以法說會逐字稿
預測半導體產業前景

研究流程

數據蒐集─
費半30歷史企業

數據處理─
詞性還原、斷詞

自編
市值加權指數

市場狀態分類─
Good/Bad/Neutral

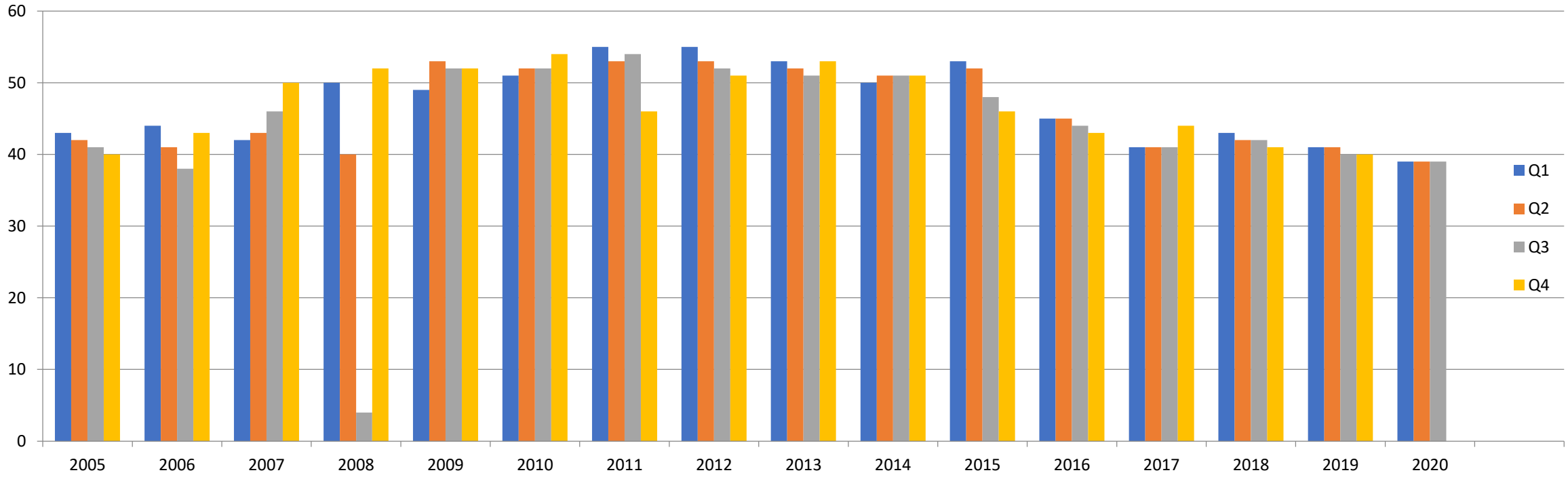
統計檢定─
將詞按照市場狀況分類

建構機器學習預測模型

半導體產業
前景預測

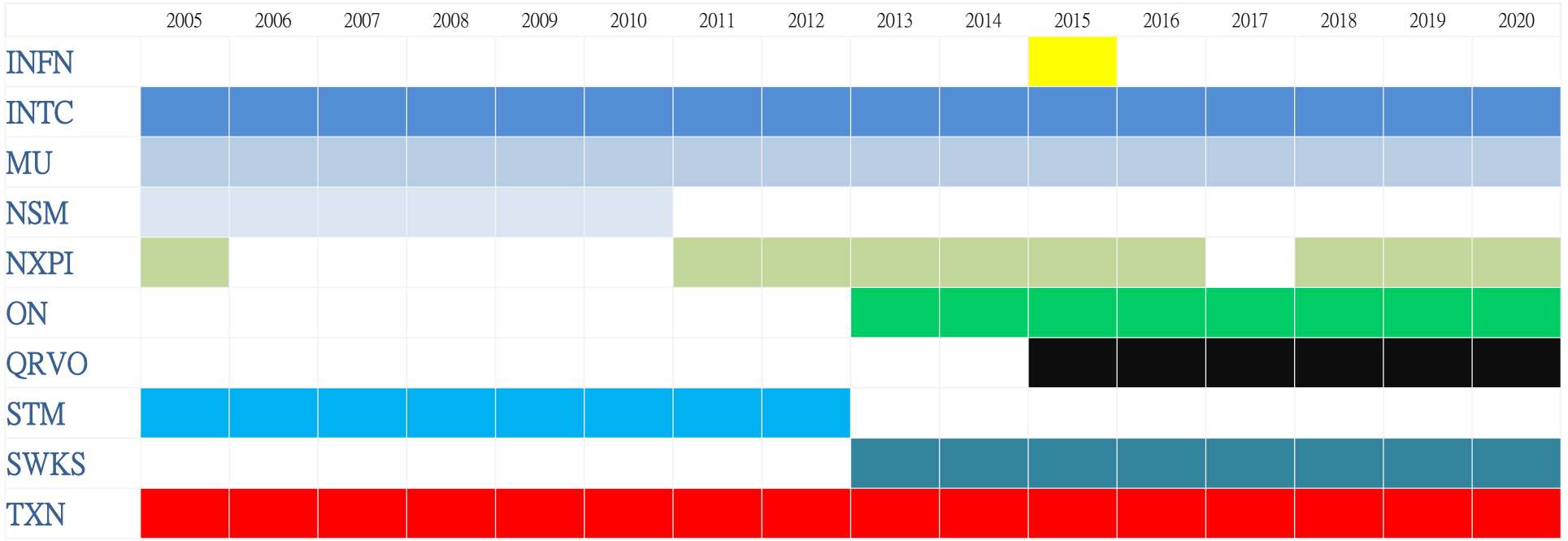
資料 – 歷史費半企業

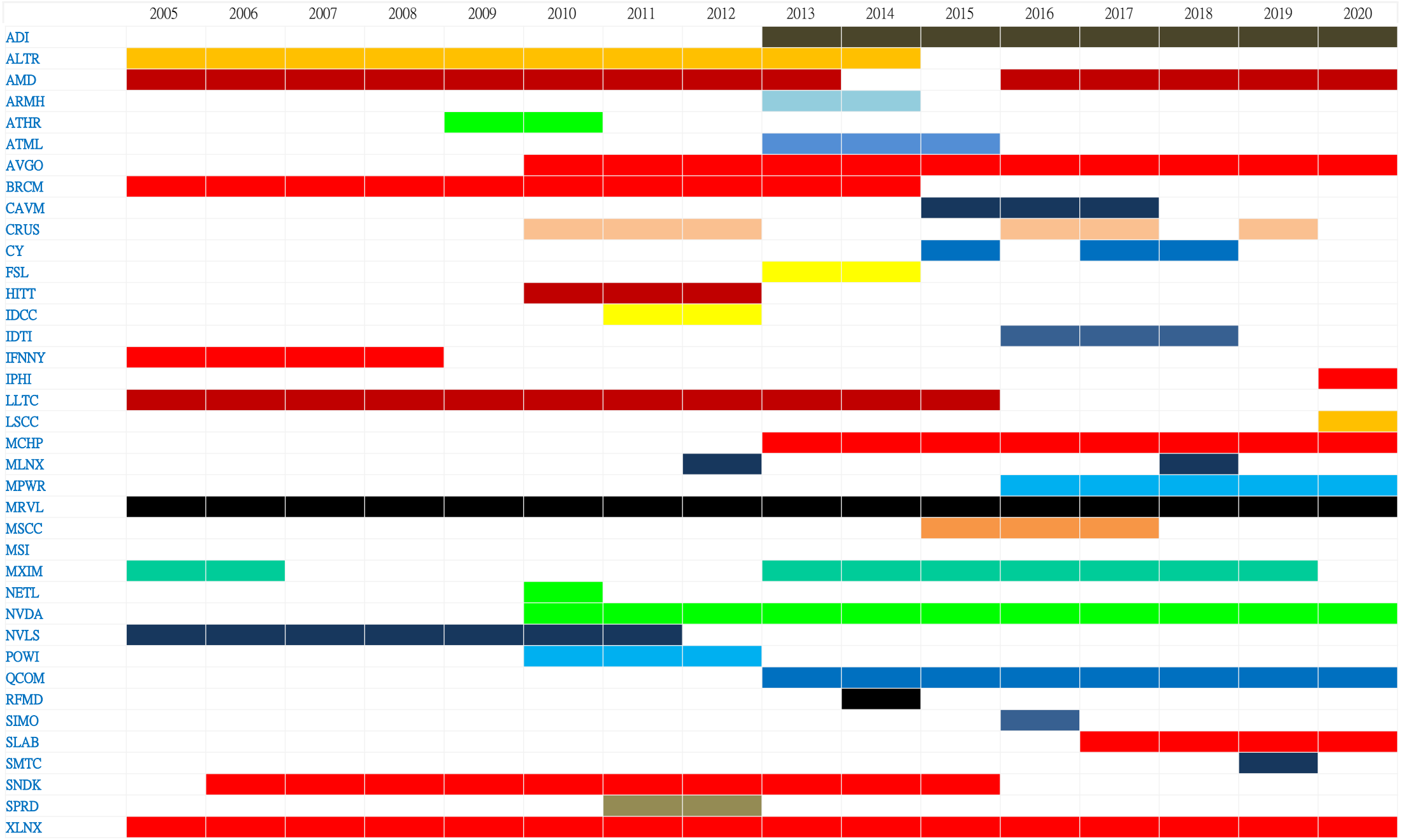
- 期間：2005/Q3 – 2020/Q3
- 總公司數：63
- 總文本數：2810

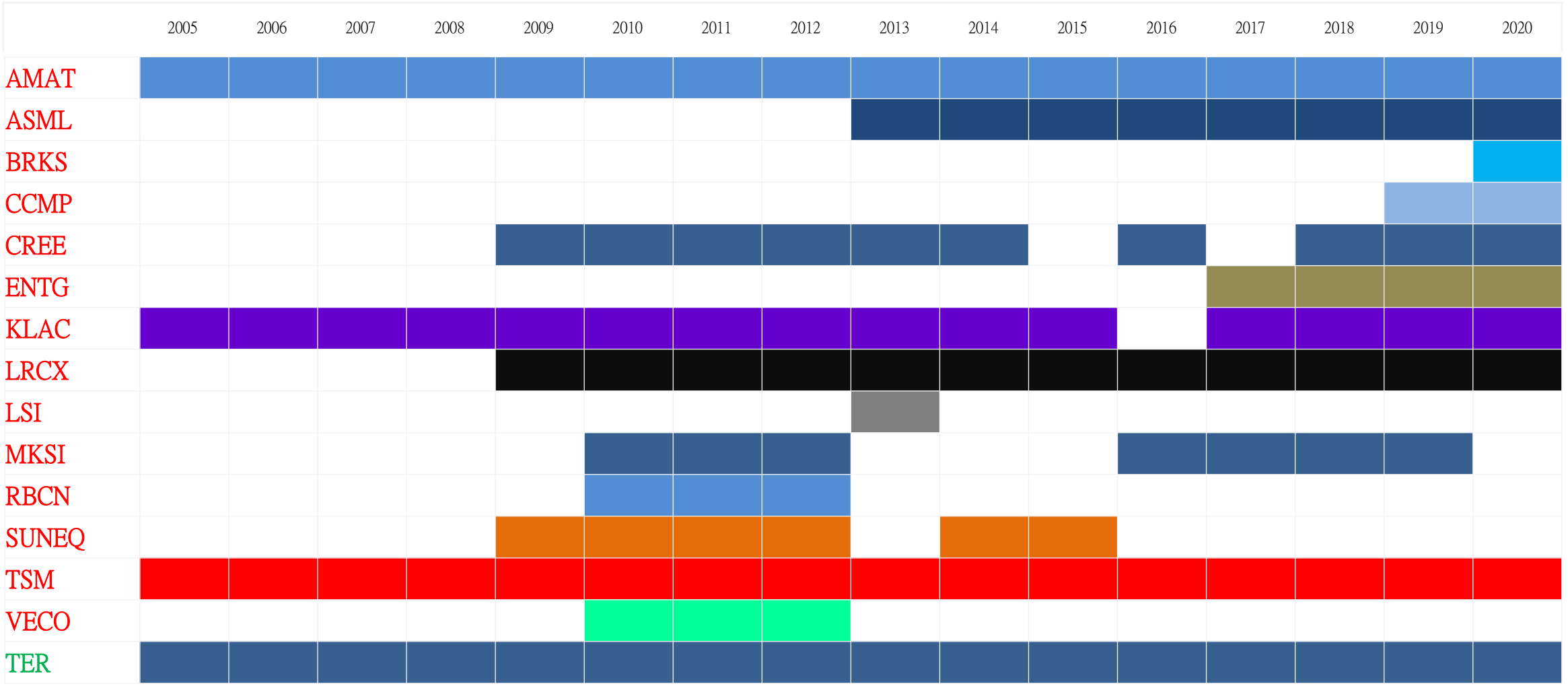


資料 – 歷史費半企業

- 上游：48
- 中游：14
- 下游：1




2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
ADI
ALTR
AMD
ARMH
ATHR
ATML
AVGO
BRCM
CAVM
CRUS
CY
FSL
HITT
IDCC
IDTI
IFNNY
IPHI
LLTC
LSCC
MCHP
MLNX
MPWR
MRVL
MSCC
MSI
MXIM
NETL
NVDA
NVLS
POWI
QCOM
RFMD
SIMO
SLAB
SMTC
SNDK
SPRD
XLNX


2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
AMAT
ASML
BRKS
CCMP
CREE
ENTG
KLAC
LRCX
LSI
MKSI
RBCN
SUNEQ
TSM
VECO
TER

數據處理—詞性還原(Lemmatization)

英文單字會因時態、單複數不同而變化，若不處理會造成文字探勘研究的偏誤，例如 the performance looks good 和 the performance is better than last year 兩句話的 good 和 better 是比較級關係，卻會被當成兩個不同的單字

- 使用套件：NLTK + Stanza(美國 Stanford大學開發之語言處理套件)
  - 以 it’ s better than before 為例
  - NLTK: it ‘s good than before
  - Stanza: it be better than before
- Stanza 無法處理形容詞之詞性還原、NLTK不夠細緻，縮寫(ex. 無法處理 It ‘s)
- 目標：

went/ goes → go
cars → car
better → good

數據處理–斷詞(Segmentation)

先進行各種文本預處理，例如透過人工標記的方式保留完整片語、去除符號及stop words，使研究更精確

- 使用套件：NLTK
- 斷詞預處理：去除符號及stopwords 後，在保留片語的前提下將句子斷成單詞
- 以 However, there are a lot of companies doing this!為例

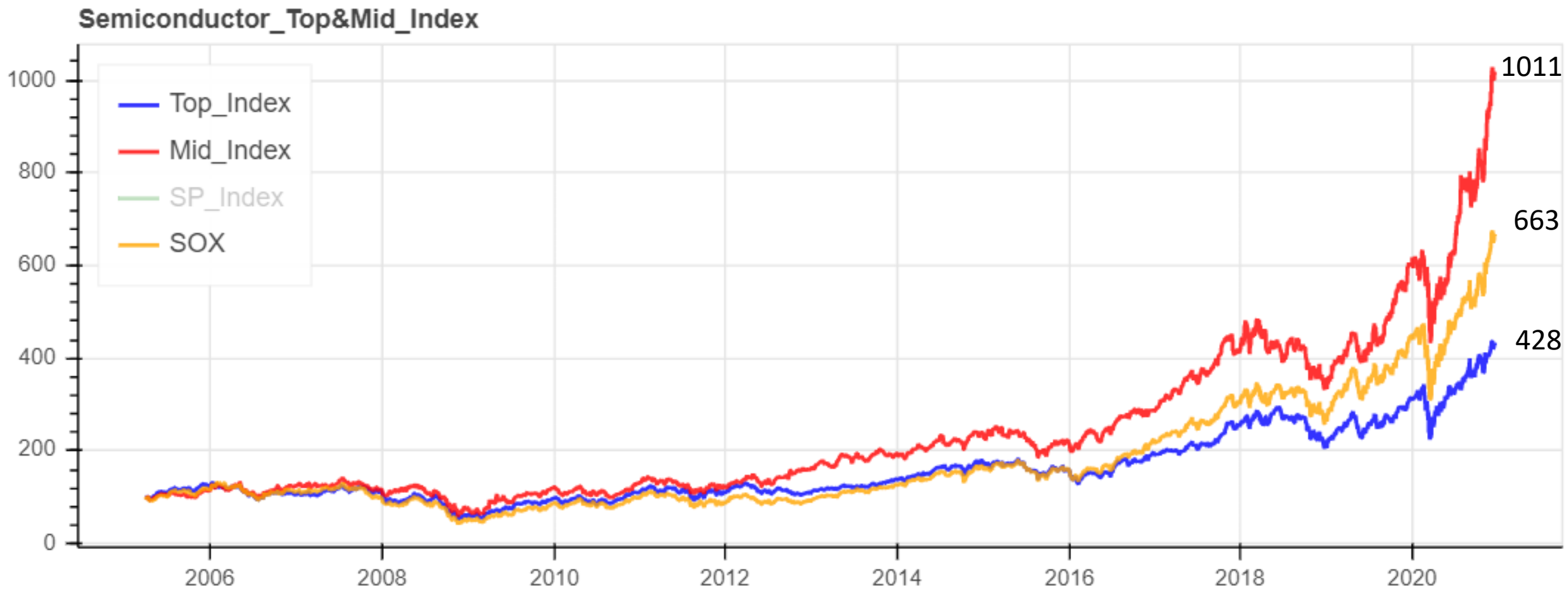
處理順序

  - 詞性還原後的句子: however, there be a lot of company do this!
  - 去除符號及 stopwords：however there a lot of company do this
  - 保留片語進行斷詞： however, there, a lot of, company, do, this
  - 若不保留片語語意會不精準：however, there, a, lot, of, company, do, this → 得到一串詞的 list 以進行後續分析

- Stop words 定義(Stanford)：some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called *stop words* .

自編市值加權指數

將半導體公司區分為上游、中游，並以建立景氣指數，用以後續判斷state



市場狀態分類—Good/Neutral/Bad

以自編加權指數作為判斷依據，依照報酬的分配區分為三個不同的狀態，分配前25%視作Good狀態，中間50%視作Neutral狀態，後25%視作Bad狀態

- 上游State結果

quarter	state
Q3_05	good
Q4_05	neutral
Q1_06	neutral
Q2_06	bad
Q3_06	neutral
Q4_06	bad
Q1_07	bad
Q2_07	good
Q3_07	neutral
Q4_07	bad

- 中游State結果

quarter	state
Q3_05	neutral
Q4_05	neutral
Q1_06	neutral
Q2_06	bad
Q3_06	neutral
Q4_06	neutral
Q1_07	neutral
Q2_07	neutral
Q3_07	neutral
Q4_07	bad

市場狀態結果—Good/Neutral/Bad

經過t-test篩選之後，將具統計顯著的詞過濾出來，再判斷詞是否在某個state時與其他state有遞減或遞增的關係，同時滿足這兩個條件，才能夠作為有意義的詞

```
t1,p1 ,t2,p2= check_words_t_test_result('growth')
print(f'good bad t-value :{t1}')
print(f'good bad p-value :{p1}')
print(f'good neu t-value :{t2}')
print(f'good neu p-value :{p2}')
```

```
<ipython-input-5-8c3172c7f9e5>:16: FutureWarning:
will perform elementwise comparison
  index = np.argwhere(words == keyword)
```

```
good mean freq :0.0035780872088622346
bad mean freq :0.0037920120814475934
neutral mean freq :0.00432032226175882
good bad t-value :-1.6045103547971702
good bad p-value :0.10860786452656342
good neu t-value :-5.722513564604276
good neu p-value :1.0538800254848262e-08
```

```
t1,p1 ,t2,p2= check_words_t_test_result('business')
print(f'good bad t-value :{t1}')
print(f'good bad p-value :{p1}')
print(f'good neu t-value :{t2}')
print(f'good neu p-value :{p2}')
```

```
<ipython-input-5-8c3172c7f9e5>:16: FutureWarning: ele
will perform elementwise comparison
  index = np.argwhere(words == keyword)
```

```
good mean freq :0.0053045445417035245
bad mean freq :0.006199974587582467
neutral mean freq :0.005744301231630468 5
good bad t-value :-4.840397604210667
good bad p-value :1.2996223000160013e-06
good neu t-value :-2.854071068585927
good neu p-value :0.004317558789745016
```

1st time BERT model (base-cased)

- 總樣本數: 98965 (pos:25747, neu:46618, neg:26601)
- 訓練集: 69276(pos:7724, neu:13986, neg:7980)
- 樣本外:pos: 20/83, neu: 44/136, neg: 61/81

改進流程

- 嘗試使用其他預訓練模型(XLNet、ELECTRA.....)
- 將State 分法修正，依照個別公司股價表現進行標註
- 取完顯著字詞後進行產業內專業字詞篩選
- 嘗試其他預測方式(Dictionary based等)