Literature Review
on Fraud Detection in Financial Statements

2019.08.16.

박지혜

jihyeparkk@dm.snu.ac.kr

Table of Contents

Humpherys, Sean L., et al.
"**Identification of fraudulent financial statements using linguistic credibility analysis.**"
Decision Support Systems 50.3 (2011): 585-594.

Agent99 Analyzer

Introduction: Fraud Detection using Text-mining

- Why text-mining?
  - Fraud detection은 회계감사 전문가들도 풀기 힘들어하는 문제이고, text-mining을 통해 전문가들의 의사결정을 도울 수 있을 것이다.
    - Since most **external auditors have low experience in detecting fraud**, finding decision aids to help auditors detect fraud is critical.
    - KPMG Fraud Survey에 따르면, 12% 정도가 external auditors에 의해 밝혀지고, 65% 정도가 internal audit department에 의해 밝혀진다.
  - Natural language processing can help **determine veracity by identifying textual cues** that indicate **the intent of the writer(s)** in an organizational reporting context.

Introduction: Agent99 Analyzer

- **Agent99 Analyzer**
  - Extracts **text-based cues for deception detection** in fraudulent and non-fraudulent financial statements.
- Performance
  - **Humans** can only successfully detect deception at an average of **54%** (There are not any studies empirically quantifying an auditor's ability to detect financial statement fraud accurately.)
  - **Agent99 Analyzer performed promisingly well with up to 67%** accuracy in discriminating between fraudulent and non-fraudulent 10-Ks.

Importance of **MD&A Section in 10-Ks**

- Data
  - **MD&A section** of 101 fraudulent and 101 non-fraudulent **10-Ks**.
    - Current **status** of the company
    - The results of operations
    - Company's **financial condition**
    - Current status of the industry
    - An analysis of the quantitative and qualitative **market risks** facing the company
    - **Forward looking statements** for the company

- Why MD&A (Management's Discussion and Analysis) section?
  - MD&A is intended to give investors a sense of **management's perspective**
  - The **most read section** of the 10-K [1]
  - There is **little research** on the language used in the **MD&A**. Many scholars have called for additional research in this area [2].

[1] L.R. Tavcar, Make the MD&A more readable, The CPA Journal 10 (January) (1998).
[2] C. Cole, Management discussion and analysis: a review and implications for future research, Journal of Accounting Literature 24 (2005).

- *Past research has already found that **deceivers use different language** than truthtellers [3].*

[3] Buller, David B., et al. "Testing interpersonal deception theory: The language of interpersonal deception." Communication theory 6.3 (1996): 268-288.

Importance of Linguistic Features

- 10-Ks에는 Fraudulent한 기업을 식별할 수 있는 요소가 존재한다.
  - Fraudulent한 기업은 **misleading**한 텍스트를 사용하여 기업의 부정적인 면을 숨기고 싶어한다.
  - 10-Ks may contain fraud in the form of intentionally **misstated** numbers and/or**misleading statements** made by the authors.

- misleading한 텍스트는 다음의 요소 등을 통해 식별될 수 있다. [4]
  - higher quantities of words
  - more non-immediate language
  - more informal language

[4] L. Zhou, J.K. Burgoon, J.F. Nunamaker Jr., D.P. Twitchell, Automating linguistics based cues for detecting deception in text based asynchronous computer mediated communication: an empirical investigation, Group Decision and Negotiation 13 (1) (2004).

Importance of Linguistic Features

Management Obfuscation Hypothesis (MOH) [5]

- **Longer sentences and longer words** are a surrogate measure for complexity and should occur **when fraud is present.**
- SMOG reading index
- sentence complexity → average sentence length
- word complexity → readibility

Interpersonal Deception Theory (IDT) [6]

- Deceivers want to minimize responsibility for their deceit if the deceit is discovered.
- Deceivers **reduce specificity, use nonimmediate language, and use inclusive terms.** These techniques will add ambiguity to statements and diffuse responsibility.

[5] Bloomfield, Robert J. "The "incomplete revelation hypothesis" and financial reporting." Accounting Horizons 16.3 (2002): 233-243.
[6] Buller, David B., and Judee K. Burgoon. "Interpersonal deception theory." Communication theory 6.3 (1996): 203-242.

Importance of Linguistic Features: Zhou et al [7]

- Zhou et al. [7] presented 9 linguistic constructs useful for detecting deception
- Fraudulent MD&As display higher
  - high (1) affect
  - high (2) complexity
  - less (3) diversity
  - high (4) expressivity
  - high (5) nonimmediacy
  - high (6) quantity
  - less (7) specificity
  - high (8) uncertainty
  - high (9) informality of language than non-fraudulent MD&As.

[7] Zhou, Lina, et al. "A comparison of classification methods for predicting deception in computer-mediated communication." Journal of Management Information Systems 20.4 (2004): 139-166.

Zhou et al.'s constructs and variable definition

(1) Affect

- Activation Ratio
  - number of activation words divided by the total number of words
- Affect Ratio
  - Total number of affect words divided by the total number of words
- Imagery
  - Number of imagery words divided by the total number of words
- **Pleasantness Ratio**
  - number of pleasantness words

(2) Complexity

- Average Sentence Length
  - Number of words divided by total number of sentences
- Average Word Length
  - Number of syllables divided by total number of words
- Pausality
  - Number of punctuation marks divided by total number of sentences

Zhou et al.'s constructs and variable definition

(3) Diversity (Less diversity \sim~ More uniqueness)

- **Content Word Diversity**
  - Percentage of unique content words (number of different content words divided by total number of content words)
  - **(cf.) BOW 기반 cosine similarity**
- Function Word Diversity
  - Number of function words divided by total number of sentences
- Lexical Diversity
  - Percentage of unique words or terms out of total words

(4) Expressivity

- Emotiveness
  - Ratio of adjective and adverbs to nouns and verbs

Zhou et al.'s constructs and variable definition

(5) Nonimmediacy

- Group References
  - First person plural pronoun count divided by total number ofverbs
- Other References
  - Count of all other singular or plural pronouns divided by total number of verbs
- Passive Verb Ratio
  - Number of passive verbs divided by total number of verbs

(6) **Quantity**

- Modifier Quantity
  - Total number of modifiers
- Sentence Quantity
  - Total number of sentences
- Verb Quantity
  - Total number of verbs
- **Word Quantity**
  - Total number of words

Zhou et al.'s constructs and variable definition

(7) Specificity

- Sensory Ratio
  - Number of words referencing five senses, divided by total number of words
- Spatial Close Ratio, Spatial Far Ratio, Temporal Immediate Ratio, and Temporal Non- immediate Ratio
  - Number of words that reference temporal or spatial information divided by total number of words

(8) Uncertainty

- Modal Verb Ratio
  - Number of modal verbs divided by the total number of verbs

Experimental Setup

[Labeling Fraudulent Companies]

- By searching for **AAER**s that included the term '10-K'.

Table 2
Selection criteria for fraudulent 10-Ks.

Count of companies identified as fraudulent by searching through AAERs	141
Count disqualified because fraud did not involve 10-Ks	(20)
Count disqualified because 10-K was not available from the SEC	(10)
Count disqualified because 10-K did not contain management discussion section	(10)
Final count of qualifying 10-Ks used in the final sample	101

Table 3
Primary types of fraud.

Type of fraud	Count of companies
Overstatement of revenues	44
Combination of overstating revenue and understating expenses	25
Disclosure issue	10
Overstatement of inventory	6
Other income increasing effects	6
Understatement of provisions for loan-loss reserves	5
Other	5

[Labeling Non-fraudulent Companies]

- By selecting companies with Standard Industrial Classification (**SIC**) codes that exactly matched the companies that filed fraudulent 10-Ks
  - By searching SEC's EDGAR (the database for online corporate financial information)
- Those in the non-fraudulent group were verified as **having no AAERs** attached to them, which suggests a history of compliance with SEC regulations.

Step1: Testing the 24-variable model

- 24 cues were analyzed with one-tailed independent sample t-tests
- Found out that all observations were independent from each other.

Table 4
Linguistic cues analyzed by Agent99 Analyzer.

Construct and variables	Results	Non-fraud		Fraud	
		M	SD	M	SD
Affect					
Activation	F>N ***	1.647	0.022	1.655	0.018
Affect ratio	F>N	0.0041	0.003	0.0044	0.002
Imagery	F>N ***	1.476	0.044	1.492	0.032
Pleasantness	F>N ***	1.801	0.019	1.807	0.015
Complexity					
Average sentence length	F>N	20.30	3.000	20.57	2.318
Average word length	F>N ***	5.393	0.171	5.481	0.123
Pausality	F>N *	3.474	0.901	3.820	1.305
Diversity					
Content word diversity	N>F ***	0.361	0.102	0.300	0.083
Function word diversity	F>N	8.738	1.426	8.835	1.087
Lexical diversity	N>F ***	0.250	0.080	0.202	0.060
Expressivity					
Emotiveness	F>N	0.230	0.038	0.233	0.027
Nonimmediacy					
Group references ratio	F>N *	0.010	0.018	0.016	0.020
Other references ratio	F>N	0.00095	0.0009	0.00105	0.0007
Passive verb ratio	F>N	0.062	0.020	0.063	0.018
Quantity					
Modifier quantity	F>N ***	529	420	898	694
Sentence quantity	F>N ***	222	164	364	264
Verb quantity	F>N ***	613	485	1020	773
Word quantity	F>N ***	4612	3707	7603	5793
Specificity					
Sensory ratio	N>F	0.0594	0.010	0.0591	0.008
Spatial close ratio	N>F	0.011	0.004	0.010	0.004
Spatial far ratio	F>N	0.0456	0.009	0.0464	0.008
Temporal immediacy ratio	N>F	0.0026	0.0012	0.0024	0.001
Temporal nonimmediacy ratio	F>N	0.0018	0.0013	0.0018	0.001
Uncertainty					
Modal verb ratio	F>N	0.038	0.025	0.043	0.030

Note. F = fraudulent 10-Ks, N = non-fraudulent 10-Ks. The results column also indicates the direction hypothesized. One-tailed t-tests; *p-value <= 0.05; ***significant at p-value <= 0.005.

Step2: Model reduction

- To achieve **greater interpretability**, data reduction techniques and the theoretical groupings of the variables were applied to the 24-variable model.
- It resulted in a **10-variable model**.

Table 5
Reduced 10-variable model.

New variables	Averaged of values from old variables	Reliability
Active language	Activation, pleasantness, imagery, modal verb ratio, active verb ratio, and group reference	.675
Diversity	Lexical diversity and content word diversity	.979
Sensory terms	Sensory ratio and spatial far ratio	.900
Syntactic complexity	Function word diversity and average sentence length	.847
Temporal/emotiveness	Emotiveness, spatial close ratio, temporal nonimmediacy ratio, and temporal immediacy ratio	.039
Quantity	Verb, modifier, sentence, and word quantities	.511
Affect ratio	Affect ratio	NA
Average word length	Average word length	NA
Other references	Other references	NA
Pausality	Pausality	NA

Note. Cronbach's alpha reliability was used.

Step3: Classification algorithms

- Locally Weighted Learning (LWL)
- simple Naive Bayes
- C4.5 decision tree (Decision Tree)
  - ‘Quantity Score'가 가장 중요한 특징으로 뽑힘.
  - by summing counts of verbs,modifiers, sentences, andwords then dividing by 4.

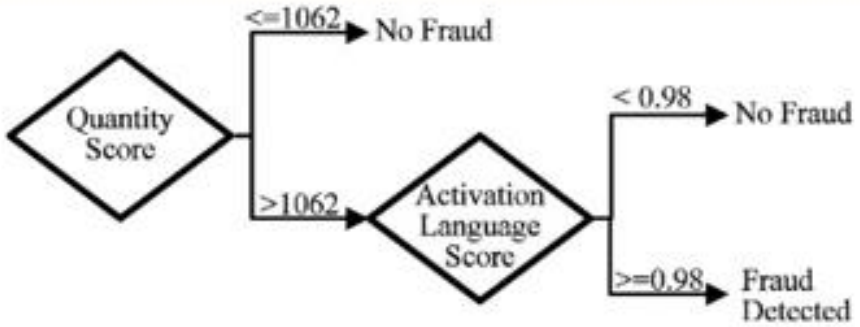


Fig. 1. Parsimonious decision tree for discriminating fraudulent 10-Ks.

Table 6
Classification accuracy of fraud/non-fraud.

Classification technique	24-variable model	10-variable model
Logistic regression	58.4%	63.4%
C4.5	64.9%	67.3%
LWL	66.3%	60.4%
Naïve Bayes	65.3%	67.3%
SVM	61.4%	65.8%

Note. 10-fold cross validation used for all tests.

24-variable 및 10-variable에 대한 Classification algorithms 결과

Evaluation Metric

- **Accuracy**
  - The total number of documents correctly classified divided by the total number of documents analyzed
- **Recall**
  - Recall of fraud
    - The ratio of number of documents correctly classified as fraudulent to the total number ofactual fraudulent documents
  - Recall of truthful
    - The ratio ofthe number ofnon- fraudulent documents correctly classified as non-fraudulent to the total number ofactual non-fraudulent documents
- **Precision**
  - Precision of fraud
    - The ratio of the number documents correctly classified as fraudulent to the total number of documents classified as fraudulent
  - Precision of truthful
    - The ratio ofthe number ofdocuments correctly classified as non-fraudulent to the total number of documents classified as non-fraudulent
- **F-measures**
  - **Combines precision and recall** into one metric using aweighted harmonic mean of each

Evaluation

Table 8
Accuracy, precision, recall, and F-measure using 10-variable model.

Classification technique	Overall accuracy	Precision		Recall		F-measure	
		Non-fraud	Fraud	Non-fraud	Fraud	Non-fraud	Fraud
C4.5	67.3%	66.7%	68.0%	66.7%	65.3%	68.0%	66.7%
LWL	60.4%	60.2%	60.6%	61.4%	59.4%	60.8%	60.0%
Jrip	67.3%	67.0%	67.7%	68.3%	66.3%	67.6%	67.0%
Naïve Bayes	67.3%	68.0%	66.7%	65.3%	69.3%	66.7%	68.0%
SVM	65.8%	67.8%	64.3%	60.4%	71.3%	63.9%	67.6%

Note. 10-fold cross validation used for all tests.

Limitation & Future Work

[Limitation]

- 정량적인 지표를 뽑을 때에 사용한 방법이 매우 단순한 counting 기반

[Future Work]

- Data-mining 기법 적용
  - Clustering?
  - Attention?

References

1. L.R. Tavcar, Make the MD&A more readable, The CPA Journal 10 (January) (1998).
2. C. Cole, Management discussion and analysis: a review and implications for future research, Journal of Accounting Literature 24 (2005).
3. Buller, David B., et al. "Testing interpersonal deception theory: The language of interpersonal deception." Communication theory 6.3 (1996): 268-288.
4. L. Zhou, J.K. Burgoon, J.F. Nunamaker Jr., D.P. Twitchell, Automating linguistics based cues for detecting deception in text based asynchronous computer mediated communication: an empirical investigation, Group Decision and Negotiation 13 (1) (2004).
5. Bloomfield, Robert J. “The “incomplete revelation hypothesis” and financial reporting.” Accounting Horizons 16.3 (2002): 233-243.
6. Buller, David B., and Judee K. Burgoon. “Interpersonal deception theory.” Communication theory 6.3 (1996): 203-242.
7. **Zhou**, Lina, et al. "**A comparison of classification methods for predicting deception in computer-mediated communication.**" Journal of Management Information Systems 20.4 (**2004**): 139-166.

Kim, Yeonkook. *"Building Financial Misstatement Detection Models using Multi-class Cost-sensitive Learning and Feature Generation from CFO survey."* Diss. 서울대학교 대학원, 2016.

Types of Financial Misstatements

Introduction: Fraud Detection using Text-mining

- Why text-mining?
  - Enron 케이스의 경우, Financial misstatements가 있었음에도 3년 후에나 그 실체가 드러났다. text-mining을 통해 financial misstatements를 가능한 한 빨리 인지할 수 있다면 좋을 것이다.

- Financial Misstatements의 2가지 종류
  - Errors
    - **Unintentional** misapplications of accounting rules
    - → Noise !
  - **Irregularities**
    - **Intentional misreporting**
    - **→ Anomaly !**

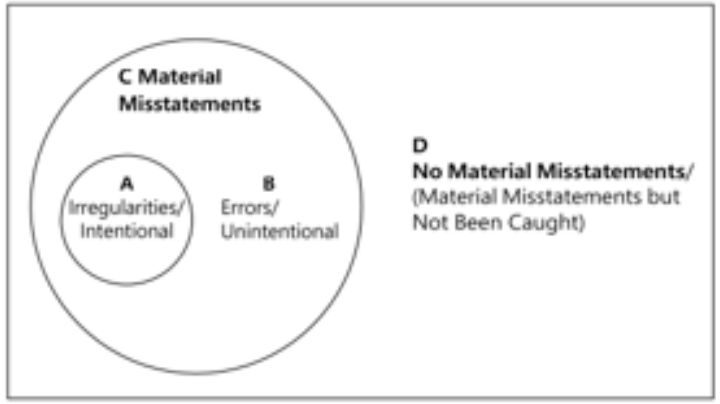


Figure 1.1: Hennes et al. (2008) classification result

앞으로의 계획

- Baseline 실험 Setup
  - "Identification of fraudulent financial statements using linguistic credibility analysis.“

1. 데이터 확보: 10-Ks MD&A Section
   - 방법 모색 후 코드 구현
   - 협업 ? ..
2. 데이터 레이블링: Fraudulent vs Non-fraudulent
   - 방법 모색 후 코드 구현
   - 협업 ? ..
3. Zhou et al의 8가지 Constructs 계산 값 도출 (대부분 Counting 기반)
   - Affect, Complexity, Deversity, Quantity
     - 기반이 되는 딕셔너리 탐색 후 선정 (Ex. pleasantness words)
     - 코드 구현

앞으로의 계획

Table 4
Linguistic cues analyzed by Agent99 Analyzer.

Construct and variables	Results
Affect	
Activation	F>N ***
Affect ratio	F>N
Imagery	F>N ***
Pleasantness	F>N ***
Complexity	
Average sentence length	F>N
Average word length	F>N ***
Pausality	F>N *
Diversity	
Content word diversity	N>F ***
Function word diversity	F>N
Lexical diversity	N>F ***
Expressivity	
Emotiveness	F>N
Nonimmediacy	
Group references ratio	F>N *
Other references ratio	F>N
Passive verb ratio	F>N
Quantity	
Modifier quantity	F>N ***
Sentence quantity	F>N ***
Verb quantity	F>N ***
Word quantity	F>N ***
Specificity	
Sensory ratio	N>F
Spatial close ratio	N>F
Spatial far ratio	F>N
Temporal immediacy ratio	N>F
Temporal nonimmediacy ratio	F>N
Uncertainty	
Modal verb ratio	F>N

(1) Affect

- Activation Ratio
  - number of activation words divided by the total number of words
- Affect Ratio
  - Total number of affect words divided by the total number of words
- Imagery
  - Number of imagery words divided by the total number of words
- **Pleasantness Ratio**
  - number of pleasantness words

(2) Complexity

- Average Sentence Length
  - Number of words divided by total number of sentences
- Average Word Length
  - Number of syllables divided by total number ofwords
- Pausality
  - Number of punctuation marks divided by total number of sentences

(3) Diversity (Less diversity \sim~ More uniqueness)

- **Content Word Diversity**
  - Percentage of unique content words (number of different content words divided by total number of content words)
  - **(cf.) BOW 기반 cosine similarity**
- Function Word Diversity
  - Number of function words divided by total number of sentences
- Lexical Diversity
  - Percentage of unique words or terms out of total words

(6) **Quantity**

- Modifier Quantity
  - Total number of modifiers
- Sentence Quantity
  - Total number of sentences
- Verb Quantity
  - Total number of verbs
- **Word Quantity**
  - Total number of words

감사합니다.

Glassdoor와 잡플래닛에 등록된 Employee Experience Data에서 얻을 수 있는 기업 관련 정보

20190816

박서영 이규진 이명원 정지수

8/16 Progress Report

1. S&P 500 Long/Short

- 2009~2018년 S&P 500 주식 데이터: *크롤링 완료*
- 2017~2018년 분기별 long/short portfolio: *구성 및 비교*

2. KOSPI 200 Long/Short

- 2014~2019년 KOSPI 200 주식 데이터: *크롤링 완료*
- 2014~2019년 KOSPI 200 잡플래닛 데이터: *크롤링 진행 중*

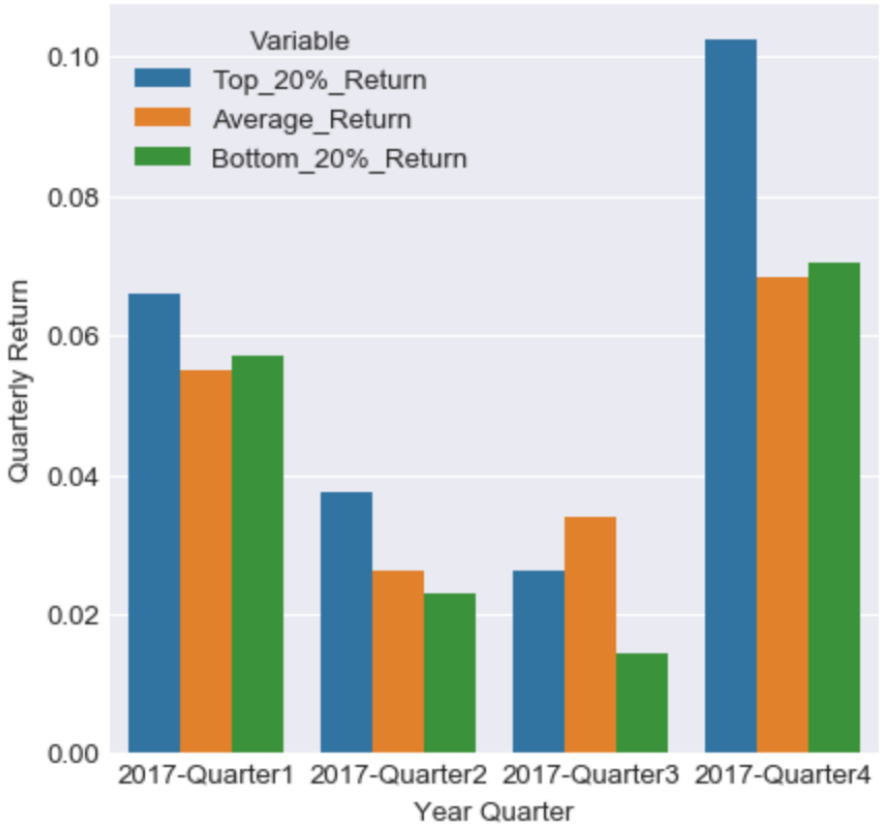
3. IPO Value Analysis

- 한국 IPO 시장의 공모가격 산정방식 분석
- 잡플래닛 데이터 활용 가능성

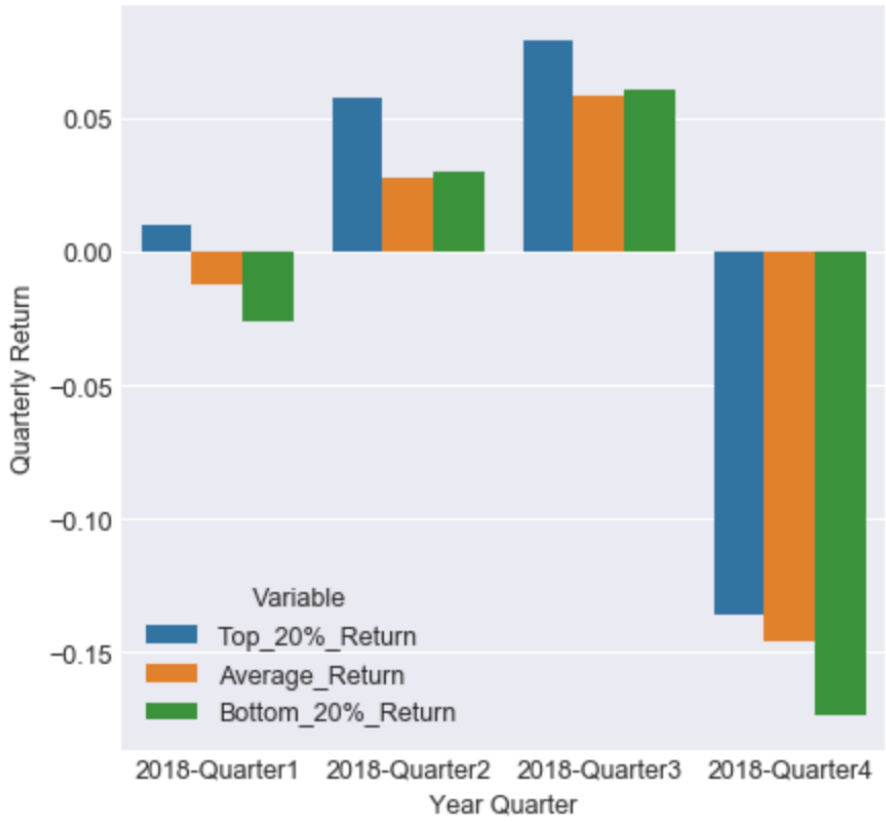
1. S&P 500 Long/Short

분기별 전체평점(1점~5점)에 따라 S&P 500 종목을 재구성

2017



2018



Top 20% Portfolio: 해당 분기에 *전체평점이 제일 높았던* 회사 90여 개

Bottom 20% Portfolio: *제일 낮았던* 회사 90여 개

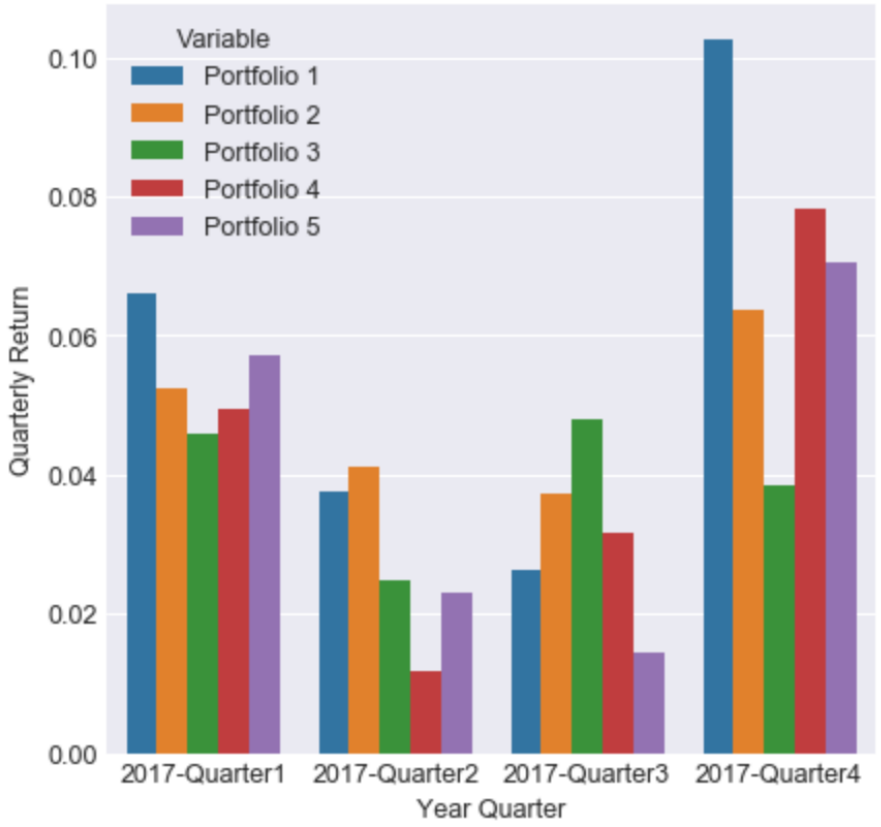
Portfolio 구성: Quarterly rebalanced, equally weighted

제외 대상: 해당 분기의 리뷰 수가 10개 미만이었던 회사

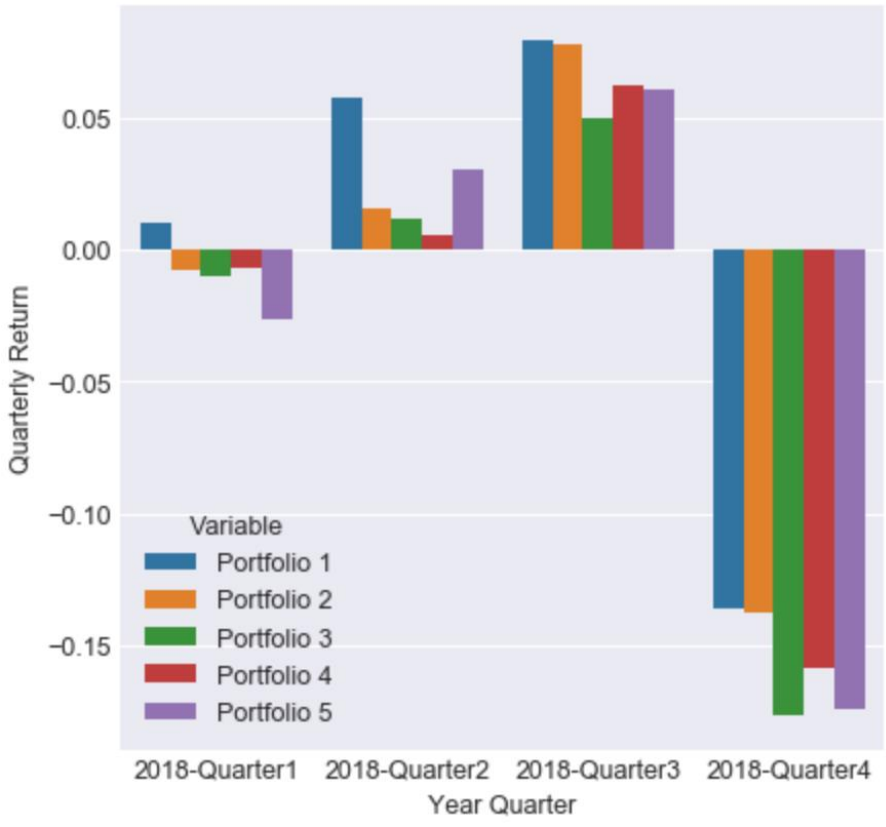
1. S&P 500 Long/Short

분기별 전체평점(1점~5점)에 따라 S&P 500 종목을 재구성

2017



2018



Portfolio 1: 해당 분기에 *전체평점이 제일 높았던* 회사 90여 개

Portfolio 5: *제일 낮았던* 회사 90여 개

Portfolio 구성: Quarterly rebalanced, equally weighted

제외 대상: 해당 분기의 리뷰 수가 10개 미만이었던 회사

1. 다음 단계

Glassdoor 데이터 의문점 해결

- 공시된 리뷰 수와 크롤한 리뷰 수의 차이 (예: 2018년에 3%가량 차이나는 기업이 10개)

overall_stats_total_number	ind_reviews_total_number	difference	Company
14416	12794	1622.0	Apple Inc.
20612	19917	695.0	The Home Depot, Inc.
32186	29680	2506.0	Target Corporation
10098	9485	613.0	UnitedHealth Group Incorporated
45313	40397	4916.0	Walmart Inc.
45770	40297	5473.0	International Business Machines Corporation
12267	9565	2702.0	United Parcel Service, Inc.
6721	5775	946.0	American Express Company
21444	20679	765.0	Wells Fargo & Company
6264	6030	234.0	DXC Technology Company

- 크

overall_stats_total_number	ind_reviews_total_number	difference	Company
NOT SCRAPED YET	NOT SCRAPED YET	NaN	Alpha Industries, Inc.
NOT SCRAPED YET	NOT SCRAPED YET	NaN	Union Electric Company
NOT SCRAPED YET	NOT SCRAPED YET	NaN	The Progressive Corporation
NOT SCRAPED YET	NOT SCRAPED YET	NaN	Norfolk & Western Railway Co.

자회사: Alpha Natural Resources, Inc.

새 이름: Ameren Corporation

자회사: Progressive Insurance

새 이름: Norfolk Southern Railway

1. 다음 단계

2008년~2016년 S&P 500에 포함되었던 기업의 Glassdoor 데이터 크롤링

- 매년 20~30개 종목이 변경

	index	sp500_from	sp500_thru	parent_name	parent_company_id	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
55	5496	7/2/2018	NaN	Copart, Inc.	27023	0	0	0	0	0	0	0	0	0	0	1	1
73	6661	6/20/2018	NaN	FleetCor Technologies, Inc.	28756	0	0	0	0	0	0	0	0	0	0	1	1
85	8423	3/19/2018	NaN	Nektar Therapeutics	29929	0	0	0	0	0	0	0	0	0	0	1	1
186	22810	12/3/2018	NaN	Maxim Integrated Products, Inc.	95673	0	0	0	0	0	0	0	0	0	0	1	1
268	36123	6/18/2018	NaN	HollyFrontier Corporation	178254	0	0	0	0	0	0	0	0	0	0	1	1
278	36682	3/19/2018	NaN	SVB Financial Group	187662	0	0	0	0	0	0	0	0	0	0	1	1
295	38632	5/31/2018	NaN	Abiomed, Inc.	247589	0	0	0	0	0	0	0	0	0	0	1	1
404	50190	11/13/2018	NaN	Jack Henry & Associates, Inc.	277617	0	0	0	0	0	0	0	0	0	0	1	1
426	55393	6/5/2018	NaN	Evergy, Inc.	283024	0	0	0	0	0	0	0	0	0	0	1	1
488	62715	10/1/2018	NaN	Rollins, Inc.	300753	0	0	0	0	0	0	0	0	0	0	1	1
592	68698	1/3/2018	NaN	Huntington Ingalls Industries, Inc.	363263	0	0	0	0	0	0	0	0	0	0	1	1
599	69092	3/19/2018	NaN	Take-Two Interactive Software, Inc.	371281	0	0	0	0	0	0	0	0	0	0	1	1
622	71765	12/24/2018	NaN	Celanese Corporation	405277	0	0	0	0	0	0	0	0	0	0	1	1
634	73636	3/7/2018	NaN	IPG Photonics Corporation	427949	0	0	0	0	0	0	0	0	0	0	1	1
673	77685	6/18/2018	NaN	Broadridge Financial Solutions, Inc.	1032802	0	0	0	0	0	0	0	0	0	0	1	1
688	79300	10/11/2018	NaN	Fortinet, Inc.	2689126	0	0	0	0	0	0	0	0	0	0	1	1

2. KOSPI 200 Long/Short

잡플래닛 크롤링 문제점

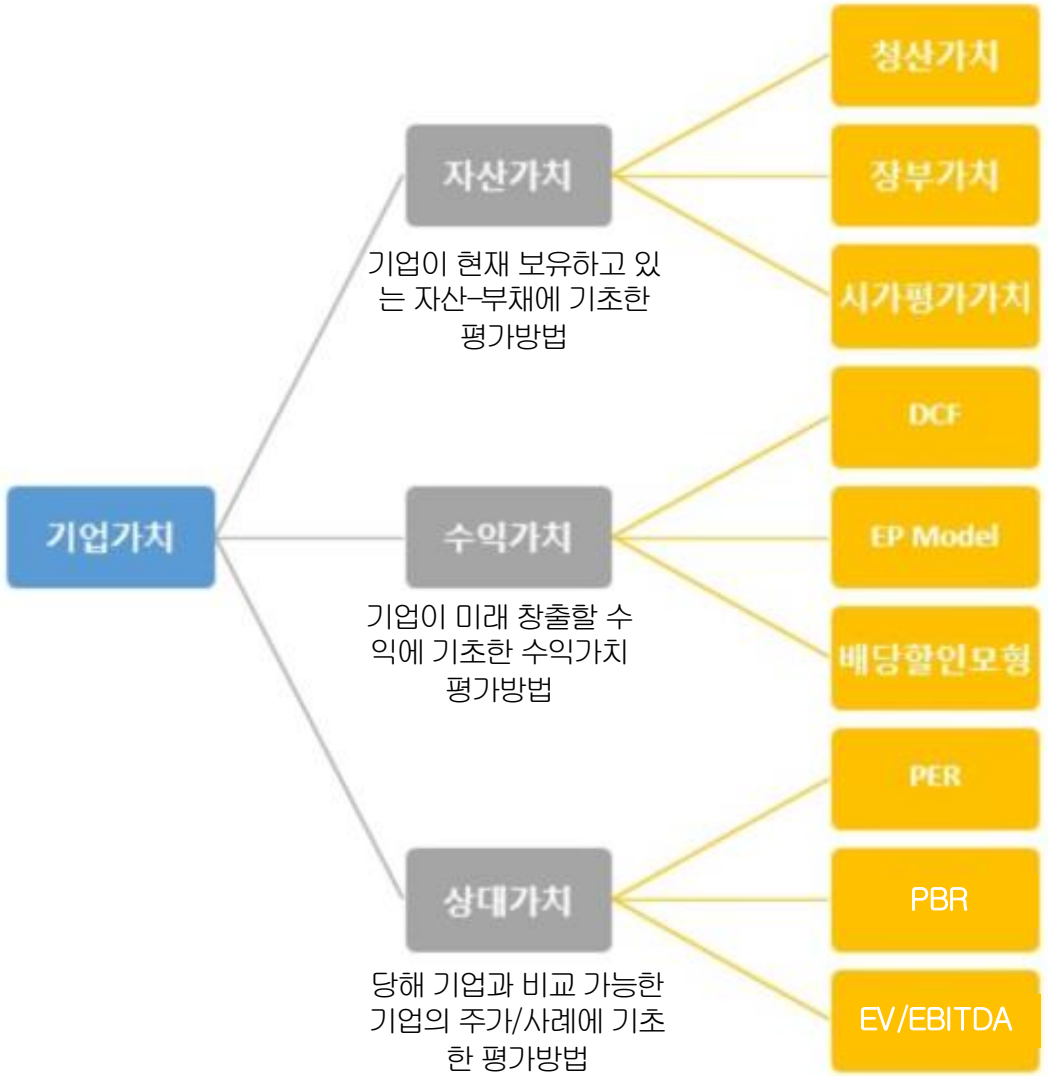
- KOPSI 200의 종목명과 잡플래닛 기재 회사명이 다름
- 잡플래닛에서 데이터가 있는 회사 위주로 검색이 됨
  - 예) SK 검색 시 SK하이닉스가 가장 상단에 등장, SK는 그 아래에 등장

해결책: Data Dictionary 구축

	A	B	C	D	E	F	G	H	I	J	K
1		종목명	종목코드	url	job korea 이름	특성	겹치는 데이터	모회사	comment		
2	190	AK홀딩스	'006840	https://www.jc	AK홀딩스(주)	데이터 없음					
3	194	BGF	'027410			모회사	4				
4	83	BGF리테일	'282330	https://www.jc	비지에프리테일(주)			027140'			
5	73	BNK금융지주	'138930			NOT EXIST					
6	89	CJ	'001040			모회사	다수				
7	150	CJ CGV	'079160	https://www.jc	씨제이씨지브이(주)			001040'			
8	92	CJ대한통운	'000120	https://www.jc	씨제이대한통운(주)			001040'			
9	64	CJ제일제당	'097950	https://www.jc	씨제이제일제당(주)			001040'			
10	56	DB손해보험	'005830	https://www.jc	디비손해보험(주)						
11	145	DB하이텍	'000990	https://www.jc	(주)디비하이텍						
12	132	GKL	'114090			NOT EXIST					
13	59	GS	'078930			모회사	다수				
14	68	GS건설	'006360	https://www.jc	지에스건설(주)			078930'			
15	98	GS리테일	'007070	https://www.jc	지에스리테일(주)			078930'			
16	140	HDC	'012630	https://www.jc	에이치디씨(주	모회사	다수				
17	90	HDC현대산업	'294870			NOT EXIST			다른 자회사는 많은데 얜 없어		
18	163	JW중외제약	'001060	https://www.jc	제이더블유중외제약(주)						
19	188	JW홀딩스	'096760	https://www.jc	제이더블유홀딩스(주)				001060'과 형제관계?		
20	9	KB금융	'105560	https://www.jc	국민은행(주)						
21	82	KCC	'002380	https://www.jc	케이씨씨(주)				얘가 거느리는 회사가 있는듯(건설, 정보통		
22	32	KT	'030200	https://www.jc	케이티(주)						

3. IPO (Initial Public Offering, 기업공개)

기업 설립 후 처음으로 외부투자자에게 주식을 공개하고, 이를 매도하는 업무



기업이 곧 청산할 것이라는 가정 아래 자산가치를 평가함
(시장에서 정상적으로 거래되는 가격보다 낮을 가능성이 높음)

공시된 재무상태표의 숫자 그대로를 기업가치로 보는 방법

(=구매가치법), 시장에서 형성된 정상적인 거래 가격으로 기업이 보유한 자산을 재평가했을 때의 가치를 의미

<수익가치 평가방법> : 미래의 이익을 현재 가치로 환원

1. 할인율 결정; 초기 기업일수록, 불확실성이 클수록, 보유하고 있는 자산이 적을수록 할인율이 높아짐.
2. 미래 이익 측정 방식
    1. DCF : 잉여현금흐름
    2. EP Model : 경제적부가가치(EVA)
    3. 배당할인 모형 : 배당현금흐름

PER(동종업체들의 주가를 주당순이익으로 나눈 값)을 대상 회사의 주당 순이익에 곱해주어 적정 주가를 구하는 방법.

PBR(동종업체들의 주가를 주당순자산(=(자산-부채)/발행주식수)을 대상 회사의 주당 순자산에 곱해주어 적정 주가를 구하는 방법

동종업체들의 EV(시가총액+순부채)를 EBITDA(이자비용, 법인세, 유무형자산 감가상각비를 반영하기 전의 이익)를 대상 회사의 EBITDA에 곱하여 기업가치를 구하는 방법

3. IPO 기업들의 공모가격 산정방식

예: 2016년 IPO 기업들의 공모가격 산정 방식

회사명	상장일	공모가 산정방식
호전실업(유)	2017-02-02	PER과 EV/EBITDA
서플러스글로벌(코)	2017-01-25	유사기업 PER 기준
유바이오로직스(코)	2017-01-24	유사기업 PER 기준
퓨전데이타(코)	2016-12-21	유사기업 PER 기준
DSC인베스트먼트	2016-12-19	유사기업 PER 기준
티에스인베스트먼트(코)	2016-12-15	유사기업 PER 기준
마이크로프랜드(코)	2016-12-12	유사기업 PER 기준
현성바이탈(코)	2016-12-09	유사기업 PER 기준 (코넥스 → 코스닥)
유니온커뮤니티(코)	2016-12-07	유사기업 PER 기준 (코넥스 → 코스닥)
애니젠(코)	2016-12-07	유사기업 PER 기준
신라젠(코)	2016-12-06	유사기업 PER 기준
핸즈코퍼레이션(유)	2016-12-02	유사기업 PER 기준
퓨쳐켐(코)	2016-12-01	유사기업 PER 기준 (코넥스 → 코스닥)
오션브릿지(코)	2016-12-01	유사기업 PER 기준
엘앤케이바이오(코)	2016-11-30	유사기업 PER 기준 (코넥스 → 코스닥)
뉴파워프라즈마(코)	2016-11-30	유사기업 PER 기준
에이치엔티(코)	2016-11-28	유사기업 PER 기준
핸디소프트(코)	2016-11-24	유사기업 PER 기준
두산밥캣(유)	2016-11-18	유사기업 PER 기준
삼성바이오로직스(유)	2016-11-10	(Growth-adjusted) EV / Sales, EV / Capacity, EV / Pipeline
클리오(코)	2016-11-09	유사기업 PER 기준
오가닉티코스메틱(코)	2016-11-04	유사기업 PER 기준
로고스바이오(코)	2016-11-03	유사기업 PER 기준
인크로스(코)	2016-10-31	유사기업 PER 기준
에이치엘사이언스(코)	2016-10-28	유사기업 PER 기준
코스메카코리아(코)	2016-10-28	유사기업 PER 기준
JW생명과학(유)	2016-10-27	PER과 EV/EBITDA
GRT(코)	2016-10-25	유사기업 PER 기준
골든센츄리(코)	2016-10-19	유사기업 PER 기준
인텔리안테크(코)	2016-10-18	유사기업 PER 기준
에이치시티(코)	2016-10-17	유사기업 PER 기준
잉글우드랩(코)	2016-10-14	유사기업 PER 기준
앤디포스(코)	2016-10-12	유사기업 PER 기준
수산아이앤티(코)	2016-10-11	유사기업 PER 기준
미투온(코)	2016-10-10	유사기업 PER 기준
화승엔터프라이즈(유)	2016-10-04	유사기업 PER 기준
제이앤티씨(코)	철회	유사기업 PER 기준
프라코(유)	철회	유사기업 PER 기준
까사미아(유)	철회	유사기업 PER 기준
LS전선아시아(유)	2016-09-22	유사기업 PER 기준
유니테크노(코)	2016-09-20	유사기업 PER 기준
자이글(코)	2016-09-06	유사기업 PER 기준
헝셩그룹(코)	2016-08-18	유사기업 PER 기준
에코마케팅(코)	2016-08-08	유사기업 PER 기준
팍스넷(코)	2016-08-01	유사기업 PER 기준
두올(유)	2016-07-29	유사기업 PER 기준
엔지스테크널러지(코)	2016-07-28	유사기업 PER 기준
우리손에프앤지(코)	2016-07-27	유사기업 PER 기준
옵토팩(코)	2016-07-20	유사기업 PER 기준 (코넥스 → 코스닥)
장원테크(코)	2016-07-15	유사기업 PER 기준
대유위니아(코)	2016-07-14	유사기업 PER 기준
한국자산신탁(유)	2016-07-13	유사기업 PER 기준
바이오리더스(코)	2016-07-07	유사기업 PER 기준 (코넥스 → 코스닥)
피앤씨테크(코)	2016-07-04	유사기업 PER 기준
로스웰(코)	2016-06-30	유사기업 PER 기준
해성디에스(유)	2016-06-24	유사기업 PER 기준
에스티팜(코)	2016-06-23	유사기업 PER 기준

〉 주로 *상대가치법*, ***특히 PER 활용***

3. 유사 회사 선정 방법

사업의 유사성

비교회사가 평가회사와 유사한 사업을 영위하고 있는지에 대하여 제품의 속성, 제품별 구성비, 주요 시장 등을 고려

규모 및 성장률

매출액이나 총자산의 규모 및 기간별 매출이나 손익의 추세를 비교하고, 차이가 클 경우 비교 대상에서 제외 또는 시가배수의 조정 등을 고려

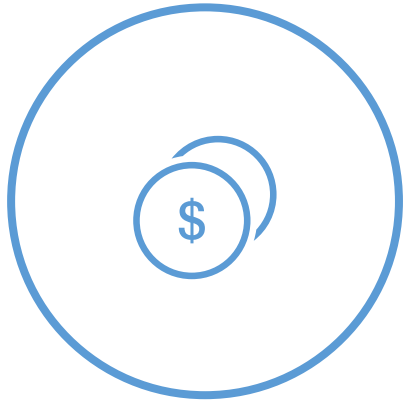
영업 및 재무 사항

시장점유율, 경쟁관계, 판매/구매처와의 관계 등 영업에 관한 사항과 각종 재무비율(수익성, 성장성, 안정성 등) 비교

객관적인 가이드라인을 제시되고 있지만,
여전히 **평가자의 주관**이 개입

(출처 : 차세대 성장동력산업의 기업가치평가 가이드라인, http://law.kofia.or.kr/service/law/lawFullScreenContent.do?seq=221&historySeq=602.)

3. Glassdoor/ 잡플래닛 데이터 활용가능성

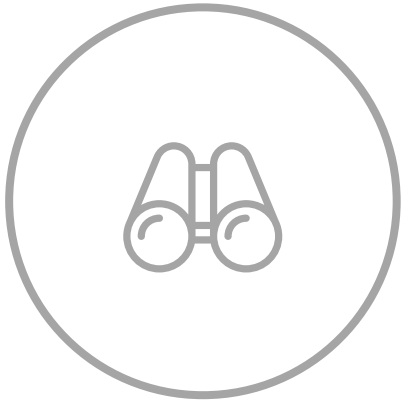


실제 헤지펀드/사모펀드에서 기업 인수나 투자에 '질적인' 자료로 Glassdoor 활용

재무제표 상 수치에서 포착되기 힘든 회사의 전망을 구인 현황이나 CEO에 대한 평가 추이, 직원들이 예측하는 회사의 미래 등을 통해 예측 가능.

기업의 가치에 유의미하게 긍정/부정적인 시그널을 주는 주제어 분석

비상장기업을 잘 아는 내부자들의 의견이 투자 결정, 혹은 할인율 결정 등에 추가 참고자료로 사용 가능.



유사 회사 선정 기준으로 활용

IPO 기업 공모가격 산정방식의 객관성에 대한 논란의 해결 방안으로, 유사 회사를 선정하는 기준으로 현재 쓰이는 지표에 더하여 기업 문화를 비교할 수 있는 Glassdoor/잡플래닛 데이터 이용 가능.

검색어 기반 연구

Aug 16, 2019

이혜진, 남상호, 나현수

0. Table of Contents

- Exploratory Data Analysis
  - Excess return to KOSDAQ & volume
  - Recent 3 months analysis
  - Related keyword
  - Market & volume & stock price

- Progress

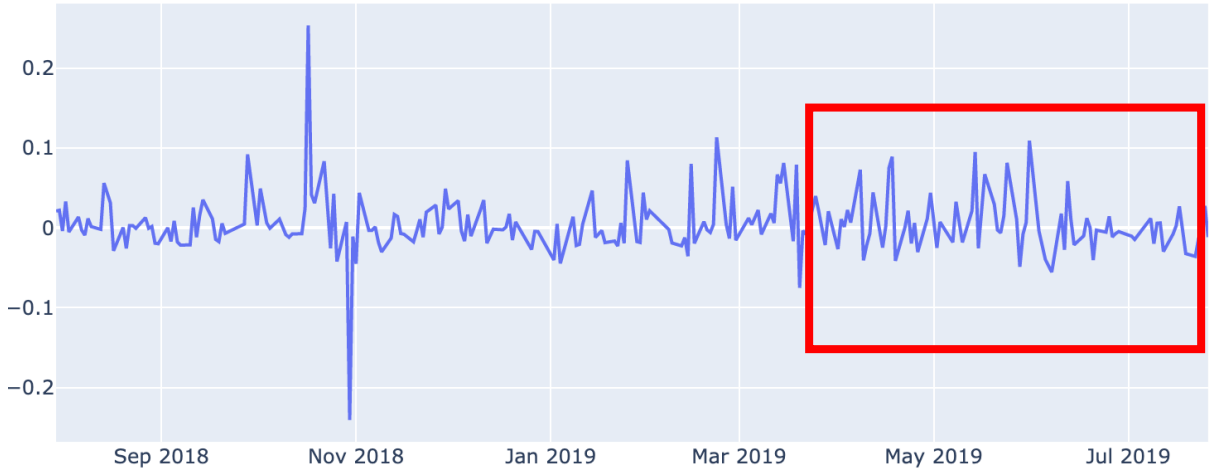
- Conclusion
  - 거래량과 키워드 : 최근 기준 거래량과 키워드 상관관계가 있어 보임, 그러나 키워드 간 linearity 가 높음.
  - 주식 가격과 거래량 : KOSDAQ 시장 상황에 따라, 거래량은 주식 가격에 선행하는 것으로 보임.

1. Exploratory Data Analysis : Excess return to KOSDAQ & Volume

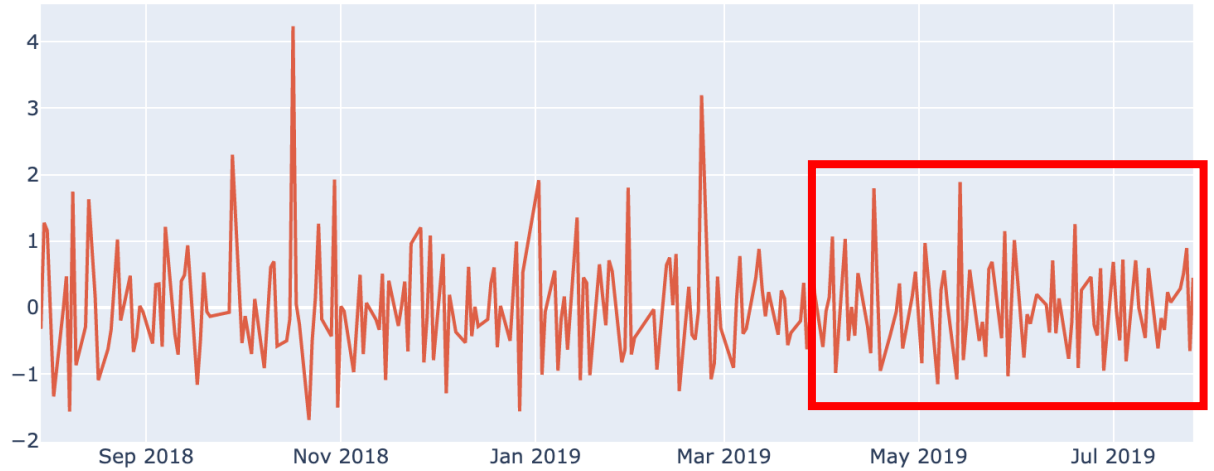
Transformation : log & yield
Period : 1 year

- **Random process 같아 보이나, 최근 거래량은 키워드와 비슷한 양상**

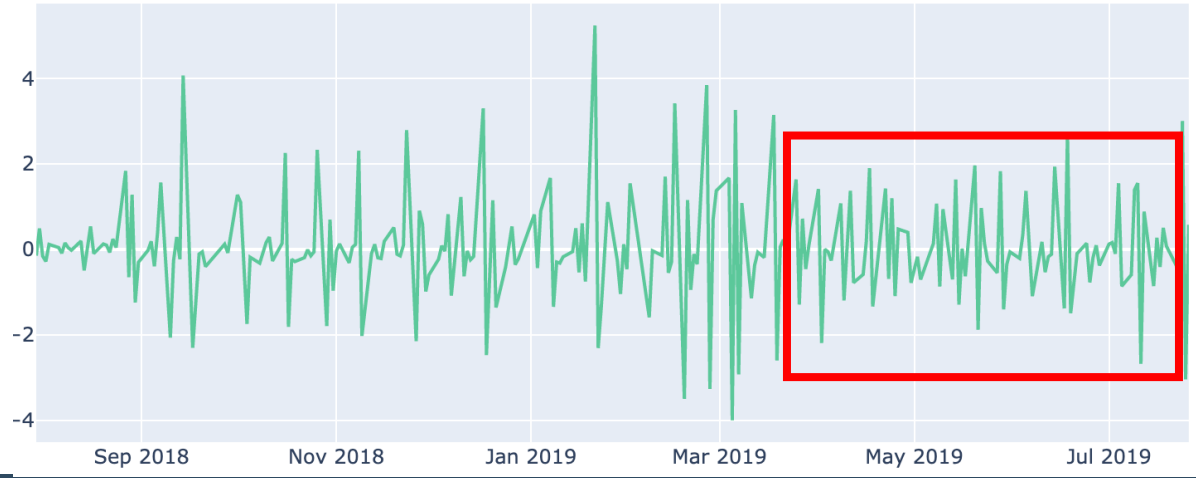
Log KOSDAQ 대비 초과 수익률



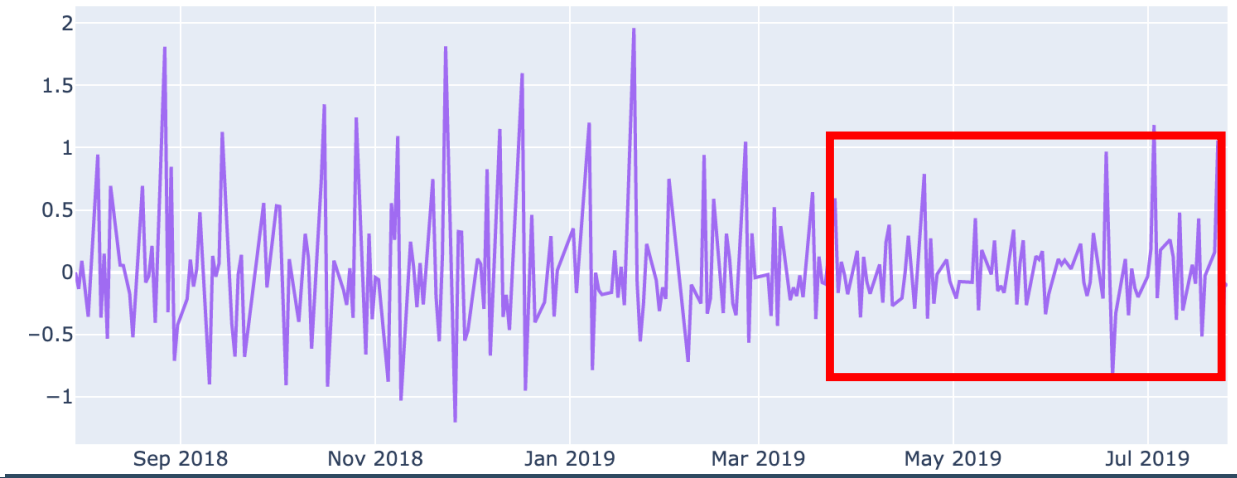
Log 거래량 변화율



Log 새싹보리 변화율



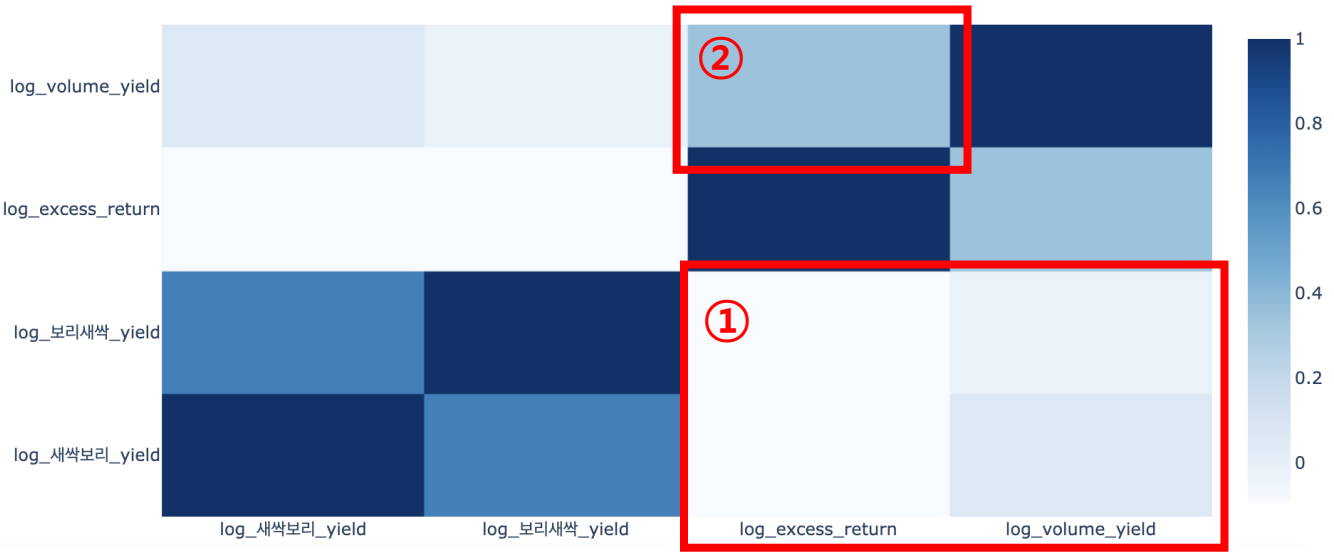
Log 보리새싹 변화율

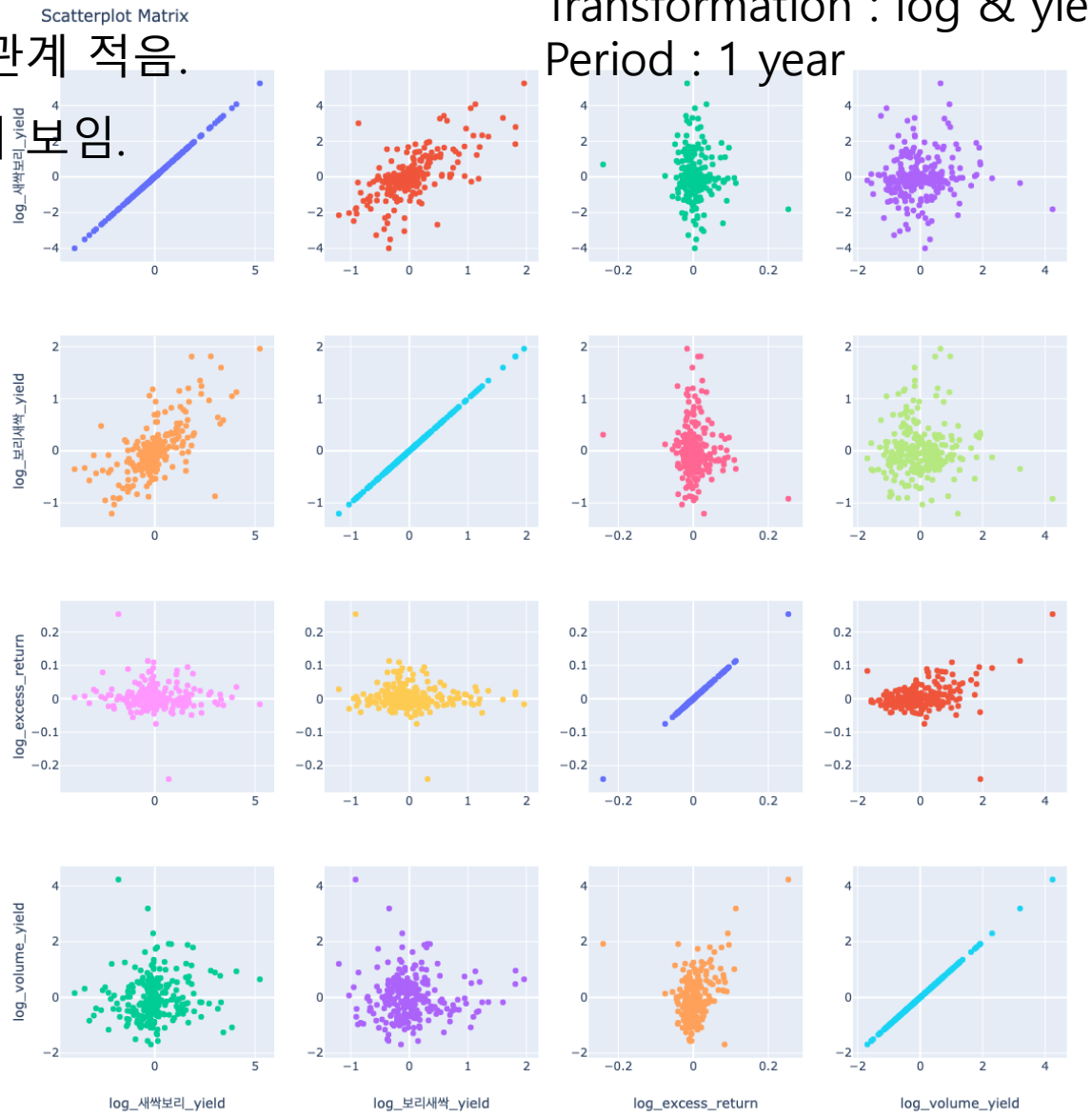


1. Exploratory Data Analysis : Excess return to KOSDAQ & Volume

- ① 1년간은 초과수익률 & 거래량 모두 키워드와 상관 관계 적음.
- ② 초과수익률과 거래량은 $\rho = 0.3$ 으로 약한 상관 관계보임.

Transformation : log & yield
Period : 1 year





1. Exploratory Data Analysis : Recent 3 months

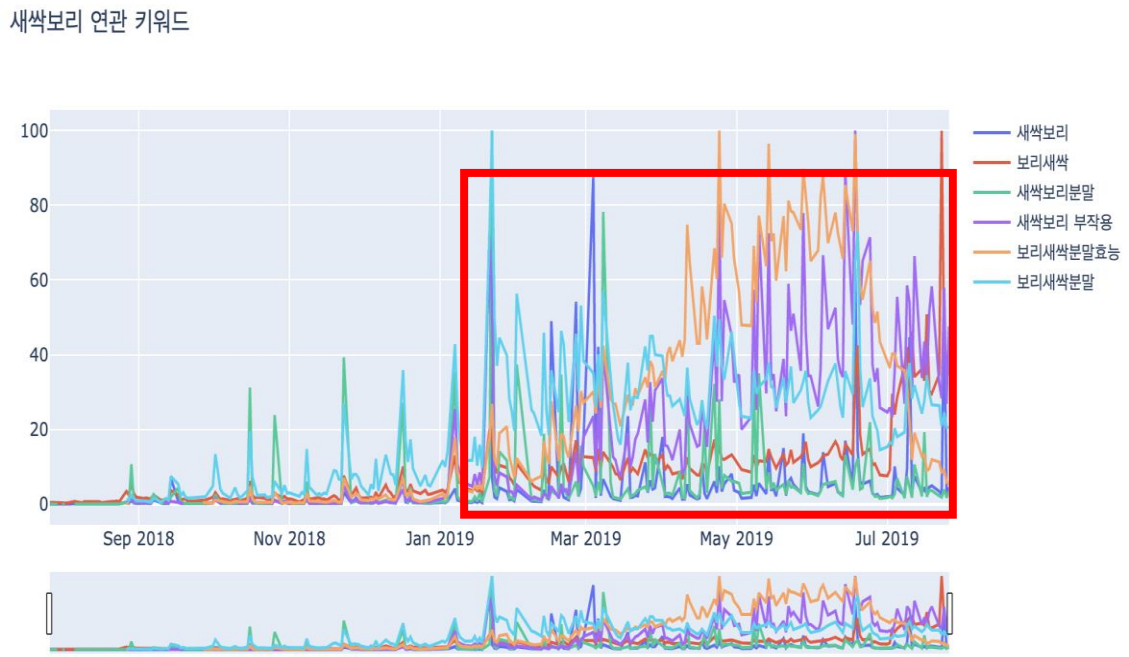
Transformation : log & yield
Period : recent 3 months

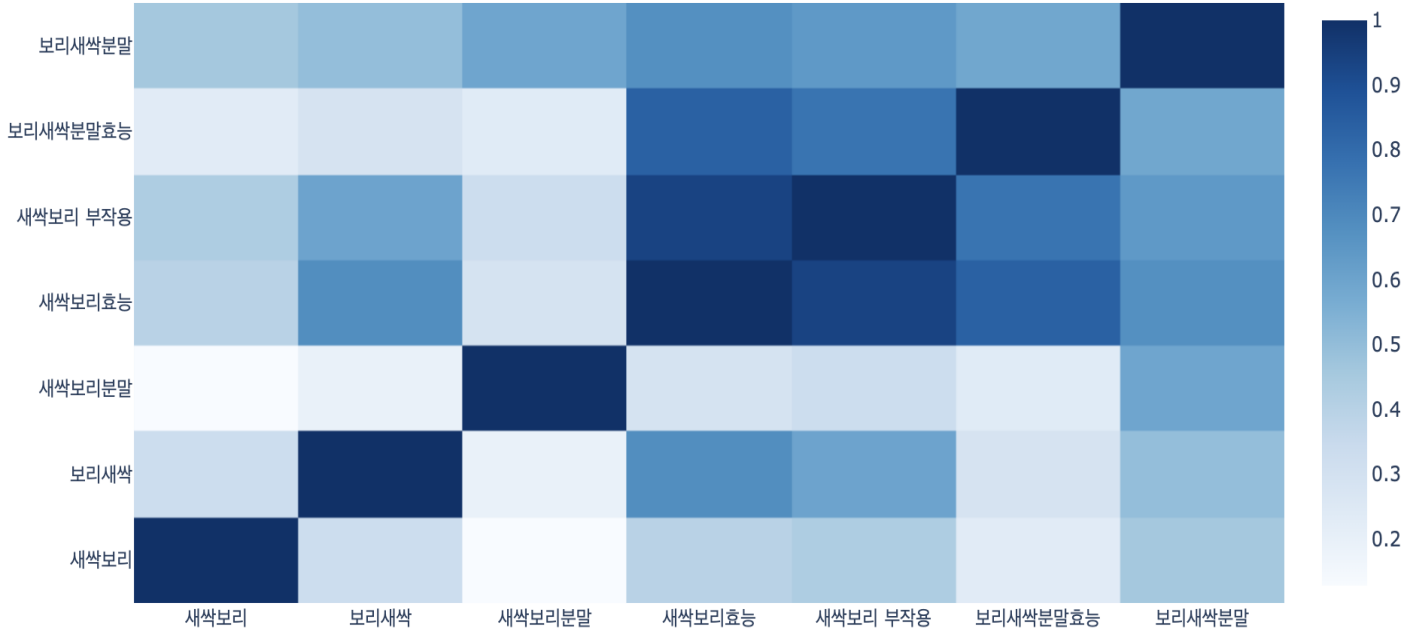
- ① **최근 3개월 간은 거래량과 새싹보리 $\rho = 0.3$ 으로 약한 상관 관계를 보임.**
- ② 초과수익률과 거래량은 역시 $\rho = 0.3$ 으로 약한 상관 관계 보임.

1. Exploratory Data Analysis : Related keyword

Transformation : Raw
Period : after 2018-11
(except for null values)

- **75% 이상 not null인 키워드 필터링. (총 17개 1차 연관 키워드 중 5개 키워드)**
- **연관 검색어인 만큼, 키워드 간 연관 관계 보임.**
- **Corr 계산 적용은 실질적으로 2018.11 이후 (5개 키워드의 null value 때문)**

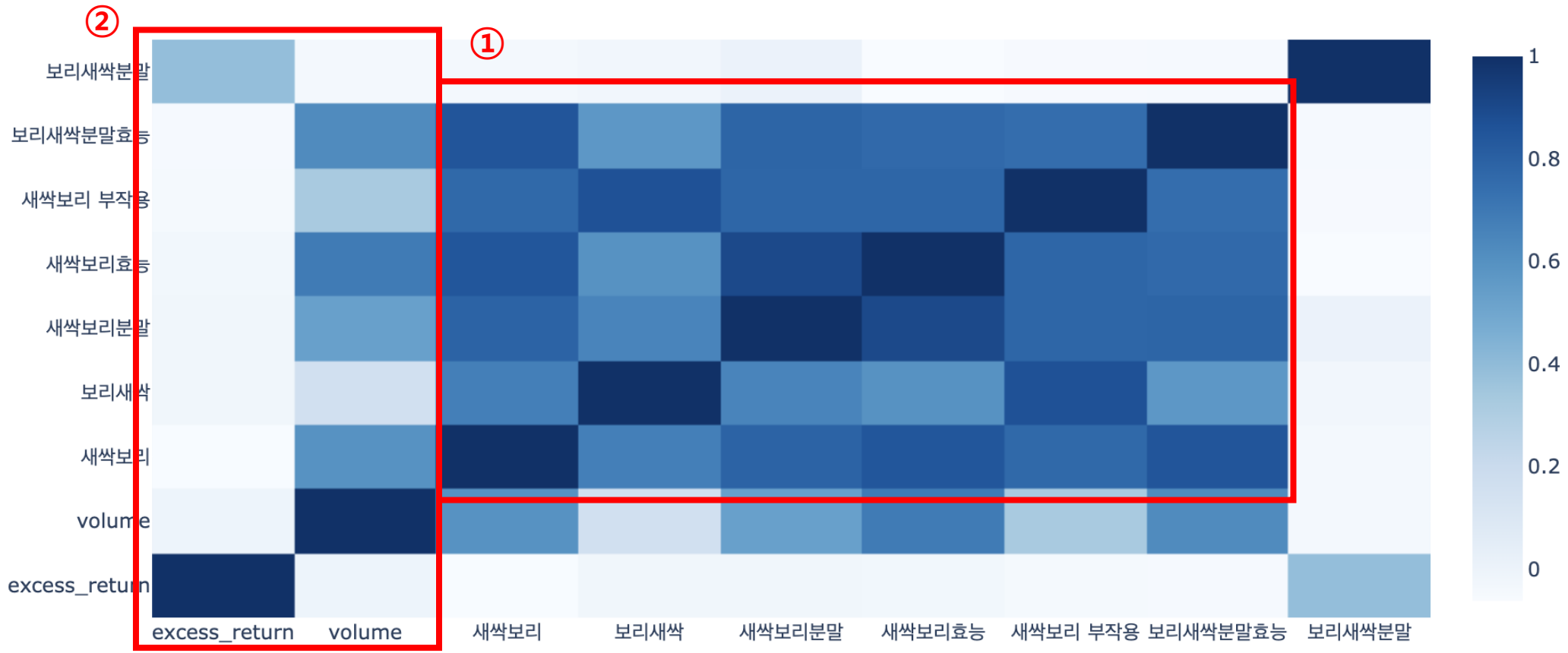




1. Exploratory Data Analysis : Related keyword

- ① **Yield & Log transformation 이후 키워드 간 연관 관계는 더 커짐.**
- ② **거래량과 키워드의 연관 관계는 보이나, 초과수익률과는 없음.**

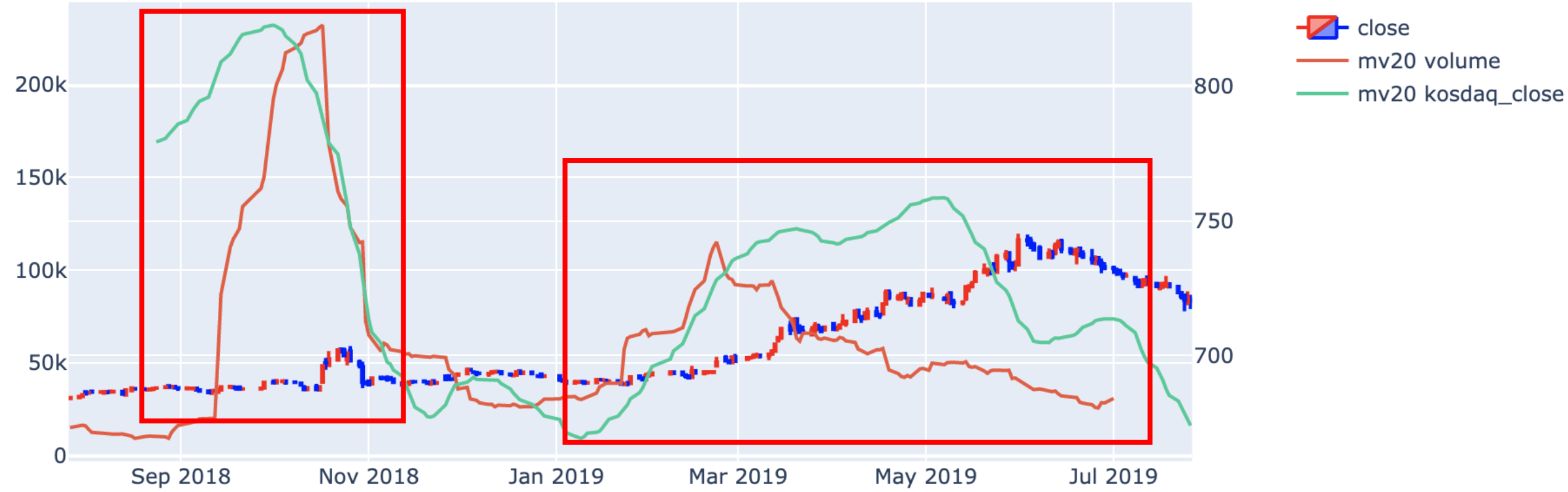
Transformation : log & yield
Period : after 2018-11
(except for null values)



1. Exploratory Data Analysis : Market & volume & stock price

Transformation : Raw / mv
Period : 1 year

- KOSDAQ 시장과 거래량의 트렌드가 주식 가격 변화에 선행하는 것으로 보임



2. Progress

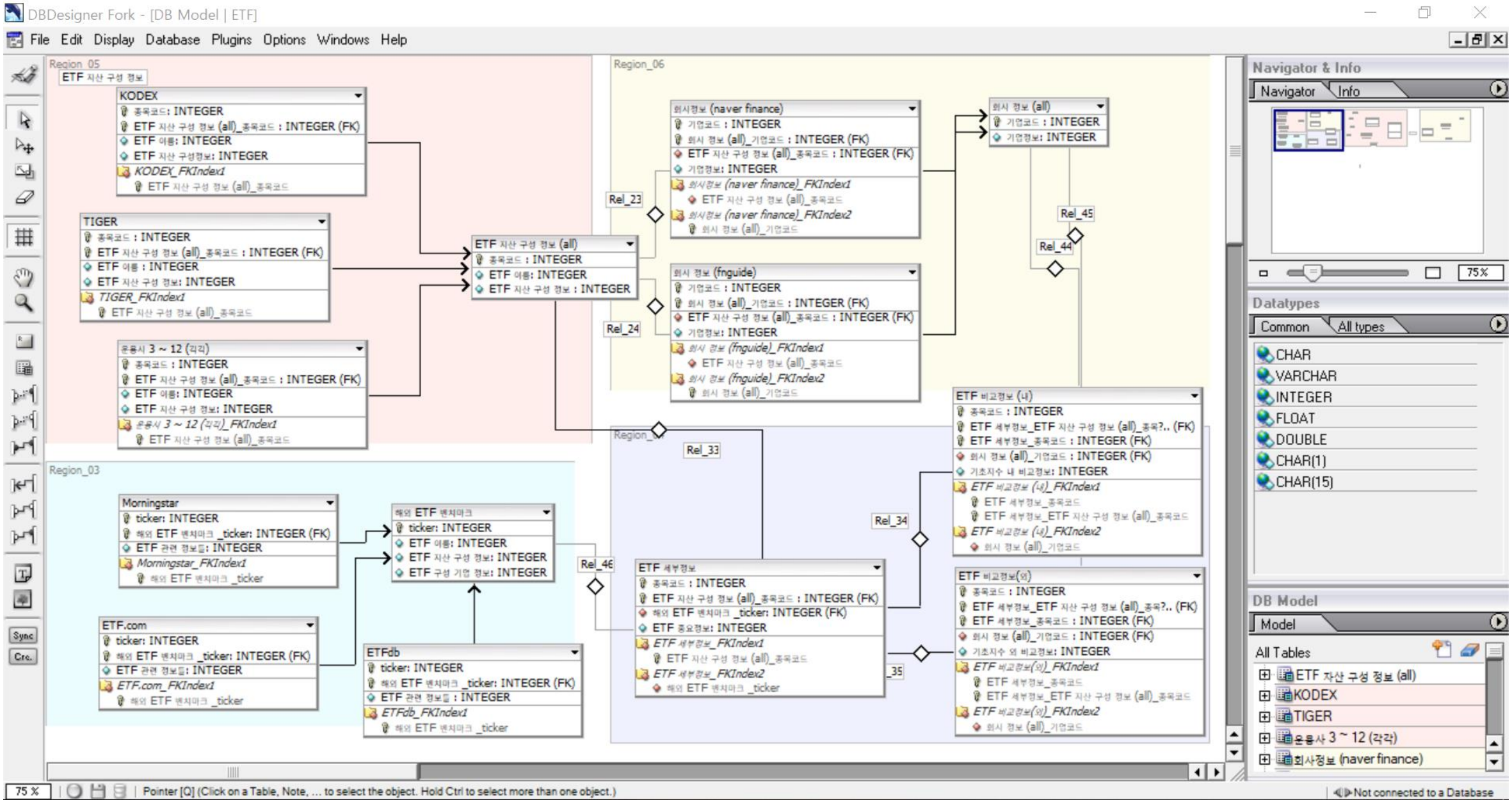
- Future works
  - Exploratory Data Analysis
    - Feature 검토 : 기존 연관 검색어와 linearity 작으면서, 검색량/null value 조건 충족
    - Time shifting : 가격에 대한 선행 지표로 거래량 (타기업 검토)
  - **Baseline modeling**
    - **Regression : 기간 반영에 따른 통계적 유의성 검토**
  - Idea
    - Sector clustering by search trend : 기업명 (종목코드) 검색 트렌드와 업종 클러스터링

FVID ETF

August 16, 2019

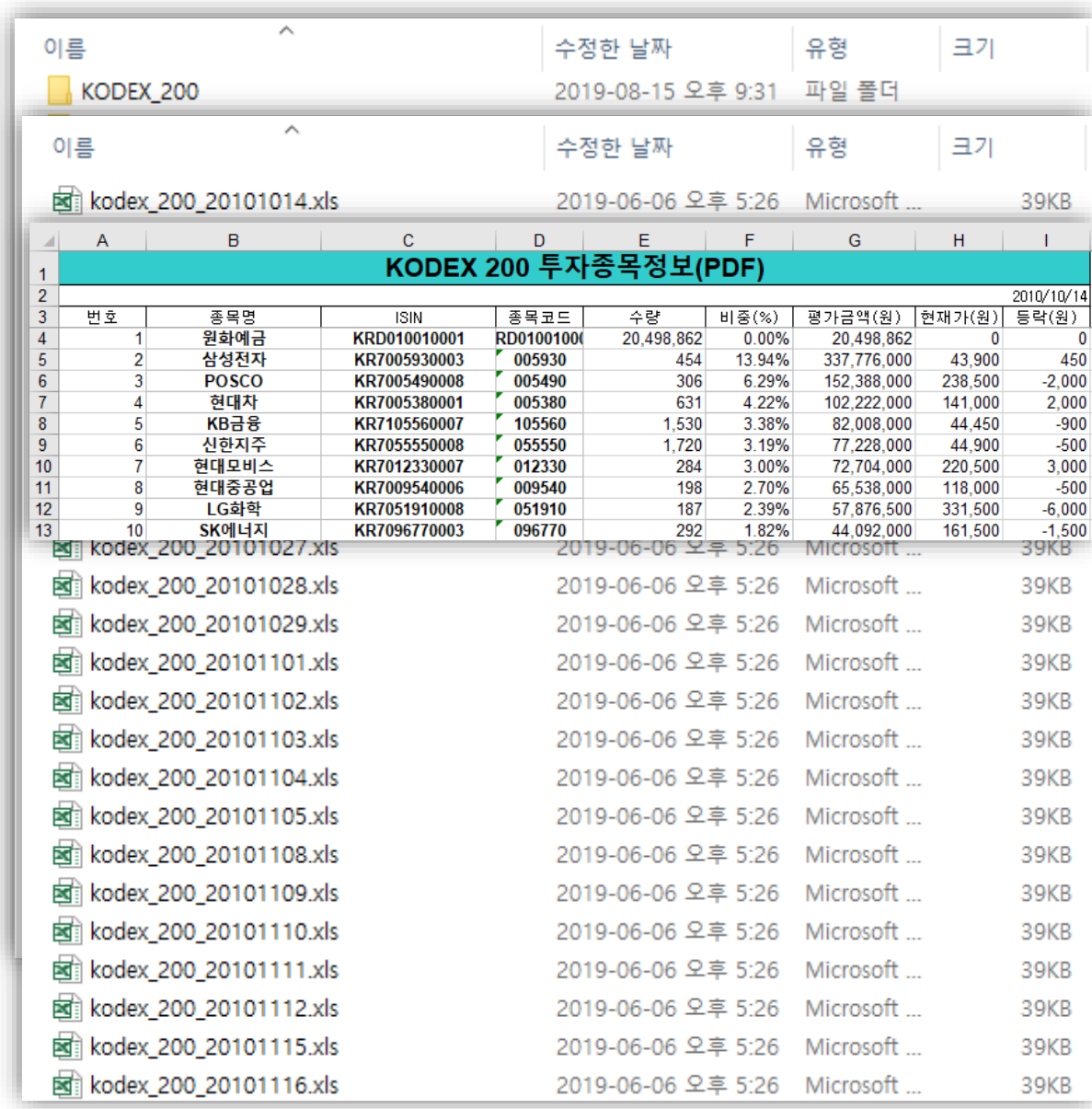
김현용, 박소정, 이명원, 이창민, 정지수

DB - Schema


DBDesigner Fork - [DB Model | ETF]
File Edit Display Database Plugins Options Windows Help
Region_05
ETF 자산 구성 정보
KODEX
종목코드: INTEGER
ETF 자산 구성 정보 (all)_종목코드 : INTEGER (FK)
ETF 이름: INTEGER
ETF 자산 구성정보: INTEGER
KODEX_FKIndex1
ETF 자산 구성 정보 (all)_종목코드
TIGER
종목코드 : INTEGER
ETF 자산 구성 정보 (all)_종목코드 : INTEGER (FK)
ETF 이름 : INTEGER
ETF 자산 구성 정보: INTEGER
TIGER_FKIndex1
ETF 자산 구성 정보 (all)_종목코드
운용사 3 ~ 12 (각각)
종목코드 : INTEGER
ETF 자산 구성 정보 (all)_종목코드 : INTEGER (FK)
ETF 이름: INTEGER
ETF 자산 구성 정보: INTEGER
운용사 3 ~ 12 (각각)_FKIndex1
ETF 자산 구성 정보 (all)_종목코드
ETF 자산 구성 정보 (all)
종목코드 : INTEGER
ETF 이름: INTEGER
ETF 자산 구성 정보 : INTEGER
Region_06
회사정보 (naver finance)
기업코드 : INTEGER
회사 정보 (all)_기업코드 : INTEGER (FK)
ETF 자산 구성 정보 (all)_종목코드 : INTEGER (FK)
기업정보: INTEGER
회사정보 (naver finance)_FKIndex1
ETF 자산 구성 정보 (all)_종목코드
회사정보 (naver finance)_FKIndex2
회사 정보 (all)_기업코드
회사 정보 (all)
기업코드 : INTEGER
기업정보: INTEGER
회사 정보 (fnguide)
기업코드 : INTEGER
회사 정보 (all)_기업코드 : INTEGER (FK)
ETF 자산 구성 정보 (all)_종목코드 : INTEGER (FK)
기업정보: INTEGER
회사 정보 (fnguide)_FKIndex1
ETF 자산 구성 정보 (all)_종목코드
회사 정보 (fnguide)_FKIndex2
회사 정보 (all)_기업코드
Rel_23
Rel_24
Rel_45
Rel_44
Region_04
Rel_33
Region_03
Morningstar
ticker: INTEGER
해외 ETF 벤치마크 _ticker: INTEGER (FK)
ETF 관련 정보들: INTEGER
Morningstar_FKIndex1
해외 ETF 벤치마크 _ticker
해외 ETF 벤치마크
ticker: INTEGER
ETF 이름: INTEGER
ETF 자산 구성 정보: INTEGER
ETF 구성 기업 정보: INTEGER
ETF.com
ticker: INTEGER
해외 ETF 벤치마크 _ticker: INTEGER (FK)
ETF 관련 정보들: INTEGER
ETF.com_FKIndex1
해외 ETF 벤치마크 _ticker
ETFdb
ticker: INTEGER
해외 ETF 벤치마크 _ticker: INTEGER (FK)
ETF 관련 정보들 : INTEGER
ETFdb_FKIndex1
해외 ETF 벤치마크 _ticker
Rel_46
ETF 세부정보
종목코드 : INTEGER
ETF 자산 구성 정보 (all)_종목코드 : INTEGER (FK)
해외 ETF 벤치마크 _ticker: INTEGER (FK)
ETF 종요정보: INTEGER
ETF 세부정보_FKIndex1
ETF 자산 구성 정보 (all)_종목코드
ETF 세부정보_FKIndex2
해외 ETF 벤치마크 _ticker
Rel_34
_35
ETF 비교정보 (내)
종목코드 : INTEGER
ETF 세부정보_ETF 자산 구성 정보 (all)_종목?.. (FK)
ETF 세부정보_종목코드 : INTEGER (FK)
회사 정보 (all)_기업코드 : INTEGER (FK)
기초지수 내 비교정보: INTEGER
ETF 비교정보 (내)_FKIndex1
ETF 세부정보_종목코드
ETF 세부정보_ETF 자산 구성 정보 (all)_종목코드
ETF 비교정보 (내)_FKIndex2
회사 정보 (all)_기업코드
ETF 비교정보(외)
종목코드 : INTEGER
ETF 세부정보_ETF 자산 구성 정보 (all)_종목?.. (FK)
ETF 세부정보_종목코드 : INTEGER (FK)
회사 정보 (all)_기업코드 : INTEGER (FK)
기초지수 외 비교정보: INTEGER
ETF 비교정보(외)_FKIndex1
ETF 세부정보_종목코드
ETF 세부정보_ETF 자산 구성 정보 (all)_종목코드
ETF 비교정보(외)_FKIndex2
회사 정보 (all)_기업코드
Navigator & Info
Navigator
Info
75%
Datatypes
Common
All types
CHAR
VARCHAR
INTEGER
FLOAT
DOUBLE
CHAR(1)
CHAR(15)
DB Model
Model
All Tables
ETF 자산 구성 정보 (all)
KODEX
TIGER
운용사 3 ~ 12 (각각)
회사정보 (naver finance)
Sync
Cre.
75 %
Pointer [Q] (Click on a Table, Note, ... to select the object. Hold Ctrl to select more than one object.)
Not connected to a Database

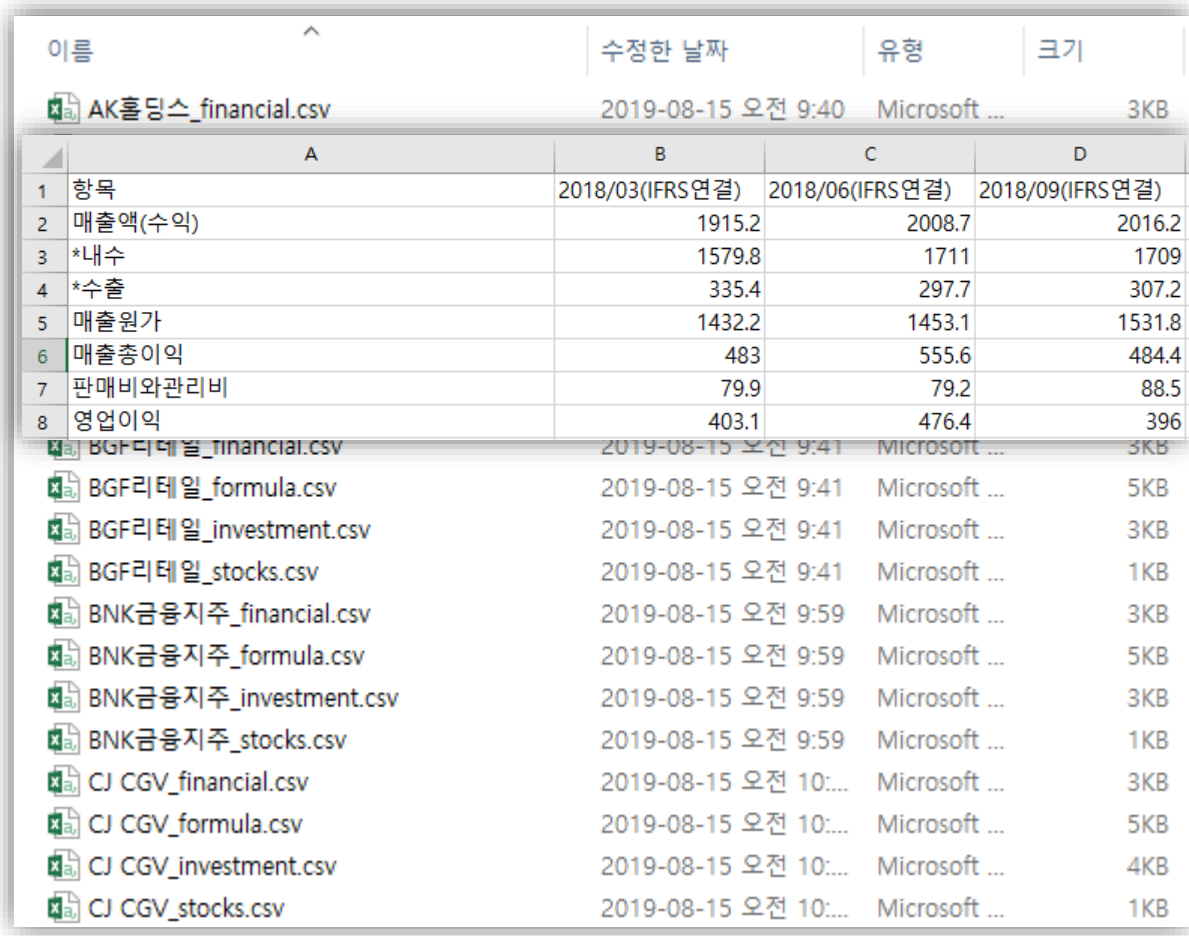
DB - ETF 구성자산

- ETF 구성자산
  - 자산운용사가 일 단위로 ETF 구성자산정보를 공개
  - 종목명, ISIN, 종목코드, 수량, 비중(%), 평가금액, 현재가, 등락 (KODEX, 삼성자산운용 기준)
  - (오른쪽 그림) ETF 브랜드별, 종목별로 그룹을 생성
  - (오른쪽 그림) ETF 구성자산정보를 XLS파일로 저장
  - KODEX, TIGER ETF 230개 종목의 정보를 수집



DB - 회사정보

- 회사정보
  - 네이버금융, 에프엔가이드 등에서 공개하는 회사정보
  - 재무분석, 투자지표, 시황 등
  - (오른쪽 그림) 회사별로 재무분석, 투자지표, 시황 등 추출
  - (오른쪽 그림) 회사정보를 분류별로 CSV파일로 저장
  - KOSPI200에 속하는 200개 회사의 정보를 수집



이름 | 수정한 날짜 | 유형 | 크기

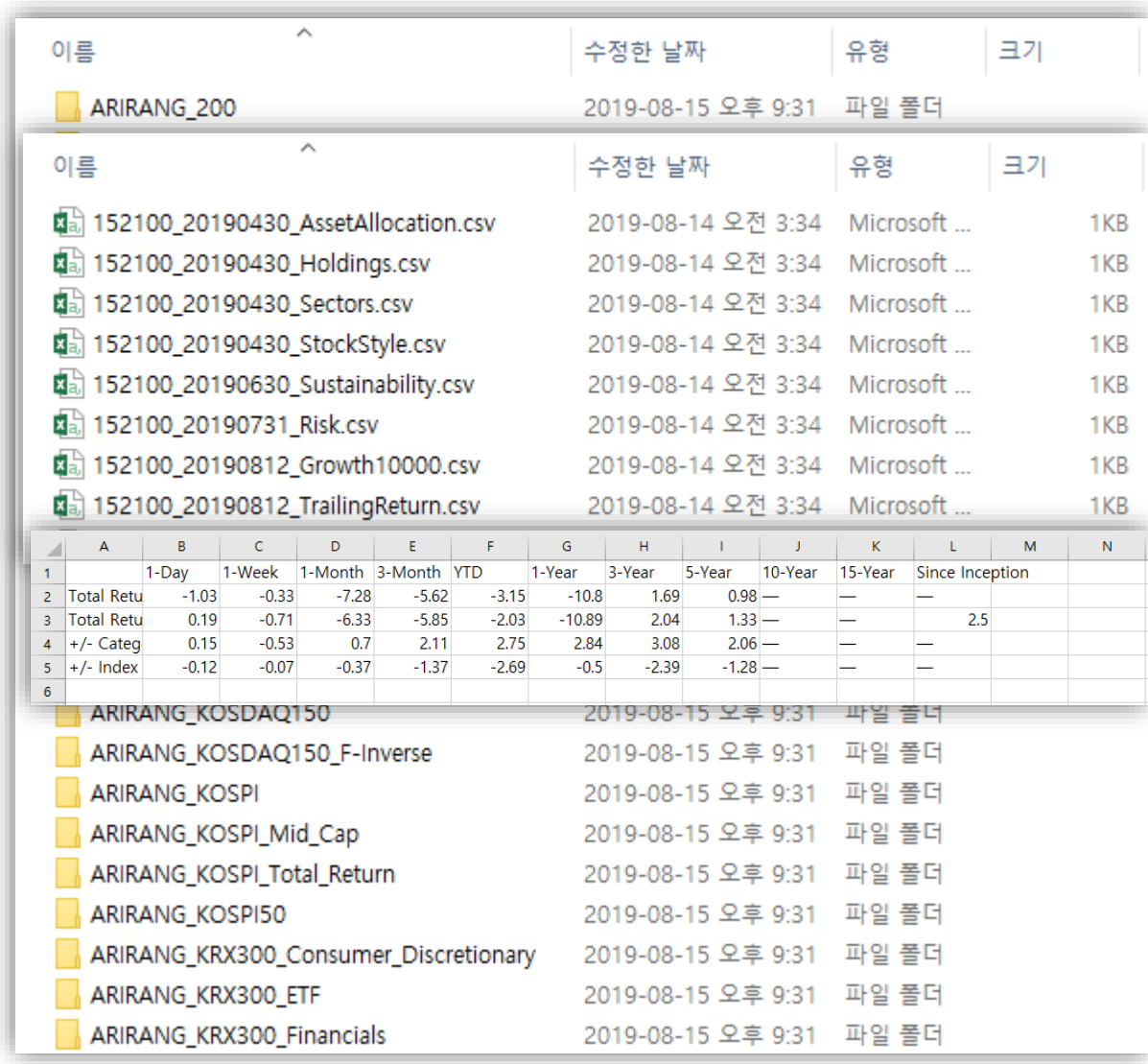
AK홀딩스_financial.csv 2019-08-15 오전 9:40 Microsoft ... 3KB

	A	B	C	D
1	항목	2018/03(IFRS연결)	2018/06(IFRS연결)	2018/09(IFRS연결)
2	매출액(수익)	1915.2	2008.7	2016.2
3	*내수	1579.8	1711	1709
4	*수출	335.4	297.7	307.2
5	매출원가	1432.2	1453.1	1531.8
6	매출총이익	483	555.6	484.4
7	판매비와관리비	79.9	79.2	88.5
8	영업이익	403.1	476.4	396

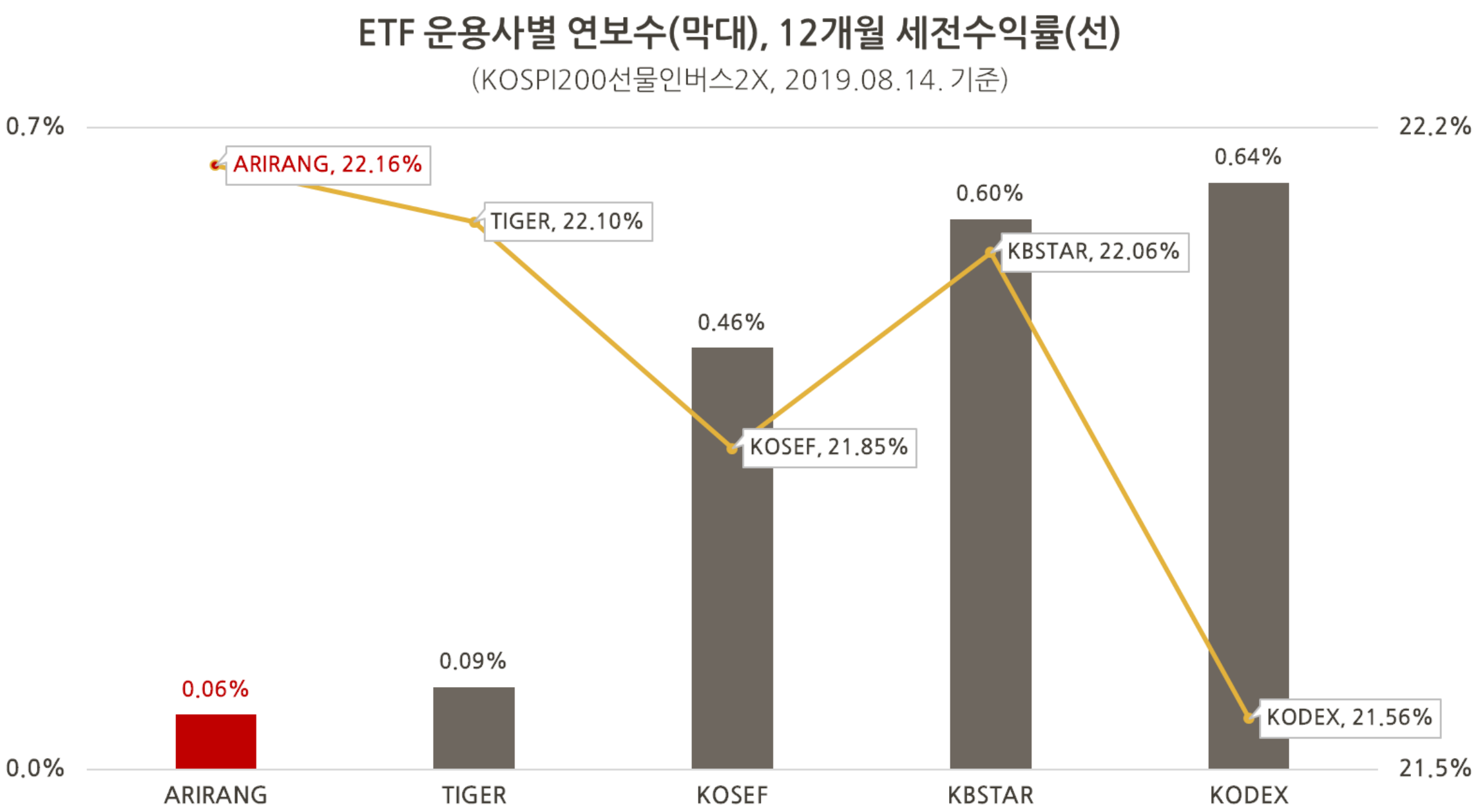
BGF리테일_financial.csv 2019-08-15 오전 9:41 Microsoft ... 3KB
BGF리테일_formula.csv 2019-08-15 오전 9:41 Microsoft ... 5KB
BGF리테일_investment.csv 2019-08-15 오전 9:41 Microsoft ... 3KB
BGF리테일_stocks.csv 2019-08-15 오전 9:41 Microsoft ... 1KB
BNK금융지주_financial.csv 2019-08-15 오전 9:59 Microsoft ... 3KB
BNK금융지주_formula.csv 2019-08-15 오전 9:59 Microsoft ... 5KB
BNK금융지주_investment.csv 2019-08-15 오전 9:59 Microsoft ... 3KB
BNK금융지주_stocks.csv 2019-08-15 오전 9:59 Microsoft ... 1KB
CJ CGV_financial.csv 2019-08-15 오전 10:... Microsoft ... 3KB
CJ CGV_formula.csv 2019-08-15 오전 10:... Microsoft ... 5KB
CJ CGV_investment.csv 2019-08-15 오전 10:... Microsoft ... 4KB
CJ CGV_stocks.csv 2019-08-15 오전 10:... Microsoft ... 1KB

DB - 해외벤치마크

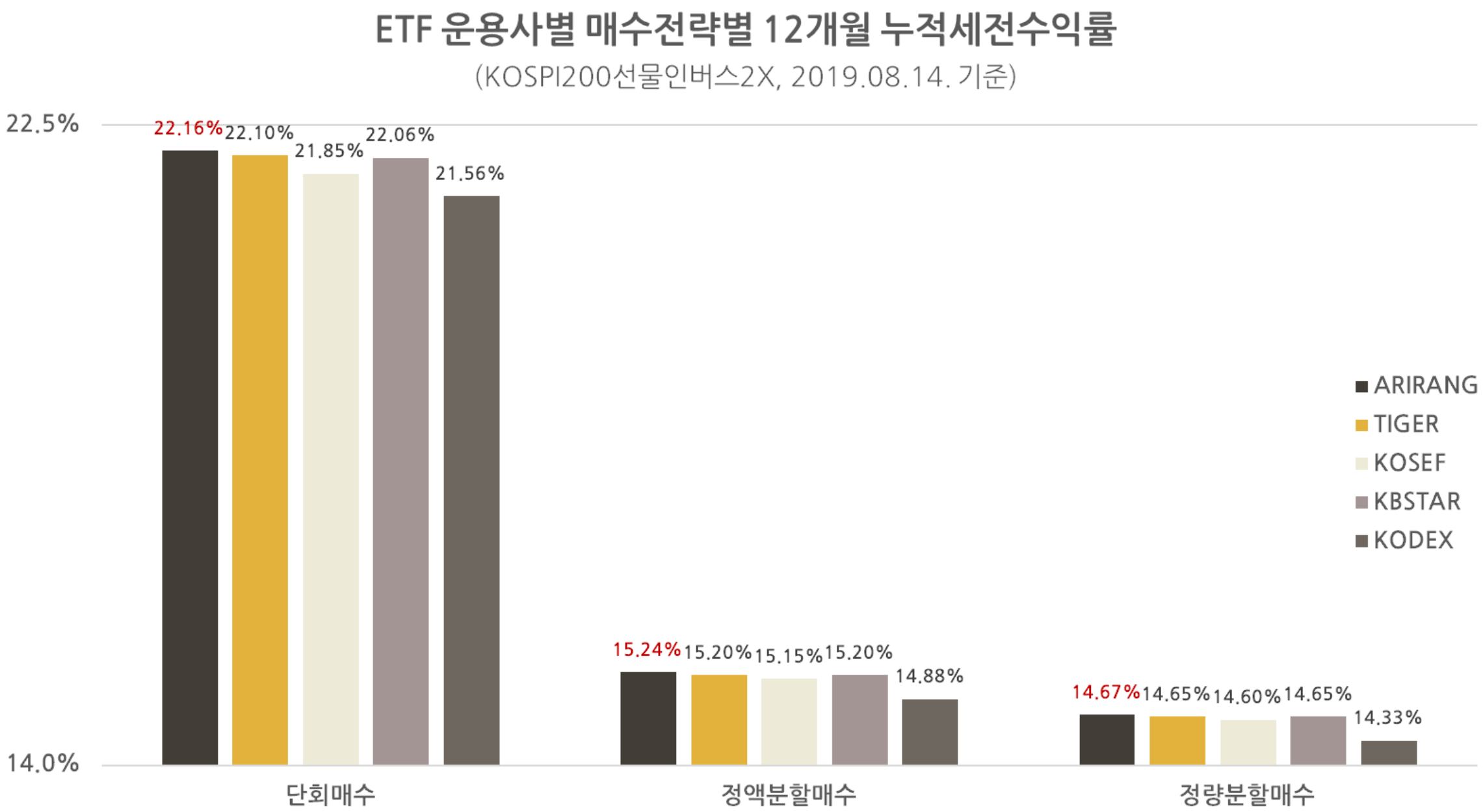
- 해외벤치마크
  - MorningStar, ETF.com, ETFdb.com 등에서 제공하는 ETF 정보
  - ETF개요, 시황, 자산구성, 수익률 등
  - (오른쪽 그림) ETF 종목별로 그룹을 생성
  - (오른쪽 그림) 제공정보를 분류별로 CSV파일로 저장
  - 국내 상장된 434개 ETF의 정보를 수집



ETF 비교정보 (1/2)


ETF 운용사별 연보수(막대), 12개월 세전수익률(선)
(KOSPI200선물인버스2X, 2019.08.14. 기준)
0.7%
22.2%
0.0%
21.5%
ARIRANG, 22.16%
TIGER, 22.10%
KOSEF, 21.85%
KBSTAR, 22.06%
KODEX, 21.56%
0.06%
0.09%
0.46%
0.60%
0.64%
ARIRANG
TIGER
KOSEF
KBSTAR
KODEX

ETF 비교정보 (2/2)


ETF 운용사별 매수전략별 12개월 누적세전수익률
(KOSPI200선물인버스2X, 2019.08.14. 기준)
22.5%
14.0%
22.16%
22.10%
21.85%
22.06%
21.56%
15.24%
15.20%
15.15%
15.20%
14.88%
14.67%
14.65%
14.60%
14.65%
14.33%
ARIRANG
TIGER
KOSEF
KBSTAR
KODEX
단회매수
정액분할매수
정량분할매수

후속연구

- 후속연구
  - 국내주식 ETF -> 국내상장 ETF -> 해외상장 ETF
  - 주식 외 자산으로 구성된 ETF 비교방법론
  - 해외상장 ETF 제도, 법률
  - 크롤링 자동화(스케쥴링, ...)
  - 크롤러 언어 변경(Python -> JavaScript)
  - 크롤링 대상 변경(근원적 정보원)
  - DB 구성(물리적, 개념적, ...)
  - ETF 관련 거래 전략

August 16, 2019

Financial Fraud Detection using Topic Modeling

조현상, 박건호, 나현수

brianhscho@dm.snu.ac.kr

Table of Contents

Data Management (Winning Regex)

```
pattern_text = [
            r"manage?ment'?s *(?:narrative *analysis|discussion *and *analysis) *of" + "(.*?)" + "(?:consolidated)?
              *financial *statement[s]? *and *supplement(?:ary|al) *(?:data|information)",

              r"manage?ment'?s *(?:narrative *analysis|discussion *and *analysis) *of" + "(.*?)" + "change[s]? *in *and
             *disagreement[s]?",

              r"manage?ment'?s *(?:narrative *analysis|discussion *and *analysis)" + "(.*?)" + "(?:consolidated)? *financial
                *statement[s]? *and *supplement(?:ary|al) *(?:data|information)",

              r"manage?ment'?s *(?:narrative *analysis|discussion *and *analysis)" + "(.*?)" + "change[s]? *in *and
                *disagreement[s]?"
              ]
```

Typos exist … (example: ‘managment’) assumes no other typos present in docs

Data Management (Statistics)

S&P500 Companies (n = 1,418)

305 unique SIC

- 1035675-ELECTRIC SERVICES-False
- 1035713-NATIONAL COMMERCIAL BANKS-False
- 1035881-RADIO TV BROADCASTING COMMUNICATIONS EQUIPMENT-False
- 1035972-MISCELLANEOUS CHEMICAL PRODUCTS-False
- 1037038-MEN'S BOYS' FURNISHINGS, WORK CLOTHING, AND ALLIED GARMENTS-False
- 1037540-REAL ESTATE INVESTMENT TRUSTS-False
- 1037646-LABORATORY ANALYTICAL INSTRUMENTS-False
- 1037868-MOTORS GENERATORS-False
- 1037949-TELEPHONE COMMUNICATIONS (NO RADIO TELEPHONE)-False
- 1038339-REAL ESTATE INVESTMENT TRUSTS-False
- 1038357-CRUDE PETROLEUM NATURAL GAS-False
- 1038914-DRILLING OIL GAS WELLS-False
- 1039101-RADIO TV BROADCASTING COMMUNICATIONS EQUIPMENT-False

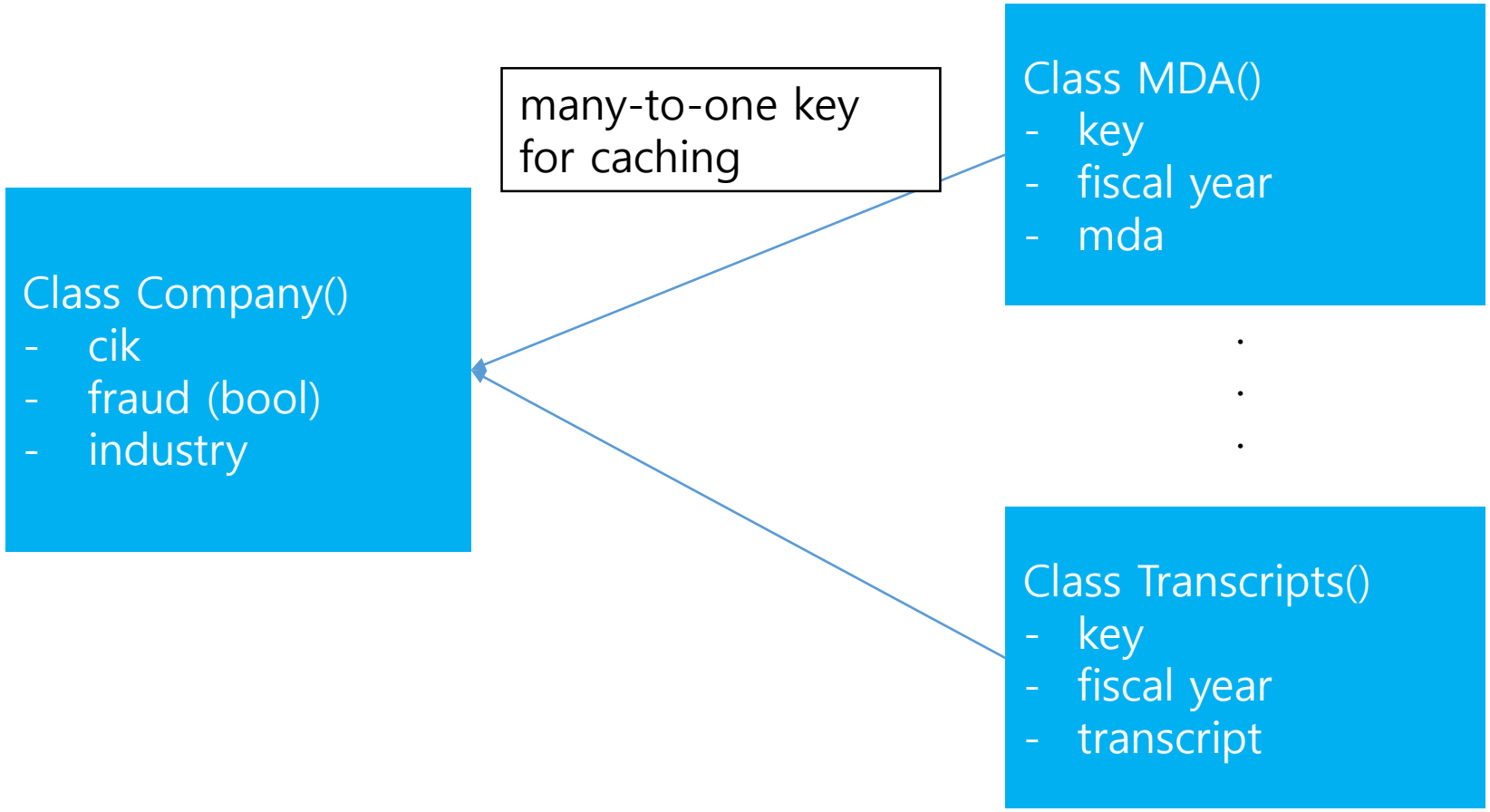
1 2 3 4 ... 14 15 1418 companys

S&P500 MDA (n = 17,330)

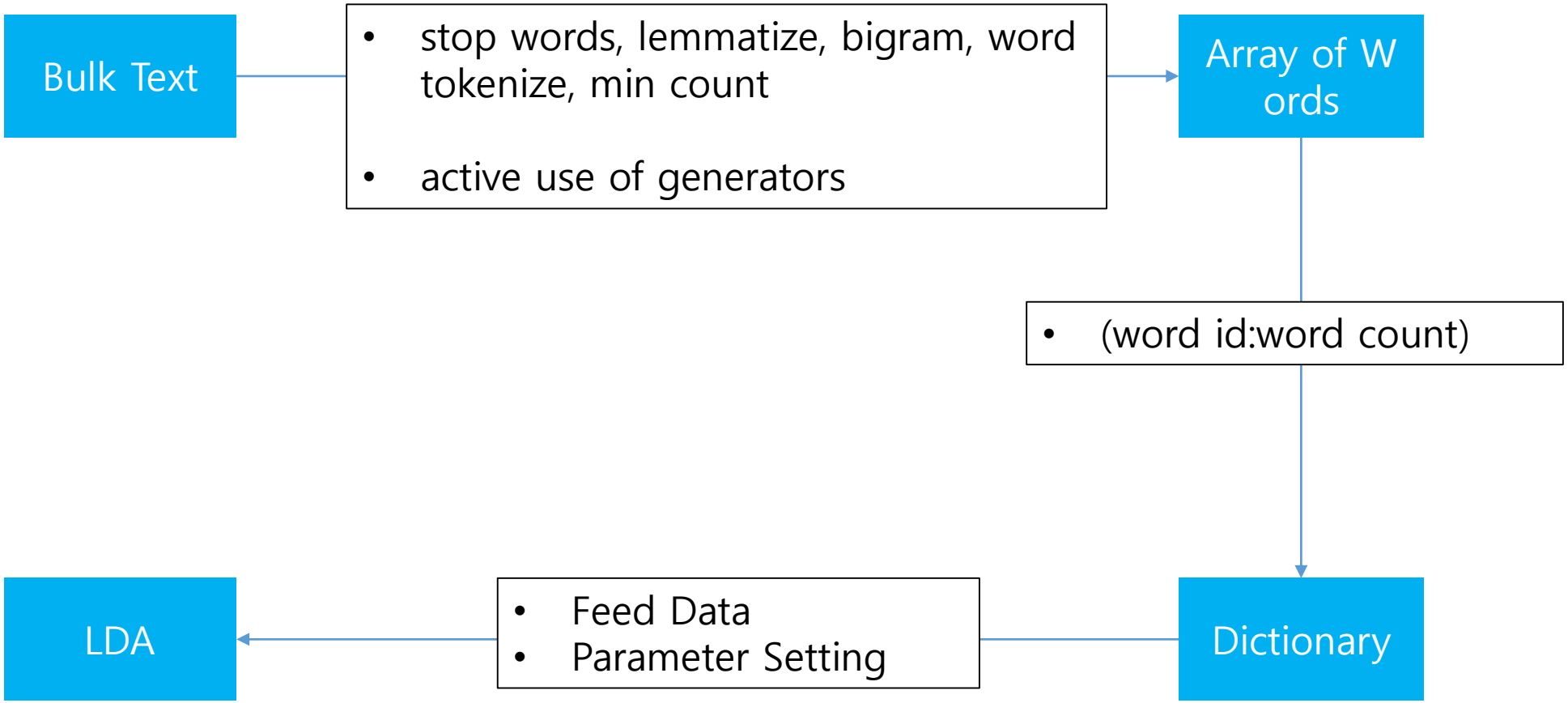
- 00-14272-PHARMACEUTICAL PREPARATIONS-False
- 00-28385-ELECTRIC SERVICES-False
- 00-757011-CONCRETE GYPSUM PLASTER PRODUCTS-False
- 00-728535-TRUCKING (NO LOCAL)-False
- 00-60519-LUMBER WOOD PRODUCTS (NO FURNITURE)-False
- 00-72162-INDUSTRIAL INORGANIC CHEMICALS-False
- 00-1074023-BLANK CHECKS-False
- 00-1011006--False
- 00-25350-CANNED, FROZEN PRESERVED FRUIT, VEG FOOD SPECIALTIES-False
- 00-73124-STATE COMMERCIAL BANKS-False
- 00-858339-HOTELS MOTELS-False
- 00-96943-SURGICAL MEDICAL INSTRUMENTS APPARATUS-False
- 00-883569-WATCHES, CLOCKS, CLOCKWORK OPERATED DEVICES/PARTS-False

1 2 3 4 ... 173 174 17330 mdas

Data Management (Django ORM)



Data Pre-processing



LDA (Selected Topics)

- **Investment**

(investment, fund, security, equity, management, share, stock, common, brokerage, refranchising)

- **Natural Resources**

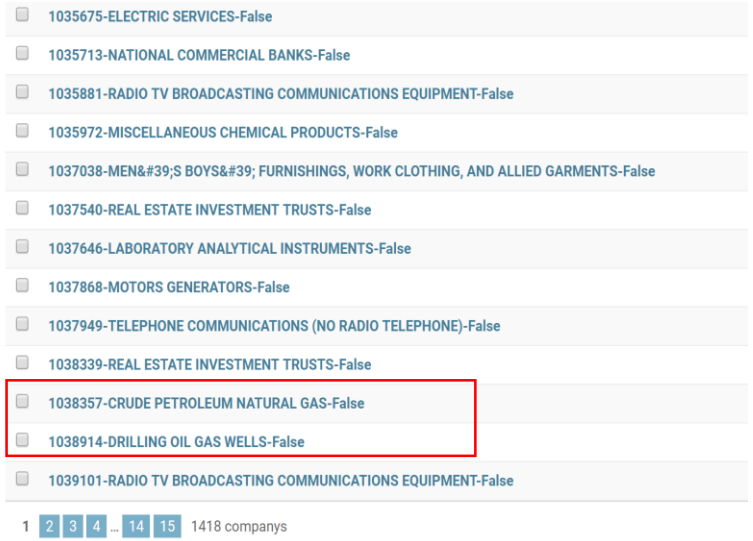
(gas, oil, production, price, natural, reserve, crude, drilling, well, per)

- **Health Care**

(product, health, care, pharmaceutical, drug, research, hca, medical, patent, patient)

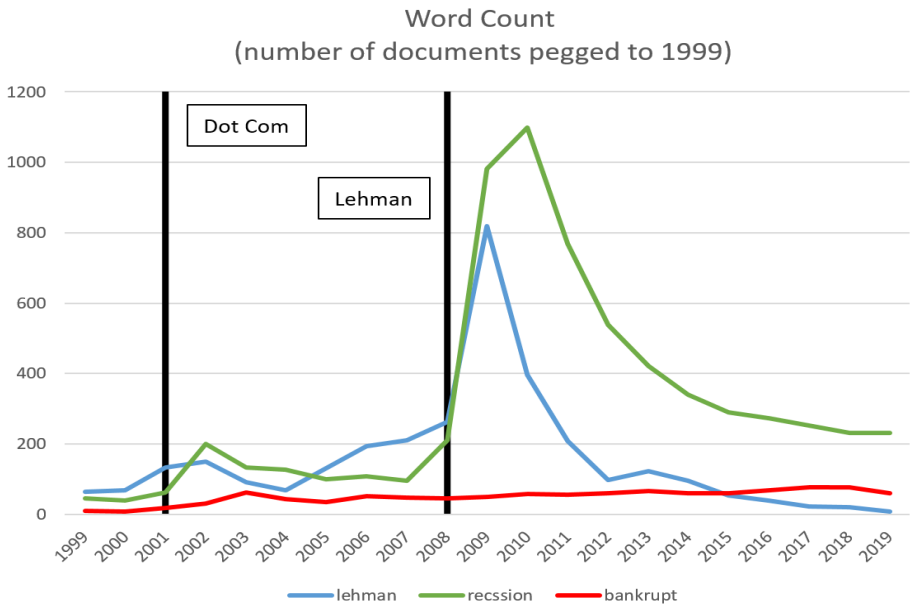
LDA (Selected Topics by Industry)

- SIC (Standard Industrial Classification) can be matched with CIK
- S&P 500 companies cover 305 SICs … Too many
- Looking for something less specific



LDA (Selected Topics by Year)

- **Trend is clear in word count**



- **2010 Topic** Visualization

- **But not in 2010 Topic?**

1. filed, court, state, year, plan, agreement, certain, claim, action, case

2. Why?

To-Dos

- **135 overlapping cases between SP500 and AAER**

  Example:

  – The Coca-Cola Company (CIK 0000021344) in SP500 Index since 1964.

  – BUT accused and finalized of fraudulent reporting.

- **Fraudulent Report Labeling**

"At no point between 1997 and 1999, however, did Coca-Cola publicly disclose to shareholders the existence of gallon pushing, the impact of gallon pushing on its current income, or the likely impact of gallon pushing on its future income."

To-Dos

- Context Embedding

Thank you

Any feedback would be appreciated!