10-K / 10-Q Text Analysis

Eric He - eh1885 - Group 21

Project Objective

1. Test pipeline and data processing capabilities by systematically downloading years of 10-K data from the SEC EDGAR database.
    a. Bonus: also download 10-Q data and special filings.
2. Can compare changes in text filings Y/Y using text distance and sentiment analysis
    a. Bonus: plug into ChatGPT to highlight and summarize differences

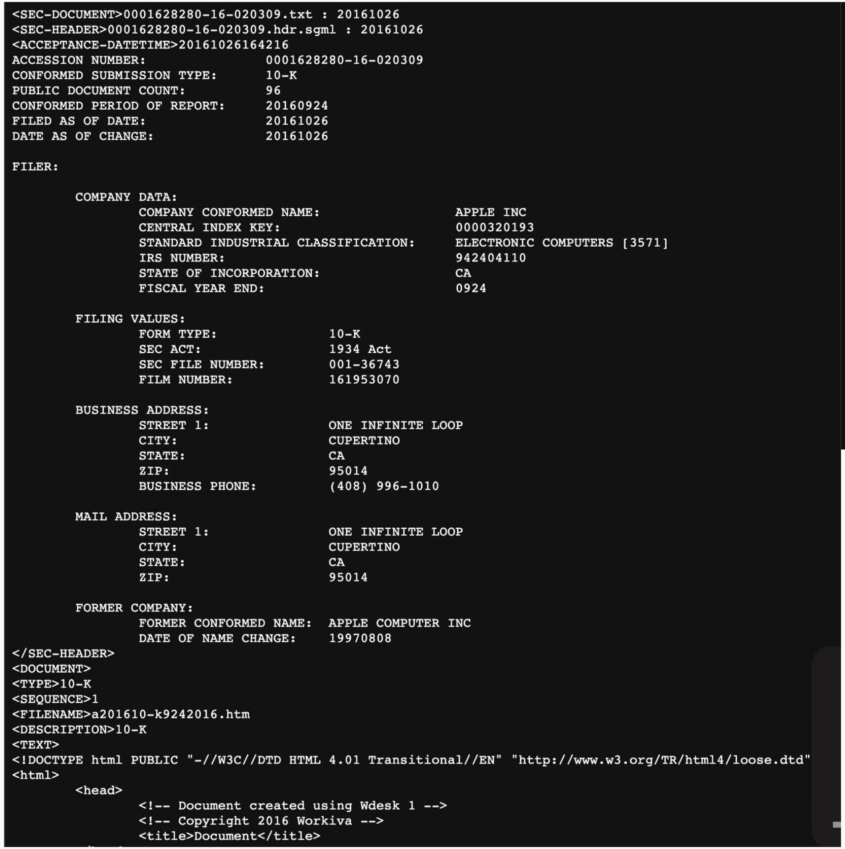
Concepts

1. Systemic data download - can we parallel process or stream process?
2. Heavy text processing - can we Cythonize?
3. Text data analysis

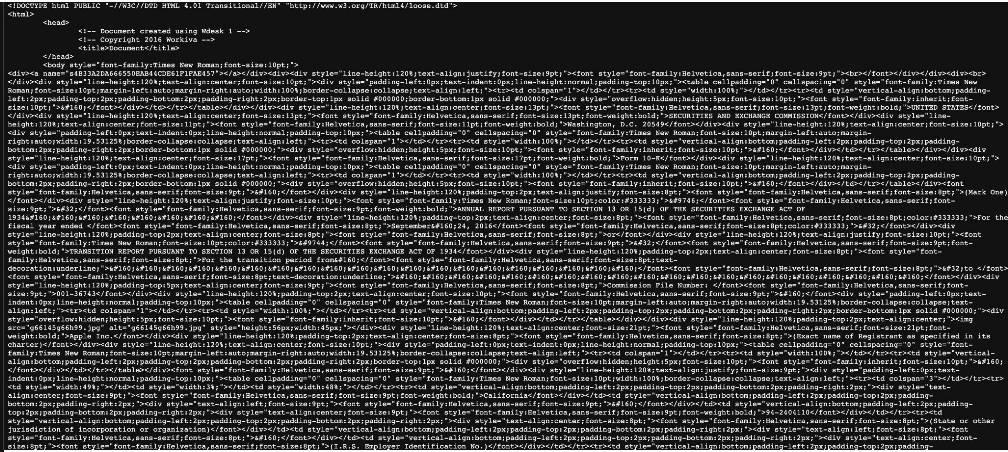
Background - EDGAR database

1. EDGAR is SEC's database for storing corporate filings. The data can be downloaded for free by anyone in the world (but if you do it wrong you can be throttled).
2. The raw data consists of raw HTML code, and requires a fairly intricate cleaning process to split 10-K or 10-Q into component sections.
3. Most of the financial statement is boilerplate, and large sections of the statement may not change Q/Q or even Y/Y. However, other times important financial data is presented in tables which are also challenging to process.
4. Downloading all the data, uncompressed, is several tens of gigabytes

Example of a 10-K filing



```
<SEC-DOCUMENT>0001628280-16-020309.txt : 20161026
<SEC-HEADER>0001628280-16-020309.hdr.sgml : 20161026
<ACCEPTANCE-DATETIME>20161026164216
ACCESSION NUMBER:		0001628280-16-020309
CONFORMED SUBMISSION TYPE:	10-K
PUBLIC DOCUMENT COUNT:		96
CONFORMED PERIOD OF REPORT:	20160924
FILED AS OF DATE:		20161026
DATE AS OF CHANGE:		20161026

FILER:

	COMPANY DATA:
		COMPANY CONFORMED NAME:			APPLE INC
		CENTRAL INDEX KEY:			0000320193
		STANDARD INDUSTRIAL CLASSIFICATION:	ELECTRONIC COMPUTERS [3571]
		IRS NUMBER:				942404110
		STATE OF INCORPORATION:			CA
		FISCAL YEAR END:			0924

	FILING VALUES:
		FORM TYPE:		10-K
		SEC ACT:		1934 Act
		SEC FILE NUMBER:	001-36743
		FILM NUMBER:		161953070

	BUSINESS ADDRESS:
		STREET 1:		ONE INFINITE LOOP
		CITY:			CUPERTINO
		STATE:			CA
		ZIP:			95014
		BUSINESS PHONE:		(408) 996-1010

	MAIL ADDRESS:
		STREET 1:		ONE INFINITE LOOP
		CITY:			CUPERTINO
		STATE:			CA
		ZIP:			95014

	FORMER COMPANY:
		FORMER CONFORMED NAME:	APPLE COMPUTER INC
		DATE OF NAME CHANGE:	19970808
</SEC-HEADER>
<DOCUMENT>
<TYPE>10-K
<SEQUENCE>1
<FILENAME>a201610-k9242016.htm
<DESCRIPTION>10-K
<TEXT>
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd"
<html>
	<head>
		<!-- Document created using Wdesk 1 -->
		<!-- Copyright 2016 Workiva -->
		<title>Document</title>
```



```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
	<head>
		<!-- Document created using Wdesk 1 -->
		<!-- Copyright 2016 Workiva -->
		<title>Document</title>
	</head>
	<body style="font-family:Times New Roman;font-size:10pt;">
<div><a name="s4B33A2DA666550EAB44CDE61F1FAE457"></a></div><div><div style="line-height:120%;text-align:justify;font-size:9pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;"><br></font></div></div><div><br>
</div><div style="line-height:120%;text-align:center;font-size:10pt;"><div style="padding-left:0px;text-indent:0px;line-height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New
Roman;font-size:10pt;margin-left:auto;margin-right:auto;width:100%;border-collapse:collapse;text-align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td style="vertical-align:bottom;padding-
left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;border-top:1px solid #000000;border-bottom:1px solid #000000;"><div style="overflow:hidden;height:5px;font-size:10pt;"><font style="font-family:inherit;font-
size:10pt;">&#160;</font></div></td></tr></table></div></div><div style="line-height:120%;text-align:center;font-size:13pt;"><font style="font-family:Helvetica,sans-serif;font-size:13pt;font-weight:bold;">UNITED STATES</font>
</div><div style="line-height:120%;text-align:center;font-size:13pt;"><font style="font-family:Helvetica,sans-serif;font-size:13pt;font-weight:bold;">SECURITIES AND EXCHANGE COMMISSION</font></div><div style="line-
height:120%;text-align:center;font-size:11pt;"><font style="font-family:Helvetica,sans-serif;font-size:11pt;font-weight:bold;">Washington, D.C. 20549</font></div><div style="line-height:120%;text-align:center;font-size:10pt;">
<div style="padding-left:0px;text-indent:0px;line-height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New Roman;font-size:10pt;margin-left:auto;margin-
right:auto;width:19.53125%;border-collapse:collapse;text-align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-
bottom:2px;padding-right:2px;border-bottom:1px solid #000000;"><div style="overflow:hidden;height:5px;font-size:10pt;"><font style="font-family:inherit;font-size:10pt;">&#160;</font></div></td></tr></table></div></div><div
style="line-height:120%;text-align:center;font-size:17pt;"><font style="font-family:Helvetica,sans-serif;font-size:17pt;font-weight:bold;">Form 10-K</font></div><div style="line-height:120%;text-align:center;font-size:10pt;">
<div style="padding-left:0px;text-indent:0px;line-height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New Roman;font-size:10pt;margin-left:auto;margin-
right:auto;width:19.53125%;border-collapse:collapse;text-align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-
bottom:2px;padding-right:2px;border-bottom:1px solid #000000;"><div style="overflow:hidden;height:5px;font-size:10pt;"><font style="font-family:inherit;font-size:10pt;">&#160;</font></div></td></tr></table></div><font
style="font-family:Helvetica,sans-serif;font-size:9pt;">&#160;</font></div><div style="line-height:120%;padding-top:2px;text-align:justify;font-size:8pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;">(Mark One)
</font></div><div style="line-height:120%;text-align:justify;font-size:10pt;"><font style="font-family:Times New Roman;font-size:10pt;color:#333333;">&#9746;</font><font style="font-family:Helvetica,sans-serif;font-
size:9pt;">&#32;</font><font style="font-family:Helvetica,sans-serif;font-size:9pt;font-weight:bold;">ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF
1934&#160;&#160;&#160;&#160;&#160;&#160;&#160;</font></div><div style="line-height:120%;padding-top:2px;text-align:center;font-size:8pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;color:#333333;">For the
fiscal year ended </font><font style="font-family:Helvetica,sans-serif;font-size:8pt;">September&#160;24, 2016</font><font style="font-family:Helvetica,sans-serif;font-size:8pt;color:#333333;">&#32;</font></div><div
style="line-height:120%;padding-top:2px;text-align:center;font-size:8pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;">or</font></div><div style="line-height:120%;text-align:justify;font-size:10pt;"><font
style="font-family:Times New Roman;font-size:10pt;color:#333333;">&#9744;</font><font style="font-family:Helvetica,sans-serif;font-size:9pt;">&#32;</font><font style="font-family:Helvetica,sans-serif;font-size:9pt;font-
weight:bold;">TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934</font></div><div style="line-height:120%;padding-top:2px;text-align:center;font-size:8pt;"><font style="font-
family:Helvetica,sans-serif;font-size:8pt;">For the transition period from&#160;</font><font style="font-family:Helvetica,sans-serif;font-size:8pt;text-
decoration:underline;">&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;</font><font style="font-family:Helvetica,sans-serif;font-size:8pt;">&#32;to </font>
<font style="font-family:Helvetica,sans-serif;font-size:8pt;text-decoration:underline;">&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;&#160;</font></div><div
style="line-height:120%;padding-top:5px;text-align:center;font-size:9pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;">Commission File Number: </font><font style="font-family:Helvetica,sans-serif;font-
size:9pt;">001-36743</font></div><div style="line-height:120%;padding-top:2px;text-align:center;font-size:10pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;">&#160;</font><div style="padding-left:0px;text-
indent:0px;line-height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New Roman;font-size:10pt;margin-left:auto;margin-right:auto;width:19.53125%;border-collapse:collapse;text-
align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;border-bottom:1px solid #000000;"><div
style="overflow:hidden;height:5px;font-size:10pt;"><font style="font-family:inherit;font-size:10pt;">&#160;</font></div></td></tr></table></div></div><div style="line-height:120%;padding-top:2px;text-align:center;"><img
src="g66145g66h99.jpg" alt="g66145g66h99.jpg" style="height:56px;width:45px;"></div><div style="line-height:120%;text-align:center;font-size:21pt;"><font style="font-family:Helvetica,sans-serif;font-size:21pt;font-
weight:bold;">Apple Inc.</font></div><div style="line-height:120%;padding-top:2px;text-align:center;font-size:8pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;">(Exact name of Registrant as specified in its
charter)</font></div><div style="line-height:120%;text-align:center;font-size:10pt;"><div style="padding-left:0px;text-indent:0px;line-height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-
family:Times New Roman;font-size:10pt;margin-left:auto;margin-right:auto;width:19.53125%;border-collapse:collapse;text-align:left;"><tr><td colspan="1"></td></tr><tr><td style="width:100%;"></td></tr><tr><td style="vertical-
align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;border-top:1px solid #000000;"><div style="overflow:hidden;height:5px;font-size:10pt;"><font style="font-family:inherit;font-size:10pt;">&#160;
</font></div></td></tr></table></div><font style="font-family:Helvetica,sans-serif;font-size:9pt;">&#160;</font></div><div style="line-height:120%;text-align:justify;font-size:9pt;"><div style="padding-left:0px;text-
indent:0px;line-height:normal;padding-top:10px;"><table cellpadding="0" cellspacing="0" style="font-family:Times New Roman;font-size:10pt;width:100%;border-collapse:collapse;text-align:left;"><tr><td colspan="3"></td></tr><tr>
<td style="width:49%;"></td><td style="width:3%;"></td><td style="width:48%;"></td></tr><tr><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;"><div style="text-
align:center;font-size:9pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;font-weight:bold;">California</font></div></td><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-
bottom:2px;padding-right:2px;"><div style="text-align:left;font-size:9pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;">&#160;</font></div></td><td style="vertical-align:bottom;padding-left:2px;padding-
top:2px;padding-bottom:2px;padding-right:2px;"><div style="text-align:center;font-size:9pt;"><font style="font-family:Helvetica,sans-serif;font-size:9pt;font-weight:bold;">94-2404110</font></div></td></tr><tr><td
style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;"><div style="text-align:center;font-size:8pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;">(State or other
jurisdiction of incorporation or organization)</font></div></td><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;"><div style="text-align:left;font-size:8pt;"><font
style="font-family:Helvetica,sans-serif;font-size:8pt;">&#160;</font></div></td><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-bottom:2px;padding-right:2px;"><div style="text-align:center;font-
size:8pt;"><font style="font-family:Helvetica,sans-serif;font-size:8pt;">(I.R.S. Employer Identification No.)</font></div></td></tr><tr><td style="vertical-align:bottom;padding-left:2px;padding-top:2px;padding-
```

What can a corporate filing tell us?

1. Management will highlight headwinds or other important updates in filings.
2. One might wonder if markets are truly reacting to the information in the filing “instantaneously” or if this market information is incorporated over a longer period of time, e.g. days, weeks or even months. Potential signals:
    a. Sentiment analysis
    b. Text vs. Numeric data
    c. Distance analysis
3. Specific language analysis may be relatively uncorrelated to other common information sources such as correlated company returns or economic data

I am not super interested in comparing to stock returns, I would rather work directly with the text data itself and treat this as a data processing exercise. 6 years ago I downloaded this data with a Google Chrome extension over the course of 4 weeks and felt I could do better.

Project Roadmap

1. [03/15/23] Proof of Concept Download database
2. [03/22/23] Proof of Concept process filings
3. [03/29/23] Proof of Concept Cosine Distance + Sentiment Analysis
4. [04/15/23] Parallelize data download with multiprocessing, compress, load into text database
5. [04/22/23] Cythonized compute of text features and condensed text summarization algos
6. [04/29/23] Summarize performance of optimization strategies
7. [Extra] Integrate with ChatGPT API to return opinions on textual differences or sentiment
8. [Extra] Do it again but for 10-Qs

Thank You

Questions?