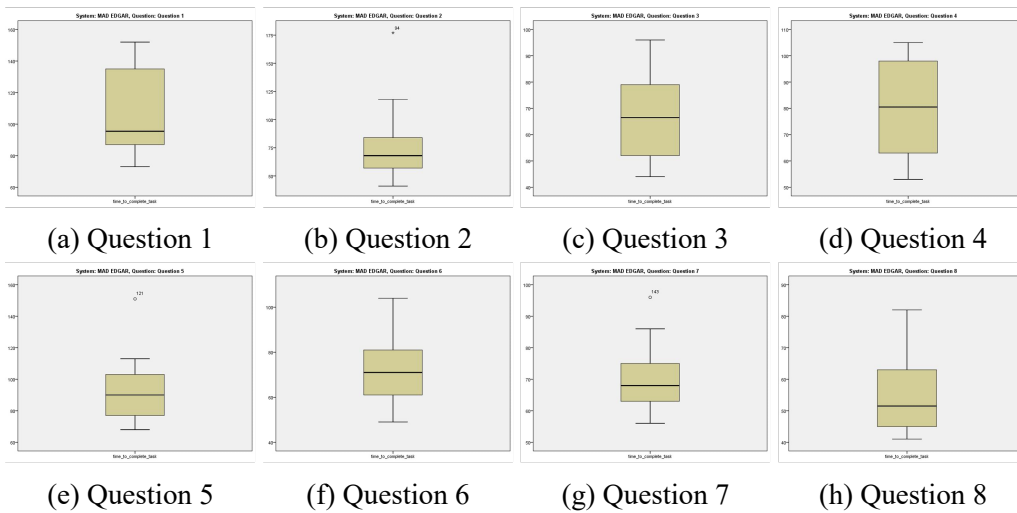


Figure 6.4: Box-Plot from the Second Task using MAD EDGAR

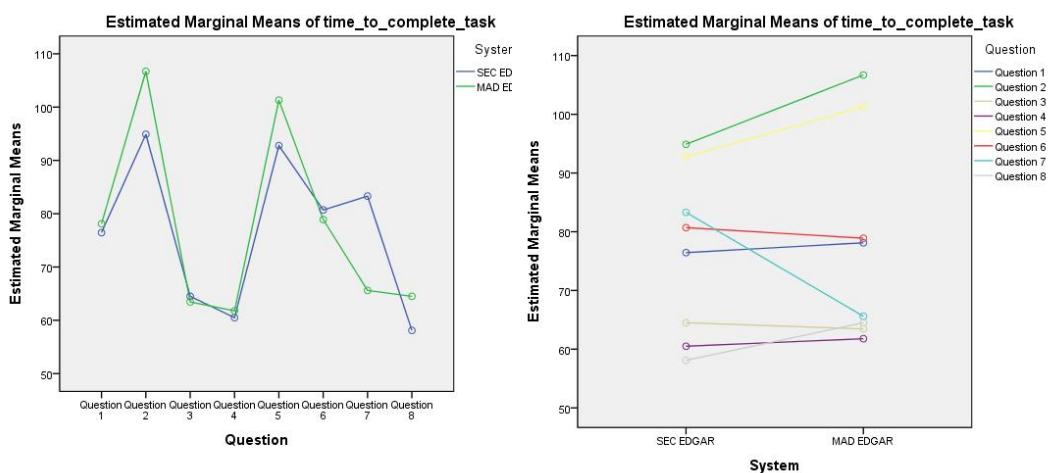
There were thirteen outliers in total where five out of thirteen were caused by unexpected interruptions during the experiment such as internet connection issues, while others were caused by unusual lookup overhead. In order to mitigate the negative effect of the outliers, we drop all of those outliers.



6.2 Summary Statistics

Table 6.1 shows a summary statistic about the time both SEC EDGAR and MAD EDGAR took to complete every question in seconds. It provides the mean and standard deviation for each combination of the groups of the independent variables. For the first task, the time it took for SEC EDGAR and MAD EDGAR to answer each question on average were 75.95 and 77.92 seconds respectively, while for the second task, the time it took for SEC EDGAR and MAD EDGAR to answer each question on average were 95.59 and 76.22 seconds respectively.

The plot of the mean *speed of financial report comprehensibility* for each combination of groups of *system* and *question* are also plotted. Figure 6.5 shows a profile plot for the first task with independent variables swapped on both axes. As it can be seen from Figure 6.5, both systems were evenly matched with respect to completing each question in the first task. No system has an obvious advantage over the other.



(a) Question-System

(b) System-Question

Figure 6.5: Profile Plot of the Mean from the First Task



First Task

Second Task

System	Question	Mean	Std.Dev.	N	System	Question	Mean	Std.Dev.	N
SEC EDGAR	Question 1	76.44	11.897	9	SEC EDGAR	Question 1	132.88	17.610	8
	Question 2	94.89	37.294	9		Question 2	88.80	21.837	10
	Question 3	64.50	22.032	10		Question 3	77.40	17.277	10
	Question 4	60.50	17.239	10		Question 4	94.11	18.238	9
	Question 5	92.78	38.545	9		Question 5	119.11	27.438	9
	Question 6	80.70	25.734	10		Question 6	94.70	21.108	10
	Question 7	83.30	42.458	10		Question 7	87.50	27.192	10
	Question 8	58.10	22.098	10		Question 8	79.90	20.388	10
	Total	75.95	30.607	77		Total	95.59	27.142	76
MAD EDGAR	Question 1	78.11	24.670	9	MAD EDGAR	Question 1	106.70	27.769	10
	Question 2	106.70	34.744	10		Question 2	70.22	22.615	9
	Question 3	63.44	19.417	9		Question 3	66.80	16.864	10
	Question 4	61.78	16.146	9		Question 4	80.50	18.253	10
	Question 5	101.30	40.186	10		Question 5	87.78	14.412	9
	Question 6	78.90	13.195	10		Question 6	74.20	17.986	10
	Question 7	65.60	19.323	10		Question 7	67.78	9.391	9
	Question 8	64.50	28.961	10		Question 8	55.50	14.378	10
	Total	77.92	30.165	77		Total	76.22	23.111	77
Pooled	Question 1	77.28	18.808	18	Pooled	Question 1	118.33	26.741	18
	Question 2	101.11	35.474	19		Question 2	80.00	23.591	19
	Question 3	64.00	20.262	19		Question 3	72.10	17.484	20
	Question 4	61.11	16.275	19		Question 4	86.95	19.057	19
	Question 5	97.26	38.560	19		Question 5	103.44	26.682	18
	Question 6	79.80	19.925	20		Question 6	84.45	21.792	20
	Question 7	74.45	33.365	20		Question 7	78.16	22.611	19
	Question 8	61.30	25.286	20		Question 8	67.70	21.248	20
	Total	76.94	30.303	154		Total	85.84	26.925	153

Table 6.1: Descriptive Statistics

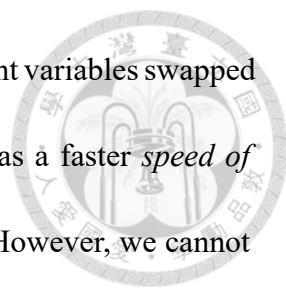
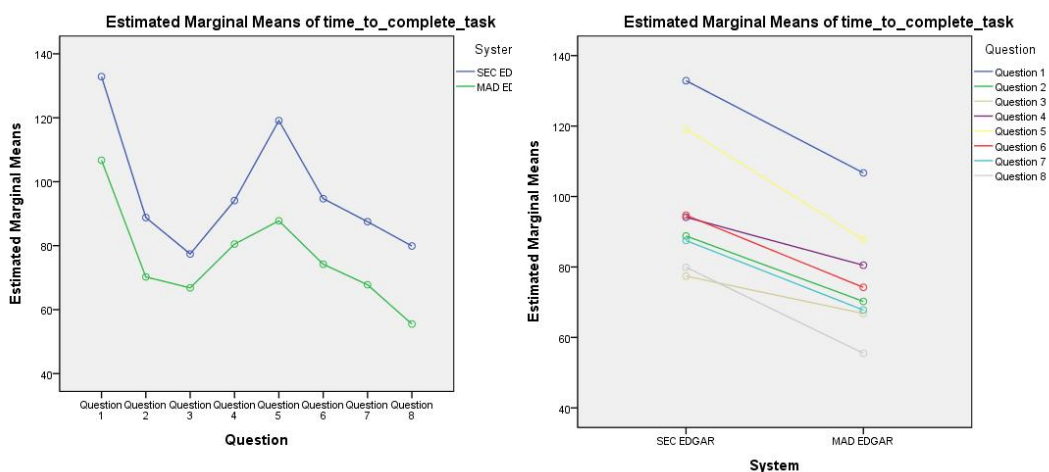


Figure 6.6 shows a profile plot for the second task with independent variables swapped on both axes. As it can be seen from Figure 6.6, MAD EDGAR has a faster *speed of financial report comprehensibility* as compared with SEC EDGAR. However, we cannot determine whether the effect of MAD EDGAR is statistical significant from the plot alone.

A formal statistical test is conducted and discussed in the next section.



(a) Question-System

(b) System-Question

Figure 6.6: Profile Plot of the Mean from the Second Task

6.3 Analysis of Variance (ANOVA)

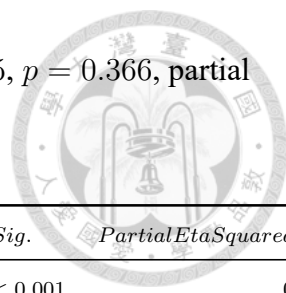
6.3.1 Result

6.3.1.1 First Task

For the first task, the result of the two-way ANOVA was presented in Table 6.2. The interaction effect between *system* and *question* on the *speed of financial report comprehensibility* was not statistically significant, $F(4, 86) = 0.155, p = 0.960, \text{partial } \eta^2 = 0.007$.

Therefore, an analysis of the main effect for *system* was performed, which also indicated

that the main effect was not statistically significant, $F(1, 86) = 0.825, p = 0.366$, partial $\eta^2 = 0.010$.



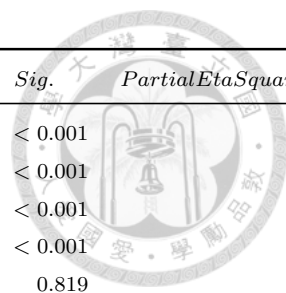
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	33001.011 ^a	9	3666.779	4.350	< 0.001	0.313
Intercept	565444.314	1	565444.314	670.745	< 0.001	0.886
System	695.683	1	695.683	0.825	0.366	0.010
Question	31238.342	4	7809.586	9.264	< 0.001	0.301
Interaction	523.767	4	130.942	0.155	0.960	0.007
Error	72498.822	86	843.01			
Total	671608	96				
Corrected Total	105499.833	95				

Table 6.2: ANOVA Table from the First Task

As mentioned in Subsection 5.1.5, for hypothesis H_1 , which addresses the difference in the first task of the *speed of financial report comprehensibility* between SEC EDGAR and MAD EDGAR, we expected there were no performance differences between both systems. With $p = 0.366$, we cannot reject the null hypothesis $H_{1,0}$ and concluded that the experiment results for the first task correspond to our expectations.

6.3.1.2 Second Task

For the second task, the result of the two-way ANOVA was presented in Table 6.3. The interaction effect between *system* and *question* on the *speed of Financial report comprehensibility* was also not statistically significant, $F(7, 137) = 0.518, p = 0.819$, partial $\eta^2 = 0.026$. Therefore, an analysis of the main effect for *system* was then performed, which indicated that the main effect for *system* was statistically significant, $F(1, 137) = 39.617, p < 0.001$, partial $\eta^2 = 0.224$.



Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	54229.116 ^a	15	3615.274	8.850	< 0.001	0.492
Intercept	1139569.596	1	1139569.596	2789.513	< 0.001	0.953
System	16184.186	1	16184.186	39.617	< 0.001	0.224
Question	38535.408	7	5505.058	13.476	< 0.001	0.408
Interaction	1482.469	7	211.781	0.518	0.819	0.026
Error	55967.119	137	408.519			
Total	1237660	153				
Corrected Total	110196.235	152				

Table 6.3: ANOVA Table from the Second Task

As mentioned in Subsection 5.1.5, for hypothesis H_2 , which addresses the difference in the second task of the *speed of financial report comprehensibility* between both systems, we expected our system, MAD EDGAR, should out-performed SEC EDGAR as it was designed specifically to answer year-over-year comparison kind of questions.

Since we just have two groups in our independent variable *system* (*SEC EDGAR* & *MAD EDGAR*), instead of performing a post-hoc analysis (e.g., all pairwise comparisons), we could grasp the difference between two systems by looking at Table 6.4.

System	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
SEC EDGAR	96.800	2.325	92.201	101.398
MAD EDGAR	76.185	2.306	71.624	80.745

Table 6.4: Estimate Table from the Second Task

The marginal means for *speed of financial report comprehensibility* were 96.800 ($SE = 2.325$) seconds for SEC EDGAR and 76.185 ($SE = 2.306$) seconds for MAD EDGAR, a statistically significant difference of 20.615 faster in second. We can reject the null hypothesis $H_{2,0}$ and conclude that MAD EDGAR has a significantly better *speed of financial report comprehensibility* than SEC EDGAR for the second task.



6.3.2 Tests of Normality

When analyzing differences between groups using parametric tests, a common assumption in these tests is that the dependent variable is approximately normally distributed for each group of the independent variable. In our experiment, Shapiro-Wilk tests have been run in both tasks for each group combination of the two independent variables: *system* and *question* to determine whether the data follows a normal distribution.

System	Question	Kolmogorov-Smirnova			Shapiro-Wilk			
		Statistic	df	Sig.	Statistic	df	Sig.	
SEC EDGAR	Question 1	Residual for time_to_complete_task	0.173	9	.200*	0.929	9	0.468
	Question 2	Residual for time_to_complete_task	0.181	9	.200*	0.906	9	0.289
	Question 3	Residual for time_to_complete_task	0.191	10	.200*	0.889	10	0.163
	Question 4	Residual for time_to_complete_task	0.189	10	.200*	0.886	10	0.153
	Question 5	Residual for time_to_complete_task	0.154	9	.200*	0.944	9	0.620
	Question 6	Residual for time_to_complete_task	0.122	10	.200*	0.981	10	0.968
	Question 7	Residual for time_to_complete_task	0.203	10	.200*	0.875	10	0.113
	Question 8	Residual for time_to_complete_task	0.228	10	0.151	0.889	10	0.167
MAD EDGAR	Question 1	Residual for time_to_complete_task	0.200	9	.200*	0.901	9	0.256
	Question 2	Residual for time_to_complete_task	0.210	10	.200*	0.878	10	0.122
	Question 3	Residual for time_to_complete_task	0.148	9	.200*	0.915	9	0.352
	Question 4	Residual for time_to_complete_task	0.186	9	.200*	0.98	9	0.964
	Question 5	Residual for time_to_complete_task	0.118	10	.200*	0.984	10	0.983
	Question 6	Residual for time_to_complete_task	0.116	10	.200*	0.957	10	0.752
	Question 7	Residual for time_to_complete_task	0.190	10	.200*	0.902	10	0.228
	Question 8	Residual for time_to_complete_task	0.215	10	.200*	0.873	10	0.108

Table 6.5: Test of Normality from the First Task

Table 6.5 and Table 6.6 show that all group combination in both tasks were normally distributed, as assessed by Shapiro-Wilk's test ($p > 0.05$). Therefore, we concluded that the dependent variable, *speed of financial report comprehensibility*, is normally distributed for each group combination of the two independent variables: *system* and *question*.

System	Question	Kolmogorov-Smirnova			Shapiro-Wilk			
		Statistic	df	Sig.	Statistic	df	Sig.	
SEC EDGAR	Question 1	Residual for time_to_complete_task	0.198	8	.200*	0.891	8	0.240
	Question 2	Residual for time_to_complete_task	0.137	10	.200*	0.976	10	0.940
	Question 3	Residual for time_to_complete_task	0.230	10	0.142	0.877	10	0.122
	Question 4	Residual for time_to_complete_task	0.207	9	.200*	0.884	9	0.174
	Question 5	Residual for time_to_complete_task	0.155	9	.200*	0.939	9	0.576
	Question 6	Residual for time_to_complete_task	0.157	10	.200*	0.949	10	0.656
	Question 7	Residual for time_to_complete_task	0.240	10	0.107	0.850	10	0.058
	Question 8	Residual for time_to_complete_task	0.165	10	.200*	0.891	10	0.176
MAD EDGAR	Question 1	Residual for time_to_complete_task	0.214	10	.200*	0.888	10	0.160
	Question 2	Residual for time_to_complete_task	0.230	9	0.187	0.91	9	0.314
	Question 3	Residual for time_to_complete_task	0.139	10	.200*	0.965	10	0.845
	Question 4	Residual for time_to_complete_task	0.167	10	.200*	0.925	10	0.402
	Question 5	Residual for time_to_complete_task	0.125	9	.200*	0.973	9	0.921
	Question 6	Residual for time_to_complete_task	0.160	10	.200*	0.919	10	0.353
	Question 7	Residual for time_to_complete_task	0.115	9	.200*	0.955	9	0.742
	Question 8	Residual for time_to_complete_task	0.186	10	.200*	0.879	10	0.128

Table 6.6: Test of Normality from the Second Task

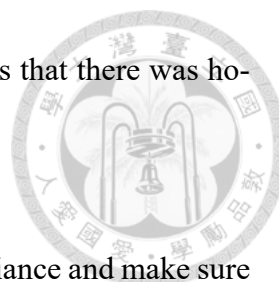
6.3.3 Homogeneity of Variances

Two-way ANOVA also assumes that the variances of the dependent variable, *speed of financial report comprehensibility*, are equal in all combinations of groups of the independent variables, *system* and *question*. In our experiment, the assumption of homogeneity of variances is tested using Levene's test for equality of variances.

First Task				Second Task			
<i>F</i>	<i>df1</i>	<i>df2</i>	<i>Sig.</i>	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>Sig.</i>
2.390	15	138	0.004	1.625	15	137	0.075

Table 6.7: Levene's Test for Equality of Variances

Table 6.7 shows that for the first task, the assumption of homogeneity of variances was violated as assessed by Levene's test for equality of variances, $p < 0.05$. For the



second task, a non-statistically significant result, $p = 0.075$, indicates that there was homogeneity of variances.

In order to meet the statistical assumption of homogeneity of variance and make sure that the distributions of the outcomes in each group are comparable and similar, we drop the question set 1, 6, and 7 in the first task for our analysis. Table 6.8 shows Levene's test result of the first task after data modifications, $p = 0.077$.

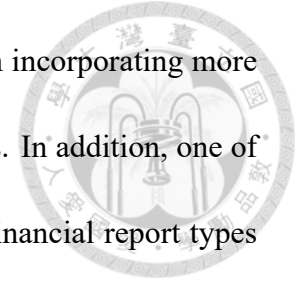
First Task (Before Dropping)				First Task (After Dropping)			
<i>F</i>	<i>df1</i>	<i>df2</i>	<i>Sig.</i>	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>Sig.</i>
2.390	15	138	0.004	1.816	9	86	0.077

Table 6.8: Levene's Test from the First Task

6.4 Conclusions

In this paper, we have introduced a new web-based information system, MAD EDGAR, to facilitate the analysis of year-over-year changes in 10-K reports. In its current form, we highlighted the differences between 10-K reports to help investors efficiently comprehend the modifications in documents. We also performed two experiments and the results indicated that our system out-performed SEC EDGAR in terms of identifying year-over-year changes.

In the future, we plan to extend the system with an emphasis on incorporating more functionality to better visualize differences between financial reports. In addition, one of our future works is to extend our system's coverage to even more financial report types such as 10-Q reports.





References

- Allee, K. D., and M. D. DeAngelis, The structure of voluntary disclosure narratives: Evidence from tone dispersion, *Journal of Accounting Research*, 53(2), 241–274, doi: 10.2139/ssrn.2375898, 2014.
- Bommarito, M., D. Katz, and E. M. Detterman, Openedgar: Open source software for sec Edgar analysis, *SSRN Electronic Journal*, doi:10.2139/ssrn.3194754, 2018.
- Brown, N. C., R. M. Crowley, and W. B. Elliott, What are you saying? using topic to detect financial misreporting, *Journal of Accounting Research*, 58(1), 237–291, doi: 10.1111/1475-679x.12294, 2020.
- Cannon, J. N., Z. Ling, Q. Wang, and O. V. Watanabe, 10-k disclosure of corporate social responsibility and firms' competitive advantages, *European Accounting Review*, 29(1), 85–113, doi:10.1080/09638180.2019.1670223, 2019.
- Cohen, L., C. J. Malloy, and Q. Nguyen, Lazy prices, *The Journal of Finance*, 75(3), 1371–1415, doi:10.2139/ssrn.1658471, 2020.
- Drake, M. S., D. T. Roulstone, and J. R. Thornock, What investors want: Evidence from

investors' use of the EDGAR database, *SSRN Electronic Journal*, doi:10.2139/ssrn.1932315, 2012.



Du, C.-H., Y.-S. Chiang, K.-C. Tsai, L.-C. Liu, M.-F. Tsai, and C.-J. Wang, Fridays: A financial risk information detecting and analyzing system, *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 9853–9854, doi:10.1609/aaai.v33i01.33019853, 2019.

Duarte-Silva, T., H. Fu, C. F. Noe, and K. Ramesh, How do investors interpret announcements of earnings delays?, *Journal of Applied Corporate Finance*, 25(1), 64–71, doi:10.1111/j.1745-6622.2013.12007.x, 2013.

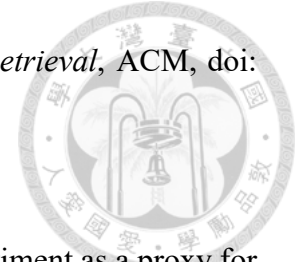
Efendi, J., J. Park, and L. Smith, Do xbrl filings enhance informational efficiency? early evidence from post-earnings announcement drift, *Journal of Business Research*, 67(6), 1099–1105, doi:10.1016/J.JBUSRES.2013.05.051, 2014.

Ege, M., J. L. Glenn, and J. R. Robinson, Unexpected sec resource constraints and comment letter quality, *Contemporary Accounting Research*, 37(1), 33–67, doi:10.1111/1911-3846.12505, 2019.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal, Management's tone change, post earnings announcement drift and accruals, *Review of Accounting Studies*, 15(4), 915–953, doi:10.1007/s11142-009-9111-x, 2009.

Feng, F., C. Luo, X. He, Y. Liu, and T.-S. Chua, Finir 2020: The first workshop on information retrieval in finance, in *Proceedings of the 43rd International ACM SI-*

GIR Conference on Research and Development in Information Retrieval, ACM, doi:
10.1145/3397271.3401462, 2020.



Gandhi, P., T. Loughran, and B. McDonald, Using annual report sentiment as a proxy for financial distress in u.s. banks, *Journal of Behavioral Finance*, 20(4), 424–436, doi: 10.1080/15427560.2019.1553176, 2019.

Han, M., Y. Liang, Z. Duan, and Y. Wang, Mining public business knowledge: A case study in sec's edgar, doi:10.1109/BDCloud-SocialCom-SustainCom.2016.65, 2016.

Hasan, M., Readability of narrative disclosures in 10-k reports: Does managerial ability matter?, *European Accounting Review*, 29(1), 147–168, doi:10.1080/09638180.2018.1528169, 2018.

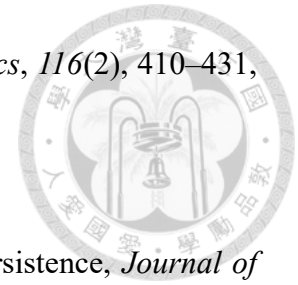
Henselmann, K., D. Ditter, and E. Scherr, Irregularities in accounting numbers and earnings management a novel approach based on SEC XBRL filings, *SSRN Electronic Journal*, doi:10.2139/ssrn.2297355, 2014.

Kim, C., K. Wang, and L. Zhang, Readability of 10-k reports and stock price crash risk, *Contemporary Accounting Research*, 36(2), 1184–1216, doi:10.1111/1911-3846.12452, 2019.

Kim, J. W., J.-H. Lim, and W. No, The effect of first wave mandatory xbrl reporting across the financial information environment, *Journal of Information Systems*, 26(1), 127–153, doi:10.2308/isys-10260, 2012.

Lee, C. M., P. Ma, and C. C. Wang, Search-based peer firms: Aggregating investor percep-

tions through internet co-searches, *Journal of Financial Economics*, 116(2), 410–431, doi:10.1016/j.jfineco.2015.02.003, 2015.



Li, F., Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics*, 45(2-3), 221–247, doi:10.2139/ssrn.887382, 2008.

Liu, Y.-W., L.-C. Liu, C.-J. Wang, and M.-F. Tsai, Fin10k: A web-based information system for financial report analysis and visualization, in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ACM, doi:10.1145/2983323.2983328, 2016.

Liu, Y.-W., L.-C. Liu, C.-J. Wang, and M.-F. Tsai, Riskfinder: A sentence-level risk detector for financial reports, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Association for Computational Linguistics, doi:10.18653/v1/n18-5017, 2018.

Lopatta, K., M. Gloger, and R. Jaeschke, Can language predict bankruptcy? the explanatory power of tone in 10-k filings, *Accounting Perspectives*, 16(4), 315–343, doi:10.1111/1911-3838.12150, 2017.

Loughran, T., and B. McDonald, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance*, 66(1), 35–65, doi:10.1111/J.1540-6261.2010.01625.X, 2010.

McMullin, J. L., B. Miller, and B. J. Twedt, Increased mandated disclosure frequency and

price formation: evidence from the 8-k expansion regulation, *Review of Accounting Studies*, 24(1), 1–33, doi:10.1007/S11142-018-9462-2, 2018.



Mitra, S., T. Al-Hayale, and M. Hossain, Does late 10k filing impact companies' financial reporting strategy? evidence from discretionary accruals and real transaction management, *Journal of Business Finance & Accounting*, 46(5-6), 569–607, doi:10.1111/jbfa.12369, 2019.

Plachouras, V., C. Smiley, H. Bretz, O. Taylor, J. L. Leidner, D. Song, and F. Schilder, Interacting with financial data using natural language, in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, ACM, doi:10.1145/2911451.2911457, 2016.

Ratcliff, J., and D. Metzener, Ratcliff-overshelp pattern recognition, *Dictionary of Algorithms and Data Structures*, 1998.

Ravula, S., Text analysis in financial disclosures, 2021.

Yen, J.-C., and T. Wang, The association between xbrl adoption and market reactions to earnings surprises, *Journal of Information Systems*, 29(3), 51–71, doi:10.2308/isys-51039, 2015.

You, H., and X.-J. Zhang, Investor under-reaction to earnings announcement and 10-k report, *SSRN Electronic Journal*, doi:10.2139/ssrn.1084332, 2007.

You, H., and X.-J. Zhang, Limited attention and stock price drift following earnings an-

nouncements and 10-k filings, *China Finance Review International*, 1(4), 358–387,
doi:10.2139/ssrn.1475479, 2011.





Appendix A — Data Collection Form



User study | 問題 回覆 10 設定 傳送

第 1 個區段, 共 20 個

An User Study for Financial Information Understanding

In this study, we are interested in knowing how different systems influence financial information understanding.

You are going to perform tasks using two systems:

1. United States Securities and Exchange Commission (SEC) Electronic Data Gathering, Analysis, and Retrieval System (EDGAR)
2. Making A Difference for EDGAR (MAD EDGAR)

Figure 1: Screenshot of SEC EDGAR (<https://www.sec.gov/edgar/search/>)

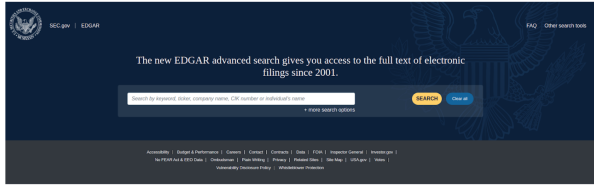



Figure 2: Screenshot of MAD EDGAR (<http://140.112.107.117:3000/>)



於區段 1 後 前往下一個區段

Figure A.7: Data Collection Form (P.1)



第 2 個區段, 共 20 個

Task 1/2: Find a specific fact in the annual report

In this task, you are going to answer a number of multiple choices questions by looking up and reading financial reports using either the SEC EDGAR or MAD EDGAR.

The questions in this task are related to each company's business overview, and you only need to focus on the [ITEM 1. BUSINESS] section.

An example of the question in task 1 will be something like:

Company: Alphabet Inc.
Ticker: GOOGL, GOOG
CIK: 1652044
Filed year: 2021

How many searches does Google serve a year?

(A) hundreds of millions
(B) hundreds of thousands
(C) trillions
(D) billions

於區段 2 後 前往下一個區段

第 3 個區段, 共 20 個

Task 1/2 Question 1/8 - SEC EDGAR

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: CORNING INC / NY
Ticker: GLW
CIK: 24741
Filed year: 2021

When was Corning established? 0 分

(A) 1856
 (B) 1851
 (C) 1855
 (D) 1853

What was Corning Incorporated's original name? 0 分

(A) Gorham Glass Works
 (B) The Corning Glass Company
 (C) Corning Glass Works
 (D) The Owens-Corning Company
 (E) The Utica Glass Company

於區段 3 後 前往下一個區段

Figure A.8: Data Collection Form (P.2)



第 4 個區段, 共 20 個

Task 1/2 Question 2/8

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: HARTFORD FINANCIAL SERVICES GROUP, INC.
Ticker: HIG, HIG-P, G, HGH
CIK: 874766
Filed year: 2021

What is the name of The Hartford's oldest subsidiary? 0 分

- (A) Hartford Fire Insurance Company
- (B) Hartford Life and Annuity Insurance Company
- (C) Hartford Life Insurance Company
- (D) Hartford Accident and Health Insurance Company

What is The Hartford working to do through investments in technology? 0 分

- (A) reduce fraud
- (B) decrease the cost of healthcare
- (C) improve customer service
- (D) increase efficiencies

於區段 4 後 前往下一個區段

第 5 個區段, 共 20 個

Task 1/2 Question 3/8

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: WEC ENERGY GROUP, INC.
Ticker: WEC
CIK: 783325
Filed year: 2021

Where are the principal executive offices located? 0 分

- (A) New York City, New York
- (B) Washington D.C.
- (C) Milwaukee, Wisconsin
- (D) Austin, Texas

How many reportable segments did the company have at December 31, 2020? 0 分

- (A) 24
- (B) 10
- (C) 8
- (D) 12
- (E) 6

於區段 5 後 前往下一個區段

Figure A.9: Data Collection Form (P.3)



第 6 個區段, 共 20 個

Task 1/2: Question 4/8

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: BERKSHIRE HATHAWAY INC
Ticker: BRK-B, BRK-A
CIK: 1067983
Filed year: 2021

Where is Berkshire Hathaway's corporate headquarters? 0 分

- (A) Omaha, Nebraska
- (B) Buffalo, New York
- (C) Milwaukee, Wisconsin
- (D) Austin, Texas

How many people did Berkshire Hathaway employ at the end of 2020? 0 分

- (A) 600,000
- (B) 360,000
- (C) 1,000,000
- (D) 100,000

於區段 6 後 前往下一個區段

第 7 個區段, 共 20 個

Task 1/2 Question 5/8 - MAD EDGAR

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: ELECTRONIC ARTS INC.
Ticker: EA
CIK: 712515
Filed year: 2021

How does Electronic Arts acquire strategic advantages? 0 分

- (A) a talented management team
- (B) unique selling proposition
- (C) superior technology
- (D) multiple business models and distribution channels
- (E) patents

What has happened to the percentage of games purchased digitally? 0 分

- (A) a significant increase
- (B) no change
- (C) a decrease

於區段 7 後 前往下一個區段

Figure A.10: Data Collection Form (P.4)



第 8 個區段, 共 20 個

Task 1/2 Question 6/8

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: ABBOTT LABORATORIES
Ticker: ABT
CIK: 1800
Filed year: 2021

When was Abbott Laboratories incorporated? 0 分

- (A) 1910
- (B) 1907
- (C) 1987
- (D) 1900
- (E) 1954

Which of the following is *not* Abbott Laboratories' important customer? 0 分

- (A) consumers
- (B) pharmacists
- (C) physicians
- (D) government agencies
- (E) private sector companies

於區段 8 後 前往下一個區段

第 9 個區段, 共 20 個

Task 1/2 Question 7/8

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: Nielsen Holdings plc
Ticker: NLSN
CIK: 1492633
Filed year: 2021

How many countries does Nielsen Holdings plc have operations in? 0 分

- (A) 90
- (B) 110
- (C) 100
- (D) 45
- (E) 25

When did Nielsen begin offering its common stock? 0 分

- (A) 1998
- (B) 2011
- (C) 1990
- (D) 1985
- (E) 1995

於區段 9 後 前往下一個區段

Figure A.11: Data Collection Form (P.5)



第 10 個區段, 共 20 個

Task 1/2 Question 8/8

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: COLGATE PALMOLIVE CO
Ticker: CL
CIK: 21665
Filed year: 2021

When was Colgate founded? 0 分

(A) 1876
 (B) 1865
 (C) 1853
 (D) 1847
 (E) 1806

When was Colgate-Palmolive Company incorporated? 0 分

(A) 1881
 (B) 1892
 (C) 1911
 (D) 1923
 (E) 1895

於區段 10 後 前往下一個區段

Figure A.12: Data Collection Form (P.6)

第 11 個區段, 共 20 個

Task 2/2: Risk Factors in the annual report

Intro: In this task, you will answer several true-false questions by looking up and reading financial reports using either the SEC EDGAR or MAD EDGAR.

The questions in this task are related to each company's risk factors, and you only need to focus on the [ITEM 1A. RISK FACTORS] section.

An example of the question in task 2 will be something like:

Company: MCDONALDS CORP
Ticker: MCD
CIK: 63908
Filed year: 2019

Whether or not the following risk factors are first included in 2019 as compared to the disclosure in 2018 (we=MCD)?

(1) Our investments to enhance the customer experience, including through technology, may not generate the expected returns.
(2) Supply chain interruptions may increase costs or reduce revenues.
(3) Challenges with respect to labor availability and cost could impact our business and results of operations.
(4) The global scope of our business subjects us to risks that could negatively affect our business.

於區段 11 後 前往下一個區段

Figure A.13: Data Collection Form (P.7)



第 12 個區段, 共 20 個

Task 2/2 Question 1/8 - SEC EDGAR

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: WYNN RESORTS LTD
Ticker: WYNN
CIK: 1174922
Filed year: 2019

Whether or not the following risk factors are first included in 2019 as compared to the disclosure in 2018 (we=WYNN)?

1. Adverse incidents or adverse publicity concerning our resorts or our corporate responsibilities could harm our brand and reputation and negatively impact our financial results. 0 分

true

false

2. Business is particularly sensitive to the willingness of our customers to travel to and spend time at our resorts. 0 分

true

false

3. Win rates for our gaming operations depend on a variety of factors, some of which are beyond our control. 0 分

true

false

4. The failure to protect the integrity and security of company employee and customer information could result in damage to reputation and/or subject us to fines, payment of damages, lawsuits or restrictions on our use or transfer of data. 0 分

true

false

5. Business could suffer if our computer systems and websites are disrupted or cease to operate effectively. 0 分

true

false

於區段 12 後 前往下一個區段

Figure A.14: Data Collection Form (P.8)



第 13 個區段, 共 20 個

Task 2/2 Question 2/8

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: L Brands, Inc.
Ticker: BBWI
CIK: 701985
Filed year: 2020

Whether or not the following risk factors are first included in 2020 as compared to the disclosure in 2019 (we=BBWI)?

1. Ability to service or refinance our debt and maintain compliance with our restrictive covenants 0 分
 true
 false

2. Business uncertainties and contractual restrictions while the VS Transaction is pending 0 分
 true
 false

3. Consumer acceptance of our products and our ability to manage the life cycle of our brands, keep up with fashion trends, develop new merchandise and launch new product lines successfully 0 分
 true
 false

4. Ability to grow through new store openings and existing store remodels and expansions 0 分
 true
 false

於區段 13 後 前往下一個區段

第 14 個區段, 共 20 個

Task 2/2 Question 3/8

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: NETFLIX INC
Ticker: NFLX
CIK: 1065280
Filed year: 2019

Whether or not the following risk factors are first included in 2019 as compared to the disclosure in 2018 (we=NFLX)?

Figure A.15: Data Collection Form (P.9)



1. Labor disputes may have an adverse effect on the Company's business. 0 分

true

false

2. Unforeseen costs and potential liability in connection with content we acquire, produce, license and/or distribute through our service. 0 分

true

false

3. Studios, content providers or other rights holders refuse to license streaming content or other rights upon terms acceptable to us, our business could be adversely affected. 0 分

true

false

4. Rely upon Amazon Web Services to operate certain aspects of our service and any disruption of or interference with our use of the Amazon Web Services operation would impact our operations and our business would be adversely impacted 0 分

true

false

於區段 14 後 前往下一個區段

Tr

Tr

?

?

Figure A.16: Data Collection Form (P.10)



第 15 個區段, 共 20 個

Task 2/2 Question 4/8

Please answer the following questions through SEC EDGAR (<https://www.sec.gov/edgar/search/>)

Company: GLOBAL PAYMENTS INC
Ticker: GPN
CIK: 1123360
Filed year: 2020

Whether or not the following risk factors are first included in 2020 as compared to the disclosure in 2019 (we=GPN)?

1. Our Business and Consumer Solutions segment relies on certain relationships with issuing banks, distributors, marketers and brand partners. The loss of such relationships, or if we are unable to maintain such relationships on terms that are favorable to us, may materially adversely affect our business, financial position, operating results or cash flows. 0 分

true
 false

2. Our business has been and will likely continue to be negatively affected by the COVID-19 pandemic. 0 分

true
 false

3. We rely on various financial institutions to provide clearing services in connection with our settlement activities. If we are unable to maintain clearing services with these financial institutions and are unable to find a replacement, our business may be adversely affected. 0 分

true
 false

4. Combining with TSYS may be more difficult, costly or time consuming than expected and we may fail to realize the anticipated benefits of the Merger. 0 分

true
 false

5. Increases in card network fees may result in the loss of customers and/or a reduction in our earnings. 0 分

true
 false

於區段 15 後 前往下一個區段

Figure A.17: Data Collection Form (P.11)



第 16 個區段, 共 20 個

Task 2/2 Question 5/8 - MAD EDGAR

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: TELEFLEX INC
Ticker: TFX
CIK: 96943
Filed year: 2020

Whether or not the following risk factors are first included in 2020 as compared to the disclosure in 2019 (we=TFX)?

1. Foreign currency exchange rate, commodity price and interest rate fluctuations may adversely affect our results. 0 分
 true
 false

2. Our results of operations and financial condition may be adversely affected by public health epidemics, including the novel coronavirus reported to have originated in Wuhan, China. 0 分
 true
 false

3. Disruptions in sterilization of our products or regulatory initiatives further restricting the use of ethylene oxide in sterilization facilities could adversely affect our results of operations and financial condition. 0 分
 true
 false

4. Health care reform may have a material adverse effect on our industry and our business. 0 分
 true
 false

5. Our substantial indebtedness could adversely affect our business, financial condition or results of operations. 0 分
 true
 false

於區段 16 後 前往下一個區段

Figure A.18: Data Collection Form (P.12)



第 17 個區段, 共 20 個

Task 2/2 Question 6/8

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: Arista Networks, Inc.
Ticker: ANET
CIK: 1596532
Filed year: 2020

Whether or not the following risk factors are first included in 2020 as compared to the disclosure in 2019 (we=ANET)?

1. We have adopted a stock repurchase program to repurchase shares of our common stock, however, any future decisions to reduce or discontinue repurchasing our common stock pursuant to our stock repurchase program could cause the market price for our common stock to decline. 0 分

true
 false

2. Because some of the key components in our products come from sole or limited sources of supply, we are susceptible to supply shortages or supply changes, which could disrupt or delay our scheduled product deliveries to our end customers and may result in the loss of sales and end customers. 0 分

true
 false

3. Our revenue and our revenue growth rate may decline. 0 分

true
 false

4. Delays in shipments could cause our revenue for the applicable period to fall below expected levels. 0 分

true
 false

5. We may not generate positive returns on our research and development investments. 0 分

true
 false

於區段 17 後 前往下一個區段

Figure A.19: Data Collection Form (P.13)



第 18 個區段, 共 20 個

Task 2/2 Question 7/8

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: DOLLAR TREE, INC.
Ticker: DLTR
CIK: 935703
Filed year: 2020

Whether or not the following risk factors are first included in 2020 as compared to the disclosure in 2019 (we=DLTR)?

1. We may stop selling or recall certain products for safety-related issues. 0 分
 true
 false

2. Our growth is dependent on our ability to increase sales in existing stores and to expand our square footage profitably. 0 分
 true
 false

3. The price of our common stock is subject to market and other conditions and may be volatile. 0 分
 true
 false

4. The continuing integration of Family Dollar's operations is not complete and may be more difficult, costly or time consuming than expected. 0 分
 true
 false

5. A downturn or adverse change in economic conditions could impact our sales or profitability. 0 分
 true
 false

於區段 18 後 前往下一個區段

Figure A.20: Data Collection Form (P.14)



第 19 個區段, 共 20 個

Task 2/2 Question 8/8

Please answer the following questions through MAD EDGAR (<http://140.112.107.117:3000/>)

Company: NORTHERN TRUST CORP
Ticker: NTRS, NTRSO
CIK: 73124
Filed year: 2019

Whether or not the following risk factors are first included in 2019 as compared to the disclosure in 2018 (we=NTRS, NTRSO)?

1. Changes in the method pursuant to which the London Interbank Offered Rate (LIBOR) or other interest rate benchmarks are determined could adversely impact our business and results of operations. 0 分

true

false

2. Failure to understand or appreciate fully the risks associated with development or delivery of new product and service offerings will affect our businesses and earnings negatively. 0 分

true

false

3. We are dependent on fee-based business for a majority of our revenues, which may be affected adversely by market volatility, a downturn in economic conditions, underperformance and/or negative trends in investment preferences. 0 分

true

false

4. Volatility levels and fluctuations in foreign currency exchange rates may affect our earnings. 0 分

true

false

於區段 19 後 前往下一個區段

第 20 個區段, 共 20 個

End of the study!

Thanks again!

Figure A.21: Data Collection Form (P.15)