國立政治大學企業管理研究所(MBA 學位學程)

碩士學位論文口試

美國聯準會會議紀要的文字分析

Text Mining on FOMC Minutes

指導教授：余清祥 博士

研究生：郭育丞

2021/07/08

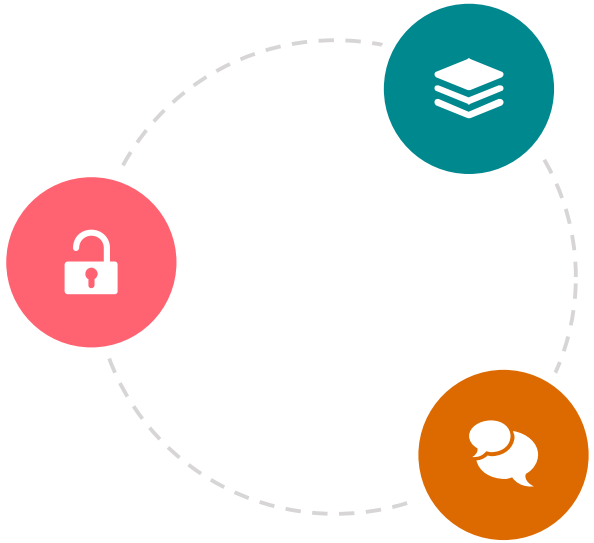
目錄

1 研究動機與目的

1-1 研究動機

更客觀地、迅速地解讀文本

以 text mining 科學化地解讀美國中央銀行 (即 Fed) 的會議紀錄。



FOMC 對市場的影響力巨大

FOMC 是 Fed 的貨幣政策決策單位，FOMC 會議上將決定貨幣政策，其中最核心的是決定聯邦基金利率 (federal funds rate) 的調整 (升息、降息、利率不變)。

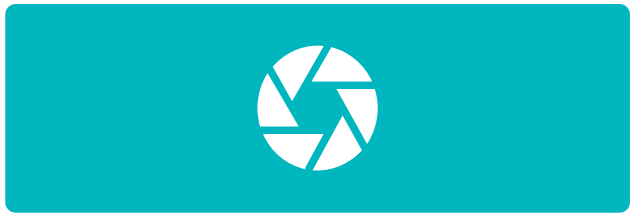
研究 FOMC minutes 的寫作風格

在升息、降息、利率不變，這三類樣本的寫作風格差異，觀察 FOMC 的態度。並以此幫助分類。

1-2 研究目的

FOMC minutes 寫作風格

透過探索性資料分析、主題模型、詞嵌入，找出FOMC minutes 在升息、降息、利率不變這三種樣本上的寫作風格差異。



聯邦基金利率特性

觀察聯邦基金利率特性，並計算 naïve model 的準確率。

FOMC minutes 文本分類

以線性降維、非線性降維、倍率指標等特徵選擇方法，搭配上 4 種分類器，計算三分類（升息、降息、利率不變）的準確率。同時，觀察最有效影響分類的字彙。

2 文獻回顧

01 FOMC minutes 的寫作風格

- Huang and Kuan (2021) 發現不同 Fed 主席任期內所關注的議題迥異，並且在每年第一次的 FOMC minutes，都會例行性的提到穩定物價這個使命。

02 FOMC minutes 的主題模型

- Boukus and Rosenberg (2006) 以 LSA 取出前 5 大主題，觀察其消長、各主題字詞、預測 10 年期國債利率的效果。
- 黃于珊 (2017) 利用 LSA 萃取出升息、降息、利率不變樣本的潛在特徵。
- Huang and Kuan (2021) 以 MAP-PLSA 區分主題，並以情緒分析，呈現 FOMC 對其三大經濟使命的正面、負面看法。

03 FOMC minutes 的文本分類

- 黃于珊 (2017) 以 LSA 處理文本並經 LDA 做三分類 (生息、降息、利率不變) 後，準確率達 75.13%。

本研究特色

01 **明確定義研究資料**

- 每篇 FOMC minutes 前後都有參與者名單、投票行為紀錄、投票者評論等較不重要的資料，因此過往文獻皆會刪除這些文字。Boukus and Rosenberg (2006)、 Huang and Kuan (2021) 給的這方面資訊過於簡略，黃于珊 (2017) 雖然較詳細的描述了，但讀者還是難以確定到底留下了那些文字。

02 **探索式資料分析**

- 過往文獻大多著墨於 CDA，較少進行 EDA。本文對 FOMC minutes、聯邦基金利率做了詳細的 EDA。

03 **文本分類特徵的維度縮減：倍率指標、非線性降維**

- 黃于珊 (2017) 僅使用 TF 前 N 大、SVD 線性降維，而本研究多使用了倍率指標、非線性降維 (kernel PCA)。

3 資料介紹與研究方法

3-1 資料介紹

1 美國聯準會會議紀要

1993年1月1日到 2020年10月1日間的美國聯準會會議紀要 (FOMC minutes)，共222篇。

FOMC 會議中將決定 Fed 的貨幣政策，會議一結束就會發布 FOMC policy statement，而在會議結束後 3 週後會發布更詳細的 FOMC minutes。

2 目標聯邦基金利率

1993年1月1日到 2020年10月1日間的目標聯邦基金利率 (targeted federal funds rate) 。一般而言，目標聯邦基金利率的調整，會在 FOMC 會議結束後馬上進行；但在特殊情況下也能臨時調整。

3-2 研究方法

(1) 資料預處理

- 定義與擷取主文
- 文本預處理
- 將處理過後的文本轉化為 DTM

(2) 探索性資料分析

- STTR

(3) 主題模型與詞嵌入

- LSA 、LDA
- Word2vec
- t-SNE

(4) 時間序列分析

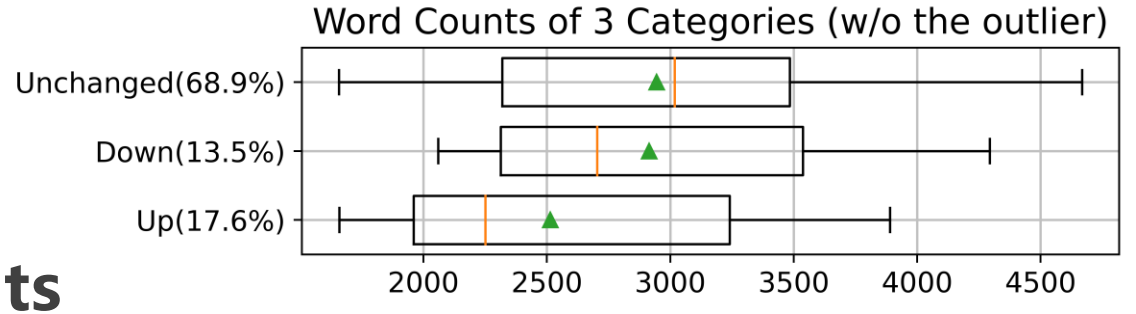
- SARIMA
- Ljung-Box test
- Naïve model

(5) 文本分類

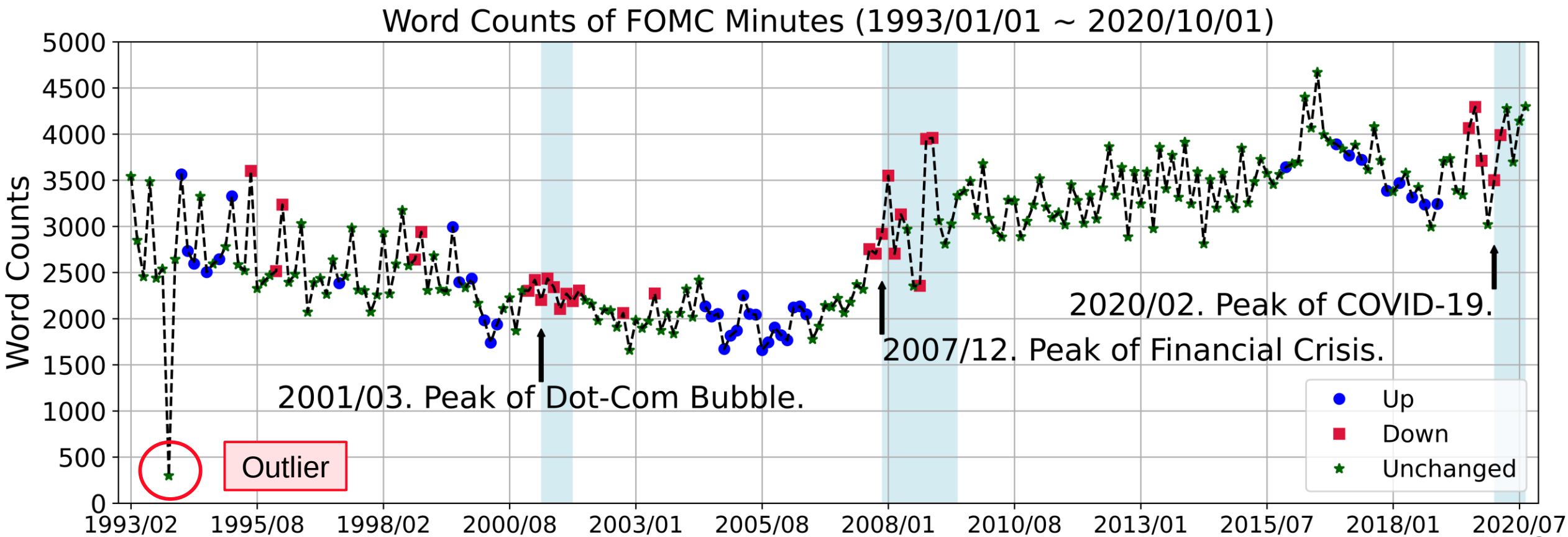
- **特徵選擇**：倍率指標、線性降維、非線性降維
- **分類器**：LR、SVM、RF、XGBoost

4 研究結果

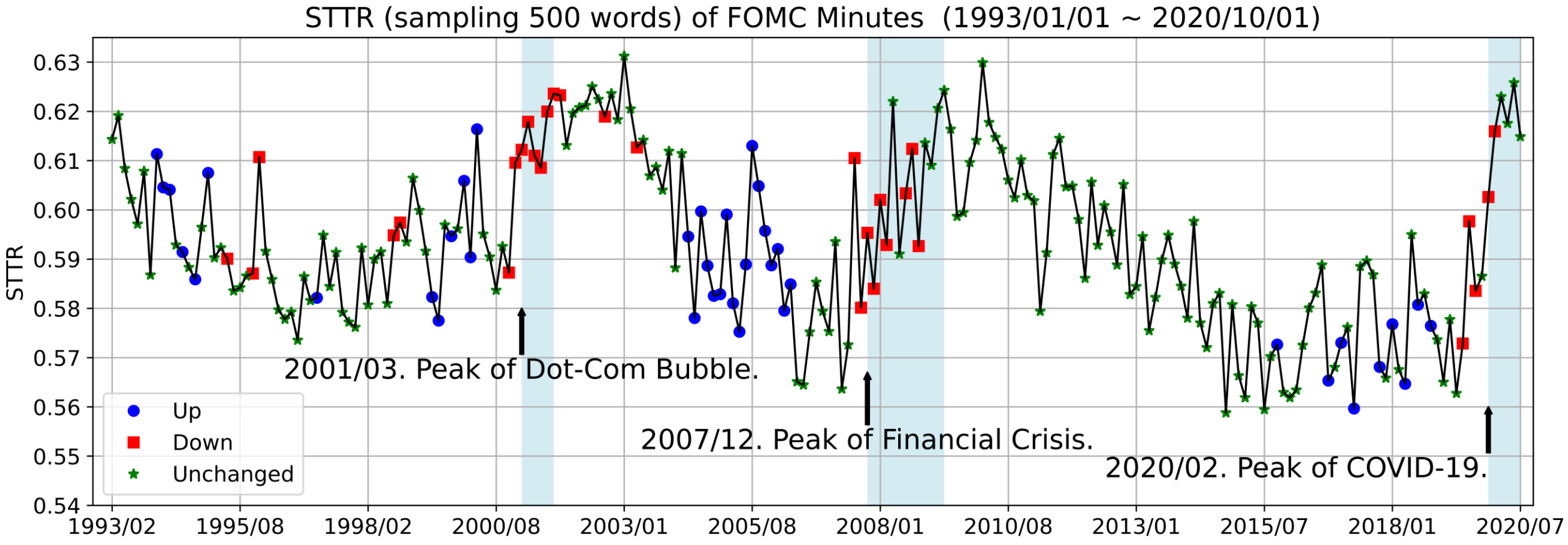
4-1 探索式資料分析 (EDA)

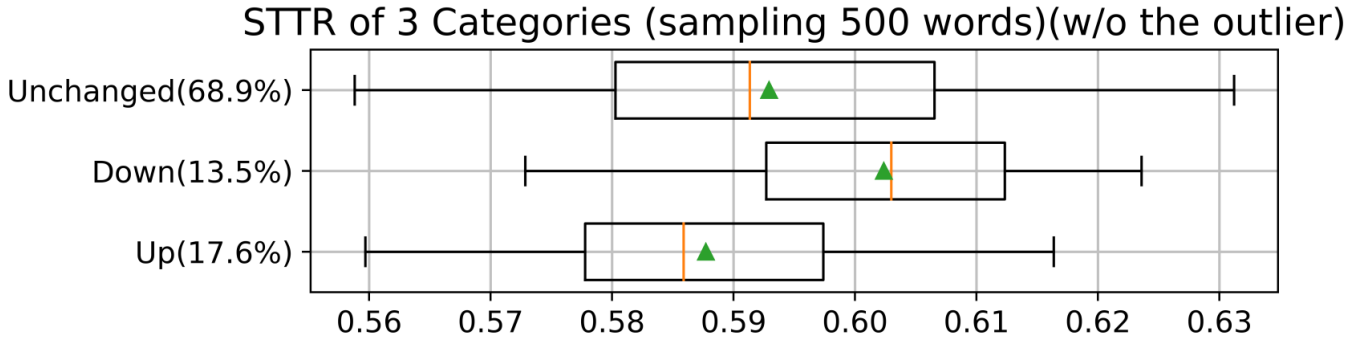


Word Counts



STTR (sampling 500 words) (w/o the outlier)





- 在自然情況下，當字數上升時，TTR 會逐漸下降。STTR 消除了這一問題。
- 發現了 STTR 是景氣循環的同時指標 (coincident indicator)。

4-2 主題模型 (topic modeling)

- Huang and Kuan (2021) 對每份 FOMC minutes 文件做 MAP-PLSA，然後依照主觀判斷，分別將前三大主題**對應到 Fed 的三大經濟使命**。
- 本研究則希望能將前三大主題，對應到 Fed 的三大經濟使命，或是對應到升息、降息、利率不變，但沒有成功。
- **失敗原因**：本研究沒有像 Huang and Kuan (2021) 統一專業術語、也沒有考慮到複合字。因此，在此，LSA、LDA 只能像黃于珊 (2017) 一樣，用於萃取重要字彙。

LSA	Topic # 01	Topic # 02	Topic # 03
Term 1	rate	growth	octob
Term 2	price	member	septemb
Term 3	inflat	expans	novemb
Term 4	market	inventori	rang
Term 5	econom	price	august
Term 6	growth	quarter	committe
Term 7	increas	sale	third
Term 8	quarter	year	percent
Term 9	particip	product	rate
Term 10	continu	demand	juli
Term 11	committe	juli	total
Term 12	year	restraint	restraint
Term 13	would	consider	monitor
Term 14	expect	economi	reserv
Term 15	remain	aggreg	veloc
Term 16	polici	industri	object
Term 17	recent	moder	lower
Term 18	declin	rang	monetari
Term 19	month	twelv	financi
Term 20	consum	direct	develop

LDA	Topic # 01	Topic # 02	Topic # 03
Term 1	april	april	octob
Term 2	particip	juli	novemb
Term 3	march	august	august
Term 4	juli	march	septemb
Term 5	june	februari	decemb
Term 6	hurrican	particip	januari
Term 7	novemb	januari	juli
Term 8	may	decemb	particip
Term 9	octob	octob	februari
Term 10	august	novemb	third
Term 11	pce	june	fourth
Term 12	agenc	guidanc	agenc
Term 13	medium	threshold	june
Term 14	septemb	restraint	twelv
Term 15	fomc	agenc	loan
Term 16	loan	program	restraint
Term 17	februari	mb	march
Term 18	decemb	loan	al
Term 19	district	recoveri	fomc
Term 20	gdp	fourth	pce

4-3 詞嵌入 (word embedding)

- Employment: unemployment, labor, growth
- Prices: inflation, price
- Interest rates: fund, interest, exchange, money

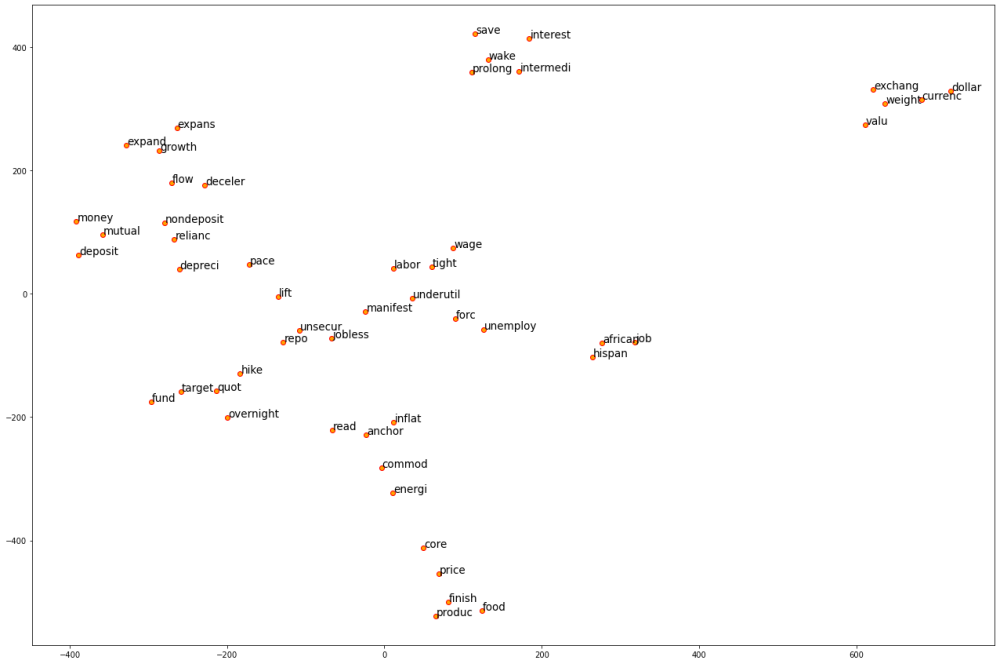
Table 1
Frequently used words for the mandates of the fed.

	Mandate: Employment		Mandate: Prices		Mandate: Interest Rates	
	Words	Counts	Words	Counts	Words	Counts
1	economic activity	1454	Inflation	3516	federal funds rate	1751
2	labor market	902	Price	1383	monetary policy	1354
3	economic growth	817	inflation expectation	1160	interest rate	776
4	unemployment rate	761	Price stability	916	Monetary	430
5	employment	735	energy price	657	foreign exchange market	286
6	output	711	equity price	306	Monetary and financial condition	202
7	labor market condition	489	core inflation	287	Monetary aggregate	184
8	maximum employment	339	consumer price	261	foreign exchange value	163
9	real GDP	300	inflation compensation	249	long term interest rate	105
10	labor cost	271	price inflation	234	market interest rate	103
11	labor	229	inflation pressure	230	foreign exchange	90
12	real GDP growth	221	commodity price	177	Mortgage interest rate	85
13	unemployment	214	house price	174	Exchange	76
14	labor compensation	201	oil price	158	Monetary policy accommodation	55
15	job	196	consumer price inflation	156	foreign monetary authority	52
16	payroll employment	145	producer price	150	real interest rate	51
17	civilian unemployment rate	136	inflationary pressure	140	Monetary growth	48
18	employ	132	food price	127	market exchange rate	47
19	potential output	120	CPI	125	monetary accommodation	41
20	labor market indicator	115	gasoline price	92	monetary base	33

如上頁，依照 Huang & Kuan (2021) 考慮到專業術語、複合字等，整理出來的 Fed 三大經濟使命各自的詞語排名，我們從中挑出 9 大關鍵字，並以這 9 大關鍵字為核心進行詞嵌入。

- Employment: unemployment, labor, growth
- Prices: inflation, price
- Interest rates: fund, interest, exchange, money

以 word2vec 計算 9 大關鍵字各自關係 (即彼此間的距離) 最近的前 5 大字彙，加上 9 大關鍵字共 54 個字彙。將這 54 個字彙的 word2vec 結果，以 t-SNE 投影至 2-D 平面，得到如右圖的結果。



Up

Cluster 1: unemployment, labor
Cluster 2: growth
Cluster 3: inflation, price
Cluster 4: fund
Cluster 5: interest
Cluster 6: exchange
Cluster 7: money

Down

Cluster 1: unemployment
Cluster 2: labor
Cluster 3: growth
Cluster 4: inflation, price
Cluster 5: fund
Cluster 6: interest
Cluster 7: exchange
Cluster 8: money

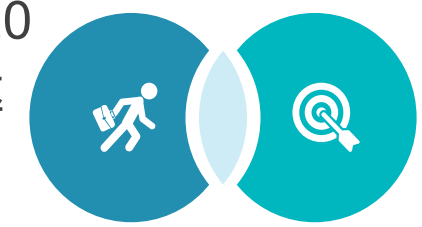
Unchanged

Cluster 1: unemployment
Cluster 2: labor
Cluster 3: growth
Cluster 4: inflation
Cluster 5: price
Cluster 6: fund
Cluster 7: interest
Cluster 8: exchange
Cluster 9: money

(1) All	**unemploy**: ['forc', 'workweek', 'hire', 'age', 'popul', 'reason', 'jobless', 'save', 'underutil', 'interest', 'hour', 'sideway', 'summari', 'quit', 'group', 'delinqu', 'addendum', 'converg', 'work', 'fill']
(2) Up	**unemploy**: ['forc', **'african'**, **'hispan'**, 'workweek', **'american'**, 'hour', 'underutil', 'men', 'job', **'white'**, 'jobless', 'claim', 'worker', **'civilian'**, 'popul', 'monthli', 'lowest', 'payrol', 'war', 'ii']
(3) Down	**unemploy**: ['popul', **'african'**, 'age', **'civilian'**, 'workweek', 'forc', 'claim', 'hour', 'group', 'steadi', **'hispan'**, **'white'**, **'american'**, 'job', 'averag', 'insur', 'hurrican', 'payrol', 'employ', 'roughli']
(4) Unchanged	**unemploy**: ['forc', 'popul', 'age', 'workweek', 'hire', 'jobless', 'delinqu', 'work', 'addendum', 'save', 'lengthi', 'weekli', 'worker', 'job', 'reason', 'suspend', 'manifest', 'interest', 'payrol', 'durat']

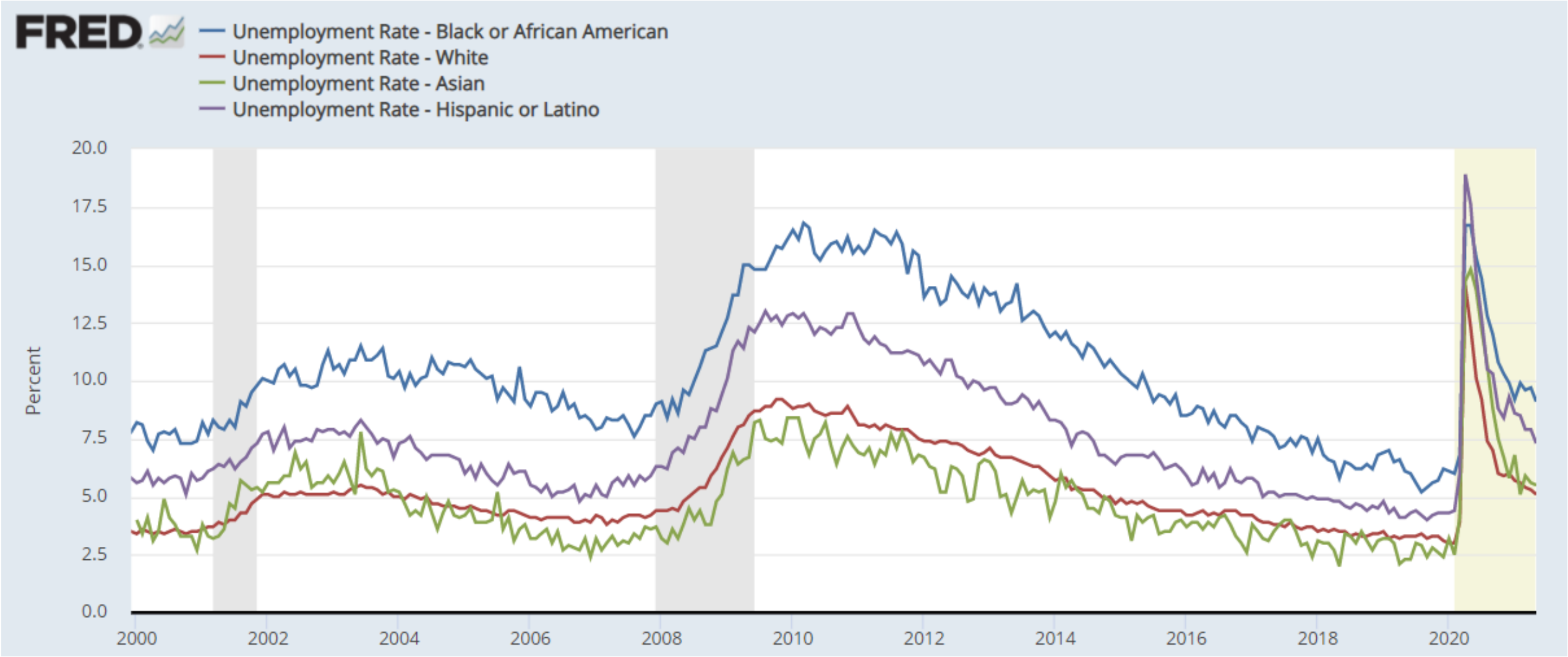
unemploy 在四類中樣本的差異特別大

- 觀察 9 大關鍵字各自最相關的前 20 大字彙在全體、升息、降息、利率不變的差異，發現 unemployment (在此為 unemploy，因為使用了 Porter stemmer) 的差異特別大 。



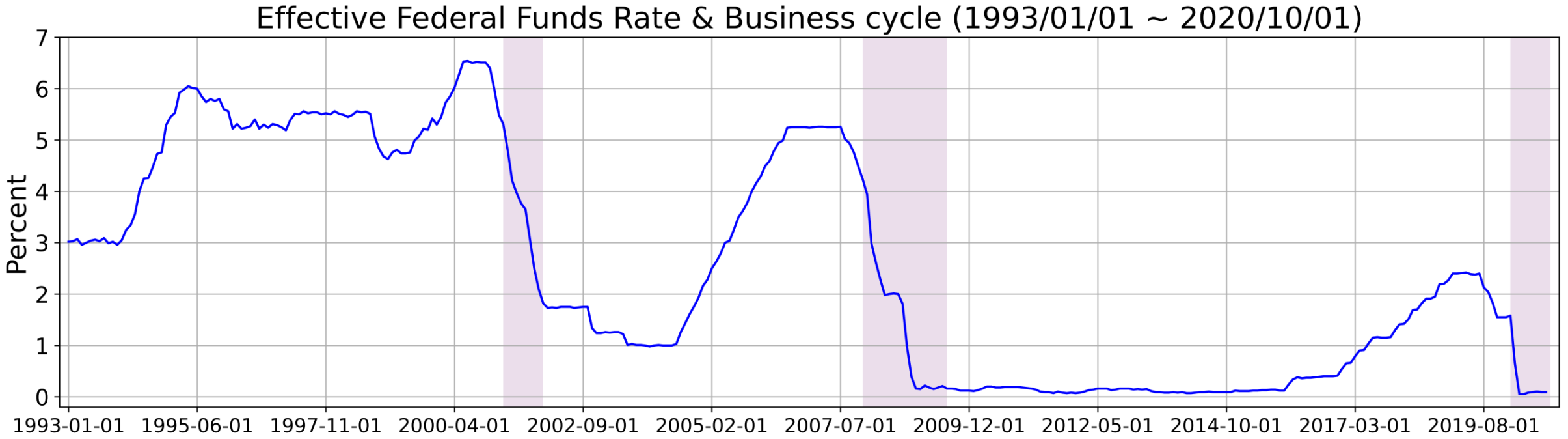
四類樣本中的共同字彙、相異字彙

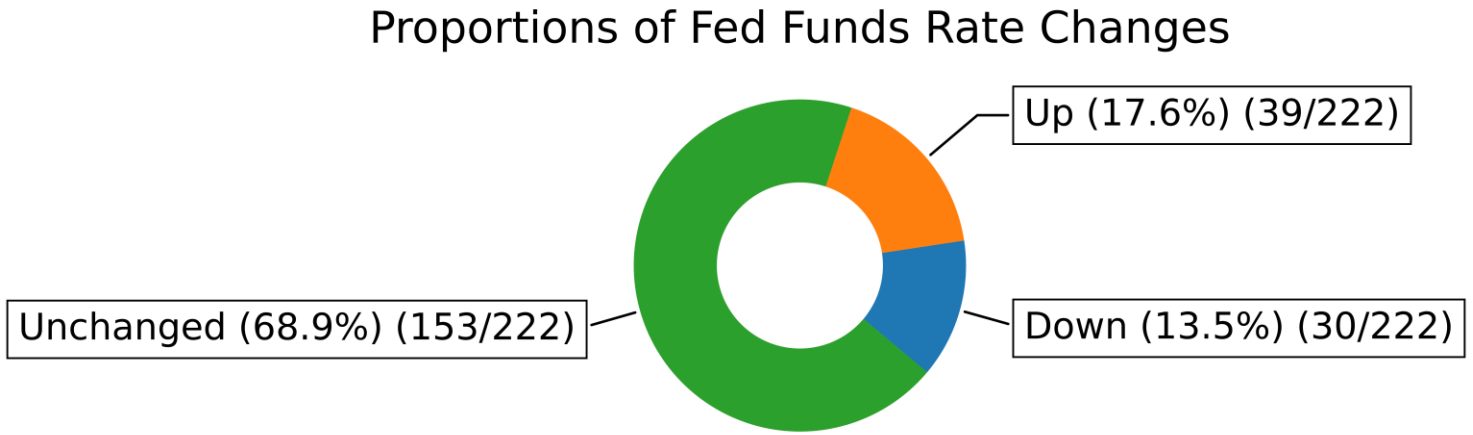
- [1] 四個種類都有出現的字彙：'forc', 'workweek', 'popul'
- [2] 只出現於升息、降息的字彙: 'african', 'hispan', 'white', 'american', 'civilian'



- 只出現於升息、降息的字彙: 'african', 'hispan', 'white', 'american', 'civilian'

4-4 時間序列分析



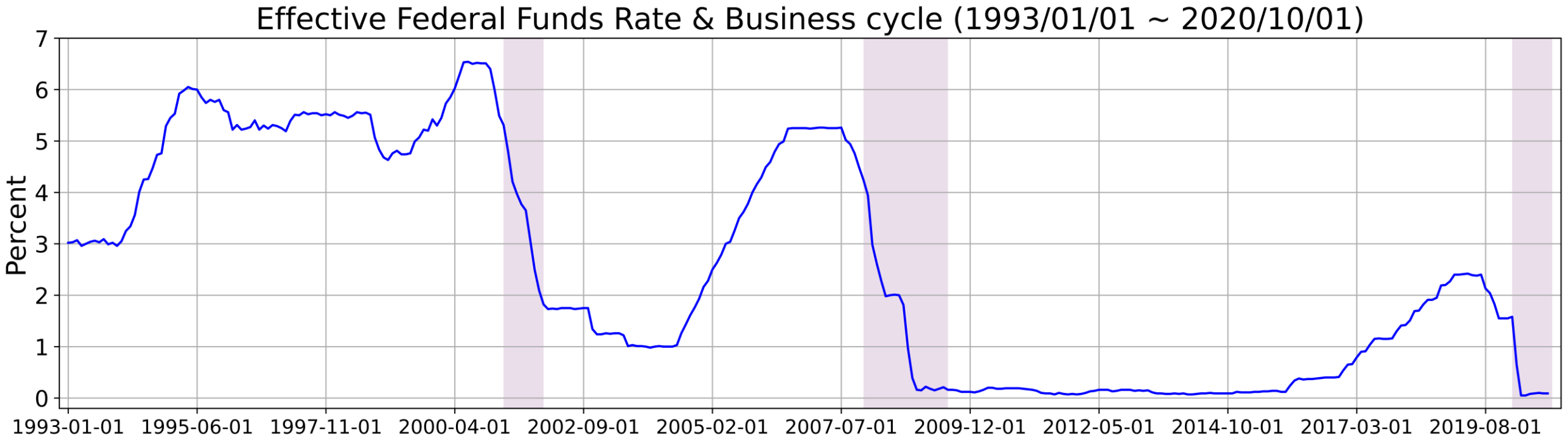


- 不平衡資料，高達 68.9% 為利率不變。
- 升息、降息、利率不變都有很強的連續性。

Up (total 39)		**Down (total 30)**		**Unchanged (total 153)**	
Successive times	counts	Successive times	counts	Successive times	counts
1	14	**1**	3	**1**	12
2	1	**2**	3	**2**	2
3	2	**3**	2	**3**	2
17	1	**6**	1	**4**	5
		9	1	**6**	1
				7	2
				8	2
				9	1
				11	1
				55	1

Up (total 39)		**Down (total 30)**	
Rate change	counts	Rate change	counts
0.25 %	33	**0.25 %**	12
0.50 %	5	**0.50 %**	12
0.75 %	1	**0.75 %**	2
		1.00 %	4

- 64.1% 的升息連續兩次以上、90.0% 的降息連續兩次以上、92.2% 的利率不變連續兩次以上。
- 只有 15.4% 的升息決策調整幅度為 2 碼 (2×0.25%) 以上，但高達 60.0% 的降息決策調整幅度為 2 碼以上。
- **升息保守、降息果斷。**



- SARIMA 最佳模型：ARIMA(1,1,2)(0,0,0)[12]
- 目標聯邦基金利率，具自相關性。

- 以天真預測法 (naïve method) 預測，得到 **74.7%** 的準確率。
- 74.7% 是以 FOMC minutes 預測未來的利率調整的基準線 (baseline)。

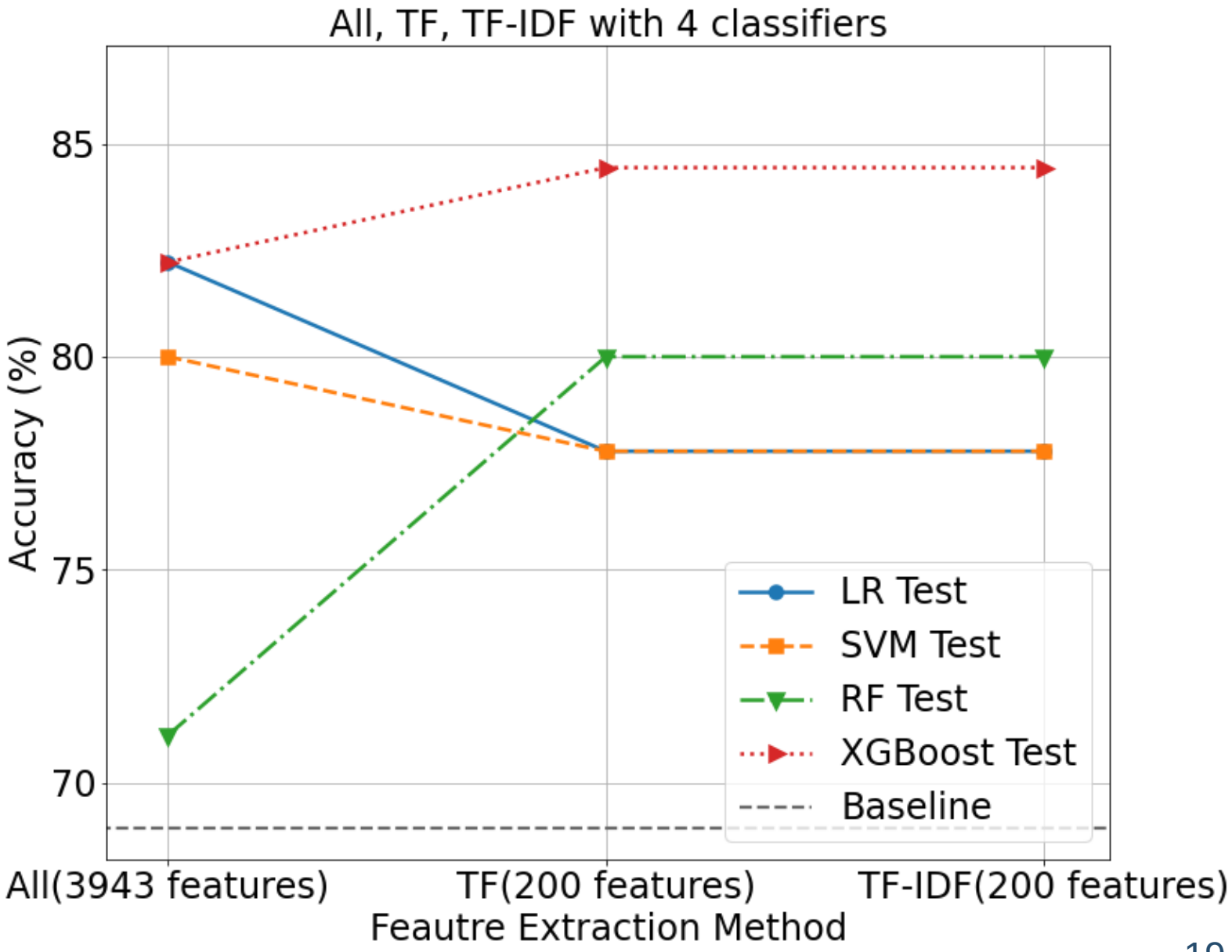
4-5 文本分類

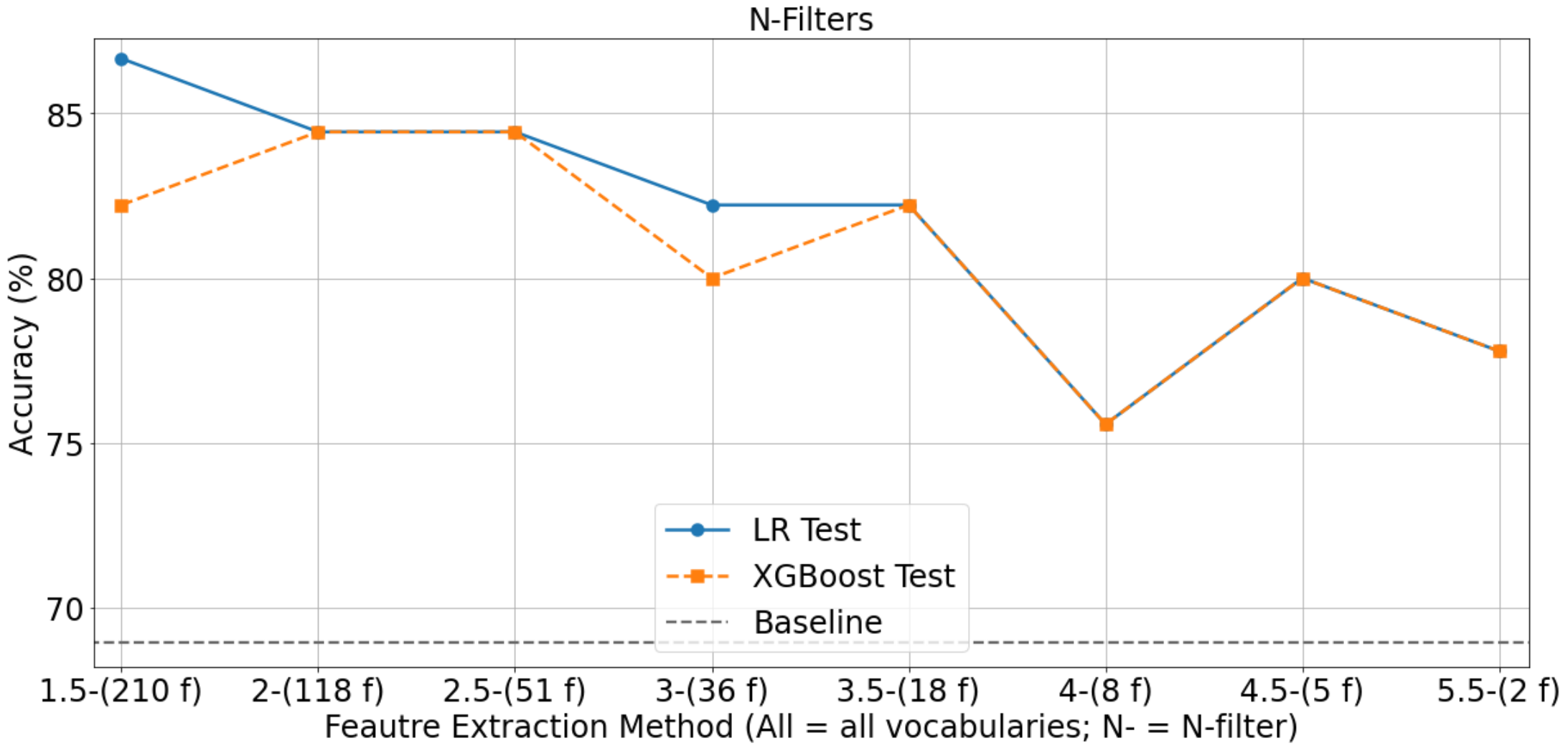
初步比較 4 種分類器

- 使用 3 種特徵選取方法：全部字彙、TF 前 200 大字彙、TF-IDF 前 200 大字彙。
- 在此，準確率最高為 84.44%，來自 XGBoost + TF 前 200 大。

➢ LR 與 SVM 二選一：選 LR。Ensemble learning 方法 RF 與 XGBoost 二選一：選 XGBoost。

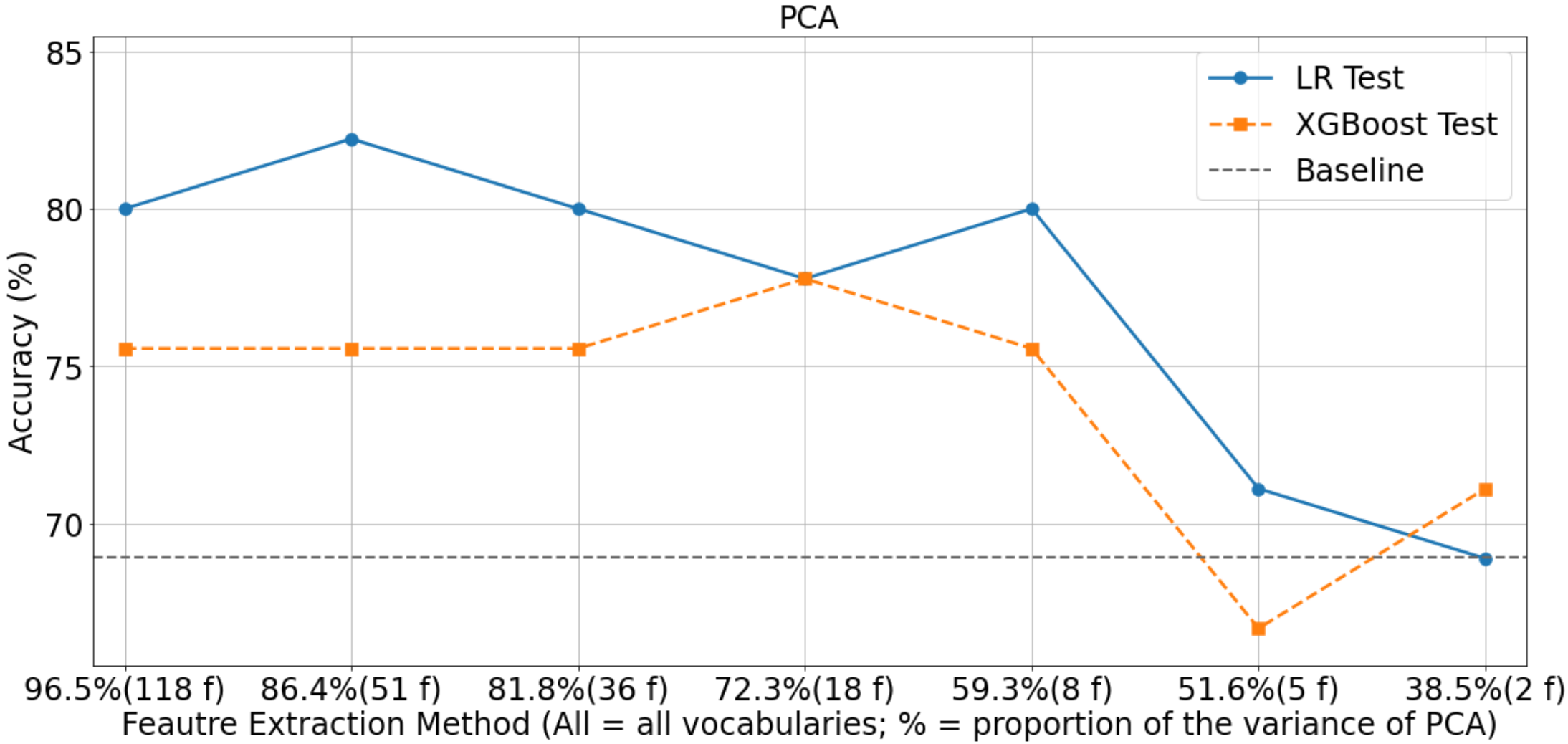
➢ 使用 LR、XGBoost 繼續接下來的分析。





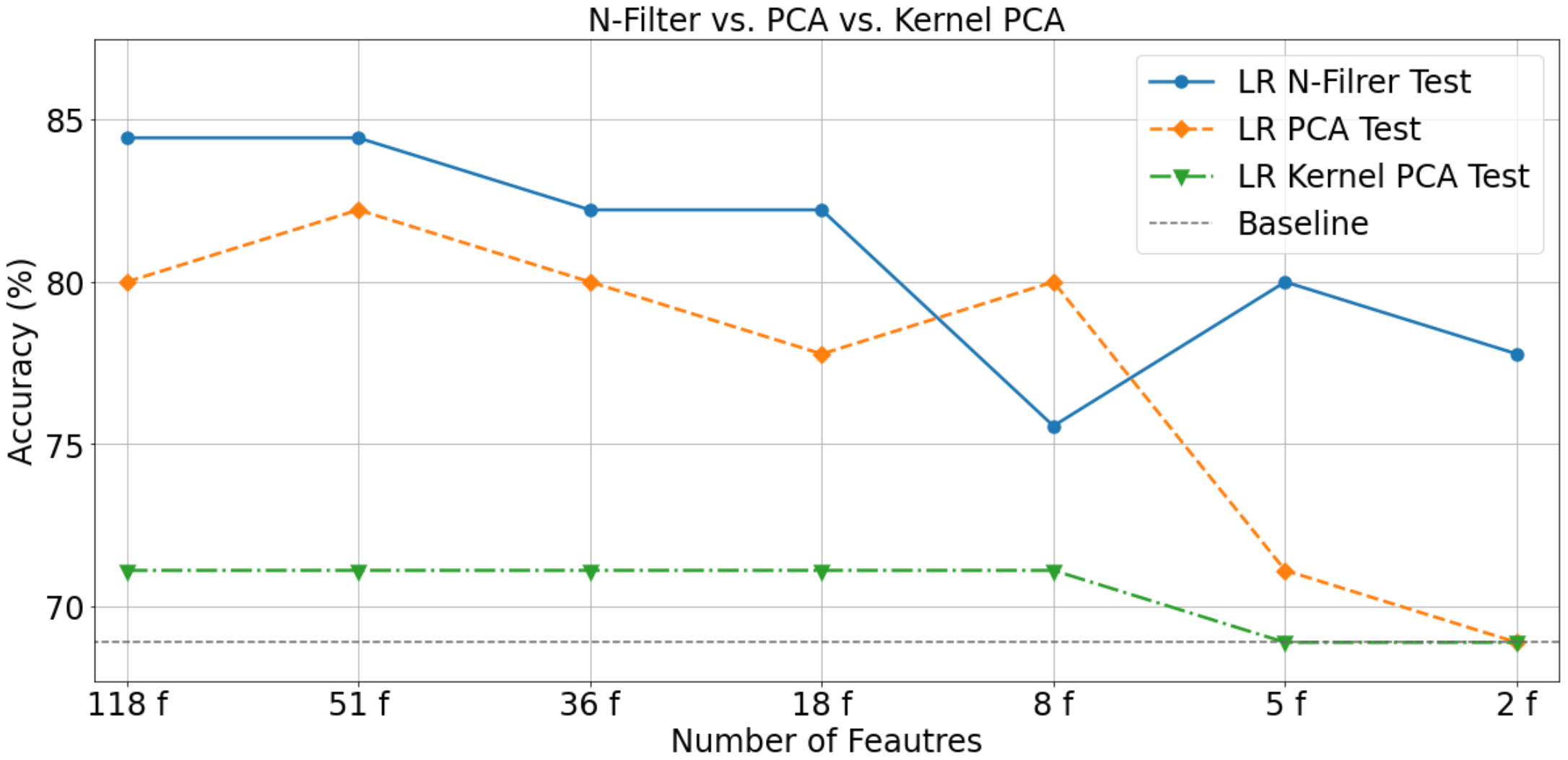
倍率指標

- ❖ 倍率指標下，LR 表現高於 XGBoost 。
- ❖ 在此，準確率最高為 86.67%，來自 LR +1.5 倍率指標。



線性降維 (PCA)

- ❖ PCA 下，LR 表現高於 XGBoost 。
- ❖ 在此，準確率最高為 82.22%，來自 LR + PCA 86.4% (51f)。



比較不同特徵選擇方法

- ❖ 整體而言，倍率指標 > PCA > kernel PCA。
- ❖ 在此，準確率最高為 84.44%，來自 LR + 2 倍率指標。

文本分類小結

降維方法準確率：倍率指標 > PCA (線性降維) > kernel PCA (非線性降維)。

分類器：大致上 LR > XGBoost。

最高準確率：86.67%，發生在 LR + 1.5 倍率指標 (210 f)。

關鍵特徵：右表為幾個表現最好的、最具代表性的「特徵選取方法 + 分類器」組合，由上到下依照「特徵數量」排列。注意到 3.5 倍率指標僅以 18 個特徵，就得到 82.22 % 的佳績，與使用全部 3943 個特徵結果一樣！

特徵選取方法	特徵數量	分類器	準確率
全部字彙 (3943 f)	**3,943**	**LR** (或 XGBoost)	**82.22 %**
1.5 倍率指標 (210 f)	210	LR	**86.67 %**
TF 前 200 大 (或 TF-IDF 前 200大)	200	XGBoost	84.44 %
2 倍率指標 (118 f) 或 2.5 倍率指標 (51 f)	118 (或 51)	LR	84.44 %
PCA 86.4% (51f)	51	LR	82.22 %
3.5 倍率指標 (18 f)	**18**	**LR**	**82.22 %**

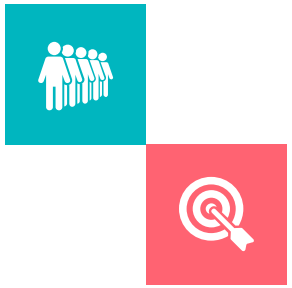
3.5 倍率指標篩選出來的 18 個極具影響力特徵：{'contract', 'turmoil', **'outbreak'**, **'weak'**, 'institut', **'eas'**, **'weaker'**, **'lend'**, 'correct', **'recoveri'**, **'downturn'**, **'weaken'**, **'rebuild'**, **'medium'**, 'remov', 'contain', **'coronaviru'**, **'hurrican'**}。

LR + 1.5 倍率指標 (210 f) 的混淆矩陣 (confusion matrix)

		Predicted		
		-1	0	1
Actual	-1	6	0	0
	0	0	30	1
	1	0	5	3

	precision	recall	f1-score	support
-1	1.00	1.00	1.00	6
0	0.86	0.97	0.91	31
1	0.75	0.38	0.50	8
accuracy			0.87	45
macro avg	0.87	0.78	0.80	45
weighted avg	0.86	0.87	0.85	45

- 最高的 86.67% 準確率，發生於 LR + 1.5 倍率指標 (210 個特徵)，所以進一步觀察其混淆矩陣。(上左圖中 +1 為升息，0 為利率不變，-1 為降息)

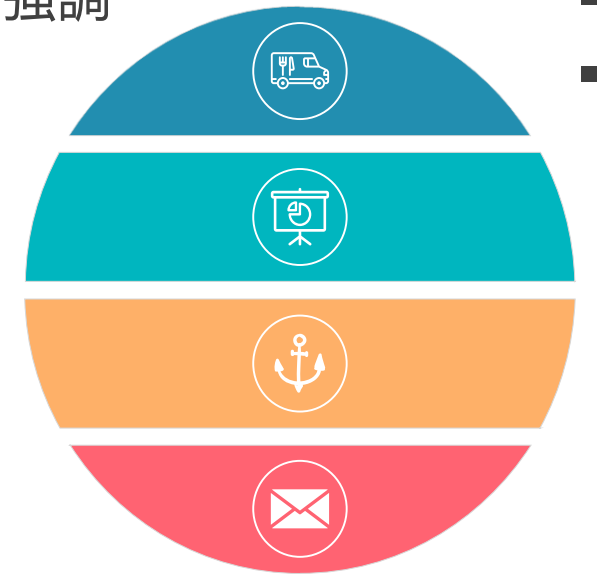


- FOMC minutes 的降息樣本，與其他兩類樣本差異甚大。

5 結論與討論

(2) 詞嵌入

- 在升息、降息時，會特別強調美國各族裔的失業率。



(1) 探索式資料分析

- STTR 是景氣循環的同時指標。

(5) 未來研究建議

- 考慮到複合字、統一專業術語。
- 只用跟 Fed 三大經濟使命相關的詞彙，觀察分類準確率。
- 以過去的 FOMC minutes 搭配 targeted federal funds rate，預測下一次 FOMC 會議將決議升息、降息、利率不變。

(3) 時間序列分析

- 升息保守、降息果斷。
- 目標聯邦基金利率具自相關性，以天真預測法，可得 74.7% 準確率。

(4) 文本分類

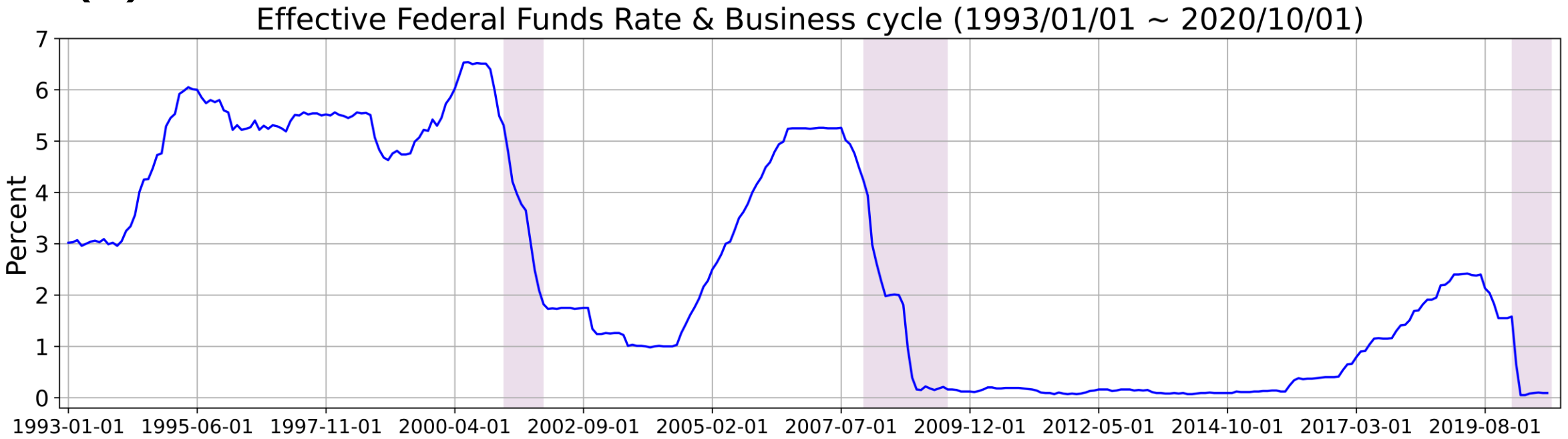
- **降維方法準確率**：倍率指標 > PCA (線性降維) > kernel PCA (非線性降維)。
- **分類器準確率**：大致上 LR > XGBoost。
- **最高準確率**：86.67%，發生在 LR + 1.5 倍率指標 (210 f)
- **關鍵特徵**：LR + 3.5 倍率指標僅以 18 個特徵，就得到 82.22 % 的佳績，與使用 LR + 全部 3943 個特徵結果一樣！

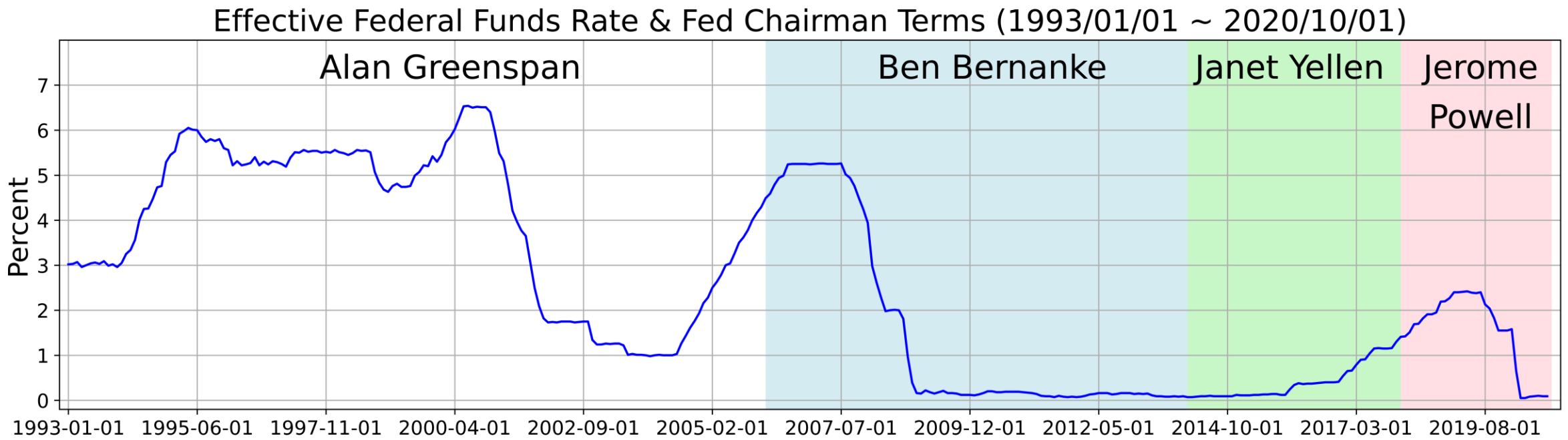
THANKS

謝謝您的觀看

附錄

(2) 聯邦基金利率變動、景氣循環、Fed 主席任期





(8) SARIMA — Auto ARMIA

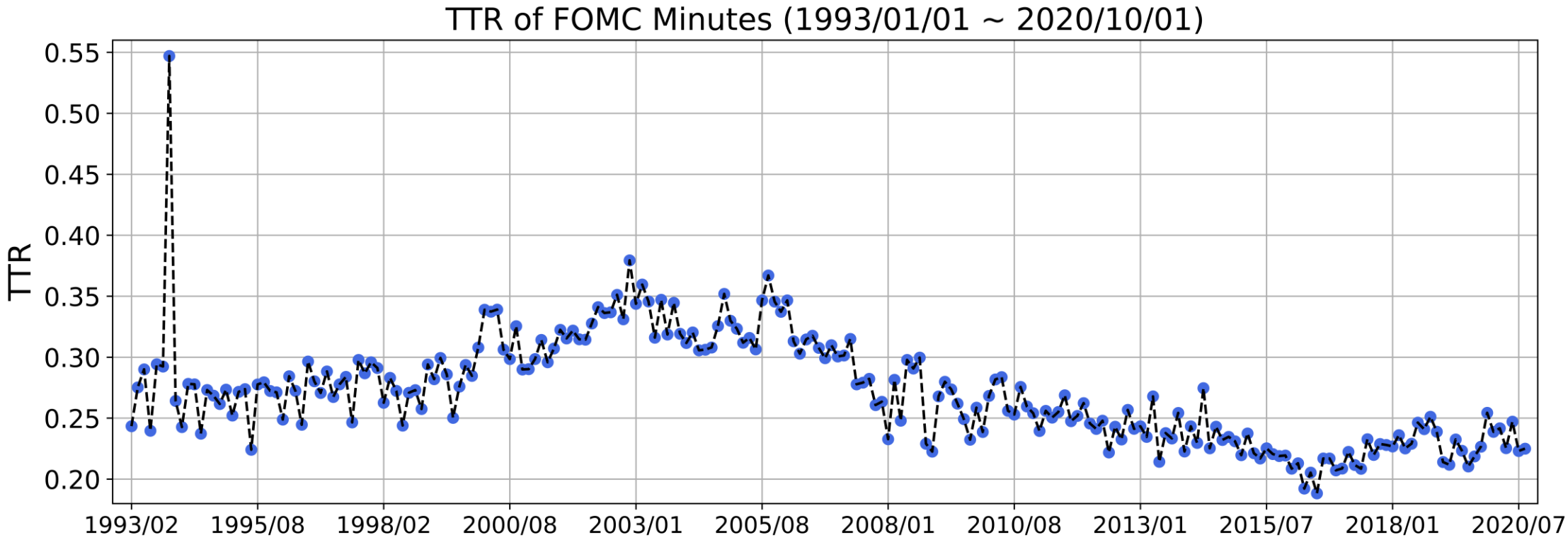
- 使用Python 的 pmdarima 套件 AutoARIMA 函式
- Information Criterion = AIC
- Best model: ARIMA(1,1,2)(0,0,0)[12]
- p-value of Ljung-Box Test: 不顯著。

Argument	Value
• p (AR)	1
• d	1
• q (MA)	2
• P	0
• D	0
• Q	0
• s	12

lag	lb_stat	lb_pvalue
1	0.003823	0.950701
2	0.035098	0.982604
3	0.333857	0.953540
4	0.339116	0.987151
5	0.339459	0.996834
6	0.977963	0.986434
7	0.980891	0.995131
8	2.23725	0.972861
9	2.239501	0.987089
10	2.249309	0.994048
11	2.834815	0.992734
12	3.089372	0.994879
13	3.123956	0.997449
14	3.278362	0.998470
15	3.285012	0.999298
16	4.259695	0.998379
17	4.570498	0.998748
18	5.182909	0.998546
19	5.242705	0.999198

(3) TTR vs. STTR — 1

- TTR (Type/Token Ratio) = vocabulary counts / word counts

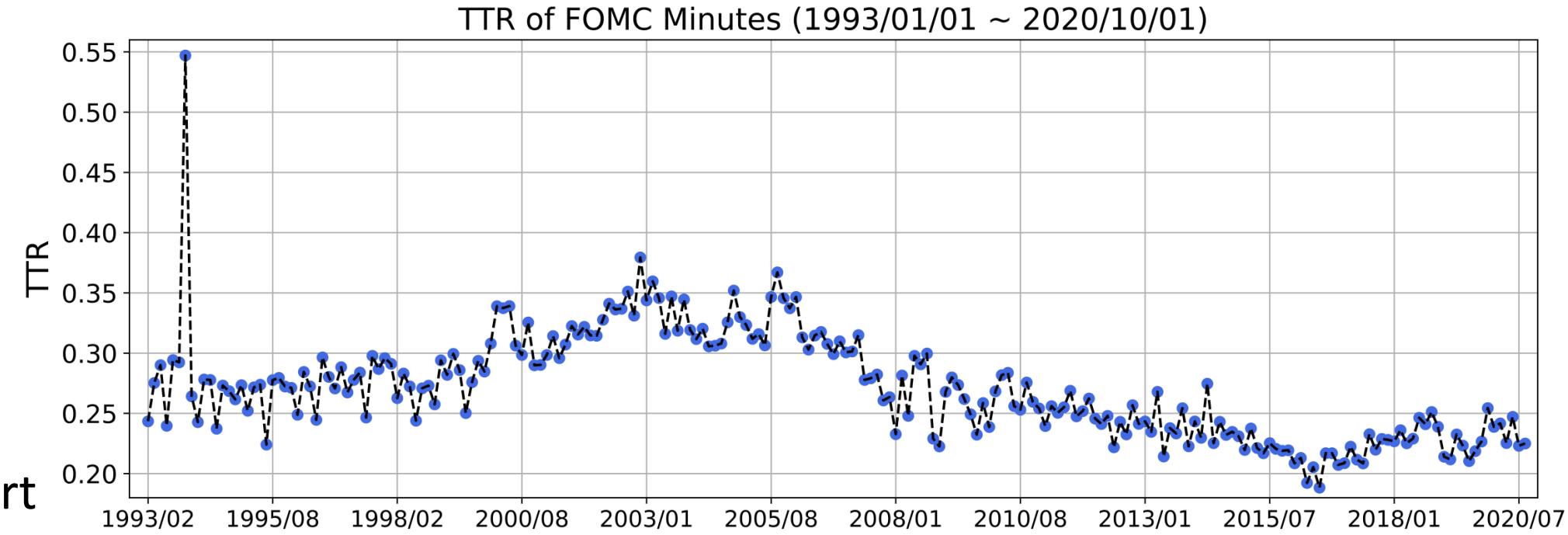


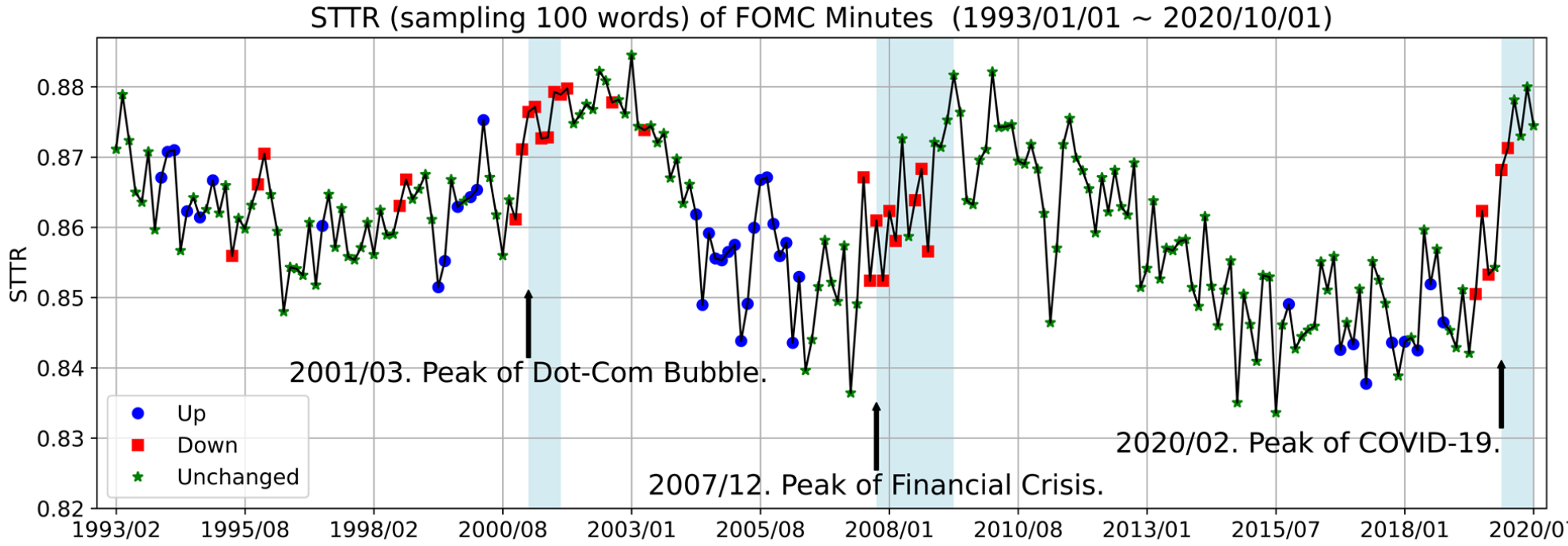
- (STTR) Standard Type/Token Ratio: It is difficult to compare the TTR of smaller against larger texts, because **as the text gets bigger, the number of new word types counts fall**. In order to remedy this, we calculate TTR based on every 100,300,500 words for 1000 times and produce an average TTR, that is, STTR.

Ref: https://www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l03_1.htm

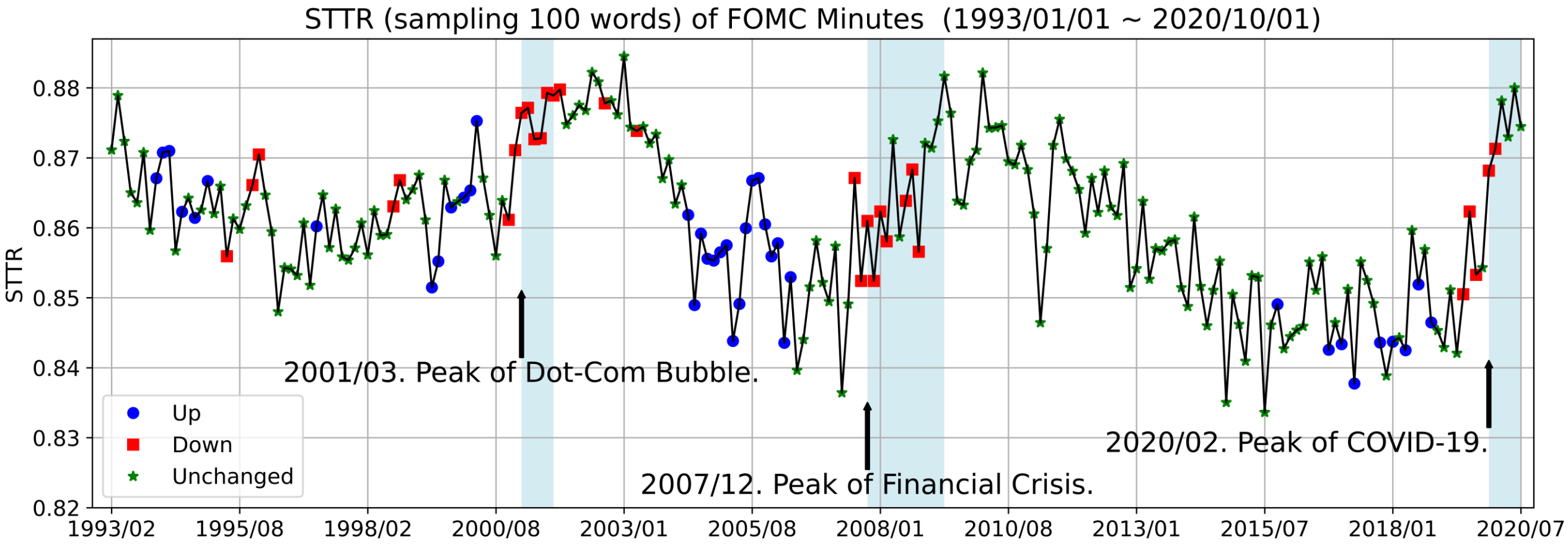
(4) TTR vs. STTR — 2

- TTR line chart is just the opposite of vocabulary counts line chart.



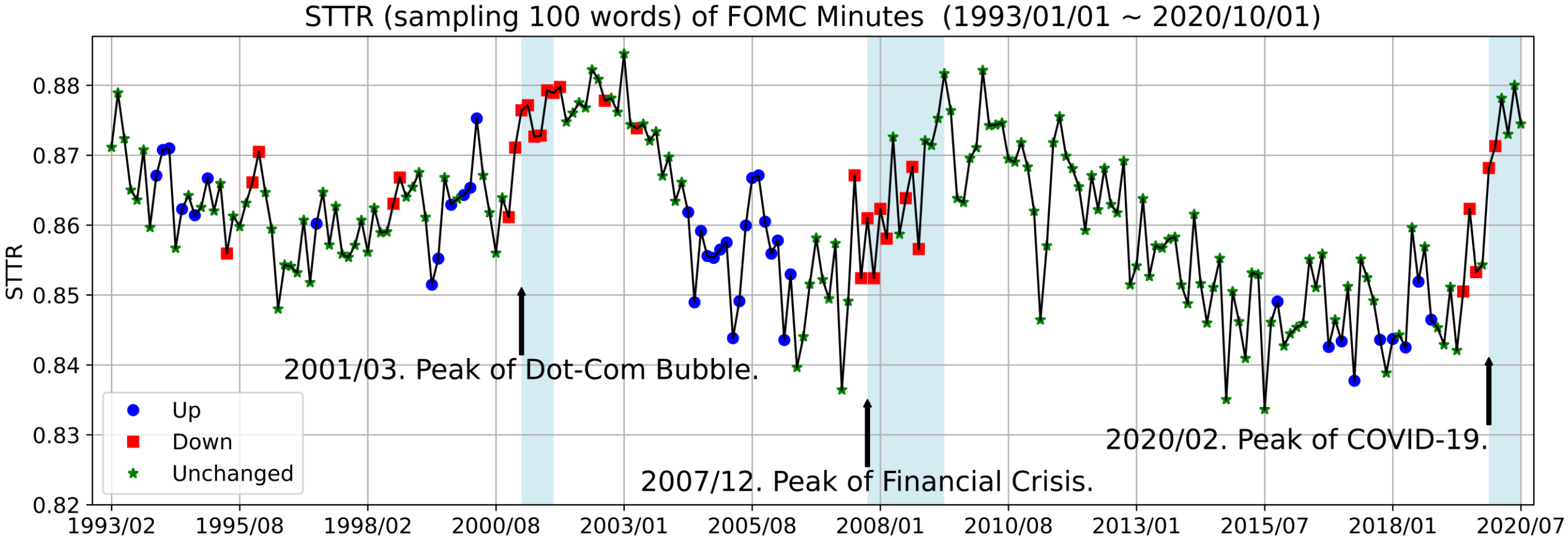


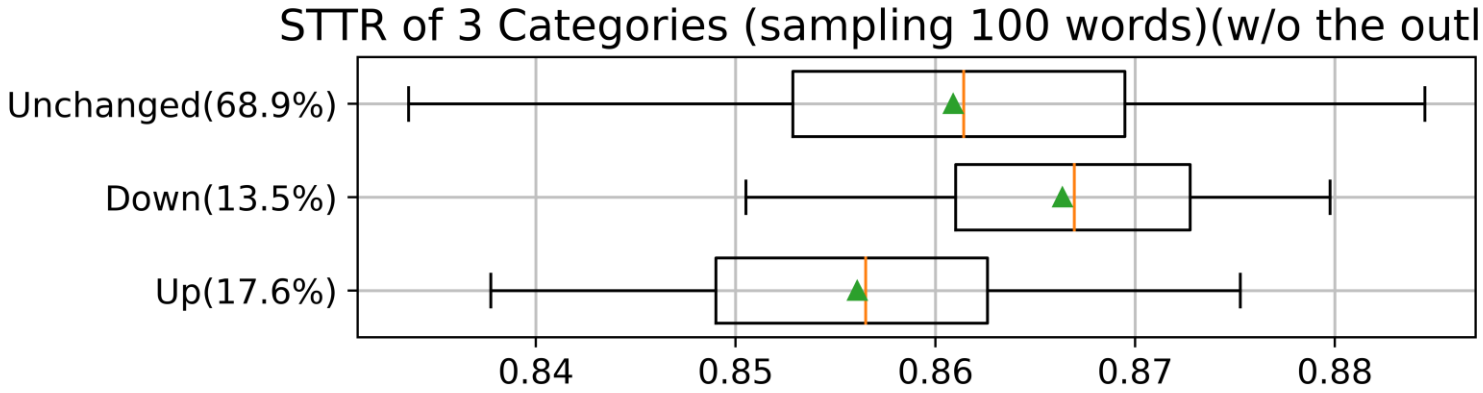
(5) STTR (sampling 100 words) (w/o the outlier) — 1



- STTR here is the average TTR of the samples withdrawn during 1000 times sampling.
- We remove the outlier in STTR line chart from now on since it drastically impacts the STTR line chart.
- Now we can see STTR looks like a lagging indicator of business cycle.

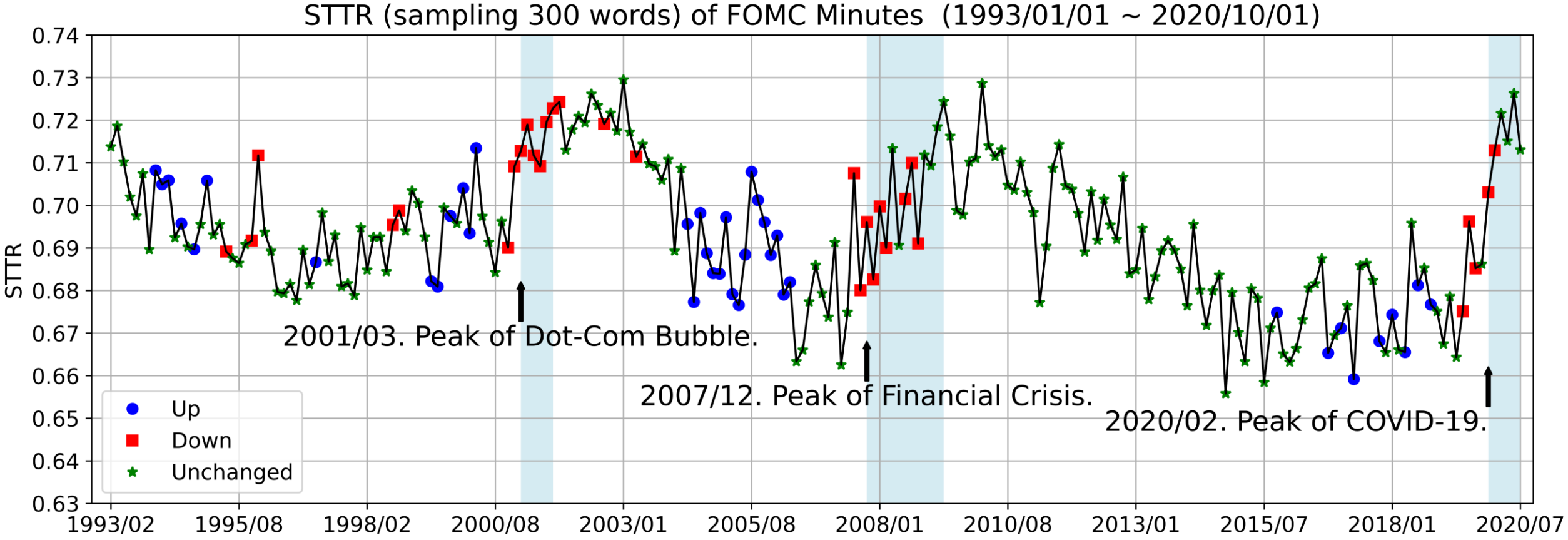
(6) STTR (sampling 100 words) (w/o the outlier) — 2

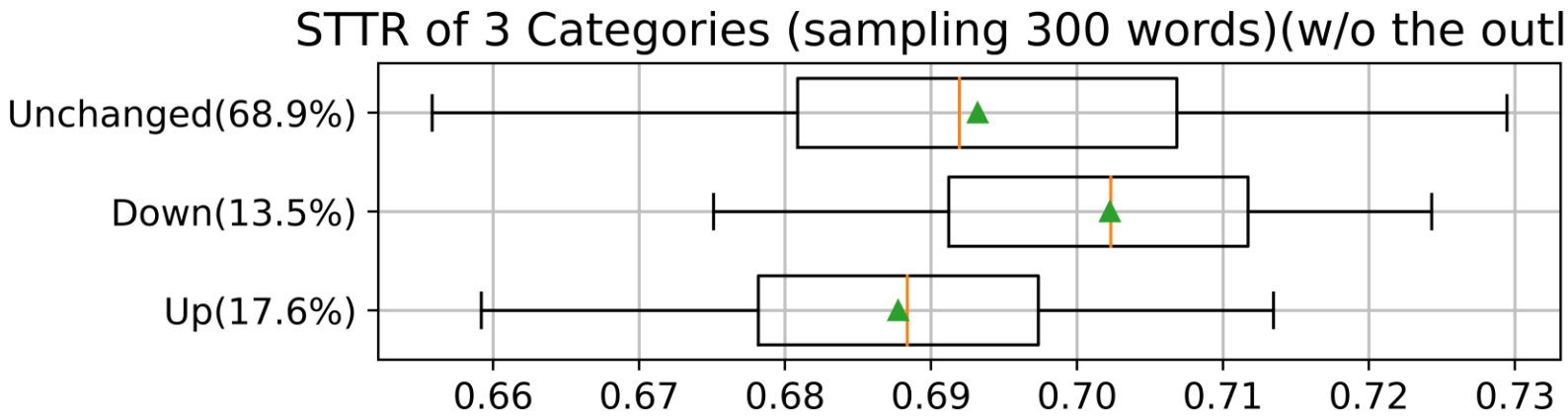




- $\frac{(\max-\min)}{(\max+\min)/2} = 0.0592$

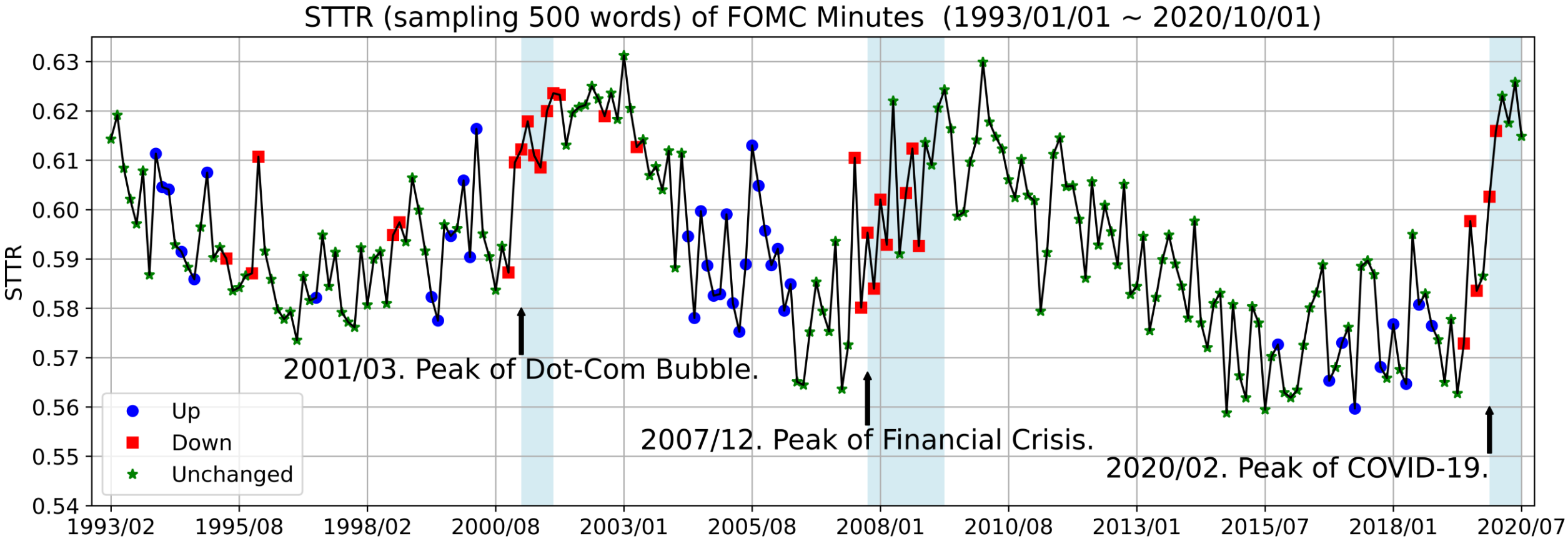
(7) STTR (sampling 300 words) (w/o the outlier)

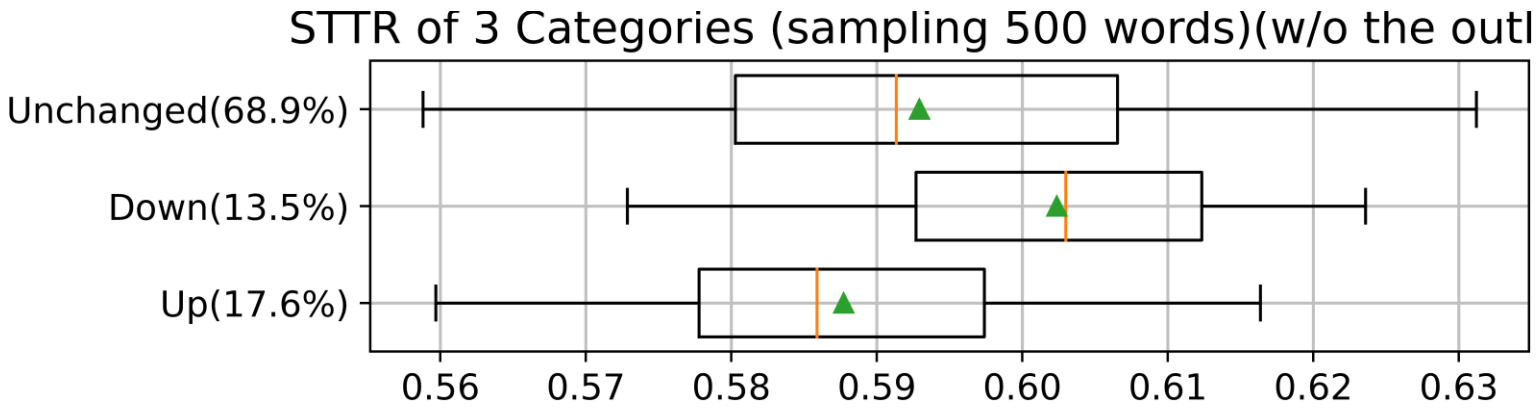




- $\frac{(\max-\min)}{(\max+\min)/2} = 0.1063$

(8) STTR (sampling 500 words) (w/o the outlier)



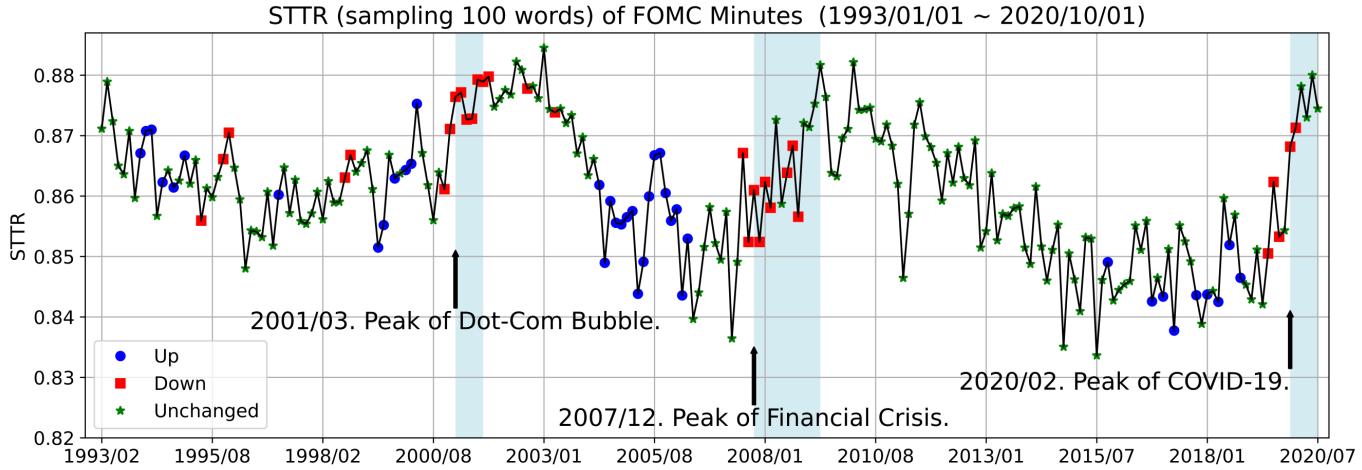


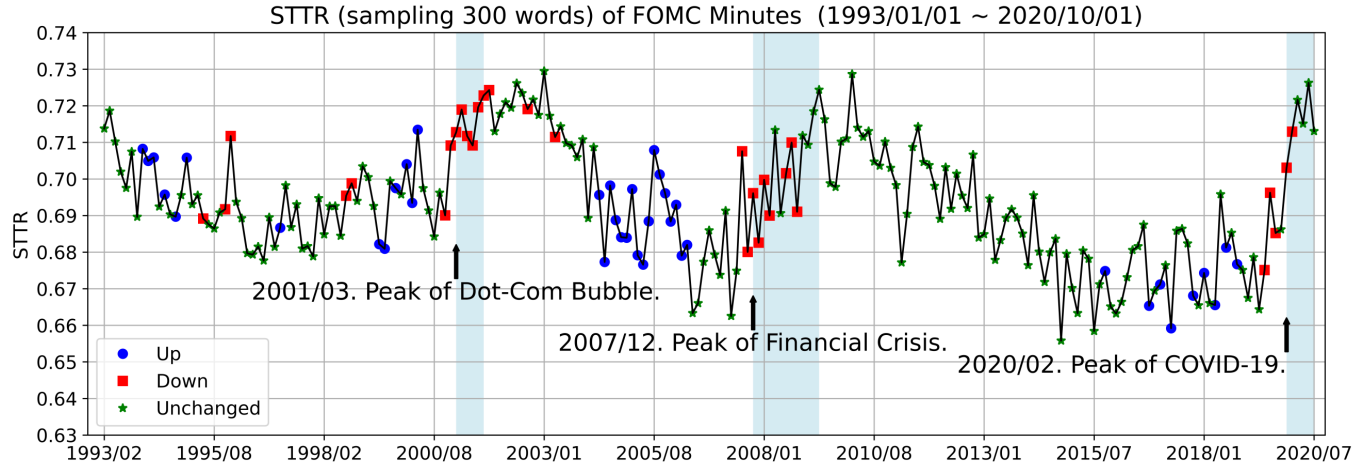
- $\frac{(\max-\min)}{(\max+\min)/2} = 0.1217$

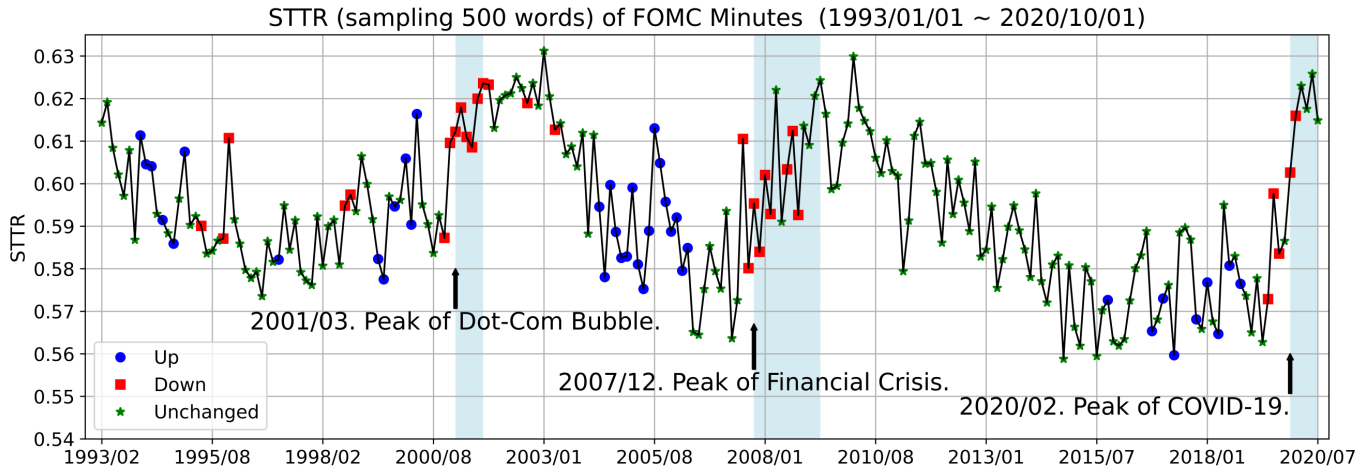
(9) Comparison of STTR

Sampling Size	$\frac{(\max - \min)}{(\max + \min)/2}$
• STTR_100	0.0592
• STTR_300	0.1063
• STTR_500	0.1217

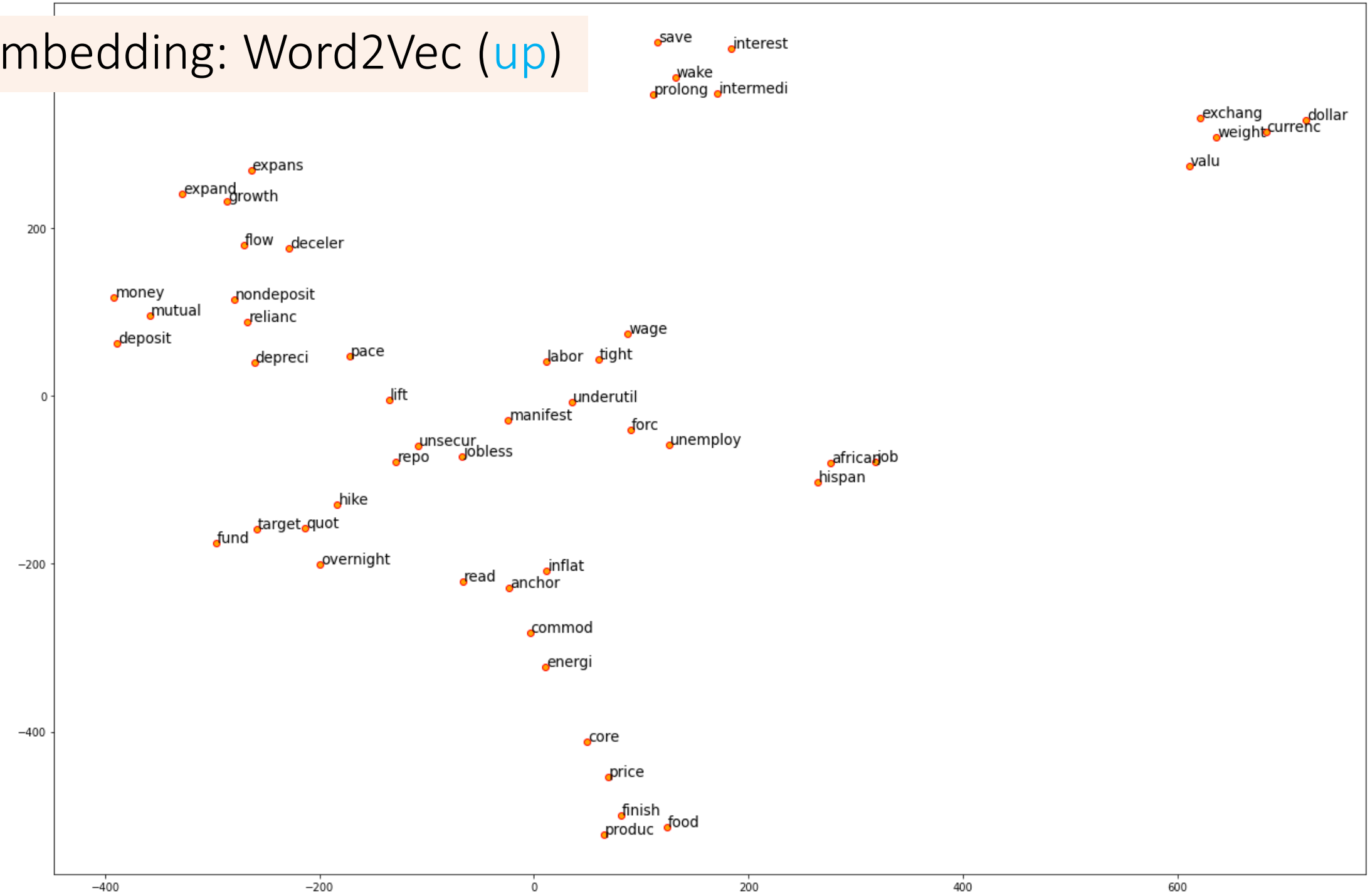
• STTR_500 has more discrimination, so we choose it to proceed time series analysis on STTR.





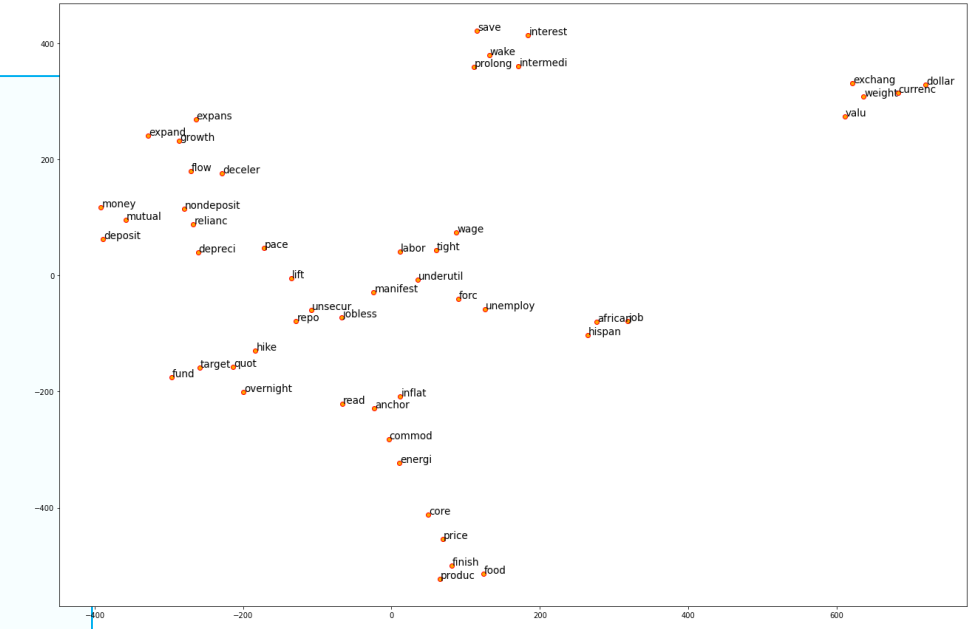


Word Embedding: Word2Vec (up)



Word Embedding: Word2Vec (up)

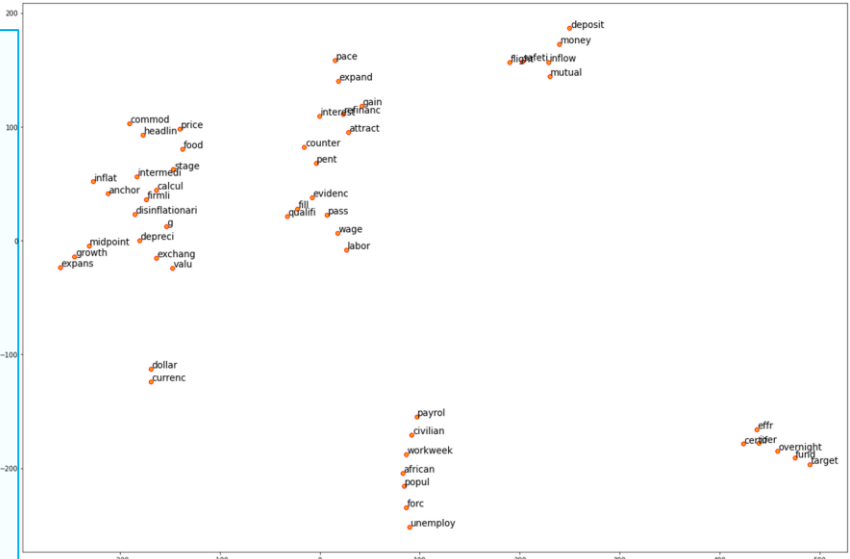
- (1) **unemploy**: ['forc', 'african', 'hispan', 'job', 'jobless']
- (2) **labor**: ['wage', 'manifest', 'job', 'underutil', 'tight']
- (3) **growth**: ['expans', 'expand', 'pace', 'flow', 'deceler']
- (4) **inflat**: ['core', 'read', 'energi', 'anchor', 'commod']
- (5) **price**: ['core', 'food', 'finish', 'produc', 'commod']
- (6) **fund**: ['target', 'quot', 'hike', 'unsecur', 'overnight']
- (7) **interest**: ['wake', 'intermedi', 'lift', 'prolong', 'save']
- (8) **exchang**: ['weight', 'dollar', 'valu', 'currenc', 'depreci']
- (9) **money**: ['deposit', 'relianc', 'mutual', 'nondeposit', 'repo']



- These words are related to the Fed's 3 mandates: **employment**, **prices**, and **interest rates**.
- We may dig into these words' similarities.

Word Embedding: Word2Vec (down)

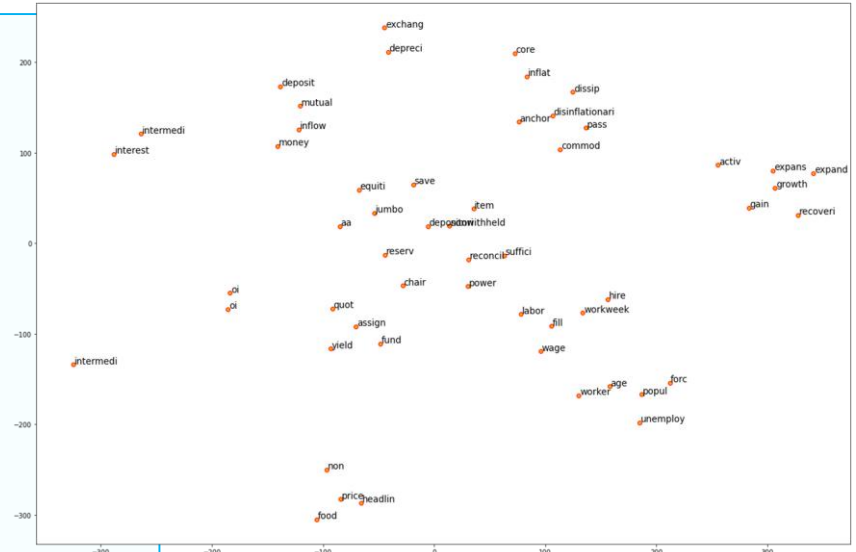
- (1) **unemploy**: ['popul', 'african', 'forc', 'workweek', 'civilian']
- (2) **labor**: ['fill', 'wage', 'payrol', 'qualifi', 'evidenc']
- (3) **growth**: ['expans', 'pace', 'expand', 'midpoint', 'gain']
- (4) **inflat**: ['disinflationari', 'calcul', 'anchor', 'firmli', 'commod']
- (5) **price**: ['food', 'headlin', 'stage', 'commod', 'pass']
- (6) **fund**: ['target', 'effr', 'overnight', 'certif', 'ioer']
- (7) **interest**: ['counter', 'refinanc', 'attract', 'intermedi', 'pent']
- (8) **exchang**: ['dollar', 'g', 'depreci', 'currenc', 'valu']
- (9) **money**: ['inflow', 'deposit', 'safeti', 'mutual', 'flight']



- These words are related to the Fed's 3 mandates: **employment**, **prices**, and **interest rates**.
- We may dig into these words' similarities.

Word Embedding: Word2Vec (unchanged)

- (1) **unemploy**: ['forc', 'popul', 'age', 'hire', 'workweek']
- (2) **labor**: ['wage', 'reconcil', 'fill', 'suffici', 'worker']
- (3) **growth**: ['expans', 'recoveri', 'gain', 'expand', 'activ']
- (4) **inflat**: ['commod', 'disinflationari', 'core', 'dissip', 'anchor']
- (5) **price**: ['food', 'headlin', 'non', 'item', 'pass']
- (6) **fund**: ['assign', 'reserv', 'quot', 'oi', 'aa']
- (7) **interest**: ['jumbo', 'yield', 'intermedi', 'oi', 'save']
- (8) **exchang**: ['depreci', 'equiti', 'intermedi', 'power', 'chair']
- (9) **money**: ['inflow', 'deposit', 'mutual', 'depositori', 'nonwithheld']



- These words are related to the Fed's 3 mandates: **employment**, **prices**, and **interest rates**.
- We may dig into these words' similarities.

Clustering the 9 keywords by word2vec + t-SNE

- [1] All
- Cluster 1: unemployment, labor
- Cluster 2: growth
- Cluster 3: inflation
- Cluster 4: price
- Cluster 5: fund, interest, exchange
- Cluster 6: money

[2] Up
Cluster 1: unemployment, labor
Cluster 2: growth
Cluster 3: inflation, price
Cluster 4: fund
Cluster 5: interest
Cluster 6: exchange
Cluster 7: money

[3] Down
Cluster 1: unemployment
Cluster 2: labor
Cluster 3: growth
Cluster 4: inflation, price
Cluster 5: fund
Cluster 6: interest
Cluster 7: exchange
Cluster 8: money

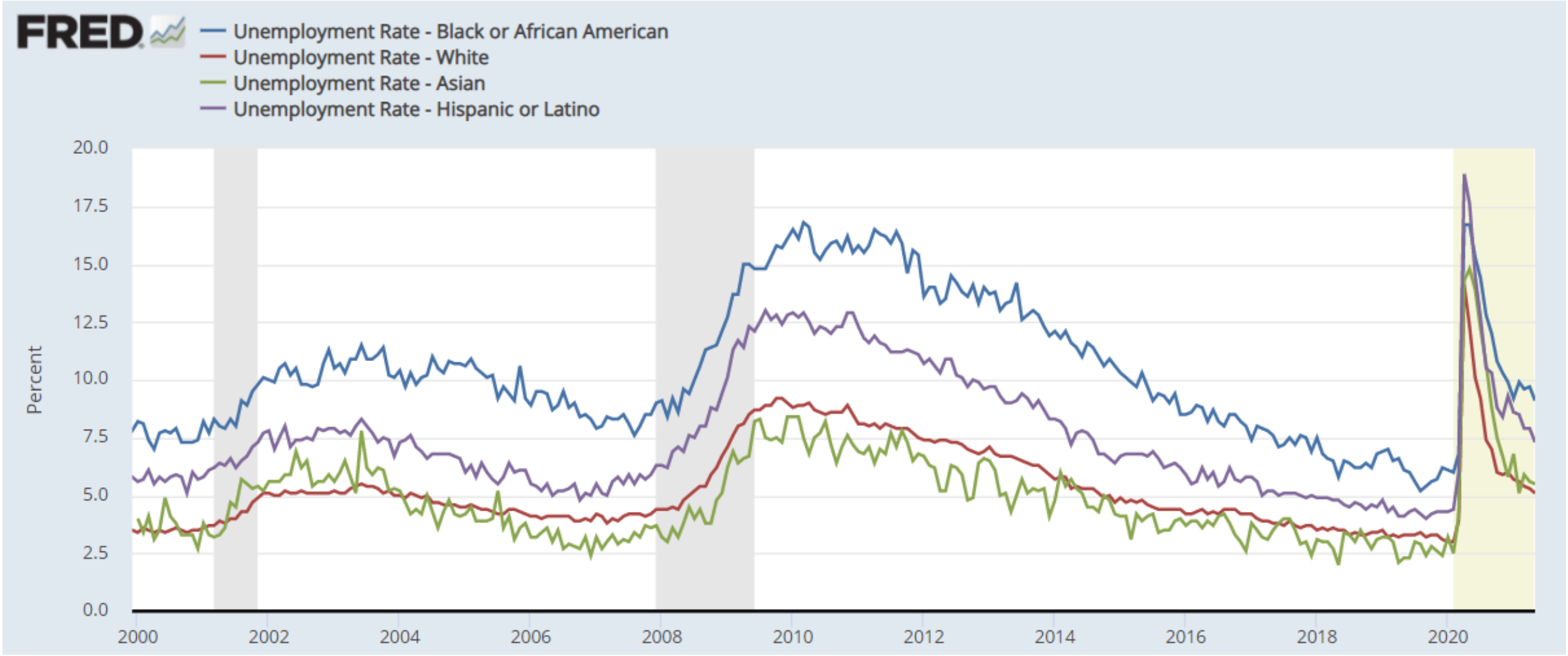
[4] Unchanged
Cluster 1: unemployment
Cluster 2: labor
Cluster 3: growth
Cluster 4: inflation
Cluster 5: price
Cluster 6: fund
Cluster 7: interest
Cluster 8: exchange
Cluster 9: money

Other Findings

(1) All	**unemploy**: ['forc', 'workweek', 'hire', 'age', 'popul', 'reason', 'jobless', 'save', 'underutil', 'interest', 'hour', 'sideway', 'summari', 'quit', 'group', 'delinqu', 'addendum', 'converg', 'work', 'fill']
(2) Up	**unemploy**: ['forc', **'african'**, **'hispan'**, 'workweek', **'american'**, 'hour', 'underutil', 'men', 'job', **'white'**, 'jobless', 'claim', 'worker', **'civilian'**, 'popul', 'monthli', 'lowest', 'payrol', 'war', 'ii']
(3) Down	**unemploy**: ['popul', **'african'**, 'age', **'civilian'**, 'workweek', 'forc', 'claim', 'hour', 'group', 'steadi', **'hispan'**, **'white'**, **'american'**, 'job', 'averag', 'insur', 'hurrican', 'payrol', 'employ', 'roughli']
(4) Unchanged	**unemploy**: ['forc', 'popul', 'age', 'workweek', 'hire', 'jobless', 'delinqu', 'work', 'addendum', 'save', 'lengthi', 'weekli', 'worker', 'job', 'reason', 'suspend', 'manifest', 'interest', 'payrol', 'durat']

- [1] Words appear in all 4 categories: 'forc', 'workweek', 'popul'
- [2] Words appear only in up & down: **'african'**, **'hispan'**, **'white'**, **'american'**, **'civilian'**

Words appear only in up & down: 'african', 'hispan', 'white', 'american', 'civilian'



- So this is where **'african'**, **'hispan'**, **'white'**, **'american'** come from.

(1) Information of Document Classification -1

- 1. ***Documents***: 222 FOMC minutes from 1993/01/01 ~ 2020/10/01.

- 2. ***Feature Extraction***: We leverage a couple of methods, including:

2-1 All vocabularies (3943 f), TF 200 (200 f), TF-IDF 200 (200 f), where “f” is “feature.”

2-2 N-filters (number of features depends)

2-3 PCA (linear dimensional reduction)

2-2 Kernel PCA (non-linear dimensional reduction)

- 3. ***Split of training & test data***: 80-20 split. We choose a random split making training data having properties similar to test data. Splitting by proportions of each groups (Up, Down, Unchanged) to solve the unbalance data problem.

Training data point may be later than the test data point.

(1) Information of Document Classification -2

- 4. ***Cross validation***: 5-fold.
- 5. ***Models***: Naïve Bayes, Logistic Regression, Linear SVM, Random Forest, Gradient Boosting Machine. All models are tuned according to training accuracy (median).
- 6. ***Evaluation criterion***: Accuracy of classifying documents into "up", "down" or "unchanged". We measure the accuracy of training data for **10 times and calculate the mean and median of them** to solve the k-fold splitting issue.
- 7. ***Baseline model***: A naïve model classifying all documents into "unchanged", yielding 68.92 % accuracy.

(2) Introduction to the Filters

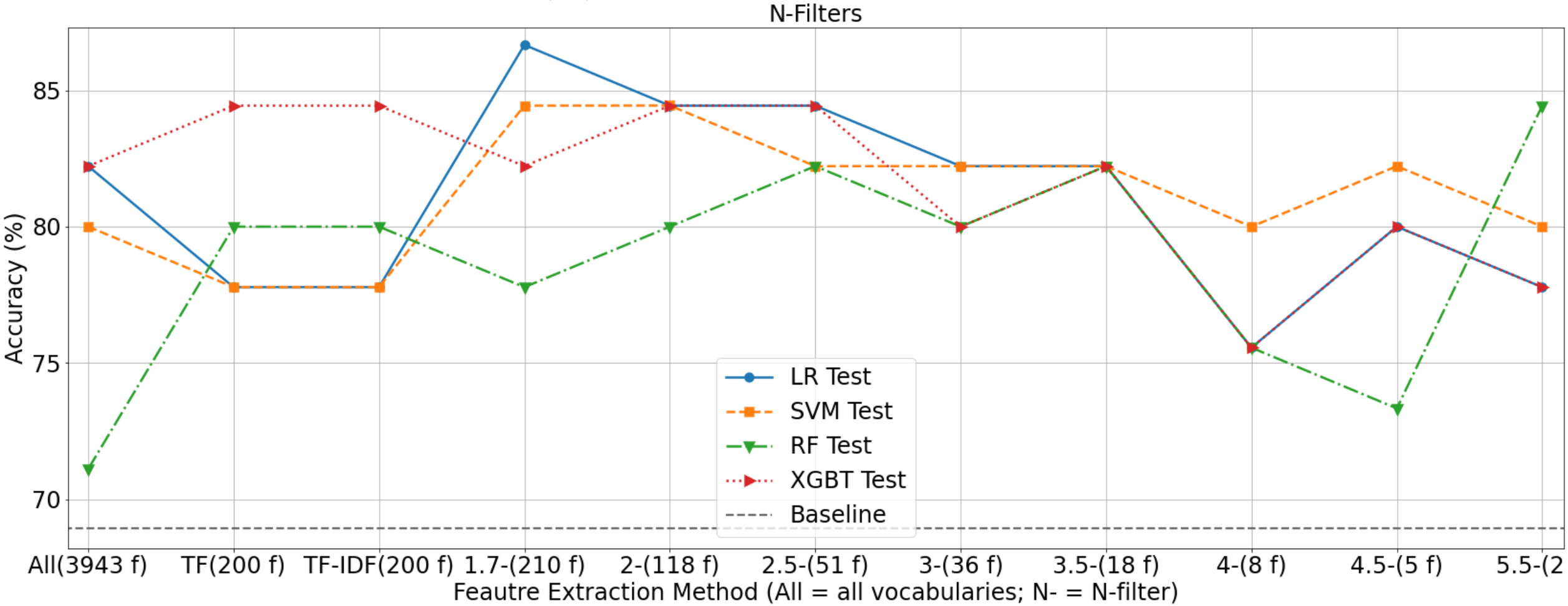
We select 3 group of words as follows to obtain the desired vocabularies:

- ✓ 1. Down / Up & Unchanged
- ✓ 2. Up / Down & Unchanged
- ✓ 3. Down / Up

Then, we choose the distinguishing words representing each groups by:

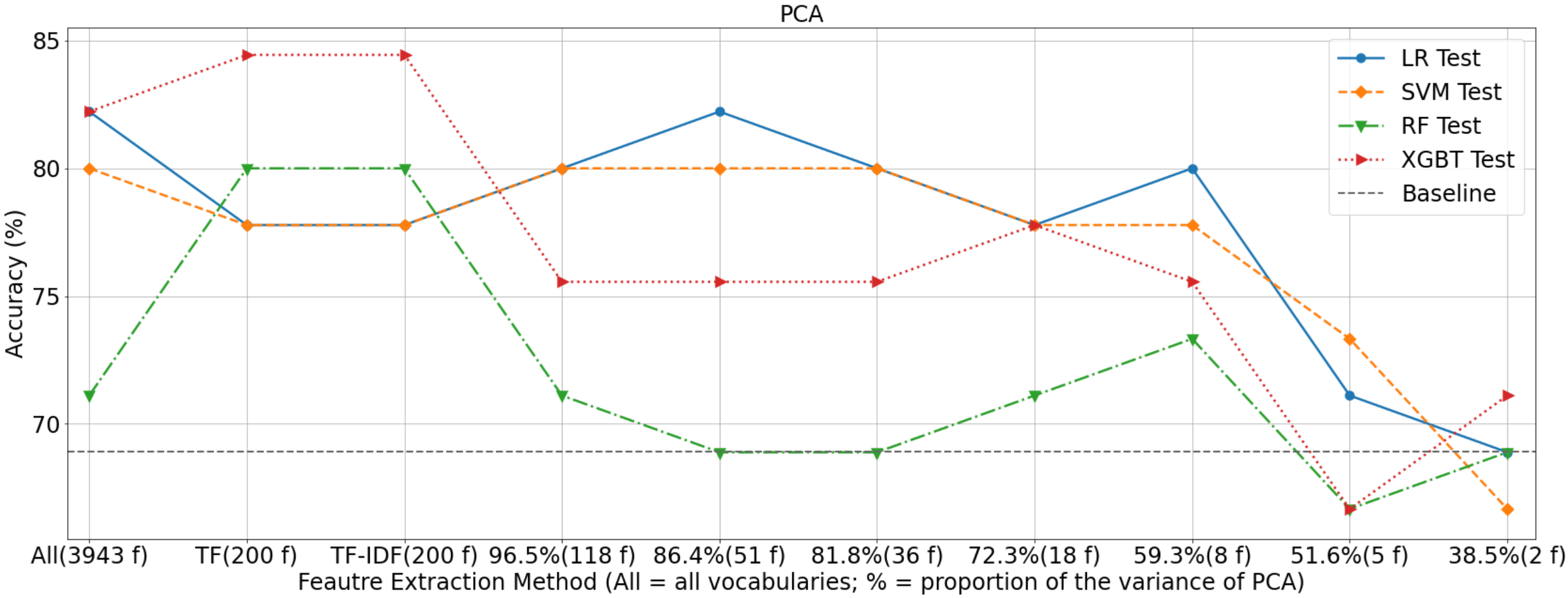
- ◆ 1. Removing the words appearing less than 20 times.
- ◆ 2. Finding the words in one group are ***N times higher than the other group*** OR ***N times lower than the other group***, where N = 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5.5.

(3) N-Filter of All Models



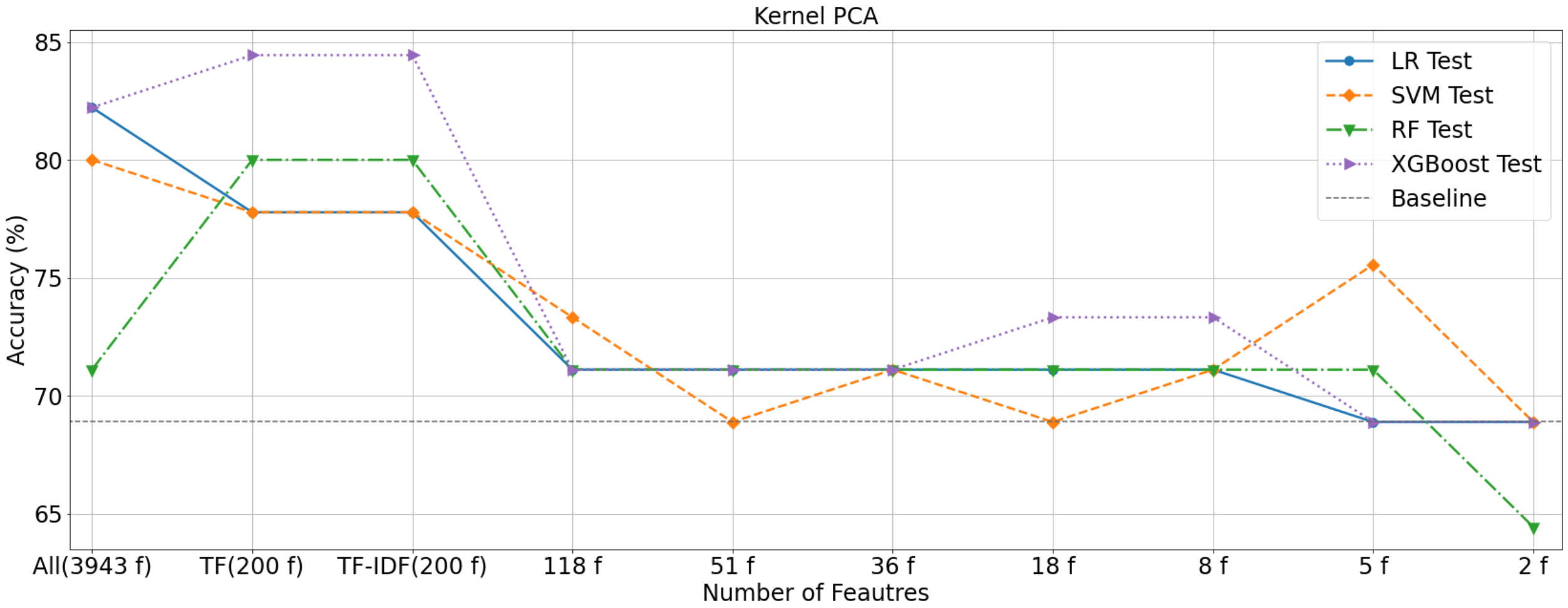
- 在 All (3943 f)、TF(200 f)、TF-IDF (200 f)，XGBoost 的表現最好；在 1.5- (210 f) 到 3.5- (36 f) 之間，LR 的表現最好；而在 4- (8 f) 到 5.5- (5f)，則是 SVM 表現最好。
- 在此，最高的分類準確率，是使用 1.5 倍率指標篩選特徵、LR 模型，只使用了 210 個特徵，卻達了 86.67%的準確率；相較之下，使用 LR 模型、全部字彙共 3943 個特徵，準確率只有 82.22%。

(6) PCA of All Models



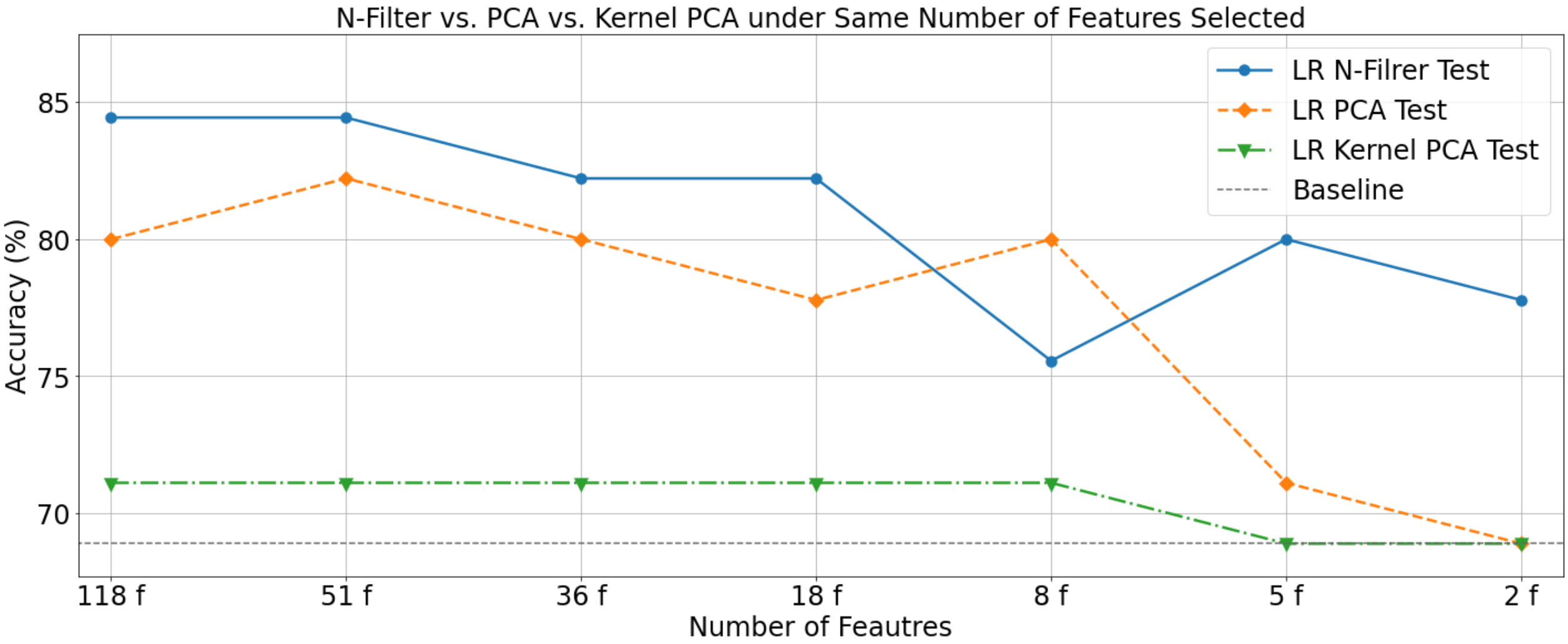
- PCA 表現比倍率指標差。
- 各模型在不同數量的特徵下的表現，和倍率指標時相似。

(7) Kernel PCA



- 表現比 PCA 差。

(8) N-Filter vs. PCA vs. Kernel PCA



- 準確率表現：倍率指標 > PCA > Kernel PCA。

(9) 最佳模型和參數選擇的混淆矩陣

- 因為使用 LR 模型、1.5 倍率指標 (共 210 個特徵) 在測試資料集上得到了最高的 86.67% 準確率，所以進一步觀察其混淆矩陣。(圖中 +1 為升息、0 為利率不變，-1 為降息)
- 發現錯誤分類的 6 個樣本中，5 個是將升息誤認為利率不變，剩下 1 個是將不變誤認為升息。這代表了降息樣本跟其他兩類差異過大，導致降息樣本不會被錯誤分類為升息或利率不變，以及升息、利率不變樣本不會被錯誤分類為降息。

		Predicted		
		-1	0	1
Actual	-1	6	0	0
	0	0	30	1
	1	0	5	3

	precision	recall	f1-score	support
-1	1.00	1.00	1.00	6
0	0.86	0.97	0.91	31
1	0.75	0.38	0.50	8
accuracy			0.87	45
macro avg	0.87	0.78	0.80	45
weighted avg	0.86	0.87	0.85	45

(10) Conclusion of Classification

- 在使用倍率指標、PCA 來選取特徵時，LR、SVM 模型表現比 RF、XGBoost，推測是因為少數特徵對 FOMC minutes 的三分類準確率有非常重要的影響，因此使用更複雜的集成學習方法的 RF、XGBoost 模型，準確率反而輸給使用簡單的 LR、SVM。

- 進一步探究，用 3.5 倍率指標篩選出來的 18 個特徵，就能使 LR 模型得到 82.22% 的準確率 ，而使用 LR 模型、全部字彙 (3943 個特徵) 得到的準確率也是 82.22% 。

- 這 18 具有很高分類影響力的特徵，羅列如下：{'contract', 'turmoil', 'outbreak', 'weak', 'institut', 'eas‘, 'weaker', 'lend', 'correct', 'recoveri', 'downturn', 'weaken', 'rebuild', 'medium', 'remov', 'contain‘, 'coronaviru', 'hurrican'}。

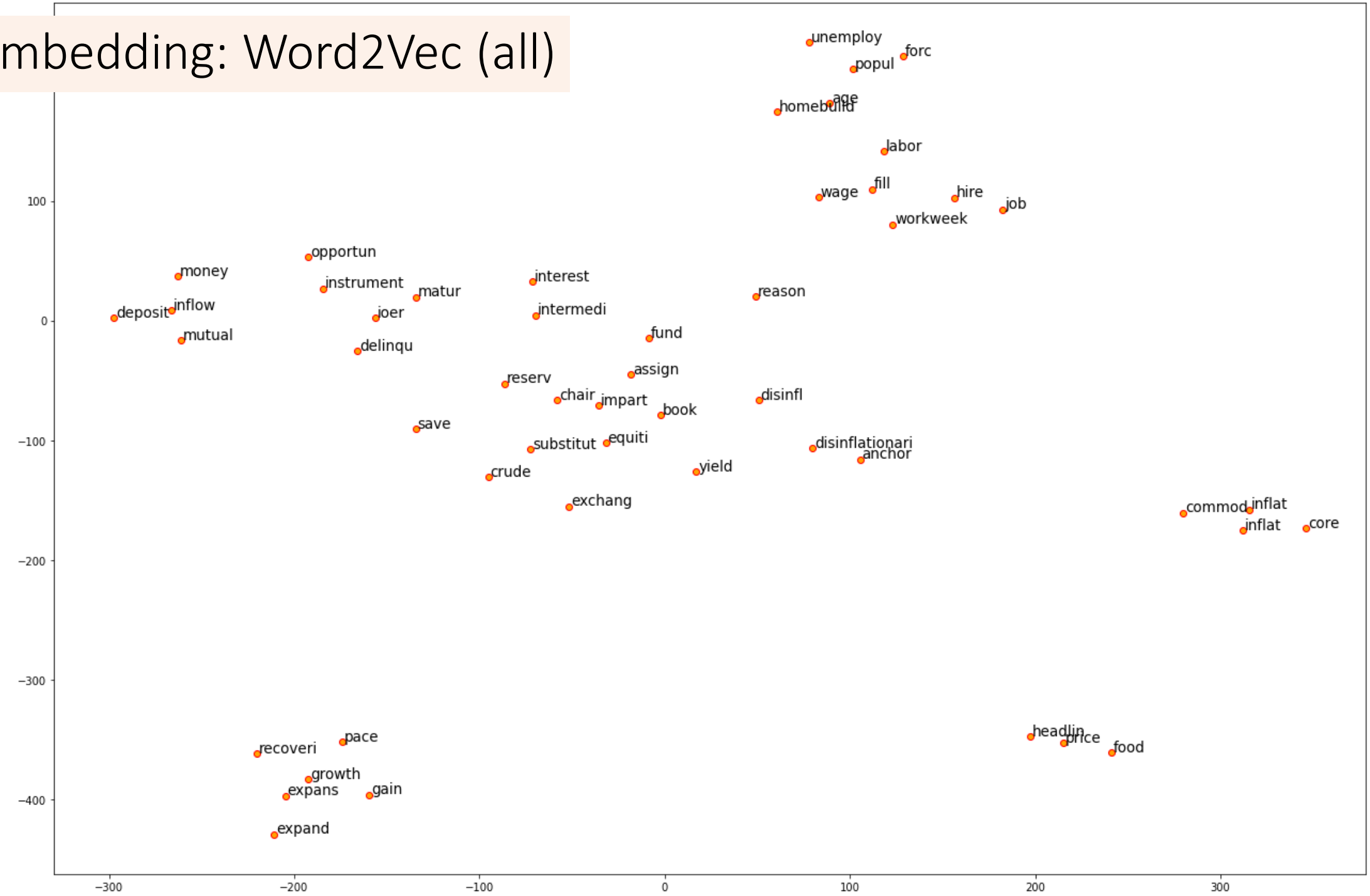
(11) Comparison with Previous Literature

表 4-1-7　全部單詞及前兩百大單詞的分三類結果比較

分三類	k=90%	k=95%	k=95% Entropy(all)	k=95% Entropy(top50)
全部單詞	69.43%	76.68%	77.20%	77.72%
TF 前兩百大	64.25%	72.54%	73.58%	75.13%
正確率差異	5.18%	4.15%	3.63%	2.59%

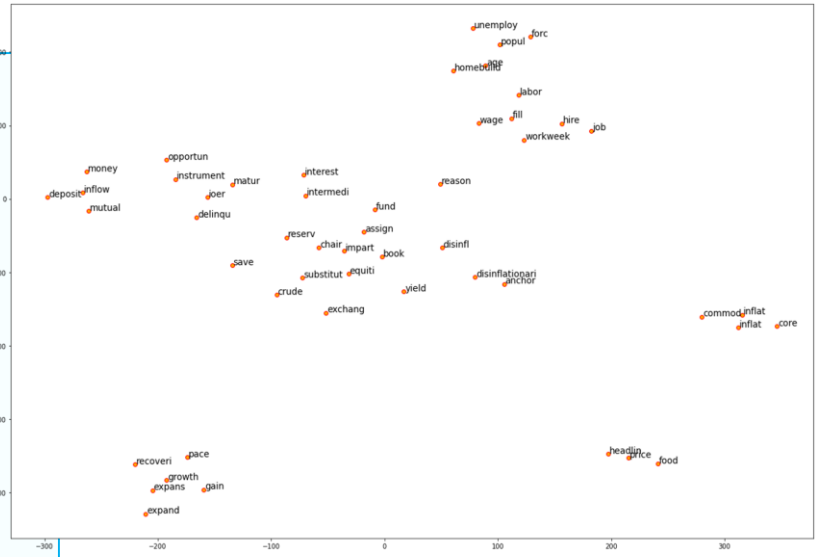
- We somehow get significant improvements in contrast to previous thesis by Huang (2017); however, we use FOMC minutes 1993/01/01 ~ 2020/10/01 whereas Huang (2017) use 1993/01/01 ~ 2017/04/01.

Word Embedding: Word2Vec (all)



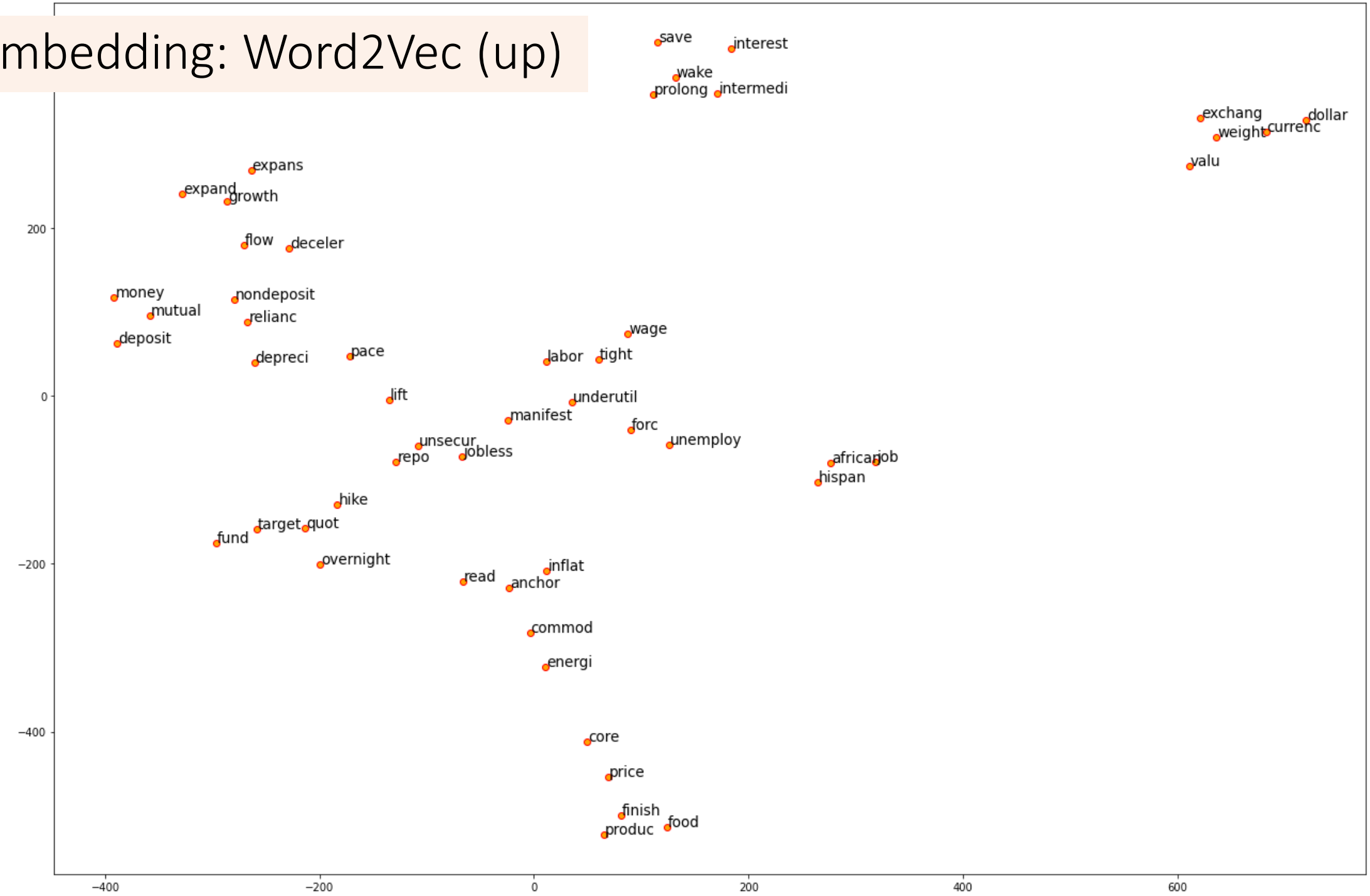
Word Embedding: Word2Vec (all)

- (1) **exchang**: ['equiti', 'chair', 'book', 'substitut', 'impart']
- (2) **fund**: ['reserv', 'assign', 'matur', 'ioer', 'interest']
- (3) **growth**: ['expans', 'recoveri', 'gain', 'pace', 'expand']
- (4) **inflat**: ['commod', 'disinflationari', 'core', 'disinfl', 'anchor']
- (5) **interest**: ['intermedi', 'yield', 'save', 'delinqu', 'homebuild']
- (6) **labor**: ['wage', 'fill', 'inflat', 'age', 'job']
- (7) **money**: ['inflow', 'deposit', 'mutual', 'instrument', 'opportun']
- (8) **price**: ['headlin', 'food', 'core', 'crude', 'inflat']
- (9) **unemploy**: ['forc', 'hire', 'popul', 'workweek', 'reason']



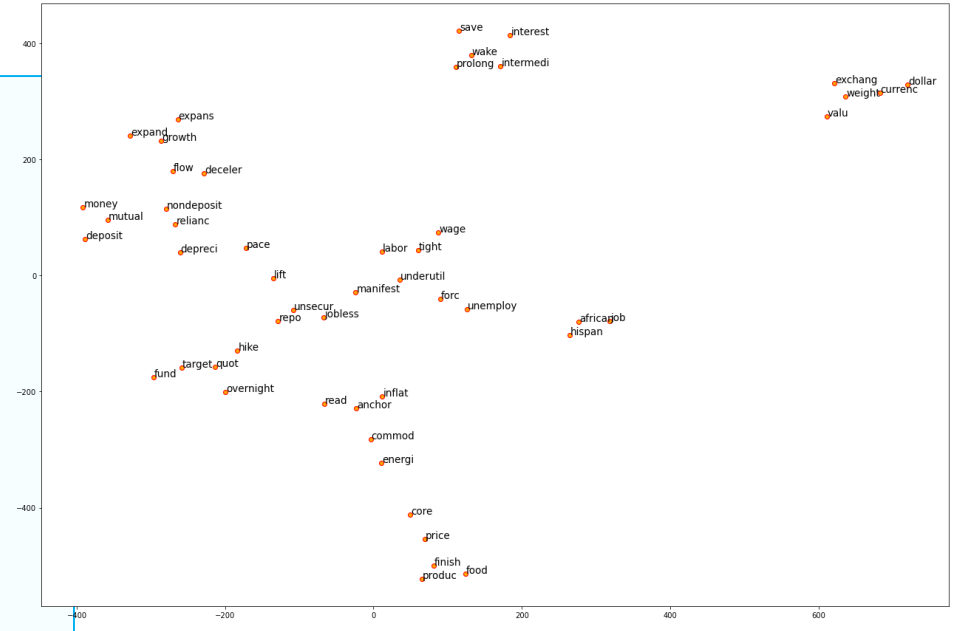
- These words are related to the Fed's 3 mandates: **employment**, **prices**, and **interest rates**.
- We may dig into these words' similarities.

Word Embedding: Word2Vec (up)



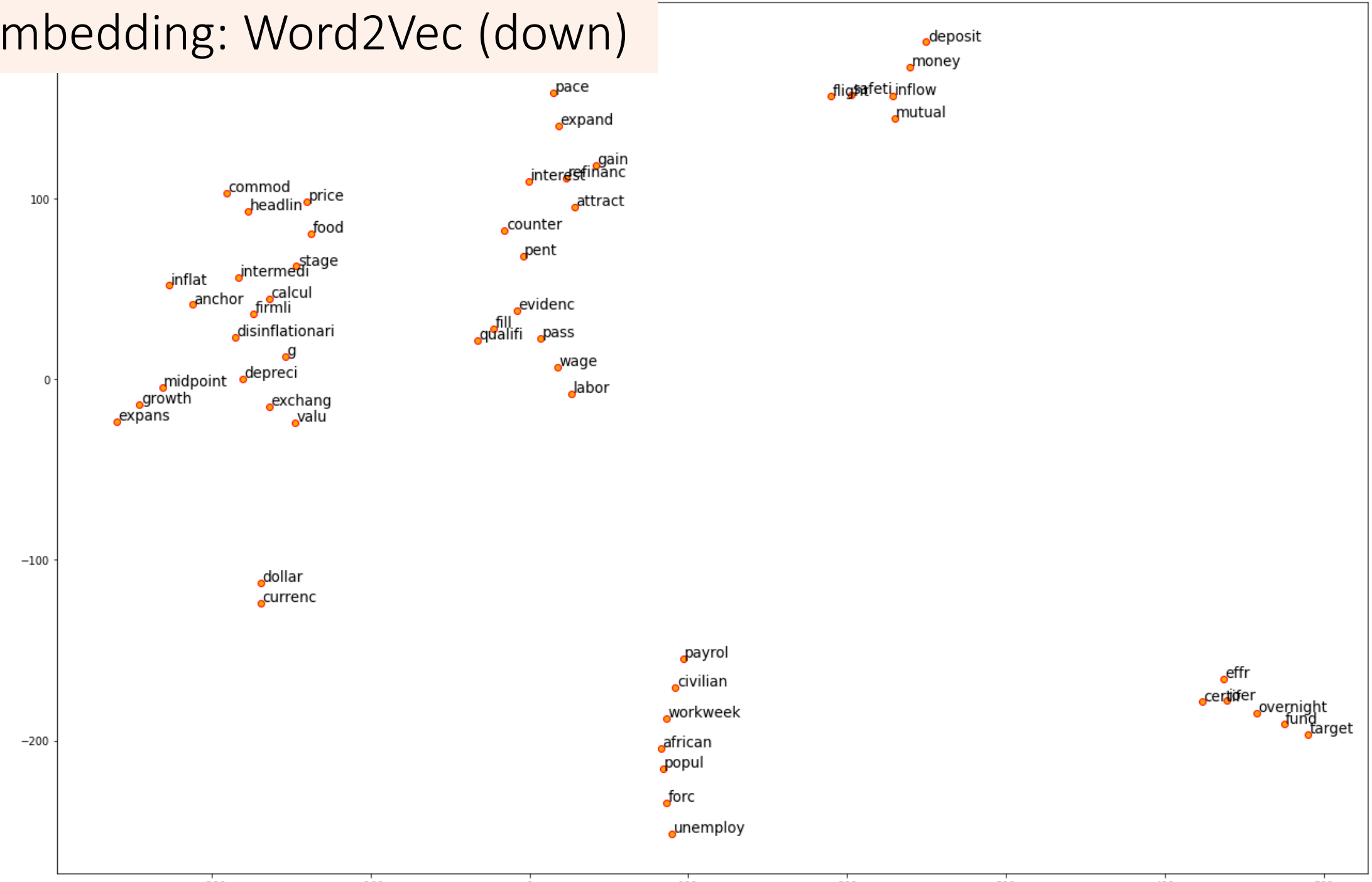
Word Embedding: Word2Vec (up)

- (1) **exchang**: ['weight', 'dollar', 'valu', 'currenc', 'depreci']
- (2) **fund**: ['target', 'quot', 'hike', 'unsecur', 'overnight']
- (3) **growth**: ['expans', 'expand', 'pace', 'flow', 'deceler']
- (4) **inflat**: ['core', 'read', 'energi', 'anchor', 'commod']
- (5) **interest**: ['wake', 'intermedi', 'lift', 'prolong', 'save']
- (6) **labor**: ['wage', 'manifest', 'job', 'underutil', 'tight']
- (7) **money**: ['deposit', 'relianc', 'mutual', 'nondeposit', 'repo']
- (8) **price**: ['core', 'food', 'finish', 'produc', 'commod']
- (9) **unemploy**: ['forc', 'african', 'hispan', 'job', 'jobless']



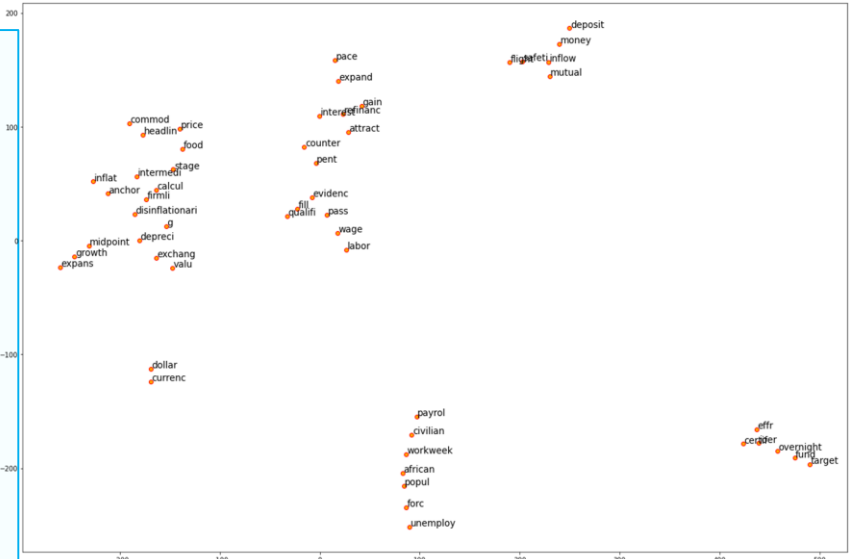
- These words are related to the Fed's 3 mandates: **employment**, **prices**, and **interest rates**.
- We may dig into these words' similarities.

Word Embedding: Word2Vec (down)



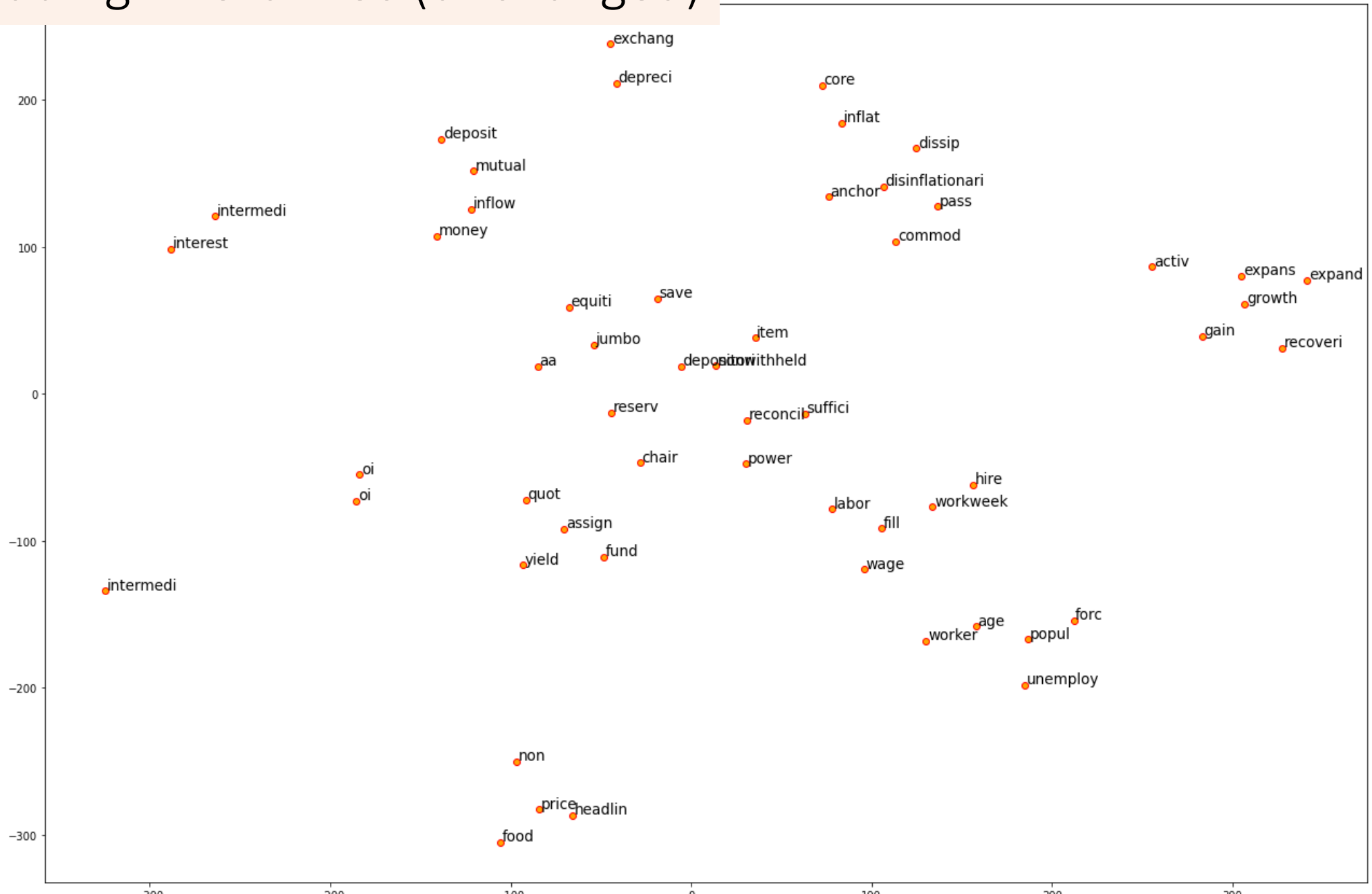
Word Embedding: Word2Vec (down)

- (1) **exchang**: ['dollar', 'g', 'depreci', 'currenc', 'valu']
- (2) **fund**: ['target', 'effr', 'overnight', 'certif', 'ioer']
- (3) **growth**: ['expans', 'pace', 'expand', 'midpoint', 'gain']
- (4) **inflat**: ['disinflationari', 'calcul', 'anchor', 'firmli', 'commod']
- (5) **interest**: ['counter', 'refinanc', 'attract', 'intermedi', 'pent']
- (6) **labor**: ['fill', 'wage', 'payrol', 'qualifi', 'evidenc']
- (7) **money**: ['inflow', 'deposit', 'safeti', 'mutual', 'flight']
- (8) **price**: ['food', 'headlin', 'stage', 'commod', 'pass']
- (9) **unemploy**: ['popul', 'african', 'forc', 'workweek', 'civilian']



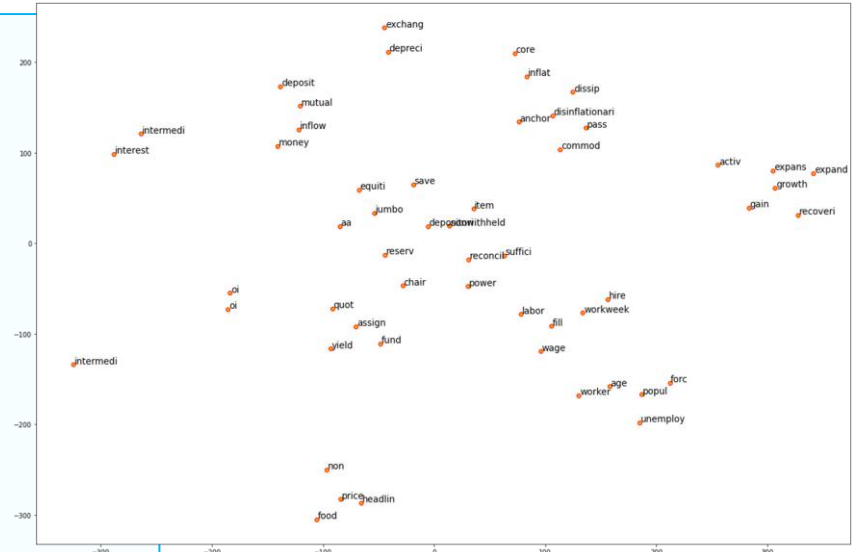
- These words are related to the Fed's 3 mandates: **employment**, **prices**, and **interest rates**.
- We may dig into these words' similarities.

Word Embedding: Word2Vec (unchanged)



Word Embedding: Word2Vec (unchanged)

- (1) **exchang**: ['depreci', 'equiti', 'intermedi', 'power', 'chair’]
- (2) **fund**: ['assign', 'reserv', 'quot', 'oi', 'aa’]
- (3) **growth**: ['expans', 'recoveri', 'gain', 'expand', 'activ’]
- (4) **inflat**: ['commod', 'disinflationari', 'core', 'dissip', 'anchor’]
- (5) **interest**: ['jumbo', 'yield', 'intermedi', 'oi', 'save’]
- (6) **labor**: ['wage', 'reconcil', 'fill', 'suffici', 'worker’]
- (7) **money**: ['inflow', 'deposit', 'mutual', 'depositori', 'nonwithheld’]
- (8) **price**: ['food', 'headlin', 'non', 'item', 'pass’]
- (9) **unemploy**: ['forc', 'popul', 'age', 'hire', 'workweek']



- These words are related to the Fed’s 3 mandates: **employment**, **prices**, and **interest rates**.
- We may dig into these words’ similarities.

(1) Input Word2Vec’s t-SNE

```
similar_words = {
    'exchang': ['clip', 'competitor', 'equiti', 'depreci', 'impart’],
    'fed': ['statement', 'descript', 'misunderstand', 'solidifi', 'sentenc’],
    'growth': ['expans', 'gain', 'expand', 'pace', 'recoveri’],
    'inflat': ['commod', 'disinflationari', 'disinfl', 'anchor’],
    'interest': ['intermedi', 'yield', 'r’],
    'labor': ['contradictori', 'nonwag', 'wage', 'reconcil', 'unambigu’],
    'money': ['inflow', 'deposit', 'mutual', 'instrument', 'depositor’],
    'price': ['headlin', 'food', 'non', 'pass', 'core’],
    'unemploy': ['forc', 'workweek', 'hire', 'ployment', 'unem']}
```

(2) Most Similar Words: fed & growth

- (1) w2v_model.wv.most_similar("fed")

[('statement', 0.4344236850738525 4),
('descript', 0.4153550267219543 5),
('misunderstand', 0.387560427188873 3),
('solidifi', 0.3779264688491821 3),
('sentenc', 0.3772746920585632 3),
('persuas', 0.351927220821380 6),
('word', 0.3511220514774322 5),
('minut', 0.3497819304466247 6),
('narrowli', 0.346691757440567),
('delet', 0.3324368000030517 6)]

- (2) w2v_model.wv.most_similar("growth")

[('expans', 0.7371159195899963),
('gain', 0.4460225403308868 4),
('expand', 0.4403714537620544 4),
('pace', 0.407827973365783 7),
('recoveri', 0.399214267730712 9),
('activ', 0.364892661571502 7),
('strive', 0.357284182705688 5),
('registr', 0.346870660781860 35),
('buildup', 0.334121257066726 7),
('nipa', 0.3148405551910400 4)]

(2) Most Similar Words : inflat & interest

- (3) w2v_model.wv.most_similar("inflat")

[('commod', 0.585535764694213 9),
('disinflationari', 0.525047898292541 5),
('core', 0.481241583824157 7),
('disinfl', 0.432153701782226 56),
('anchor', 0.426322013139724 73),
('energi', 0.418815076351165 77),
('transitori', 0.413706392049789 43),
('price', 0.39166384935379 03),
('benign', 0.37420865893363 95),
('gasolin', 0.37243175506591797)]

- (4) w2v_model.wv.most_similar("interest")

[('intermedi', 0.436071336269378 66),
('yield', 0.42109403014183044),
('r', 0.34883520007133484),
('disentangl', 0.34690099954605 1),
('save', 0.34402811527252 2),
('jumbo', 0.33732223510742 19),
('premium', 0.33215617777252 2),
('ioer', 0.31185999512672424),
('score', 0.30941528081893 92),
('inexpens', 0.30903333425521 85)]

(3) Most Similar Words : labor & money

- (5) w2v_model.wv.most_similar("labor")

[('contradictori', 0.4790265858173370 4),
('nonwag', 0.3961641490459442),
('wage', 0.3721073865890503),
('reconcil', 0.3551770746707916),
('unambigu', 0.3529078066349029 5),
('involuntari', 0.3427288830280304),
('recruit', 0.3414557278156280 5),
('entic', 0.3284064829349518),
('men', 0.3271397948265075 7),
('conjunctur', 0.3249199092388153)]

- (6) w2v_model.wv.most_similar("money")

[('inflow', 0.6170750856399536),
('deposit', 0.5829259157180786),
('mutual', 0.5383480191230774),
('instrument', 0.5042903423309326),
('depositor', 0.4492363035678863 5),
('nonwithheld', 0.4289240241050 72),
('opportun', 0.4233032763004303),
('bulg', 0.4125939011573791 5),
('depositori', 0.3971582949161529 5),
('heavier', 0.3866728544235229 5)]

(3) Most Similar Words : price & unemploy

- (7) w2v_model.wv.most_similar("price")

[('headlin', 0.5150044560432434),
('food', 0.455824613571167),
('non', 0.442389070987701 4),
('pass', 0.4265880584716797),
('core', 0.4244893193244934),
('inflat', 0.3916638493537903),
('administ', 0.369436115026474),
('crude', 0.3612046241760254),
('er', 0.3569049239158 6304),
('valu', 0.3555984795093 5364)]

- (8) w2v_model.wv.most_similar("unemploy")

[('forc', 0.514600396156311),
('workweek', 0.4797388911247 2534),
('hire', 0.4521238207817 0776),
('ployment', 0.4509002566337 5854),
('unem', 0.420595884323 1201),
('age', 0.4154482185840 6067),
('men', 0.4144760370254 5166),
('reason', 0.3815283179283142),
('popul', 0.3700640499591 8274),
('women', 0.3414238691329956)]