RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Use Cases for Text Mining in the Audit

Dr. Kevin C. Moffitt

Use Cases for Text Mining in the Audit

- Source Code Analysis
- Contract Analysis
- Risk Factor Analysis

SOURCE CODE ANALYSIS

Detecting Irregularities in COBOL Code

- In many large banks and financial institutions there are millions of unaudited lines of COBOL Code handling transactions
- I proposed a method for automating the detection of fraudulent or irregular COBOL Code
- Fraudulent code can give deep discounts, eliminate red flags, and give other unauthorized benefits to individual account holders
- My system will automatically identify suspicious COBOL code, and flag COBOL files that are similar to files with known fraud

Fraud Risk Examples

Fees set to Zero for certain commercial superintendents

```
238000          SUPERINTENDENTE COMERCIAL NEGOCIOS UPJ              ILBH92
238100              IF  WCOD-CRGO    EQUAL  '0001612' OR            ILYH13
238200                  WCOD-CRGO    EQUAL  '0001616' OR            ILYH13
238300                  WCOD-CRGO    EQUAL  '0004880'               ILYH13
238400                  MOVE  ZEROS         TO  WTARIFAN            ILBH92
```

Fraud Risk Examples

Social security numbers are hard coded

```
136600 IF CPFTIT-WGEX1S EQUAL '00028353453886' OR '00015123152814'  ILCP63
136700                     OR '00022004618876' OR '00030293008892'  ILCP63
136800                     OR '00018471105845' OR '00027212537861'  ILCP63
136900                     OR '00006077892807' OR '00012963495862'  ILCP63
.
.
.
.
142400           GO  TO RTSC4CTX                                    ILCP34
142500           END-IF.                                            ILCP34
```

Fraud Risk Examples

Specific contractor will not be charged for a specific product

```
028010      IF  PROD EQUAL 1199787 AND                          CAL195B
028020          NUMCTROR EQUAL 000562800031                     CAL195B
028030          MOVE 'B'          TO INIBECO-CAWOPCA            CAL195B
```

Some More Risks

- Tax ID numbers hard coded
- Account numbers, Policy numbers, Credit card numbers hard coded
- Hard coded values in various other fields to provide fixed rates and fees, and tax exemptions
- Employee ID is hard coded
- Username and Password information is set in source code
- Combinations of variable names related to the above data types and COBOL key words

Solution – Part 1

- Use Regular Expressions to automatically identify suspicious patterns in the code

- Match 14 digit CPF which is equivalent to our SSN

```
(?<!\d)\d{14}(?!\d)
```

- Match lines of code with hard coded account numbers

```
.* conta[ -].*?[EQUAL|=].*?\d{3,}.*
```

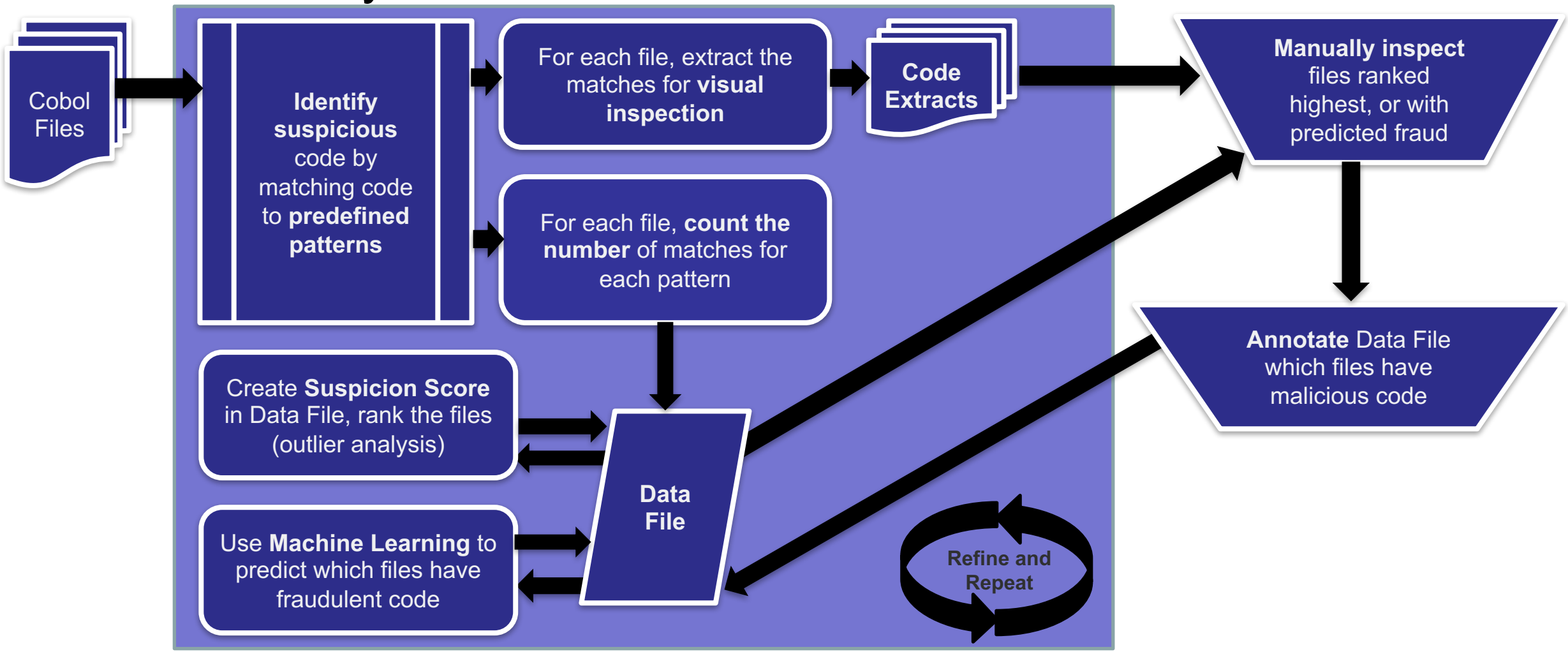
- Match hard coded fee values

```
.* move .*\d{1,}.*?tarifa.*
```

Solution – Part 2

- Create a record of each COBOL file with attributes related to fraud risk
  - Process tens of thousands of COBOL files
  - Create a risk score for each file based on the number of matches to our regular expressions

COBOL Analyzer: Flowchart



	A	
1	001490* * - VERIFICAR AGENCIA CONTA GESTORA SOMENTE *	CAL132F
2	007400 03 WCONTA PIC 9(08).	CAL185B
3	007610 03 WK-CONTA-AUX PIC X(005) VALUE '12181'.	CAL190
4	011450 01 WCONTA-B0 PIC 9(12).	CAL169
5	011460 01 FILLER REDEFINES WCONTA-B0.	CAL169
6	011510 01 WCONTA-AD PIC 9(15).	CAL157A
7	011520 01 FILLER REDEFINES WCONTA-AD.	CAL157A
8	035080 MOVE ZEROS TO WCONTA	
9	042000 PERFORM RT-AGCONTA	
10	047030* ROTINA DE MOVIMENTACAO DA CONTA CORRENTE	
11	047050 RT-AGCONTA SECTION.	
12	070700 MOVE CDCTACDN-CAWOPCA TO WCONTA-B0	
13	079420 MOVE NUMCNTRR-CAWTIO TO WCONTA-AD	
14	102540* * - INIBE COBRANCA P/ AD.DEPOSIT C/ CONTA	
15	104750* * CONTA ENCERRADA	
16	108830* * CONTA GESTORA);	
17		
18		
19		
20		
21		
22		
23		
24		
25		

... Susp. Numbers | CPFs | Agencia | **Conta** | Move CNPJ | Funcional

Microsoft Visual Basic for Applications - COBOL_Analyzer_v4.4b.xlsm - [CodeControl_Module (Code)]

File Edit View Insert Format Debug Run Tools Add-Ins Window Help

Ln 1, Col 1

(General) findCodePatterns

```
Public Sub findCodePatterns()
    ' This is the main sub that executes the code. The button is linked to this sub
    Dim ws_sheet As String: ws_sheet = "Control"

    ' Delete the previous results' sheets
    ' ===================================
    ' In the following logic, everything that isn't the control
    ' sheet is a result sheet.
    Application.DisplayAlerts = False       ' Disables the "Are you sure?" prompt
    On Error Resume Next
        Worksheets(ws_sheet).Move before:=Worksheets(1)
        For m = 2 To Worksheets.Count
            Worksheets(2).Delete
        Next m
    Application.DisplayAlerts = True        ' Re-enables the application alerts

    ' Open the file specified in the Index!Name Range
    ' ===============================================
    ' A name was given to a cell, letting us to call its value
    ' without the absolute address
    On Error GoTo FileErrorHandle              ' Goto FileError handle if the file do
        Dim filename As String: filename = Worksheets(ws_sheet).Range("_file").Value
        fileListing = ImportTextFile(filename)
    On Error GoTo 0 ' Disable the FileError exception handle

    ' Execute the program
    ' ===================
    'Dim startTime: startTime = Now
    ifStatements = getNestedIfs(fileListing)
    'Dim endTime: endTime = Now
    'MsgBox (Format(endTime - startTime, "hh:mm:ss"))

    suspNumbersStatements = getSuspNumbersMatches(fileListing)
    cpfStatements = getCpfMatches(fileListing)
    agenciaStatements = getAgenciaMatches(fileListing)
    contaStatements = getContaMatches(fileListing)
    moveCnpjStatements = getMoveCnpjMatches(fileListing)
```

Example Data File

FileName	SuspicionScore	KnownFraud	numSuspiciousNumbers	WCOD	MOVE_Tarifa	Funcional	Apolices	Agencia	Password_UserName	Susp_Numbers
COBOL_FILE1.txt	0.636		28	134	139	33	9	64	93	37
COBOL_FILE2.txt	0.866	1	29	26	10	84	57	50	70	68
COBOL_FILE3.txt	0.215		30	112	118	72	52	142	37	106
COBOL_FILE4.txt	0.669		31	57	55	107	93	118	26	118
COBOL_FILE5.txt	0.247		32	76	137	115	23	52	30	14
COBOL_FILE6.txt	0.770		33	131	38	70	31	41	84	47
COBOL_FILE7.txt	0.007		34	91	29	142	150	49	150	77
COBOL_FILE8.txt	0.988		35	128	73	10	139	11	106	44
COBOL_FILE9.txt	0.013		36	37	44	133	72	40	111	25
COBOL_FILE10.txt	0.068		37	32	6	103	48	91	145	113
COBOL_FILE11.txt	0.483		38	135	25	84	56	55	14	62
COBOL_FILE12.txt	0.676	1	39	6	22	141	23	57	146	76
COBOL_FILE13.txt	0.794	1	40	49	128	131	101	1	26	111
COBOL_FILE14.txt	0.745		41	93	92	84	108	118	91	133
COBOL_FILE15.txt	0.067		42	112	49	143	126	81	79	6
COBOL_FILE16.txt	0.120		43	129	78	131	107	69	11	80
COBOL_FILE17.txt	0.260		44	73	25	137	74	40	40	122
COBOL_FILE18.txt	0.968		45	118	27	44	91	81	140	90
COBOL_FILE19.txt	0.263		46	48	25	65	74	7	78	11
COBOL_FILE20.txt	0.819		47	150	7	131	142	118	53	47
COBOL_FILE21.txt	0.892	1	48	62	124	146	56	73	37	81

Conclusion

- Increase the efficiency and effectiveness of source code audits
  - Create a risk score for every COBOL file
  - Direct attention of internal auditors to highly suspicious files
  - Within those files display the suspicious lines of code
- This work generalizes to other organizations' COBOL files

CONTRACT ANALYSIS

Contract Analysis

- In this project, we propose a framework that utilizes text mining techniques to audit the whole population of contracts.

Framework

1. Determine what contracts are similar and group them together
   - Proposed method: TF-IDF score and cosine similarity
     - They are one of the most commonly used algorithms for determining the similarity between documents (Pang-Ning et al., 2006).
   - Hierarchical clustering
2. Extract audit related variables from contracts
   - 2 types of variables
     - Values (e.g. $100,000; 10/11/2011; NJAZ003467)
     - Text (e.g. "Vendor is responsible for shipping and return cost")
   - Proposed method: Regular Expression
   - Perform audit procedures on collected variables
     - Match other records for completeness, valuation, and accuracy
     - Data analytic tools
3. Identify the contract template and perform anomaly checks at the sentence level

Methodology: TF-IDF score and Cosine Similarity

Term Frequency(TF) = (Number of times term *t* appears in a document (w_t)) / (Total number of terms in the document (W_t))

Inverse Document Frequency(IDF) = log(Total number of documents (N) / Number of documents with term *t* in it (n_t))

$$\mathrm{TF}_t = \frac{w_t}{W_t} \qquad IDF_t = \log\frac{N}{n_t}$$

TF-IDF = TF * IDF

Methodology: TF-IDF score and Cosine Similarity

- The similarity between any two contracts (i, j) can be calculated as:

$$\mathrm{Sim}_{i,j} = \cos(\vec{t_i}, \vec{t_j}) = \frac{\vec{t_i} \cdot \vec{t_j}}{|\vec{t_i}| \times |\vec{t_j}|} = \frac{\sum_a (d_{ai} \cdot d_{aj})}{\sqrt{\sum_a (d_{ai})^2} \times \sqrt{\sum_a (d_{aj})^2}}$$

- The cosine of the angle between vector $\vec{t_i}$ and vector $\vec{t_j}$ represents the degree of similarity between documents *i* and *j*.

Methodology: Regular Expression

- A regular expression is a sequence of characters that define a search pattern
  - E.g. \$[\d,]* will match monetary value like “$54, 323, 266” or “$54323266” or “$32”

- Using regular expression, we can extract variables identified by the auditor from the similar contracts
  - [[A-Z]{1}\d{1,2}[A-Z]{2}\d{8}]? will match policy number of “A4DB12012013”, “A11DB12012013”, “M0DB12012013”, etc.

Samples

- 10 reinsurance contracts from KPMG Cayman Islands
  - Every contracts has around 29,530 words in 64 pages
  - Most contents and provisions are identical
  - PDF format
  - 3 additional contracts
    - unrelated contract
    - reinsurance agreement example from mbia.com
    - sample reinsurance agreement from the SEC website
- PDFs are converted to text files
  - Error Issues
    - “11/11/2000” is recognized as “11/112000”
    - “indemnitee” is recognized as “in-dem n itee”

Preprocessing

- The standard steps of text retrieval described by Baeza-Yates and Ribeiro-Neto (1999) are used to preprocess the samples

- Remove Stopwords
  - E.g. ['i', 'me', 'my', 'ourselves', 'you', 'your', 'yours', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', ', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very‘, ‘will', 'just', 'don', 'should', 'now']
  - The list is retrieved from Glasgow Information Retrieval Group

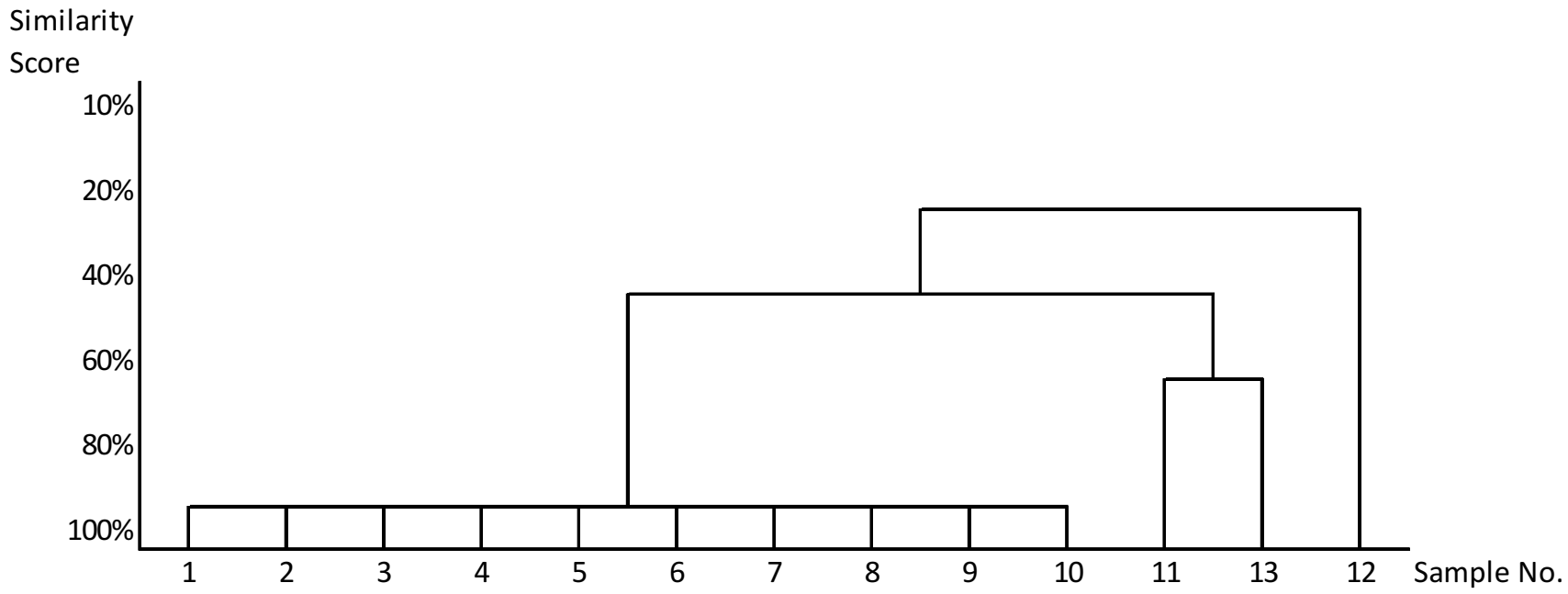
- Stem words
  - E.g. transfer “run, ran, running” into “run”
  - Reduce computation complexity

Cosine Similarity Score

Cosine Similarity Score Between 13 Contracts

Contract No.	1	2	3	4	5	6	7	8	9	10	11	12	13
1	100%	99.95%	99.95%	99.95%	99.85%	99.77%	99.76%	99.77%	99.77%	99.76%	43.53%	12.45%	18.07%
2		100%	99.95%	99.95%	99.86%	99.77%	99.76%	99.77%	99.77%	99.76%	43.54%	12.45%	18.06%
3			100%	99.95%	99.85%	99.77%	99.76%	99.77%	99.77%	99.76%	43.43%	12.42%	18.02%
4				100%	99.85%	99.78%	99.76%	99.77%	99.77%	99.77%	43.40%	12.41%	18.02%
5					100%	99.67%	99.71%	99.70%	99.70%	99.68%	43.40%	12.43%	17.99%
6						100%	99.91%	99.91%	99.91%	99.91%	43.46%	12.34%	18.03%
7							100%	99.95%	99.95%	99.93%	43.49%	12.35%	18.05%
8								100%	99.96%	99.93%	43.54%	12.37%	18.05%
9									100%	99.94%	43.44%	12.36%	18.02%
10										100%	43.47%	12.37%	18.06%
11											100%	27.82%	59.92%
12												100%	24.08%
13													100%

Hierarchical Clustering



Extracted Variables

Variables of Contracts																
		Period of Insurance		Limit of Liability								Retention			Retroactive date	
					Professional Liability		General Liability									
Contract Number	Policy Number	From	To	Policy Aggregate	Each claim	Each Location Aggregate	Each claim	Each Location Aggregate	Products Completed Operations Aggregate	Personal and Advertising Injury Limit	Damage To Premises Rented to You Limit	Professional Liability	General Liability	Gross premium USD$	Professional Liability	General Liability
1	A4DB12012013	12/1/2013	12/1/2014	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	11/1/2000	11/2/2000
2	M0DB12012013	12/1/2013	12/1/2014	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	5/1/2004	5/1/2004
3	A11DB12012013	12/1/2013	12/1/2014	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	11/1/2000	11/1/2000
4	A12DB12012013	12/1/2013	12/1/2014	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	11/1/2000	11/1/2000
5	A17DB12012013	12/1/2013	12/1/2014	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	6/1/2001	6/1/2001
6	NA	12/1/2013	12/1/2014	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	11/112000	11/1/2000
7	NA	12/1/2013	12/1/2014	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	5/1/2004	5/1/2004
8	NA	NA	NA	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	11/1/2000	11/1/2000
9	NA	NA	NA	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$100,000	$0	$0	NA	11/1/2000	11/1/2000

Benefits to the Audit

- Analyze 100% of the population of contracts
- Increase the efficiency and effectiveness of the audit
- Bypass materiality constraints
- Detect outliers and anomalies

RISK FACTOR ANALYSIS

Risk Factor Analysis

- Since 2006 the SEC requires that annual reports include the disclosure of material risks

- Our project seeks to identify risk profiles for companies based on disclosed risks.
  - Compare companies to themselves over time
  - Compare companies to companies in the same industry
  - Enhance the planning stage of the audit

Data Sample

- 1. Download 10-Ks for the retail industry (sic starting with 52-59), and extract item 1a
- 2. Split item 1a into risk factors and extract each into a separate text file (we call them risk factor files)
- 3. Because of the long computation time, we use risk factor files of three companies (Walmart, Target, and Home Depot) for a pilot study

Research Method

- Approach 1: based on Financial Times lexicon
- Financial Times lexicon (http://lexicon.ft.com/) : provides definitions of 12629 words and phrases selected by Financial Times editors.

Approach 1

- Download the Financial Times lexicon into a csv with terms and their definitions (12,629 unique terms with definitions)
- Extract the first sentence (if it has more than 10 words) or the first two sentences of each risk factor file
- Calculate the similarity score (using tfidf to weigh each non stop word, and then calculate the cosine similarity of two texts) between the first one or two sentences of the risk factor files and the definition of each term-get a similarity matrix containing the similarity score for each pair of term and risk factor
- Delete the terms for which the maximum value of similarity scores is smaller than 0.1, and 4,387 terms remain
- Conduct factor analysis (set minimum eigenvalue as 1) on the 4,387 terms to identify similar risks, and factor rotation (orthogonal varimax) is also conducted

Factor Analysis Result

Walmart (blue) Target (red) Home Depot (orange)

Year	Factor1-Factor4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11	Factor 12	Factor 13	Factor 14	Factor 15	Factor 16	Factor 17	Factor 18	Factor 19	Factor 20	Factor 21	Factor 22	Factor 23	Factor 24	Factor 25	Factor 26	Factor 27	Factor 28	Factor 29	Factor 30	Factor 31	Factor 32	Factor 33	Factor 34	Factor 35
2007			X				X				X	X				X		X			X										
2008			X				X				X	X					X	X			X										
2009			X				X				X	X				X	X	X			X										
2010			X				X				X	X				X	X	X			X										
2011			X				X				X					X	X	X			X										
2012			X				X				X					X	X	X			X						X				
2013			X				X	X								X	X						X				X				
2014			X				X	X								X	X						X				X				
2015			X		X					X		X			X	X		X	X							X	X	X				
2009			X												X			X										X				
2010			X											X	X			X										X				
2011			X											X	X		X	X								X		X				
2012			X											X	X		X	X								X	X	X				
2013			X							X		X			X	X	X	X								X	X					
2014			X							X		X			X	X	X	X	X							X	X	X				
2015			X		X					X		X			X	X		X	X							X	X	X				
2007		X			X			X										X				X			X						X	X
2009		X	X		X			X				X						X		X		X			X	X					X	X
2010		X			X			X										X			X	X			X	X					X	
2011		X			X											X		X			X	X				X					X	
2012		X			X											X		X			X	X			X	X					X	
2013		X			X											X		X			X	X			X	X					X	
2014		X			X											X		X			X	X			X	X					X	
2015		X			X			X							X	X					X	X				X	X				X	

Approach 2

- Extract the first sentence (if it has more than 10 words) or the first two sentences of each risk factor file, and extract all the noun phrases (in the form of NBAR: {<NN.*|JJ>*<NN.*>}, or <NBAR><IN><NBAR>)
- Calculate the tfidf of each noun phrase extracted for each risk factor file (regardless of the stop words), and keep the top two phrases highest tfidf as the important phrases for the file

Approach 2 Result-An example Walmart Result

	# of similar risk terms from previous year	# of new risk terms	list of new risk terms
2008	15	5	computer systems consumer trends market share products transactions
2009	20	0	
2010	20	0	
2011	20	0	
2012	20	0	
2013	20	5	on-going FCPA matter other adverse consequences impediments expansion changes in climate
2014	24	1	retail offerings
2015	22	4	digital retail benefit cost increases in wage