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

1. Identify suspicious code by matching code to predefined patterns
2. For each file, extract the matches for visual inspection
3. For each file, count the number of matches for each pattern
4. Create Suspicion Score in Data File, rank the files (outlier analysis)
5. Use Machine Learning to predict which files have fraudulent code
6. Manually inspect files ranked highest, or with predicted fraud
7. Annotate Data File which files have malicious code
8. Refine and Repeat
VERIFICAR AGÊNCIA CONTA GESTORA SOMENTE

007400 03 WCONTA PIC 9(08).
007610 03 WK-CONTA-AUX PIC X(005) VALUE '12181'.
011450 01 WCONTA-B0 PIC 9(12).
011460 01 FILLER REDEFINES WCONTA-B0.
011510 01 WCONTA-AD PIC 9(15).
011520 01 FILLER REDEFINES WCONTA-AD.

035080 MOVE ZEROS TO WCONTA
042000 PERFORM RT-AGCONTA

047030* ROTINA DE MOVIMENTAÇÃO DA CONTA CORRENTE
047050 RT-AGCONTA SECTION.
070700 MOVE CDCTACDN-CAWOPCA TO WCONTA-B0
079420 MOVE NUMCNTR-CANTIO TO WCONTA-AD

102540* INIBE COBRANÇA P/ AD.DEPOSIT C/ CONTA
104750* CONTA ENCERRADA
108830* CONTA GESTORA);
## Example Data File

<table>
<thead>
<tr>
<th>FileName</th>
<th>SuspicionScore</th>
<th>KnownFraud</th>
<th>numSuspiciousNumbers</th>
<th>WCOD</th>
<th>MOVE_Tarifa</th>
<th>Funcional</th>
<th>Apolices</th>
<th>Agencia</th>
<th>Password_UserName</th>
<th>Susp_Numbers</th>
</tr>
</thead>
</table>
| COBOL_FILE1.txt | 0.636           | 28 | 134 | 139 | 33 | 9 | 64 | 93 | 37
| COBOL_FILE2.txt | 0.866           | 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          | 39 | 6 | 22 | 141 | 23 | 57 | 146 | 76
| COBOL_FILE13.txt | 0.794          | 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          | 48 | 62 | 124 | 146 | 56 | 73 | 37 | 81

Example Data File
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) = \( \log \left( \frac{\text{Total number of documents (N)}}{\text{Number of documents with term } t \text{ in it (n_t)}} \right) \)

\[
\text{TF}_t = \frac{w_t}{W_t} \quad \text{IDF}_t = \log \frac{N}{n_t}
\]

TF-IDF = TF * IDF
Methodology: TF-IDF score and Cosine Similarity

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

\[
\text{Sim}_{i,j} = \cos(t_i, t_j) = \frac{\overrightarrow{t_i} \cdot \overrightarrow{t_j}}{|\overrightarrow{t_i}| \times |\overrightarrow{t_j}|} = \frac{\sum_a (d_{ai} \cdot d_{aj})}{\sqrt{\sum_a (d_{ai})^2} \times \sqrt{\sum_a (d_{aj})^2}}
\]

• The cosine of the angle between vector \(\overrightarrow{t_i}\) and vector \(\overrightarrow{t_j}\) 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 contract 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 Between 13 Contracts

<table>
<thead>
<tr>
<th>Contract No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>99.95</td>
<td>99.95</td>
<td>99.95</td>
<td>99.97</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.76</td>
<td>43.53</td>
<td>12.45</td>
<td>18.07</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>99.95</td>
<td>99.95</td>
<td>99.95</td>
<td>99.97</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.76</td>
<td>43.54</td>
<td>12.45</td>
<td>18.06</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>99.95</td>
<td>99.95</td>
<td>99.95</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.76</td>
<td>43.43</td>
<td>12.42</td>
<td>18.02</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>99.85</td>
<td>99.78</td>
<td>99.76</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.77</td>
<td>99.76</td>
<td>43.40</td>
<td>12.41</td>
<td>18.02</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>99.91</td>
<td>99.91</td>
<td>99.91</td>
<td>99.91</td>
<td>99.91</td>
<td>99.91</td>
<td>99.91</td>
<td>99.91</td>
<td>99.91</td>
<td>43.49</td>
<td>12.35</td>
<td>18.05</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>99.95</td>
<td>99.95</td>
<td>99.95</td>
<td>99.93</td>
<td>99.93</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.93</td>
<td>43.49</td>
<td>12.35</td>
<td>18.05</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>99.96</td>
<td>99.93</td>
<td>99.93</td>
<td>99.93</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.93</td>
<td>43.54</td>
<td>12.37</td>
<td>18.05</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>99.94</td>
<td>43.44</td>
<td>12.36</td>
<td>18.02</td>
</tr>
<tr>
<td>11</td>
<td>100</td>
<td>99.82</td>
<td>99.82</td>
<td>99.82</td>
<td>99.82</td>
<td>99.82</td>
<td>99.82</td>
<td>99.82</td>
<td>99.82</td>
<td>99.82</td>
<td>27.82</td>
<td>59.92</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>100</td>
<td>99.08</td>
<td>99.08</td>
<td>99.08</td>
<td>99.08</td>
<td>99.08</td>
<td>99.08</td>
<td>99.08</td>
<td>99.08</td>
<td>99.08</td>
<td>24.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical Clustering

Similarity Score

100% 10% 20% 40% 60% 80% 100%

Sample No.

1 2 3 4 5 6 7 8 9 10 11 12
<table>
<thead>
<tr>
<th>Contract Number</th>
<th>Policy Number</th>
<th>Period of Insurance</th>
<th>Policy Aggregate</th>
<th>Professional Liability</th>
<th>General Liability</th>
<th>Retention</th>
<th>Gross premium USD$</th>
<th>Retroactive date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A4DB12012013</td>
<td>12/1/2013 - 12/1/2014</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>11/1/2000</td>
</tr>
<tr>
<td>2</td>
<td>M0DB12012013</td>
<td>12/1/2013 - 12/1/2014</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>5/1/2004</td>
</tr>
<tr>
<td>3</td>
<td>A1DB12012013</td>
<td>12/1/2013 - 12/1/2014</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>11/1/2000</td>
</tr>
<tr>
<td>4</td>
<td>A12DB12012013</td>
<td>12/1/2013 - 12/1/2014</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>11/1/2000</td>
</tr>
<tr>
<td>5</td>
<td>A17DB12012013</td>
<td>12/1/2013 - 12/1/2014</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>6/1/2001</td>
</tr>
<tr>
<td>6</td>
<td>NA</td>
<td>12/1/2013 - 12/1/2014</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>11/1/2000</td>
</tr>
<tr>
<td>7</td>
<td>NA</td>
<td>12/1/2013 - 12/1/2014</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>11/1/2000</td>
</tr>
<tr>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>5/1/2004</td>
</tr>
<tr>
<td>9</td>
<td>NA</td>
<td>NA</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$100,000</td>
<td>$0</td>
<td>$0</td>
<td>11/1/2000</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Year</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
<th>Factor 8</th>
<th>Factor 9</th>
<th>Factor 10</th>
<th>Factor 11</th>
<th>Factor 12</th>
<th>Factor 13</th>
<th>Factor 14</th>
<th>Factor 15</th>
<th>Factor 16</th>
<th>Factor 17</th>
<th>Factor 18</th>
<th>Factor 19</th>
<th>Factor 20</th>
<th>Factor 21</th>
<th>Factor 22</th>
<th>Factor 23</th>
<th>Factor 24</th>
<th>Factor 25</th>
<th>Factor 26</th>
<th>Factor 27</th>
<th>Factor 28</th>
<th>Factor 29</th>
<th>Factor 30</th>
<th>Factor 31</th>
<th>Factor 32</th>
<th>Factor 33</th>
<th>Factor 34</th>
<th>Factor 35</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
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 \textbf{NBAR}: \{<NN.*|JJ>*<NN.*>\}, or \textbf{<NBAR><IN><NBAR>} )

• Calculate the \textbf{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
<table>
<thead>
<tr>
<th>Year</th>
<th># of similar risk terms from previous year</th>
<th># of new risk terms</th>
<th>list of new risk terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>15</td>
<td>5</td>
<td>computer systems, consumer trends, market share, products, transactions</td>
</tr>
<tr>
<td>2009</td>
<td>20</td>
<td>0</td>
<td>on-going FCPA matter, other adverse consequences, impediments, expansion, changes in climate</td>
</tr>
<tr>
<td>2010</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>20</td>
<td>5</td>
<td>retail offerings, digital retail, benefit, cost, increases in wage</td>
</tr>
<tr>
<td>2014</td>
<td>24</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>22</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>