Clustering Analysis of Investors & Trending Topic Detection

Capstone Project

RUI BAI, XINYI LIU, YICHI LIU, YUCHEN PEI, YUJING SONG

CAPSTONE PLAN 2020

Group Introduction

Student Group:
- Rui Bai (rb3454)
- Xinyi Liu (xl2904)
- Yichi Liu (yl4327)
- Yuchen Pei (yp2533)
- Yujing Song (ys3251)

Instructor:
- Adam S. Kelleher

Guide Group:
- Shadi Fadaee
- Flora Huang
- Stephen Lawrence
- Maria Colangelo
- Min Fang

Project Motivation & Objectives

Motivation

- Clustering analysis of institutional investors remained less explored.
- Investors start to act differently than the others, which reflected on their portfolio holding changes.

Objective

- Create a clustering model for institutional investors. Obtain the cluster that is dissimilar to Vanguard.
- Detect significant trending topics:
    - Identify the features of investors that are significant
    - Identify the cluster of investors that tend to drive the trend

01

Data Exploration

1 Data Exploration - Refinitiv Database

- **Investor List**:
  - 225 institutional investors short-listed by Vanguard

- **Data Scope**
  - Clustering: 2016Q2 - 2020Q2
  - Trending topics detection: 2010Q2 - 2020Q2

- **Data Manipulation**
  - Merged 6 tables from 2 data schemes to get **13F** holding details
  - Merged 6 tables to get employees' information including their investment style
  - Calculated market cap of instruments by multiplying their price and outstanding shares
  - Queried industry information in 5 tables across 2 data schemes
  - Queried total assets, turnover rate, number of positions in the database
  - Explored asset allocation and return rate of investors
  - Glanced at return information of fund-level investors

02 Features Extraction

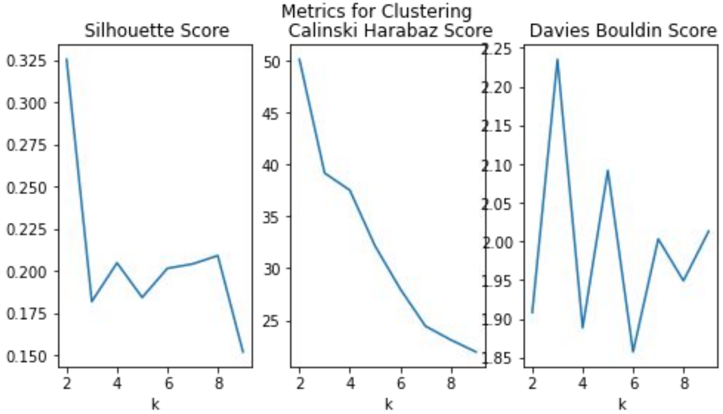
2 7 Features of Investors

		Feature	Meaning
Time-series Features	01	percentage of portfolio by aggregating instruments on market capitalization	Investment tendency towards instruments with different market cap size
	02	percentage of portfolio by aggregating instruments on industry	Investment tendency towards instruments with different industry
	03	percentage of portfolio in top 20% instruments	Investment concentration
	04	Quarterly turnover rate	Investment activeness
	05	Number of instruments	Investment diversity
	06	Total assets	the size of the company
Static Features	07	Investment style distribution of employees	the interested investment style of companies.

03 Clustering

3.1 Clustering Model

Features	Similarity measurement	Model	Evaluation
7 features defined above	✓ Euclidean distance	✓ K-means model	✓ Silhouette Score
	DTW distance	Gaussian Mixture Model	✓ Calinski Harabasz Score
		Spectral clustering	✓ Davies Bouldin Score

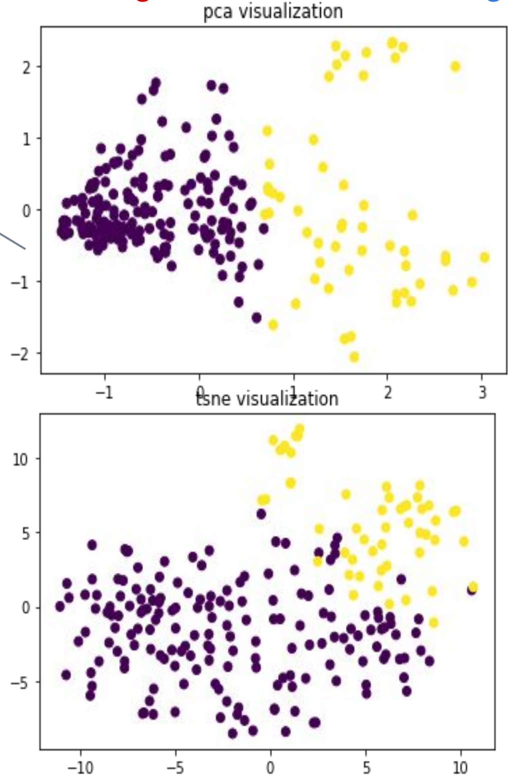


→ Best number of cluster is 2

3.2 Descriptive Analysis for Clustering Results

Cluster Visualization

With Vanguard   Without Vanguard



Cluster A:
171 investors
Similar to Vanguard

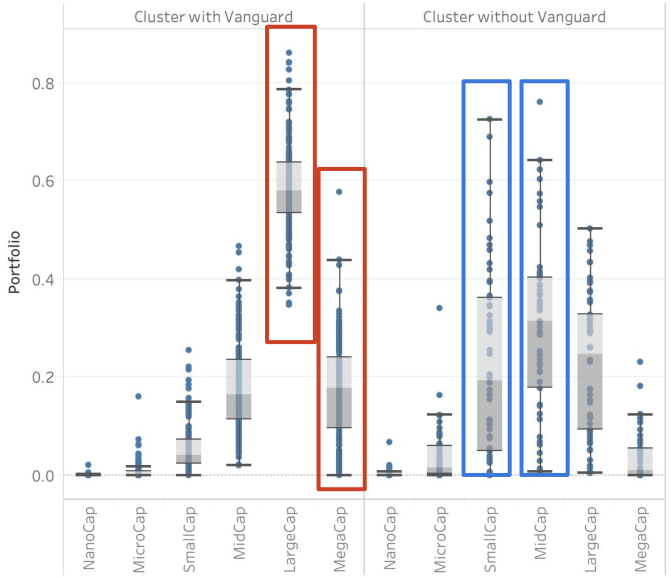
- JP Morgan Asset Management
- UBS Financial Services
- Goldman Sachs
- The Vanguard Group, Inc.
- BlackRock Institutional

Cluster B:
54 investors
Dissimilar to Vanguard

- Heartland Advisors, Inc.
- Angelo, Gordon & Co.
- New York Life Investment Management, LLC
- Discovery Capital Management, LLC
- King Street Capital Management, L.P.

3.2 Descriptive Analysis for Clustering Results

1 Tendency towards instruments with different market cap size



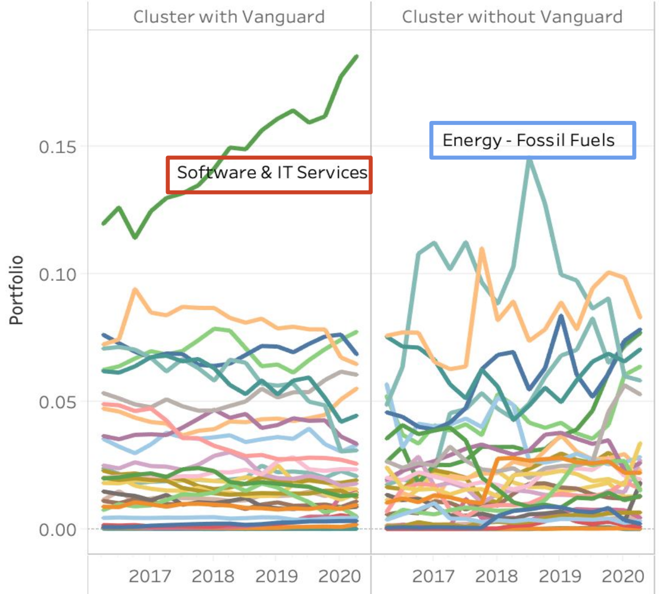
Cluster with Vanguard:

Invest more in instruments with large & mega market cap

Cluster without Vanguard:

Invest more in instruments with small & middle market cap

2 Tendency towards instruments with different industry



Cluster with Vanguard:

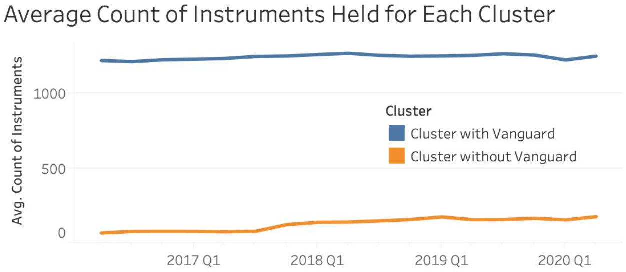
Invest more in the Software and IT service industry

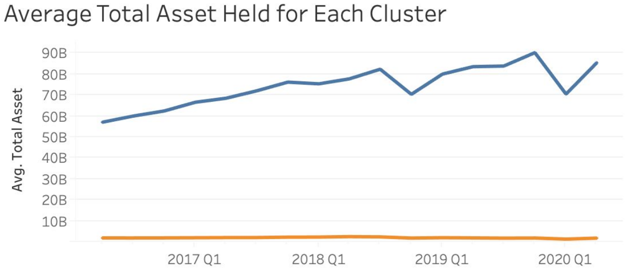
Cluster without Vanguard:

Invest more in the Energy industry

3.2 Descriptive Analysis for Clustering Results

3 Size of the company





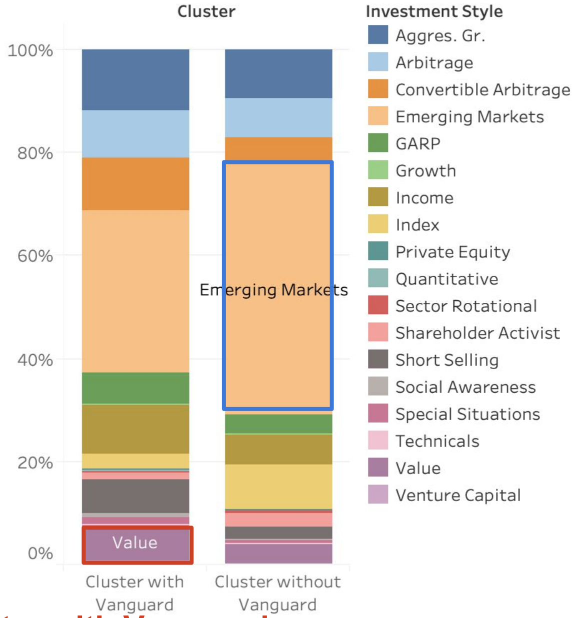
Cluster with Vanguard:

More total assets and more number of instruments

Cluster without Vanguard:

Less total assets and less number of instruments

4 Employees' investment style



Cluster with Vanguard:

Investment style has more focus on value

Cluster without Vanguard:

Investment style has more focus on emerging market

3.2 Descriptive Analysis for Clustering Results

Summary of the difference

Cluster with Vanguard:

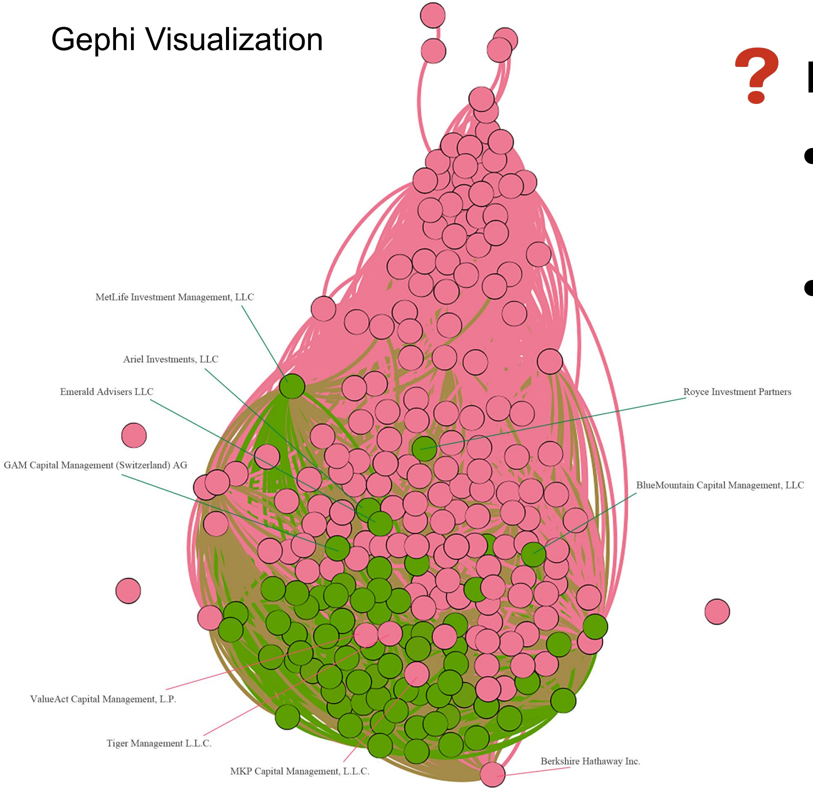
- Invest more in instruments with large & mega market capitalizations
- Invest more in the Software and IT service industry
- More total assets and more number of instruments
- Investment style has more focus on value

Cluster without Vanguard:

- Invest more in instruments with small & middle market capitalizations
- Invest more in the Energy industry
- Less total assets and less number of instruments
- Investment style has more focus on emerging market

3.3 Network Analysis for Clustering Results

Gephi Visualization



? Is clustering result correct

- Distance is measured by the similarity of features

$$Sim_i' = D_{max} - D_i + D_{min}$$

- Clusters in K-means positioned in different parts of the graph

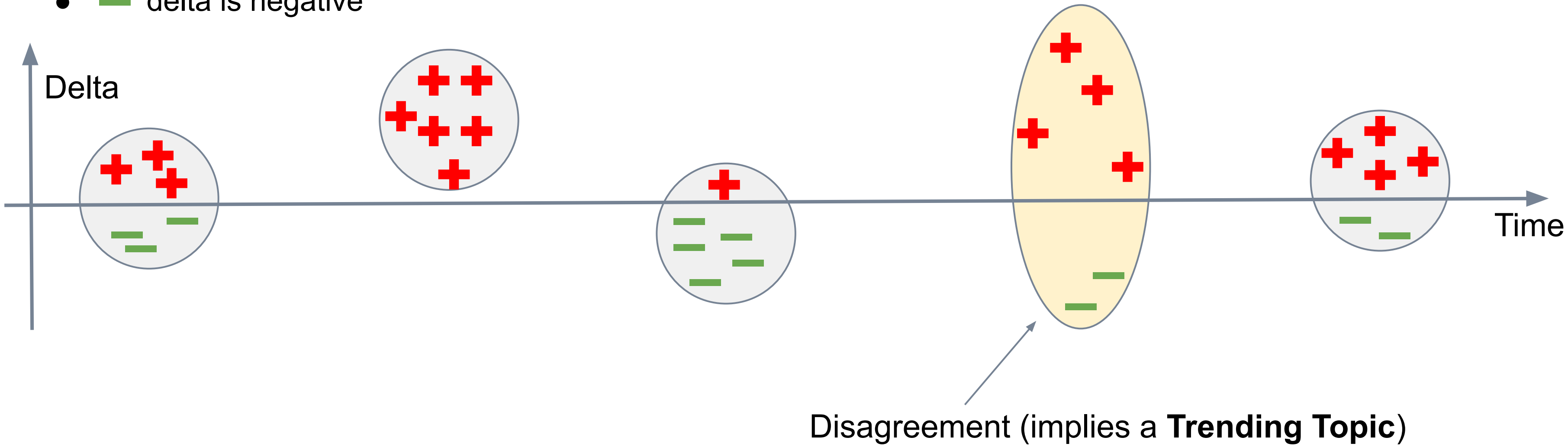
04

Trending Topics

4.1 Trending Topics Definition & Metrics

In an industry, using **delta** to represent the change of percentage in an investor's portfolio holdings

- ➕ delta is positive
- ➖ delta is negative



Large dispersion of deltas → Disagreement among the investors → **Trending topic** (some are optimistic, while others are pessimistic towards this industry)

4.1 Trending Topics Definition & Metrics

Define 3 metrics to measure this **disagreement**:

For each instrument j at time t:

Metric 1:
Variance

$$V_{jt} = Var(\Delta_{ijt})$$

- Measuring the dispersion of investors' percentage portfolio changes
- Sensitive to extreme values

Metric 2:
Quantile Range

$$IQR_{jt} = Q_3(\Delta_{ijt}) - Q_1(\Delta_{ijt})$$

- Measuring where the middle 50% of the delta sit
- Not sensitive to extreme values (more resistant)

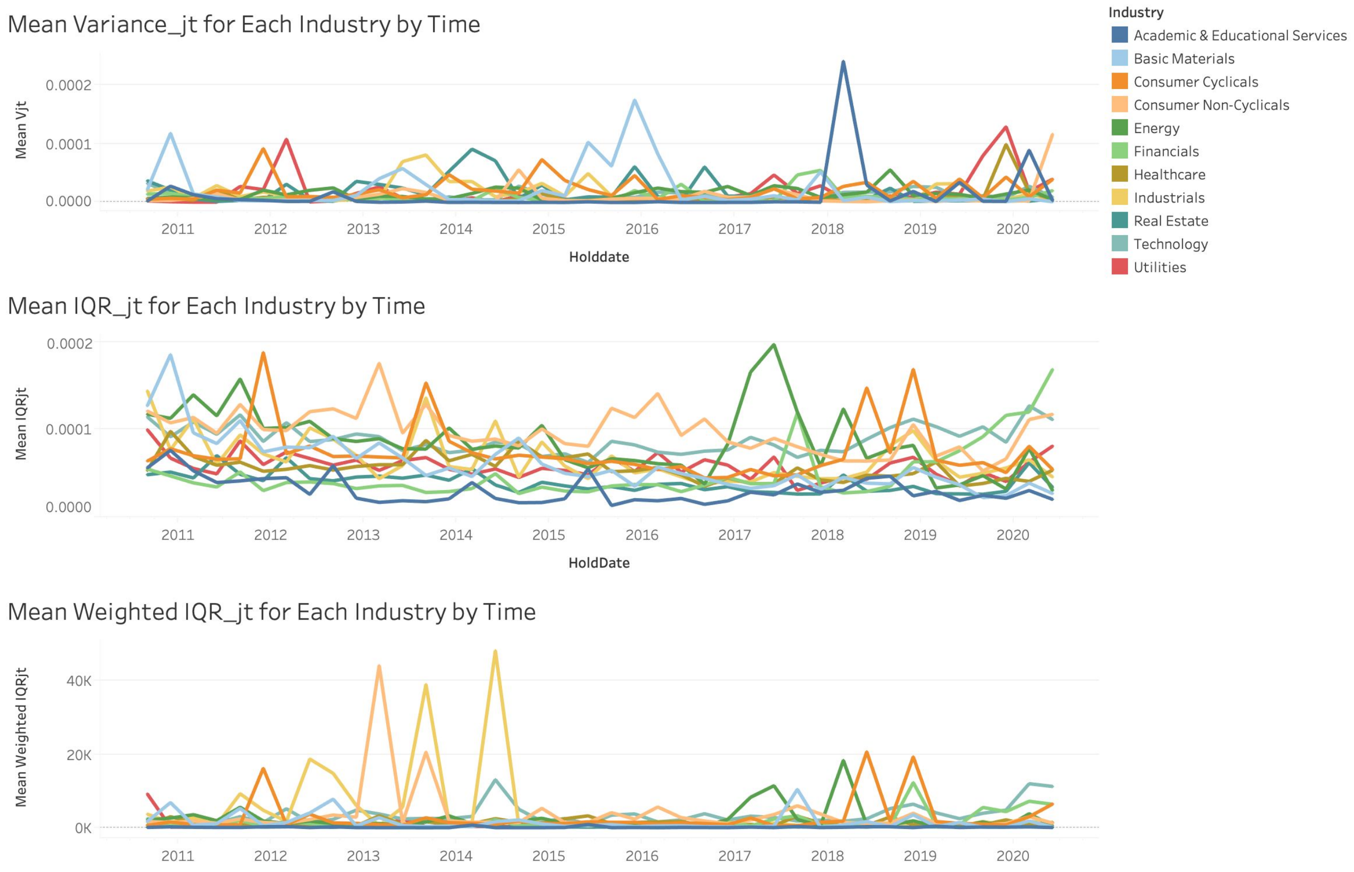
Metric 3:
Quantile Range with Weights

$$Weighted\ IQR_{jt} = Weight_{ijt} \times \{Q_3(\Delta_{ijt}) - Q_1(\Delta_{ijt})\}$$

- Adjusting Metric 2 by actual transaction value
- Not sensitive to extreme values (more resistant)

Where Δ_{ijt} is the percentage of portfolio change in instrument j of investor i at time t, and $Weight_{ijt}$ is the dollar value change in Δ_{ijt}.

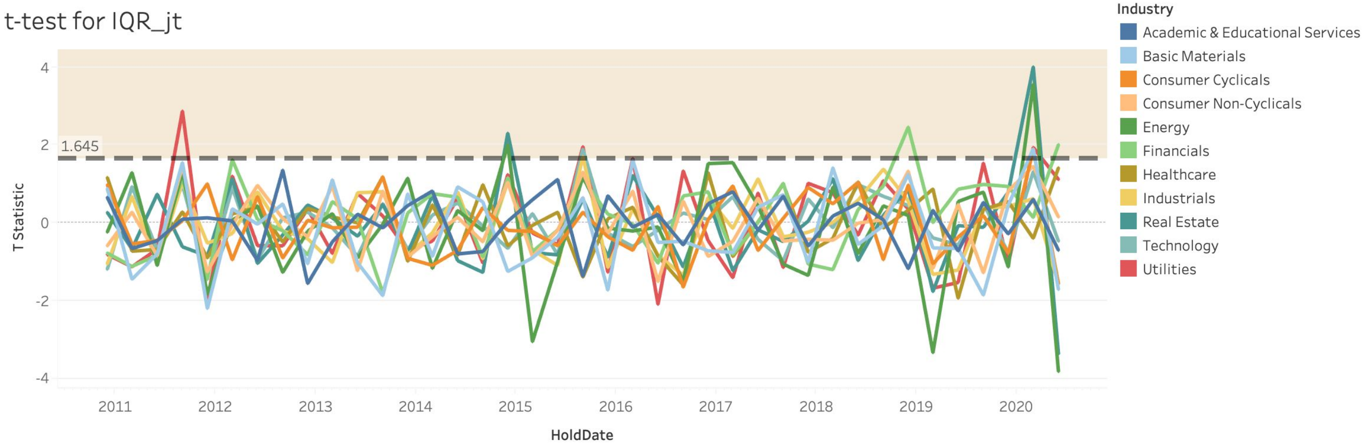
4.1 Trending Topics Definition & Metrics

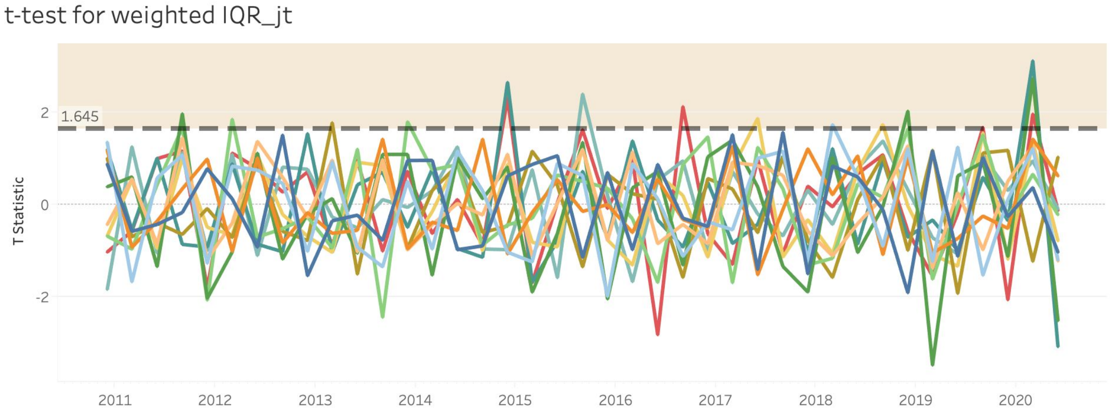


Instruments are aggregated into 11 industries

Spikes in each plot: Investors disagree with one another in a particular industry & timepoint.

4.2 Detection of Significant Trending Topics





Purpose:
To recognize the significant spikes from the plots of the two IQR metrics (previous slide)

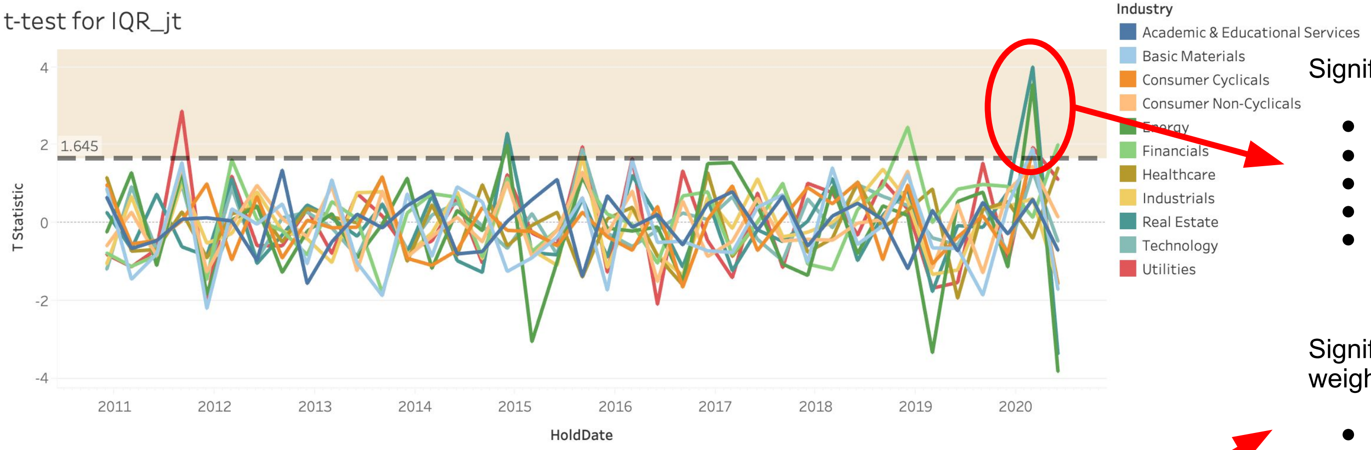
Null hypothesis:
Given a particular industry and a fixed time point t, the mean of IQR_jt for all investors = the mean of IQR_j(t-1) for all investors. (Significance level = 0.05)

Test results:
Reject the null hypothesis when t-statistic is larger than 1.645 (above the horizontal lines of 1.645), we can claim that its corresponding spike in the metric plot is significant, and thus a significant trend is detected.

Reference:
Sean, C. (2020, August 11). Global commercial real estate markets feel impact of COVID-19. https://www.theinvestor.jll/news/world/others/global-commercial-real-estate-markets-feel-impact-of-covid-19/

4.3 Closer Look at 2020 Q1 (COVID-19 related)

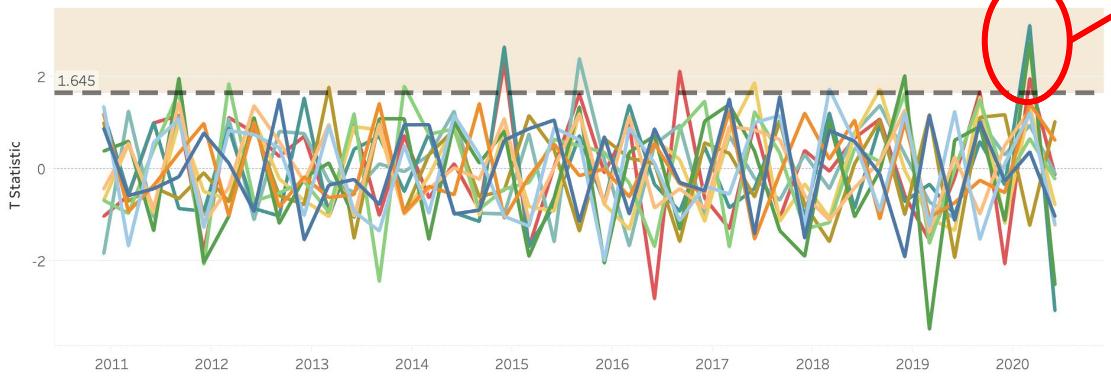
t-test for IQR_jt



Significant trending topics from IQR_jt:

- Real Estate
- Energy
- Utilities
- Basic Materials
- Consumer Cyclicals

t-test for weighted IQR_jt



Significant trending topics from weighted IQR_jt:

- Real Estate
- Energy
- Utilities

Example: Real Estate industry @ 2020 Q1
Spike Interpretation: Investors acted differently at 2020 Q1 and them similarly after 2020 Q1
Possible cause: Some investors sensed the negative effect of COVID-19 (e.g. the widespread lockdowns and travel restrictions) earlier than the others.
Proof by reference: According to data from Jones Lang LaSalle Incorporated (Sean, 2020), the investment in commercial real estate fell almost 30% globally in the first six months of 2020.

4.4 Regression Analysis for Features

- **Data we used**
  - Significant spikes detected from 2016Q2 to 2020Q2, measured by quantile range
  - 7 pairs of (time point, industry)
- **Definition of variables**
  - Independent variables: features of investors (excluding pct of portfolio in each industry)
  - Control variable: industry
  - Dependent variable: deviation of an investor's action from the average market action

For each investor i on time point t, when we are considering about one specific industry D, the dependent variable y_i follows the equation:

$$\Delta_{it} = \frac{1}{n} \sum_{j \in \mathrm{D}} \Delta_{ijt}$$

$$y_i = |\Delta_{it} - \bar{\Delta}_t|$$

n = # of instruments that investor i invested in industry D

4.4 Regression Analysis for Features

Positive correlation

- percentage of portfolio investment in mid-cap instruments
- concentration on top 20% of its instruments
- percentage of employees with the investment style of Emerging Market

The **larger** these features of one investor are, the more likely it is to drive a trending topic.

Negative correlation

- number of instruments
- percentage of employees with the investment style of Arbitrage / Index / Private Equity / Social Awareness

The **smaller** these features of one investor are, the more likely it is to drive a trending topic.

We applied **LASSO** regression to select significant features.
Results are shown in Appendix 1.

4.5 Regression Analysis for Clustering Results

Cluster B (dissimilar to Vanguard) tends to lead the trending topic.

Investors in group B tend to **disagree** with the overall investment market (all the 225 investors) and "not following the crowd". This group contains 54 investors, which we believe need the attention in tracking their real-time changes in the portfolio holdings since they are likely to **give signals** of trending topics.

Results are shown in Appendix 2.

05 Conclusion

5 Conclusion

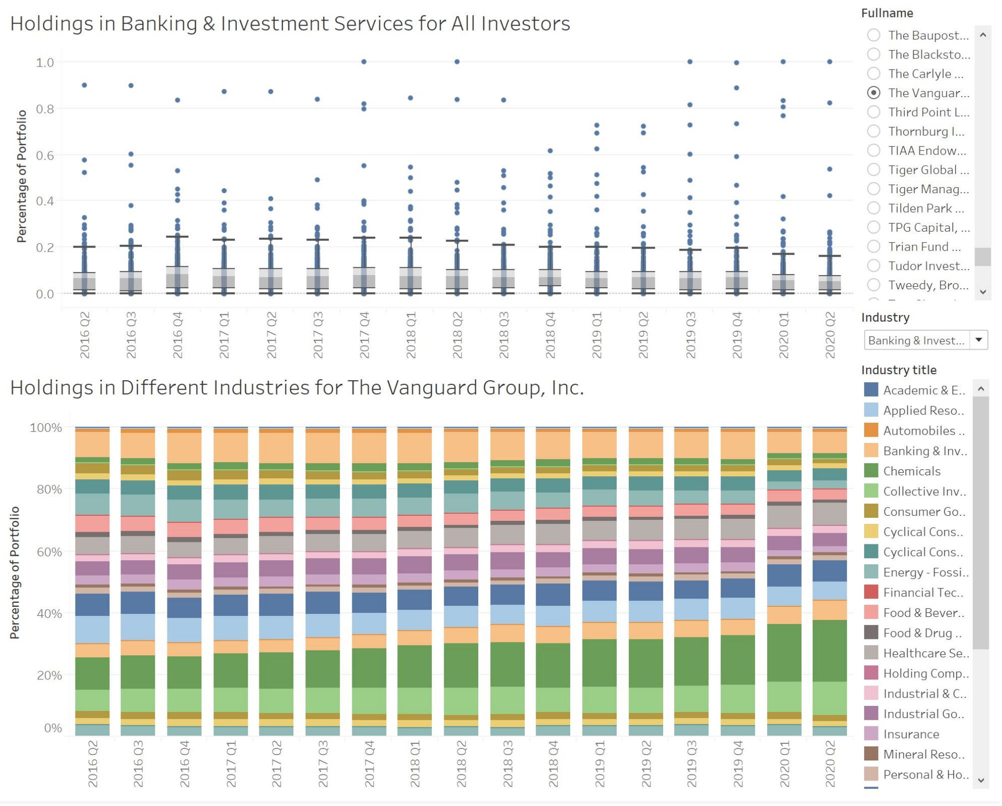
01	Obtained 2 clusters for 225 investors using 7 features	• The cluster with Vanguard:
		  ○ More: Invest in instruments with large & mega market cap in Software and IT industry, total assets, number of instruments, focusing on value
		• The cluster without Vanguard:
		  ○ More: Invest in instruments with small & middle market cap in the Energy industry, focusing on emerging market
		  ○ Less: total assets, number of instruments

02	Detected trending topics in the investment market from 2010-2020	• 2011 Q3 Utilities • 2014 Q4 Energy, Real Estate • 2015 Q3 Industrials, Technology, Utilities • 2016 Q1 Utilities • 2018 Q4 Financials • 2020 Q1 Energy, Basic Materials, Consumer Cyclicals, Utilities, Real Estate • 2020 Q2 Financials
03	Identified features & cluster of an investor that decides whether it tend to drive the trending topics	• High percentage portfolio investment in mid-cap instruments • High concentration on top 20% of its instruments • More employees with the investment style of Emerging Market • Fewer employees with the investment style of Index • Holding fewer positions • The cluster that is dissimilar to Vanguard

06

Dashboard Demo

6 Demo Sample Page



* Full demo is in the video

07 Appendix

Appendix 1 LASSO regression results for features

	s0
(Intercept)	0.0154277889
Turnover_Max	.
PctOfLargeCap	.
PctOfMegaCap	.
PctOfMicroCap	.
PctOfMidCap	0.0115340719
NumberOfInstr_Max	-0.0234034281
TOP20%SHSHLDVALpct	0.0027260264
TotalAssets_Max	.
InvStyle_Arbitrage	-0.0005510147
InvStyle_Convertible Arbitrage	.
InvStyle_Emerging Markets	0.0042905696
InvStyle_GARP	.
InvStyle_Income	.
InvStyle_Index	-0.0068695217
InvStyle_Private Equity	-0.0191547034
InvStyle_Quantitative	.
InvStyle_Sector Rotational	.
InvStyle_Shareholder Activist	.
InvStyle_Short Selling	.
InvStyle_Social Awareness	-0.0031062604
InvStyle_Special Situations	.
InvStyle_Value	.
InvStyle_Venture Capital	.
HierarchicalId_50	0.0020045683
HierarchicalId_51	-0.0015019441
HierarchicalId_53	-0.0004349306
HierarchicalId_55	0.0009340004
HierarchicalId_59	0.0003624638

Positive correlation

Negative correlation

Control variables

Appendix 2 Regression results for clustering results

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.045
Model:	OLS	**Adj. R-squared:**	0.041
Method:	Least Squares	**F-statistic:**	10.80
Date:	Wed, 02 Dec 2020	**Prob (F-statistic):**	9.09e-12
Time:	23:52:04	**Log-Likelihood:**	3235.1
No. Observations:	1369	**AIC:**	-6456.
Df Residuals:	1362	**BIC:**	-6420.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.0013	0.002	0.731	0.465	-0.002	0.005
Cluster	0.0108	0.002	7.085	0.000	0.008	0.014
HierarchicalId_50	0.0055	0.002	2.344	0.019	0.001	0.010
HierarchicalId_51	-0.0011	0.002	-0.445	0.657	-0.006	0.004
HierarchicalId_53	0.0011	0.002	0.458	0.647	-0.003	0.006
HierarchicalId_55	0.0041	0.002	2.050	0.041	0.000	0.008
HierarchicalId_59	0.0032	0.002	1.318	0.188	-0.002	0.008

Positive correlation

Control variables

Omnibus:	2850.817	**Durbin-Watson:**	1.910
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8503711.723
Skew:	16.866	**Prob(JB):**	0.00
Kurtosis:	387.631	**Cond. No.**	7.87