UCLAAnderson
SCHOOL of MANAGEMENT

MGMTMFE 431:

Data Analytics and Machine Learning

Topic 7: Further Topics in Textual Analysis

Sentiment, attention, and topic modeling

Spring 2025

Professor Lars A. Lochstoer

Unstructured Data and Introduction to Textual Analysis

a. EDGAR and financial reports

b. Homework 6: 10-Ks and word sentiment

c. Homework 6: Mapping unstructured data to numerical signals

d. Lazy Prices: Investor Attention and Financial Reports

e. Topic Modelling and LDA: FOMC minutes

a. The EDGAR database

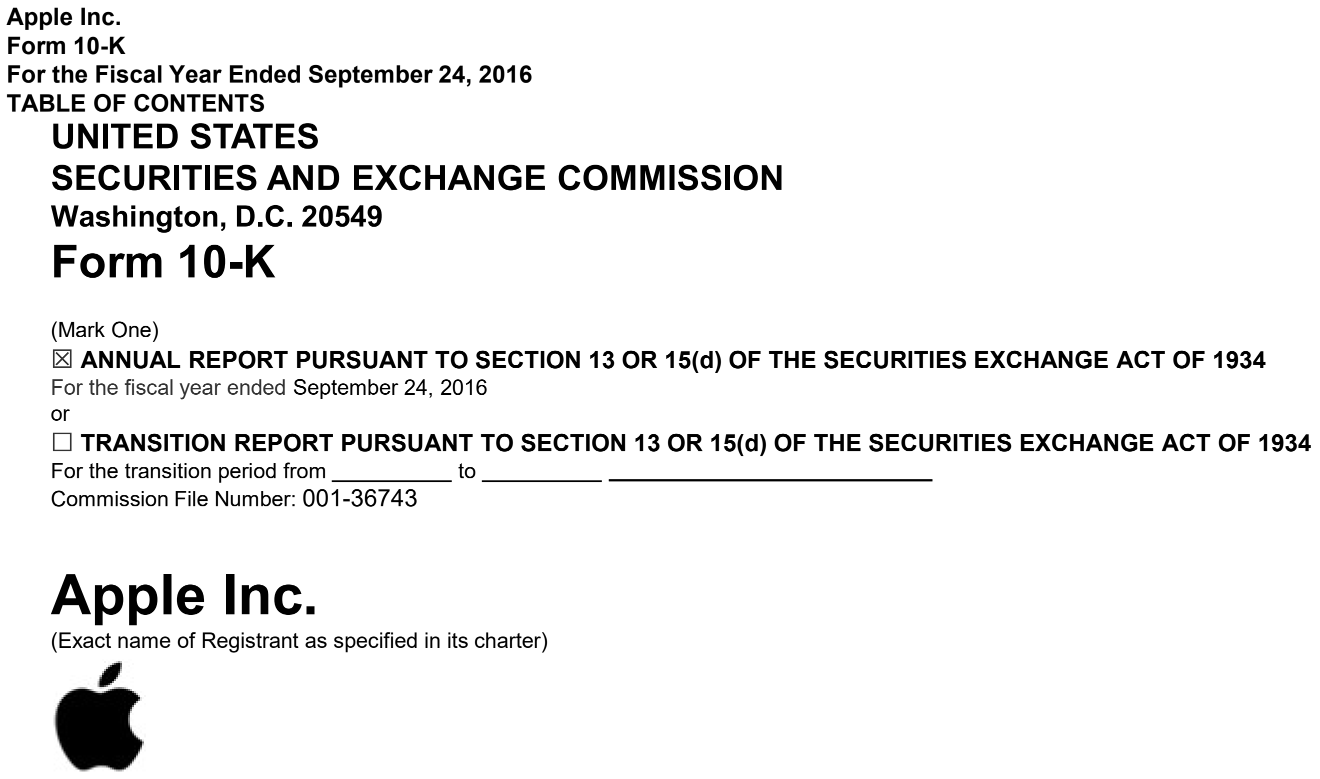
https://www.sec.gov/edgar/searchedgar/companysearch.html

Companies are required to report various activities to the SEC

Examples include forms like:

1. 10-Ks and 10-Qs
2. 13-F: institutional portfolio holdings
3. 3, 4, and 5: Insider trading

a. An example of an annual report



Apple Inc.
Form 10-K
For the Fiscal Year Ended September 24, 2016
TABLE OF CONTENTS

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

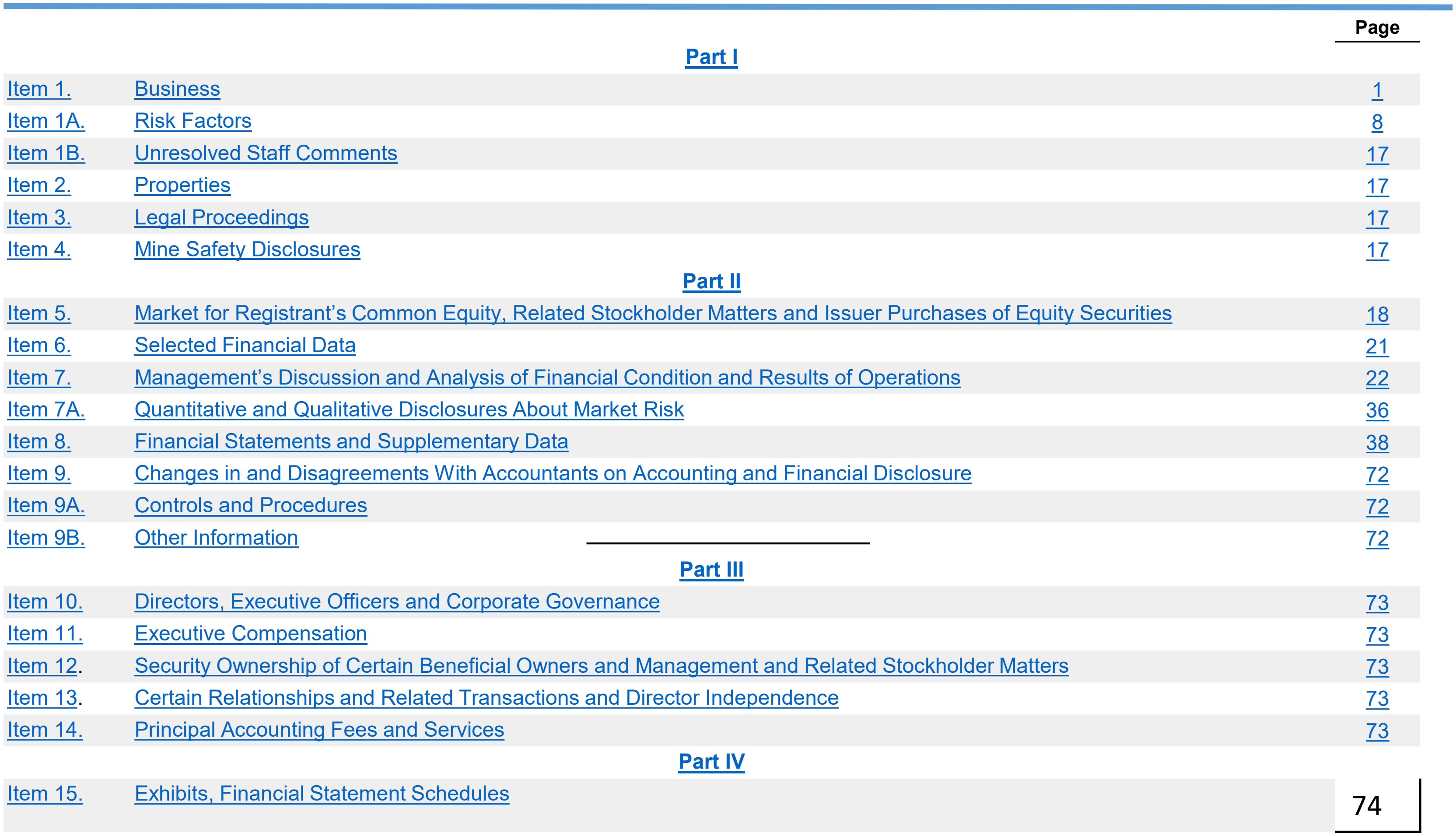
Form 10-K

(Mark One)

☒ **ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934**
For the fiscal year ended September 24, 2016
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 Number: 001-36743

Apple Inc.
(Exact name of Registrant as specified in its charter)

a. An example of an annual report



Page

Part I

Part II

Part III

Part IV

Note: over 70 pages of dense financial information. And that’s just one report…

b. AAPL 10-Ks and Homework 6

See Homework solution

c. Mapping to signal: Growth expectations

See homework 6, creating of sentiment index

d. It's all about the idea:
Text Similarity and Investor Inattention

- Linking text to trading strategies typically starts with a good idea (as opposed to blind data-mining)

- Investor inattention is a well-established behavioral bias
  - Intuitive: investors do not have mental capacity to keep track of all information and markets and so markets are not fully informationally efficient

- Example of inattention:
  - Predictable changes in health care sector revenues due to aging population not fully priced in stock returns, which lead to subsequent positive "alpha" for such stocks (Dellavigna and Pollet)
  - Anytime publicly available information is not impounded in prices as investors simply have not discovered the link

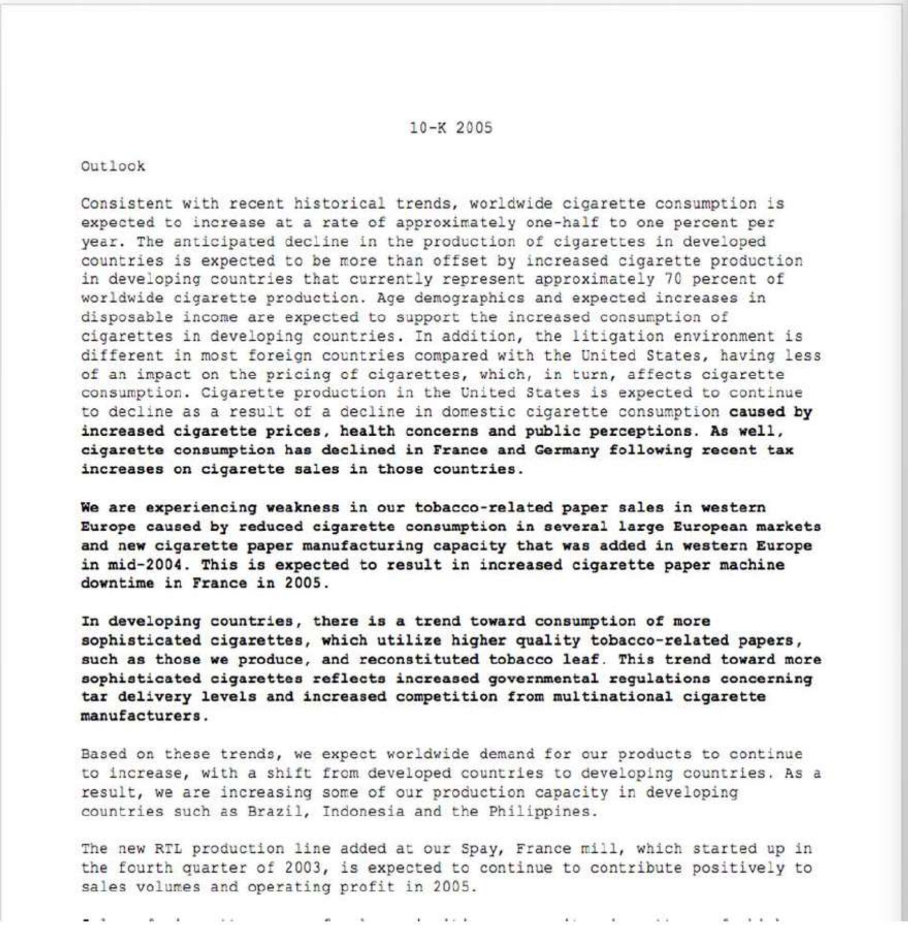
d. Lazy Prices

- Cohen, Malloy, and Ngyun (2015) argue that investors are inattentive to certain information in quarterly/annual reports
- In particular, firms typically repeat almost all information in their reports from their last report (copy-paste)
- Using data from EDGAR, 1994-2014, they show that active changes to the wording is associated with negative alpha of up to 22% p.a.
- The reporting changes are concentrated in the MD&A section (management discussion).
- Changes in language referring to executive team (CEO and CFO), or regarding litigation, are especially informative.
- Authors use textual analysis tools to execute their study

UCLAAnderson
SCHOOL of MANAGEMENT

d. Lazy Prices: Motivating Example

This table shows the first few paragraphs that are taken from Item 7, "Management's Discussion and Analysis of Financial Condition and Results of Operations", for Schweitzer-Mauduit International's (NYSE:SWM) 2004 and 2005 10-K reports. The new discussion in the 2005 10-K is highlighted.



10-K 2005

Outlook

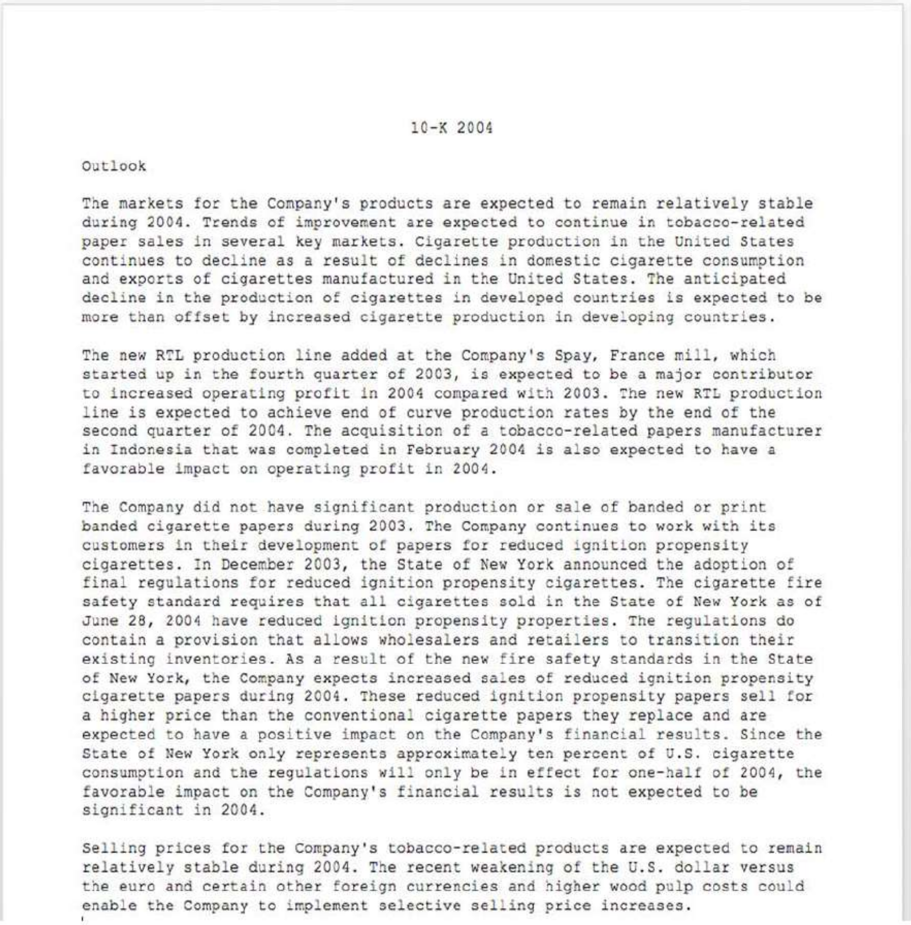
Consistent with recent historical trends, worldwide cigarette consumption is expected to increase at a rate of approximately one-half to one percent per year. The anticipated decline in the production of cigarettes in developed countries is expected to be more than offset by increased cigarette production in developing countries that currently represent approximately 70 percent of worldwide cigarette production. Age demographics and expected increases in disposable income are expected to support the increased consumption of cigarettes in developing countries. In addition, the litigation environment is different in most foreign countries compared with the United States, having less of an impact on the pricing of cigarettes, which, in turn, affects cigarette consumption. Cigarette production in the United States is expected to continue to decline as a result of a decline in domestic cigarette consumption **caused by increased cigarette prices, health concerns and public perceptions. As well, cigarette consumption has declined in France and Germany following recent tax increases on cigarette sales in those countries.**

We are experiencing weakness in our tobacco-related paper sales in western Europe caused by reduced cigarette consumption in several large European markets and new cigarette paper manufacturing capacity that was added in western Europe in mid-2004. This is expected to result in increased cigarette paper machine downtime in France in 2005.

In developing countries, there is a trend toward consumption of more sophisticated cigarettes, which utilize higher quality tobacco-related papers, such as those we produce, and reconstituted tobacco leaf. This trend toward more sophisticated cigarettes reflects increased governmental regulations concerning tar delivery levels and increased competition from multinational cigarette manufacturers.

Based on these trends, we expect worldwide demand for our products to continue to increase, with a shift from developed countries to developing countries. As a result, we are increasing some of our production capacity in developing countries such as Brazil, Indonesia and the Philippines.

The new RTL production line added at our Spay, France mill, which started up in the fourth quarter of 2003, is expected to continue to contribute positively to sales volumes and operating profit in 2005.



10-K 2004

Outlook

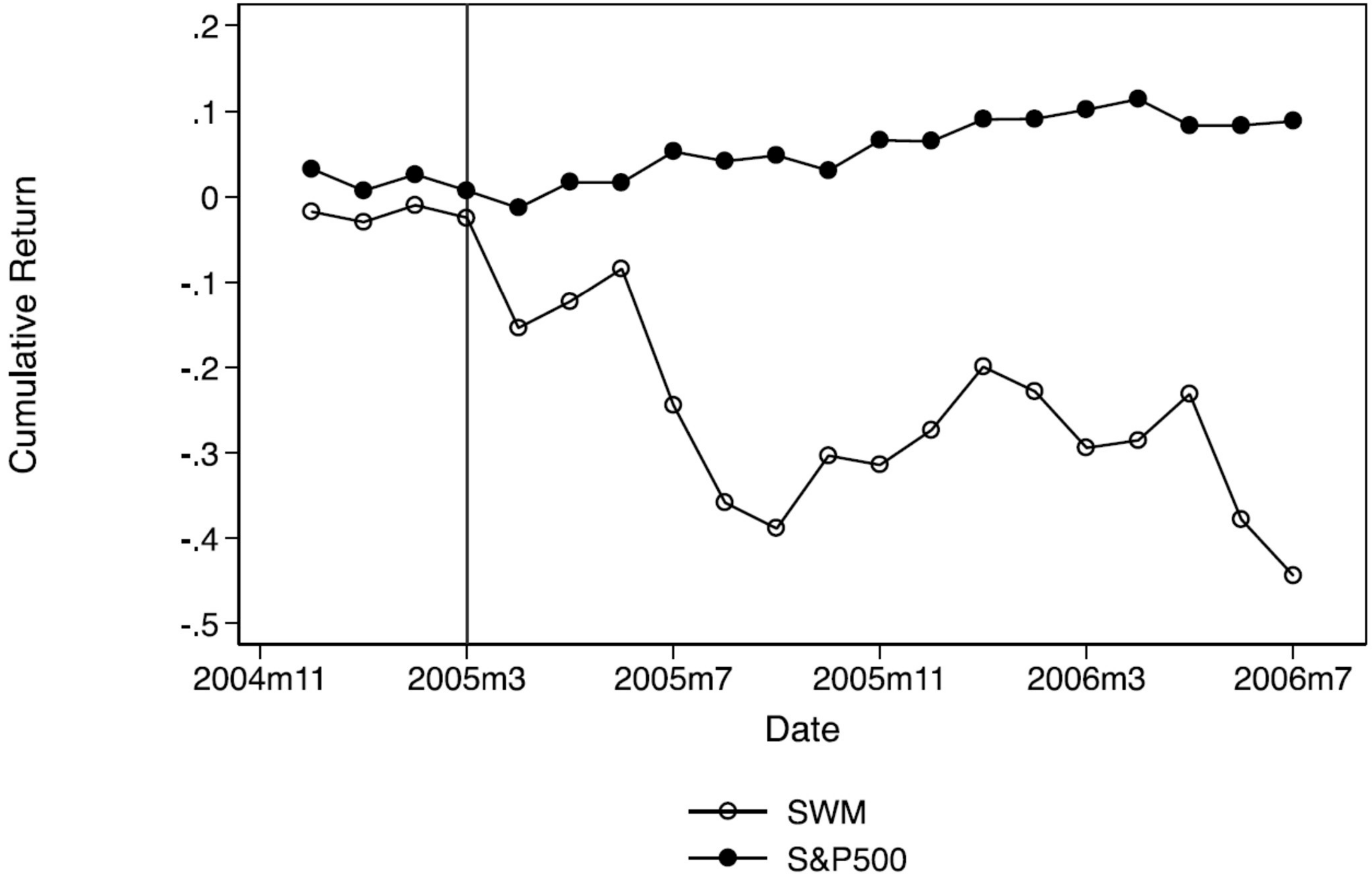
The markets for the Company's products are expected to remain relatively stable during 2004. Trends of improvement are expected to continue in tobacco-related paper sales in several key markets. Cigarette production in the United States continues to decline as a result of declines in domestic cigarette consumption and exports of cigarettes manufactured in the United States. The anticipated decline in the production of cigarettes in developed countries is expected to be more than offset by increased cigarette production in developing countries.

The new RTL production line added at the Company's Spay, France mill, which started up in the fourth quarter of 2003, is expected to be a major contributor to increased operating profit in 2004 compared with 2003. The new RTL production line is expected to achieve end of curve production rates by the end of the second quarter of 2004. The acquisition of a tobacco-related papers manufacturer in Indonesia that was completed in February 2004 is also expected to have a favorable impact on operating profit in 2004.

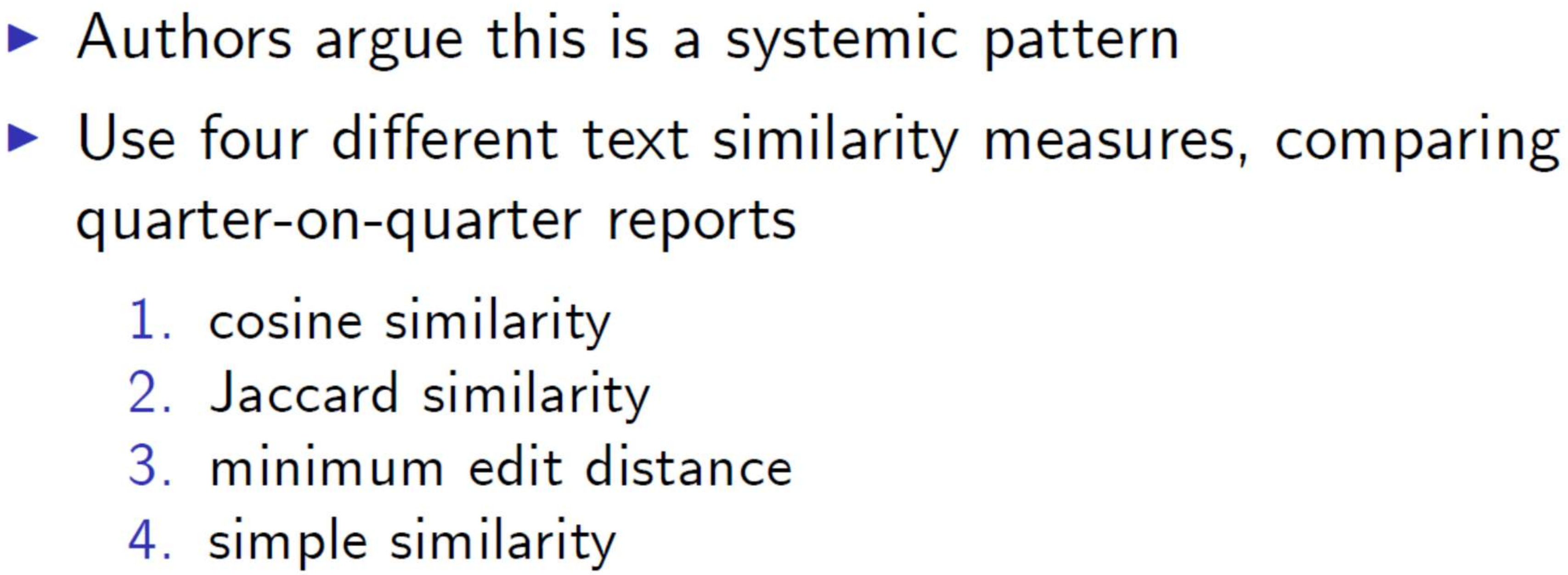
The Company did not have significant production or sale of banded or print banded cigarette papers during 2003. The Company continues to work with its customers in their development of papers for reduced ignition propensity cigarettes. In December 2003, the State of New York announced the adoption of final regulations for reduced ignition propensity cigarettes. The cigarette fire safety standard requires that all cigarettes sold in the State of New York as of June 28, 2004 have reduced ignition propensity properties. The regulations do contain a provision that allows wholesalers and retailers to transition their existing inventories. As a result of the new fire safety standards in the State of New York, the Company expects increased sales of reduced ignition propensity cigarette papers during 2004. These reduced ignition propensity papers sell for a higher price than the conventional cigarette papers they replace and are expected to have a positive impact on the Company's financial results. Since the State of New York only represents approximately ten percent of U.S. cigarette consumption and the regulations will only be in effect for one-half of 2004, the favorable impact on the Company's financial results is not expected to be significant in 2004.

Selling prices for the Company's tobacco-related products are expected to remain relatively stable during 2004. The recent weakening of the U.S. dollar versus the euro and certain other foreign currencies and higher wood pulp costs could enable the Company to implement selective selling price increases.

d. Lazy Prices: Motivating Example



d. Text Similarity



- Authors argue this is a systemic pattern
- Use four different text similarity measures, comparing quarter-on-quarter reports
  1. cosine similarity
  2. Jaccard similarity
  3. minimum edit distance
  4. simple similarity

d. Cosine Similarity

- Let D_{S1} and D_{S2} be the set of terms in documents D_1 and D_2, respectively
- Define T as the union of words in D_{S1} and D_{S2} and let t_i be the i'th element of T
- Define the *term frequency vectors*

$$\begin{aligned} D_1^{TF} &= \left[nD_1 \left(t_1 \right), nD_1 \left(t_2 \right), ..., nD_1 \left(t_N \right) \right], \\ D_2^{TF} &= \left[nD_2 \left(t_1 \right), nD_2 \left(t_2 \right), ..., nD_2 \left(t_N \right) \right] \end{aligned}$$

where $nD_j \left(t_i \right)$ is the number of occurences of term t_i in document D_j

- Cosine similarity is defined as

$$\text{Sim\_Cosine} = \frac{\left(D_1^{TF} \right)' D_2^{TF}}{\left\| D_1^{TF} \right\| \times \left\| D_2^{TF} \right\|}$$

d. Cosine Similarity (cont'd)

- To see how this works, consider the following (short) documents

$$\begin{aligned} D_A &: \quad \text{We expect demand to increase} \\ D_B &: \quad \text{We expect worldwide demand to increase} \\ D_C &: \quad \text{We expect weakness in sales} \end{aligned}$$

- Clearly, D_A and D_B are more similar than, say, D_A and D_C
- Let's calculate the cosine measure between A and B:

$$T\left(D_A, D_B\right) = [\text{we, expect, worldwide, demand, to, increase}]$$

d. Cosine Similarity (cont'd)

- From last slide

$$T\left(D_A, D_B\right) = [\text{we, expect, worldwide, demand, to, increase}]$$

- Then

$$\begin{aligned} D_1^{TF} &= [1, 1, 0, 1, 1, 1] \\ D_2^{TF} &= [1, 1, 1, 1, 1, 1] \end{aligned}$$

- So cosine similarity is:

$$\frac{\left(D_1^{TF}\right)' D_2^{TF}}{\|D_1^{TF}\| \times \|D_2^{TF}\|}$$

$$\begin{aligned} &= \frac{1 \times 1 + 1 \times 1 + 0 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 1}{\sqrt{1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2} \times \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2}} \\ &= 0.91 \end{aligned}$$

UCLA Anderson
SCHOOL of MANAGEMENT

d. Cosine Similarity (cont'd)

- Next, let's do D_A and D_C

  $T\left(D_A, D_C\right) = [\text{we, expect, demand, to, increase, weakness, in, sales}]$

- Then

$$\begin{aligned} D_1^{TF} &= [1,1,1,1,1,0,0,0] \\ D_2^{TF} &= [1,1,0,0,0,1,1,1] \end{aligned}$$

- And so:

$$\frac{\left(D_1^{TF}\right)' D_2^{TF}}{\|D_1^{TF}\| \times \|D_2^{TF}\|}$$

$$\begin{aligned} &= \frac{1\times 1 + 1\times 1 + 1\times 0 + 1\times 0 + 1\times 0 + 0\times 1 + 0\times 1 + 0\times 1}{\sqrt{1^2+1^2+1^2+1^2+1^2} \times \sqrt{1^2+1^2+1^2+1^2+1^2}} \\ &= 0.40 \end{aligned}$$

d. Jaccard Similarity

- Definition:

$$\text{Sim\_Jaccard} = \left| D_1^{TF} \cap D_2^{TF} \right| / \left| D_1^{TF} \cup D_2^{TF} \right|$$

  Interaction divided by size of union

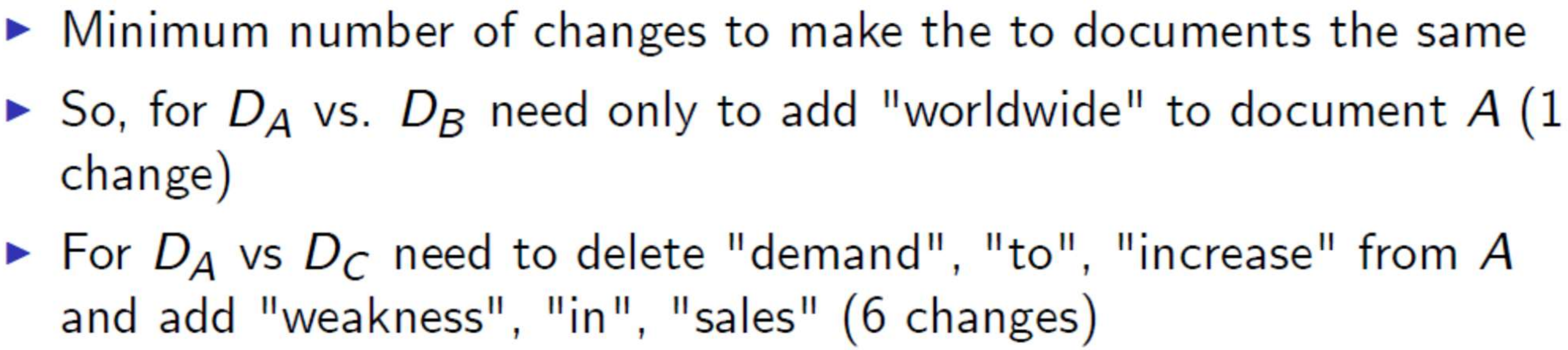
- Then

$$\begin{aligned} \text{Sim\_Jaccard}\,(D_A, D_B) &= |\{\text{we,expect,demand,to,increase}\}| \, / \\ &\quad |(\text{we,expect,worldwide,demand,to,increase})| \\ &= 5/6 \end{aligned}$$

- And

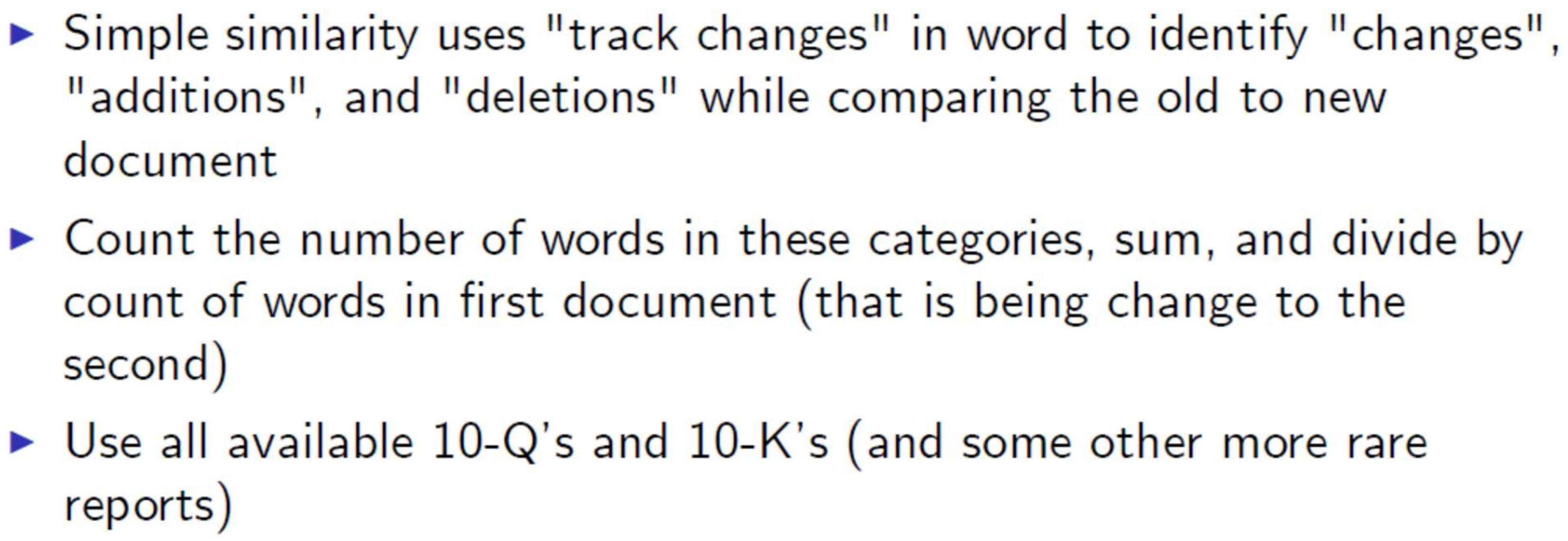
$$\begin{aligned} \text{Sim\_Jaccard}\,(D_A, D_C) &= |\{\text{we,expect}\}| \, / \\ &\quad |(\text{we,expect,demand,to,increase,weakness,in,s} \\ &= 2/8 \end{aligned}$$

d. Min-Edit Similarity



- Minimum number of changes to make the to documents the same
- So, for D_A vs. D_B need only to add "worldwide" to document A (1 change)
- For D_A vs D_C need to delete "demand", "to", "increase" from A and add "weakness", "in", "sales" (6 changes)

d. Simple Similarity



- Simple similarity uses "track changes" in word to identify "changes", "additions", and "deletions" while comparing the old to new document
- Count the number of words in these categories, sum, and divide by count of words in first document (that is being change to the second)
- Use all available 10-Q's and 10-K's (and some other more rare reports)

d. Results: FMB regressions 1994 - 2014

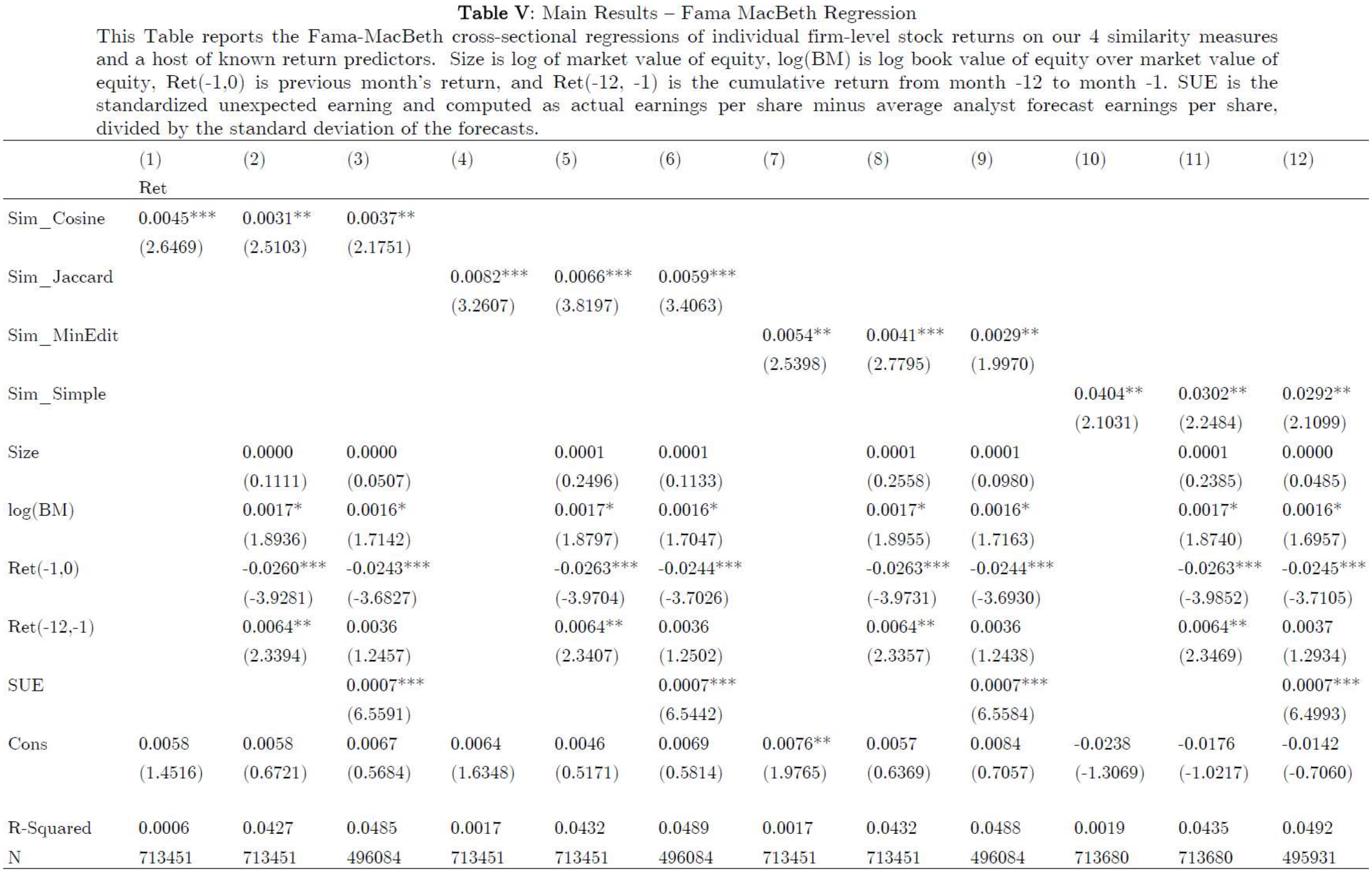


Table V: Main Results – Fama MacBeth Regression

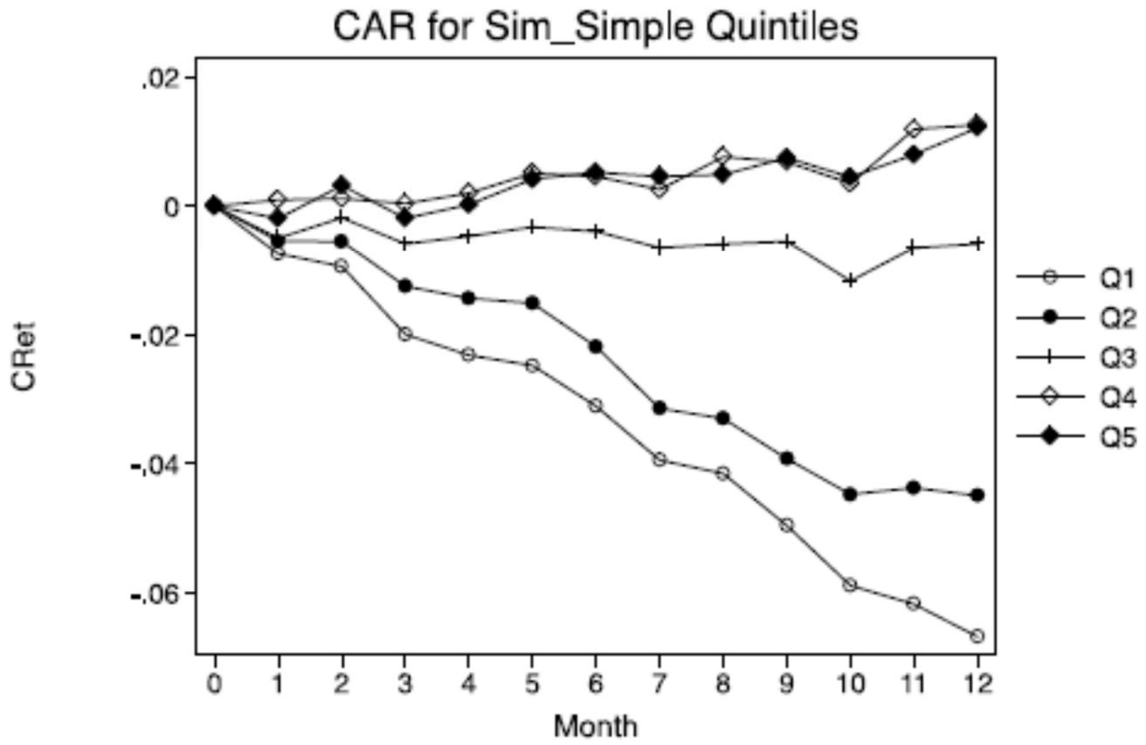
This Table reports the Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our 4 similarity measures and a host of known return predictors. Size is log of market value of equity, log(BM) is log book value of equity over market value of equity, Ret(-1,0) is previous month's return, and Ret(-12, -1) is the cumulative return from month -12 to month -1. SUE is the standardized unexpected earning and computed as actual earnings per share minus average analyst forecast earnings per share, divided by the standard deviation of the forecasts.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Ret											
Sim_Cosine	0.0045***	0.0031**	0.0037**									
	(2.6469)	(2.5103)	(2.1751)									
Sim_Jaccard				0.0082***	0.0066***	0.0059***						
				(3.2607)	(3.8197)	(3.4063)						
Sim_MinEdit							0.0054**	0.0041***	0.0029**			
							(2.5398)	(2.7795)	(1.9970)			
Sim_Simple										0.0404**	0.0302**	0.0292**
										(2.1031)	(2.2484)	(2.1099)
Size		0.0000	0.0000		0.0001	0.0001		0.0001	0.0001		0.0001	0.0000
		(0.1111)	(0.0507)		(0.2496)	(0.1133)		(0.2558)	(0.0980)		(0.2385)	(0.0485)
log(BM)		0.0017*	0.0016*		0.0017*	0.0016*		0.0017*	0.0016*		0.0017*	0.0016*
		(1.8936)	(1.7142)		(1.8797)	(1.7047)		(1.8955)	(1.7163)		(1.8740)	(1.6957)
Ret(-1,0)		-0.0260***	-0.0243***		-0.0263***	-0.0244***		-0.0263***	-0.0244***		-0.0263***	-0.0245***
		(-3.9281)	(-3.6827)		(-3.9704)	(-3.7026)		(-3.9731)	(-3.6930)		(-3.9852)	(-3.7105)
Ret(-12,-1)		0.0064**	0.0036		0.0064**	0.0036		0.0064**	0.0036		0.0064**	0.0037
		(2.3394)	(1.2457)		(2.3407)	(1.2502)		(2.3357)	(1.2438)		(2.3469)	(1.2934)
SUE			0.0007***			0.0007***			0.0007***			0.0007***
			(6.5591)			(6.5442)			(6.5584)			(6.4993)
Cons	0.0058	0.0058	0.0067	0.0064	0.0046	0.0069	0.0076**	0.0057	0.0084	-0.0238	-0.0176	-0.0142
	(1.4516)	(0.6721)	(0.5684)	(1.6348)	(0.5171)	(0.5814)	(1.9765)	(0.6369)	(0.7057)	(-1.3069)	(-1.0217)	(-0.7060)
R-Squared	0.0006	0.0427	0.0485	0.0017	0.0432	0.0489	0.0017	0.0432	0.0488	0.0019	0.0435	0.0492
N	713451	713451	496084	713451	713451	496084	713451	713451	496084	713680	713680	495931

d. Results: Quintile portfolio returns

- No reversal
- Effect is that on average changes are bad news

This figure shows the average cumulative abnormal return for each quintile portfolio sorted based on firms' similarity score, for 1 month to 12 months after portfolio formation.



UCLAAnderson
SCHOOL of MANAGEMENT

d. Which sections are important?

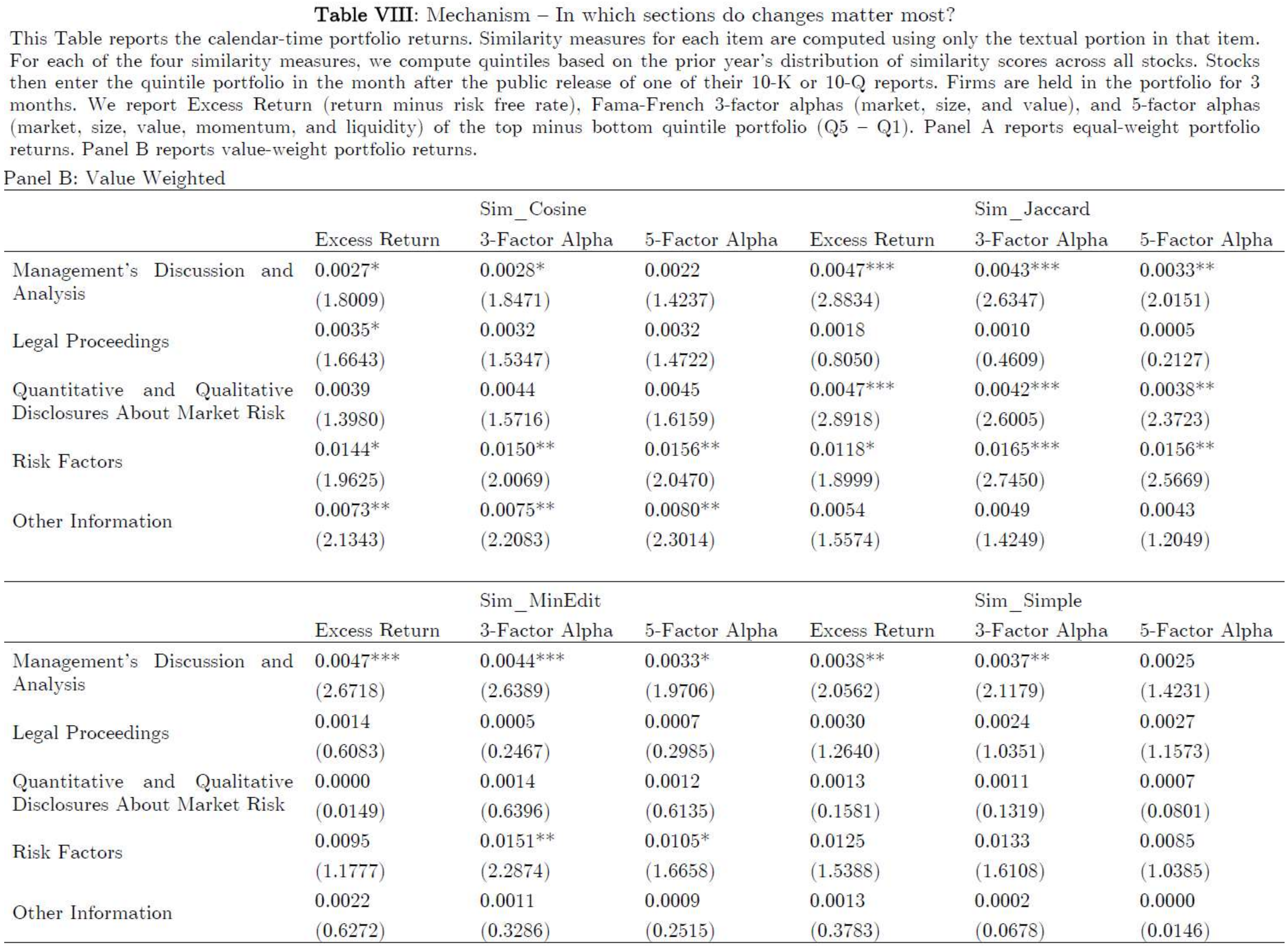


Table VIII: Mechanism – In which sections do changes matter most?

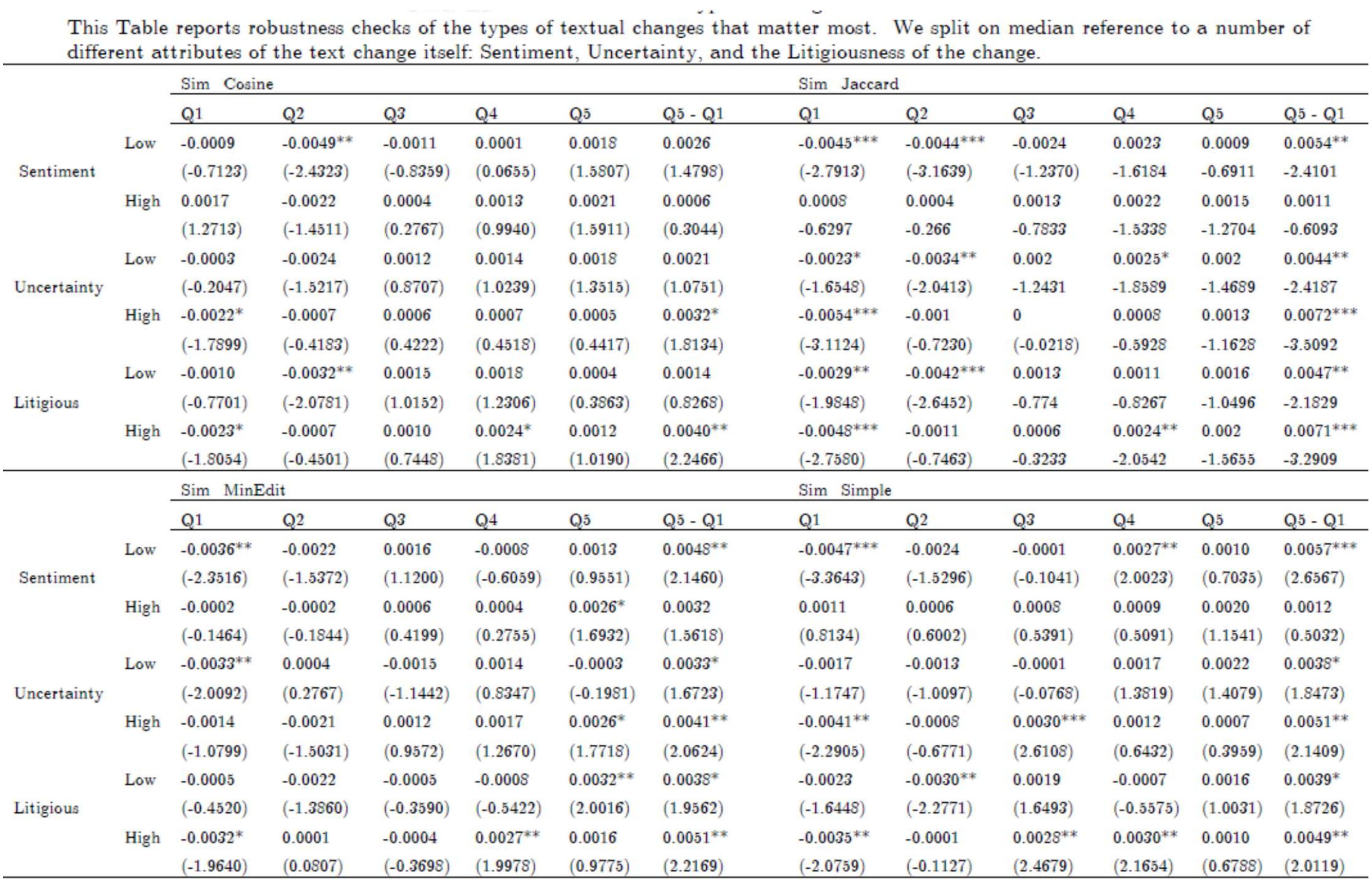
This Table reports the calendar-time portfolio returns. Similarity measures for each item are computed using only the textual portion in that item. For each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity scores across all stocks. Stocks then enter the quintile portfolio in the month after the public release of one of their 10-K or 10-Q reports. Firms are held in the portfolio for 3 months. We report Excess Return (return minus risk free rate), Fama-French 3-factor alphas (market, size, and value), and 5-factor alphas (market, size, value, momentum, and liquidity) of the top minus bottom quintile portfolio (Q5 – Q1). Panel A reports equal-weight portfolio returns. Panel B reports value-weight portfolio returns.

Panel B: Value Weighted

		Sim_Cosine			Sim_Jaccard	
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0027*	0.0028*	0.0022	0.0047***	0.0043***	0.0033**
	(1.8009)	(1.8471)	(1.4237)	(2.8834)	(2.6347)	(2.0151)
Legal Proceedings	0.0035*	0.0032	0.0032	0.0018	0.0010	0.0005
	(1.6643)	(1.5347)	(1.4722)	(0.8050)	(0.4609)	(0.2127)
Quantitative and Qualitative Disclosures About Market Risk	0.0039	0.0044	0.0045	0.0047***	0.0042***	0.0038**
	(1.3980)	(1.5716)	(1.6159)	(2.8918)	(2.6005)	(2.3723)
Risk Factors	0.0144*	0.0150**	0.0156**	0.0118*	0.0165***	0.0156**
	(1.9625)	(2.0069)	(2.0470)	(1.8999)	(2.7450)	(2.5669)
Other Information	0.0073**	0.0075**	0.0080**	0.0054	0.0049	0.0043
	(2.1343)	(2.2083)	(2.3014)	(1.5574)	(1.4249)	(1.2049)

		Sim_MinEdit			Sim_Simple	
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0047***	0.0044***	0.0033*	0.0038**	0.0037**	0.0025
	(2.6718)	(2.6389)	(1.9706)	(2.0562)	(2.1179)	(1.4231)
Legal Proceedings	0.0014	0.0005	0.0007	0.0030	0.0024	0.0027
	(0.6083)	(0.2467)	(0.2985)	(1.2640)	(1.0351)	(1.1573)
Quantitative and Qualitative Disclosures About Market Risk	0.0000	0.0014	0.0012	0.0013	0.0011	0.0007
	(0.0149)	(0.6396)	(0.6135)	(0.1581)	(0.1319)	(0.0801)
Risk Factors	0.0095	0.0151**	0.0105*	0.0125	0.0133	0.0085
	(1.1777)	(2.2874)	(1.6658)	(1.5388)	(1.6108)	(1.0385)
Other Information	0.0022	0.0011	0.0009	0.0013	0.0002	0.0000
	(0.6272)	(0.3286)	(0.2515)	(0.3783)	(0.0678)	(0.0146)

d. What type of language changes are most important?



This Table reports robustness checks of the types of textual changes that matter most. We split on median reference to a number of different attributes of the text change itself: Sentiment, Uncertainty, and the Litigiousness of the change.

		Sim Cosine						Sim Jaccard					
		Q1	Q2	Q3	Q4	Q5	Q5 - Q1	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Sentiment	Low	-0.0009	-0.0049**	-0.0011	0.0001	0.0018	0.0026	-0.0045***	-0.0044***	-0.0024	0.0023	0.0009	0.0054**
		(-0.7123)	(-2.4323)	(-0.8359)	(0.0655)	(1.5807)	(1.4798)	(-2.7913)	(-3.1639)	(-1.2370)	-1.6184	-0.6911	-2.4101
	High	0.0017	-0.0022	0.0004	0.0013	0.0021	0.0006	0.0008	0.0004	0.0013	0.0022	0.0015	0.0011
		(1.2713)	(-1.4511)	(0.2767)	(0.9940)	(1.5911)	(0.3044)	-0.6297	-0.266	-0.7833	-1.5338	-1.2704	-0.6093
Uncertainty	Low	-0.0003	-0.0024	0.0012	0.0014	0.0018	0.0021	-0.0023*	-0.0034**	0.002	0.0025*	0.002	0.0044**
		(-0.2047)	(-1.5217)	(0.8707)	(1.0239)	(1.3515)	(1.0751)	(-1.6548)	(-2.0413)	-1.2431	-1.8589	-1.4689	-2.4187
	High	-0.0022*	-0.0007	0.0006	0.0007	0.0005	0.0032*	-0.0054***	-0.001	0	0.0008	0.0013	0.0072***
		(-1.7899)	(-0.4183)	(0.4222)	(0.4518)	(0.4417)	(1.8134)	(-3.1124)	(-0.7230)	(-0.0218)	-0.5928	-1.1628	-3.5092
Litigious	Low	-0.0010	-0.0032**	0.0015	0.0018	0.0004	0.0014	-0.0029**	-0.0042***	0.0013	0.0011	0.0016	0.0047**
		(-0.7701)	(-2.0781)	(1.0152)	(1.2306)	(0.3863)	(0.8268)	(-1.9848)	(-2.6452)	-0.774	-0.8267	-1.0496	-2.1829
	High	-0.0023*	-0.0007	0.0010	0.0024*	0.0012	0.0040**	-0.0048***	-0.0011	0.0006	0.0024**	0.002	0.0071***
		(-1.8054)	(-0.4501)	(0.7448)	(1.8381)	(1.0190)	(2.2466)	(-2.7580)	(-0.7463)	-0.3233	-2.0542	-1.5655	-3.2909

		Sim MinEdit						Sim Simple					
		Q1	Q2	Q3	Q4	Q5	Q5 - Q1	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Sentiment	Low	-0.0036**	-0.0022	0.0016	-0.0008	0.0013	0.0048**	-0.0047***	-0.0024	-0.0001	0.0027**	0.0010	0.0057***
		(-2.3516)	(-1.5372)	(1.1200)	(-0.6059)	(0.9551)	(2.1460)	(-3.3643)	(-1.5296)	(-0.1041)	(2.0023)	(0.7035)	(2.6567)
	High	-0.0002	-0.0002	0.0006	0.0004	0.0026*	0.0032	0.0011	0.0006	0.0008	0.0009	0.0020	0.0012
		(-0.1464)	(-0.1844)	(0.4199)	(0.2755)	(1.6932)	(1.5618)	(0.8134)	(0.6002)	(0.5391)	(0.5091)	(1.1541)	(0.5032)
Uncertainty	Low	-0.0033**	0.0004	-0.0015	0.0014	-0.0003	0.0033*	-0.0017	-0.0013	-0.0001	0.0017	0.0022	0.0038*
		(-2.0092)	(0.2767)	(-1.1442)	(0.8347)	(-0.1981)	(1.6723)	(-1.1747)	(-1.0097)	(-0.0768)	(1.3819)	(1.4079)	(1.8473)
	High	-0.0014	-0.0021	0.0012	0.0017	0.0026*	0.0041**	-0.0041**	-0.0008	0.0030***	0.0012	0.0007	0.0051**
		(-1.0799)	(-1.5031)	(0.9572)	(1.2670)	(1.7718)	(2.0624)	(-2.2905)	(-0.6771)	(2.6108)	(0.6432)	(0.3959)	(2.1409)
Litigious	Low	-0.0005	-0.0022	-0.0005	-0.0008	0.0032**	0.0038*	-0.0023	-0.0030**	0.0019	-0.0007	0.0016	0.0039*
		(-0.4520)	(-1.3860)	(-0.3590)	(-0.5422)	(2.0016)	(1.9562)	(-1.6448)	(-2.2771)	(1.6493)	(-0.5575)	(1.0031)	(1.8726)
	High	-0.0032*	0.0001	-0.0004	0.0027**	0.0016	0.0051**	-0.0035**	-0.0001	0.0028**	0.0030**	0.0010	0.0049**
		(-1.9640)	(0.0807)	(-0.3698)	(1.9978)	(0.9775)	(2.2169)	(-2.0759)	(-0.1127)	(2.4679)	(2.1654)	(0.6788)	(2.0119)

d. Big Data and EDGAR Web Crawling

- An important skill is the ability to obtain new (non-standard) data efficiently and execute your idea

- In “Code Snippets 7 – Edgar Download.py”, I provide code for downloading data directly from EDGAR using the sec_edgar_download package. Such automated downloads enable you to do large-scale textual analysis and is often used by, e.g., hedge funds

```
from sec_edgar_downloader                import Downloader

# Initialize a downloader instance. If no argument is passed
# to the constructor, the package will download filings to
# the current working directory.
dl = Downloader("/Documents/Repos/Lars/data/data")

# Example: get the 20 most recent 10-K filings for Apple
dl.get("10-K", "AAPL", amount=20)
```

e. Text Topic Modelling

Text Topic Modelling is an unsupervised learning algorithm that tries to extract the main topics from a set of underlying documents (the Corpus)

Suppose you want to classify documents according to the relevance of what you want to analyze

- For instance, you are interested in newspaper articles related to investments and growth of business
- But, you don't want to pre-specify words that define such documents either (a) because you are not sure exactly what words that would be and/or (b) because you don't want to bias the analysis with your prior views

LDA helps you extract the text topics that best 'fits' your corpus

- With the topics in hand, you can then perform classification of each document

e. Latent Dirichlet Allocation (LDA): Intuition

- Suppose you have the following set of sentences:
  - I like to eat broccoli and bananas.
  - I ate a banana and spinach smoothie for breakfast.
  - Chinchillas and kittens are cute.
  - My sister adopted a kitten yesterday.
  - Look at this cute hamster munching on a piece of broccoli.

- What is latent Dirichlet allocation? It's a way of automatically discovering ***topics*** that these sentences contain. For example, given these sentences and asked for 2 topics, LDA might produce something like
  - **Sentences 1 and 2**: 100% Topic A
  - **Sentences 3 and 4**: 100% Topic B
  - **Sentence 5**: 60% Topic A, 40% Topic B
  - **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
  - **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

The question, of course, is: how does LDA perform this discovery?

e. The LDA Model

- In more detail, LDA represents documents as ***mixtures of topics*** that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document, you:

1. Decide on the number of words N the document will have (say, according to a Poisson distribution).

2. Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics). For example, assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

3. Generate each word *w_i* in the document by:
   i. First picking a topic (according to the multinomial distribution that you sampled above; for example, you might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability).
   ii. Using the topic to generate the word itself (according to the topic's multinomial distribution). For example, if we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on.

4. Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

e. The LDA Model: Example

Consider the following example:

- According to the above process, when generating some particular document D, you might

    1. Pick 5 to be the number of words in D.
    2. Decide that D will be 1/2 about food and 1/2 about cute animals.
    3. Pick the first word to come from the food topic, which then gives you the word “broccoli”.
    4. Pick the second word to come from the cute animals topic, which gives you “panda”.
    5. Pick the third word to come from the cute animals topic, giving you “adorable”.
    6. Pick the fourth word to come from the food topic, giving you “cherries”.
    7. Pick the fifth word to come from the food topic, giving you “eating”.

- So the document generated under the LDA model will be “broccoli panda adorable cherries eating”

UCLAAnderson
SCHOOL *of* MANAGEMENT

e. The LDA Model

- In the LDA model, we use Bayesian methods to estimate
- The priors over topics and words within each topic obey the Dirichlet distribution:
  - We use the symmetric version of this distribution (α the same across *x*'s)
  - Think of the below *x*'s (words, topics) as probabilities. The *x*'s are between zero and one and sum to one..

$$f(x_1, \dots, x_K; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha-1}, \quad B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha)}{\Gamma\left(\prod_{i=1}^{K} \alpha\right)}$$

- The prior distribution over topics is governed by α. The higher α is, the more likely each document is to contain a mixture of most of the topics instead of any single topic.
- The prior distribution of words within each topic is also governed by a Dirichlet distribution as above, but to distinguish we here instead denote alpha as η
  - A higher value of η denotes that each topic is likely to contain a mixture of most of the words and not any word specifically.
- Finally, we update beliefs based on Bayes' rule:

$$Posterior = \frac{likelihood \times prior}{marginal\ likelihood}$$

UCLAAnderson
SCHOOL of MANAGEMENT

e. The LDA Model: Learning

- So now suppose you have a set of documents. You've chosen some fixed number of K topics to discover, and want to use LDA to learn the topic representation of each document and the words associated to each topic. How do you do this? One way (known as collapsed Gibbs sampling) is the following:

- Go through each document, randomly assign each word in the document to one of the K topics.

- Notice that this random assignment already gives you both topic representations of all the documents and word distributions of all the topics (albeit not very good ones).

- So to improve on them, for each document *d*…
  - Go through each word *w* in *d*…
    - And for each topic t, compute two things: 1) p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and 2) p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w. Reassign *w* a new topic, where we choose topic t with probability p(topic t | document d) * p(word w | topic t) (according to our generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current word's topic with this probability). (Also, I'm glossing over a couple of things here, in particular the use of priors in these probabilities.)
    - In other words, in this step, we're assuming that all topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how documents are generated.

- After repeating the previous step a large number of times, you'll eventually reach a rough steady state where your assignments are pretty good. So use these assignments to estimate the topic mixtures of each document (by counting the proportion of words assigned to each topic within that document) and the words associated to each topic (by counting the proportion of words assigned to each topic overall).

e. The LDA Model: Worked example

- The full background maths of the procedure is in "LDA White Paper", posted on BruinLearn

- Let's consider the FOMC minutes
  - Federal Open Market Committee
  - Sets federal funds rate (short term interest rate, effectively)
  - https://www.federalreserve.gov/monetarypolicy/fomc.htm

- Let's find the main topics in these minutes

- We will use the additional Python package "gensim"
  - https://radimrehurek.com/gensim/

UCLAAnderson
SCHOOL of MANAGEMENT

e. Reading in Fed Minutes

- The zipped folder FOMC_minutes.zip has the minutes for FOMC meetings in 2007 through 2010.
- Alternatively, one can scrape Federal Reserve data directly using https://pypi.org/project/Fedtools/
- Example code (partial, full code on BruinLearn):

```
from FedTools import FederalReserveMins

fed_mins = FederalReserveMins(
            main_url = 'https://www.federalreserve.gov',
            calendar_url ='https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm',
            start_year = 1995,
            historical_split = 2010,
            verbose = True,
            thread_num = 10)
```

UCLAAnderson
SCHOOL of MANAGEMENT

e. Cleaning Fed Minutes

- We clean the data using standard methods (StopWords, Lemmatization, Tokenization, etc.) See BruinLearn for code

- We save the data in two pickle files:
  - local_fomc_df.pkl for the manually downloaded
  - server_fomc_df.pkl for downloaded using FedTools

- We also take out a custom set of Stop Words:

```
# Words that appear in minutes frequently without carrying significant contributions in understanding context, I remove
them manually.
custom_stopwords = {'federal', 'committee', 'open', 'would', 'could', 'also', 'somewhat', 'however', 'month',
'year', 'quarter', 'participant', 'm1', 'm2', 'first', 'second', 'statement',
'twelve', 'tuesday', 'wednesday', 'unanimous', 'unanimously', 'percentage', 'vote', 'take', 'thus', 'far', 'otherwise',
'voted', 'minute', 'accordance', 'conclusion', 'discussion', 'notation', 'manager', 'back', 'instructed', 'agreed',
'encompassed'}
```

e. LDA model estimation

- As usual, many tuning parameters!
  - Look up documentation in genism and code on BruinLearn
  - Note: this is a “guided” LDA, we pay more attention to certain words: purchase, asset, inflation, growth
  - Generally, this is quite helpful if you are looking for certain topics to be identified

```
# Setting custom eta for keywords
base_eta = 0.1 # Normal eta value for most words

# Base eta with a low default value for all bi, trigrams
custom_eta = np.full(len(dictionary), base_eta)

# Define words to emphasize in multigrams and their weights
key_words_emphasis = {'asset': 10, 'purchase': 20} # Emphasize 'purchase' but not 'repurchase'
other_key_words = {'inflation': 5, 'growth': 5} # Other keywords to emphasize

# Update eta for bigrams containing key words
for token_id, token in dictionary.iteritems():
if 'repurchase' not in token: # Exclude bigrams containing 'repurchase'
for key_word, weight in key_words_emphasis.items():
if key_word in token:
custom_eta[token_id] = weight
for key_word, weight in other_key_words.items():
if key_word in token:
custom_eta[token_id] = weight

# Create the LDA model
lda_model = models.LdaModel(corpus, id2word=dictionary, passes=passes, iterations=iterations,
num_topics=num_topics, eta=custom_eta, alpha=alpha, random_state=200)
```

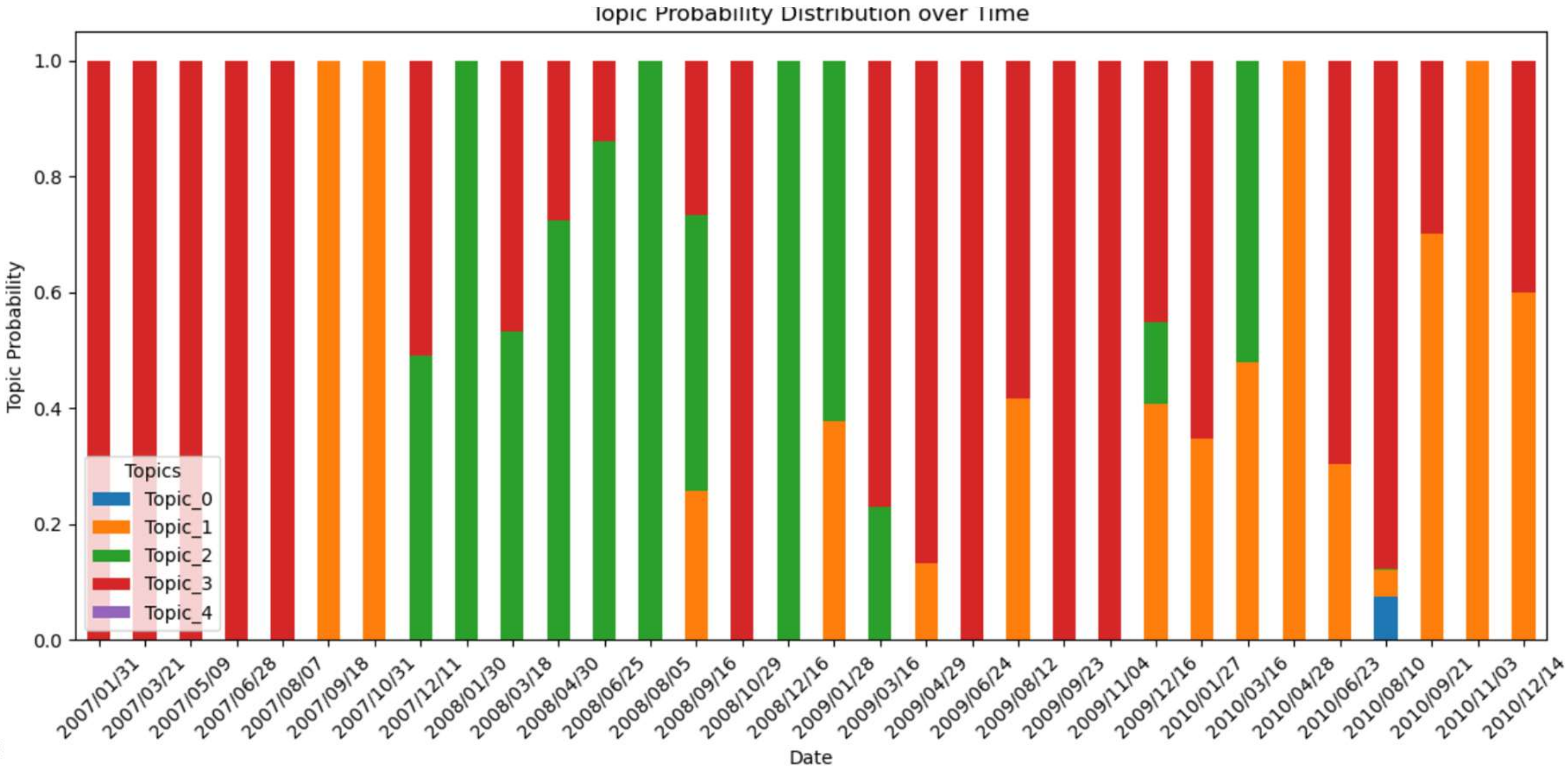
e. Estimated topics and probabilities

- A growth and inflation topic, an asset purchase topic, two less clear topics but related to nominal versus real yield curve and technicalities on asset purchases

```
# Evaluate the topics
topics = lda_model.print_topics(num_words=num_words)
for topic in topics:
print(topic)
(0, '0.012*"outright_purchase" + 0.012*"security_may_purchased"+0.012*"time_purchase_sell"')
(1, '0.006*"economic_activity" + 0.005*"treasury_security" + 0.004*"asset_purchase"')
(2, '0.004*"unemployment_rate" + 0.004*"interest_rate" + 0.004*"inflation_expectation"')
(3, '0.007*"financial_market" + 0.007*"economic_growth" + 0.005*"economic_activity"')
(4, '0.012*"asset_purchase" + 0.012*"purchase_program" + 0.012*"purchase_large_quantity"')
```

e. Time series of topic probabilities

- Plot time series of probabilities of a given FOMC minutes belonging to each topic
  - Topic 3 dominates early (red: economic growth), then as unemployment starts to tick up it focuses on Topic 2 (green: unemployment and inflation), Topic 1 (orange: asset purchases show up in fall of 2008 and then quite a bit after that as asset purchases increase)



e. Epilogue

- It is often useful to classify data into a lower dimensional space
  - E.g., Principal Components Analysis, text topics
- This can help make sense of what otherwise may look like just noise
- Data-cleaning is always very important
- Also, you will need to work on the exact topic specifications to get useful results.
  - You have to ‘get your hands dirty’ with whatever data you are analyzing
- With persistence, building a carefully tuned model along these lines can be a useful addition to quantitative analysis