Table 1: The 20 most frequent words presented in the business description of sample firms with different industries in 2013. The 20 most frequently used words are described. The number of the occurrence frequency of each word is the value in parentheses after the word. For example, SANDISK CORP uses word "memory" as 67 times when describing their business and products at the business description part of the 10-K annual report in 2013.

Firm 1: SANDISK CORP (SIC code: 3572)

Business: Flash Memory Storage

Core words: memory(67), product(52), technology(44), storage(36), market(31), device(31), solution(28), NAND(26), flash(24), drive(20), manufacturer(19), design(19), corporation(18), venture(18), card(18), president(17), data(16), wafer(16), cost(15), year(15)

Firm 2: SCHEIN (HENRY) INC (SIC code: 5047)

Business: Healthcare Distribution

Core words: health(102), product(89), care(65), service(62), state(47), customer(47), law(44), practice(41), president(40), business(40), distribution(37), sale(35), drug(33), act(30), vice(30), practitioner(26), officer(25), technology(24), order(23), management(23)

Firm 3: IMPAC MORTGAGE HOLDINGS INC (SIC code: 6162)

Business: Long-term Portfolio

Core words: mortgage(174), loan(159), origination(53), portfolio(49), service(45), estate(34), operation(33), channel(33), Mae(30), interest(30), correspondent(29), lending(27), rate(27), credit(21), security(21), broker(20), sale(20), borrower(19), seller(18), act(17)

Firm 4: TENGASCO INC (SIC code: 1311)

Business: Oil & Gas

Core words: company(200), gas(96), methane(44), production(43), well(37), oil(36), agreement(33), hoactzin(30), pipeline(28), management(27), program(26), interest(25), sale(25), property(24), operation(23), swan(23), report(22), project(21), price(21), field(21)

Firm 5: ACCELRYS INC (SIC code: 7372)

Business: Software Development

Core words: product(82), software(54), customer(46), platform(34), development(32), service(29), data(27), solution(26), process(24), market(24), acquisition(24), research(24), enterprise(23), industry(23), organization(22), quality(19), informatics(19), system(18), management(18), information(17)

Table 2: The clustering result of the sample firms allocated in different SIC and Fama-French classification code ranges. The first case (Case 1) is in an industry related to healthcare, medical equipment, and drugs. TEAM HEALTH HOLDINGS INC and WELLCARE HEALTH PLANS INC are classified as a "Money" and "Others" sectors respectively in terms of the Fama-French classification system. The proposed method clusters the two firms into same industry or sector. The second case (Case 2) is in an industry related to oil, gas, and coal extraction and products. GENESIS ENERGY LP and CROSSTEX ENERGY LP are classified as a sector of "Shops" in terms of the Fama-French classification system. The proposed method clusters the two firms as the same industry with energy-related firms.

Firm name	SIC code	Fama-French 12	Clustered code				
		Classification code					
Case 1 - Healthcare, Medical Equipment, and Dru	igs						
TEAM HEALTH HOLDINGS INC	7363	12 Others	11				
WELLCARE HEALTH PLANS INC	6324	11 Money	11				
SELECT MEDICAL HOLDINGS	8069	10 Hlth	11				
CORP SYMBION INC TN	8011	10 Hlth	11				
LHC GROUP INC	8082	10 Hlth	11				
LIFEPOINT HOSPITALS INC	8062	10 Hlth	11				
TENET HEALTHCARE CORP	8062	10 Hlth	11				
AMN HEALTHCARE SERVICES INC	8090	10 Hlth	11				
HCA HOLDINGS INC	8062	10 Hlth	11				
Case 2 - Oil, Gas, and Coal Extraction and Products							
GENESIS ENERGY LP	5171	9 Shops	4				
CROSSTEX ENERGY LP	5172	9 Shops	4				
HOLLY ENERGY PARTNERS LP	4613	12 Others	4				
GULFPORT ENERGY CORP	1311	4 Energy	4				
CONTINENTAL RESOURCES INC	1311	4 Energy	4				
UNIT CORP	1311	4 Energy	4				
MID CON ENERGY PARTNERS LP	1311	4 Energy	4				
CHEVRON CORP	2911	4 Energy	4				

Table 3: The most overlapped words used in a set of business descriptions textin the same industry. The 9 firms related to the healthcare, medical equipment, and drugs (Case 1 in Table2) uses words "hospital", "billing", "Medicare" and etc. as 9 times to describe their business and product in the 10-K annual report. The word "abuse", "therapy", "HIPPA" is occurred 8 times out of the 9 firms.

Unique words out of 2000 words in the bag-of-words	The number of occurrences	
Case 1 - Healthcare, Medical Equipment, and Drugs		
hospital, billing, physician, Medicaid, productivity, patient, submission, reimbursement, Medicare, referral	9 times out of 9 documents	
beneficiary, methodology, recruitment, length, abuse, accountability, authorization, accreditation, CM, associate, prohibition, utilization, therapy, transition, employer, sanction, eligibility, safeguard, notification, fraud, worker, HIPPA(Health Insurance Portability and Accountability Act), spending, portability, admission, antikickback, update	8 times out of 9 documents	
Case 2 - Oil, Gas, and Coal Extraction and Products		
crude, commodity, pipeline, hydrocarbon, petroleum, transport	8 times out of 8 documents	
proximity, carrier, cleanup, liquid, pollution, barrel, index, discharge, mile, tank, basin, emergency, exploration, drilling, commerce, injection, FERC(Federal Energy Regulatory Commission), shale, formation, greenhouse, dioxide, emission, gathering, fuel	7 times out of 8 documents	

 ${\it Table 4: Within and Across industry \ variations \ by \ different \ industry \ classification \ systems.}$

Industry classification systems	# of	Weighed	Weighed	Weighed		
	industries	OI/asset	OI/sales	OI/asset		
A. Across-Industry Standard Deviations: Firm-size weighted Result						
1. SIC 3-digit industries	245	0.390	0.066	0.741		
2. GICS sub-industries	157	3.455	0.136	0.619		
3. TNIC 300 fixed industries	300	4.493	0.139	0.809		
4. Autoencoder + SKmenas	300	10.819	0.150	0.924		
5. Original Intransitive TNIC	300	0.125	19.081	0.678		
6. Autoencoder + Intransitive TNIC	300	0.115	20.103	0.703		
B. Within-Industry Standard Deviations: Industry-size weighted Result						
1. SIC 3-digit industries	245	0.126	18.296	0.884		
2. GICS sub-industries	157	0.143	13.823	0.803		
3. TNIC 300 fixed industries	300	0.130	10.243	0.980		
4. Autoencoder + SKmenas	300	0.113	5.857	0.856		
5. Original Intransitive TNIC	300	0.124	8.655	1.055		
6. Autoencoder + Intransitive TNIC	300	0.132	4.250	0.996		

Table 5: The summary statistics of cosine similarity measures in terms of the original TNIC and the applied autoencoder of the TNIC system.

Industry classification systems	Mean	Standard	Min	Max
		Deviation		
Original intransitive TNIC	0.073	0.063	0.000	0.904
${\bf Autoencoder+IntransitiveTNIC}$	0.521	0.173	0.011	0.988