Special Topics in Accounting – Decoding of Textual Corporate Communications

Fall 2017

Professor Kevin Moffitt
100 Rockefeller RD, 4153
Piscataway, NJ 08854-8054
kc9m7@scarletmail.rutgers.edu
Office hours Newark: CARLAB 9th Floor 1PM-2PM on the day of class

Course Description
In Decoding of Textual Corporate Communications, you will learn the literature and techniques needed to incorporate unstructured text into your research.

Academic Integrity
Academic dishonesty will not be tolerated in class and could result in a failing grade. Please read the complete Rutgers policy on academic integrity at:
http://academicintegrity.rutgers.edu/

Harassment and Discrimination
Harassment and discrimination of any sort will not be tolerated in this class. If you feel you are a victim of harassment or discrimination, or if you are a witness to such behavior, please contact the instructor immediately. Please see the complete Rutgers policy on harassment and discrimination at:
http://policies.rutgers.edu/PDF/Section60/60.1.12-current.pdf

Point Allocation
Attendance (140)
Participation (260)
Final Research Proposal (200)

Attendance and Participation Assignments
Attendance will be taken every day. You will not earn these points if you are absent. You will not lose points for University excused absences that are also reported through the University systems at https://sims.rutgers.edu/ssra/. If you do miss a day of class due to an excused absence you will still be responsible for turning in the required assignments by the following Monday at noon.

Students are required to read all papers assigned for each class and prepare a 40-minute PowerPoint presentation on one of the papers for each class (these will be assigned in class). PowerPoint files must be turned in before class starts. From time to time coding assignment will also be given and will count toward the participation grade.
Final Research Proposal

A minimum 2000-word research proposal (excluding the required references) will be turned in at the end of the semester. More details will be given during the sixth week of class.

Required Book

Make your own Neural Network E-Book (Rashid, 2016)

Course Schedule

The course schedule is subject to change according to our progress in class.

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Papers</th>
<th>Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overview of Text Mining Research</td>
<td></td>
<td>Install Anaconda environment. Create Python 3.2 instance with NLTK</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Python basics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Program control flow</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Read files</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Write to .txt files</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Write to .csv files</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Accessing Data sources</td>
</tr>
<tr>
<td>2</td>
<td>Dictionaries</td>
<td>When is a Liability not a Liability. (Loughran and McDonald, 2011)</td>
<td>Dictionary building</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Words</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Phrases</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower case</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stemming and lemmatization</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Removing stop words</td>
</tr>
<tr>
<td>3</td>
<td>Readability</td>
<td>Annual report readability, current earnings, and earnings persistence (Feng Li, 2008)</td>
<td>Sentence length</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Word length</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Syllable counting</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FOG, Flesch-Kincaid, ARI</td>
</tr>
<tr>
<td>4</td>
<td>Regular Expressions</td>
<td>Automated contract analysis in auditing. (Yan, Moffitt, Titera, 2017...On Blackboard)</td>
<td>Regular Expressions Part 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RegEx Basics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RegEx in Python</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Regular expression exercise</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
</tbody>
</table>
|   | Textual risk disclosures and investors’ risk perceptions. (Kravet, Todd, Muslu, 2013) | Use regular expressions to count and write to CSV  
Use regular expressions to parse a document  
Use regular expressions to count phrases, permutations of words, etc...  
RegEx Parsing an XBRL document |
| 5 | Regular Expressions | Regular Expressions Part 2  
RegEx Advanced Topics  
RegEx Parsing HTML 10-K  
RegEx Implementing Dictionary |
| 6 | POS Tagging | POS Tagging Techniques  
Penn Treebank  
Expressivity  
Parsing Sentences by Tense  
Passive Voice  
Noun Phrase Identification |
| 7 | Document Similarity | TF-IDF  
Cosine Similarity |
| 8 | MYSQL | Writing results to MySQL database  
Install MySQL  
Connect to MySQL  
Build conceptual model  
Query using MySQL  
Query using Python |
| 9 | Topic Modeling | Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures (Bao and Datta, 2014)  
The information content of mandatory risk factor disclosures in corporate filings. (Campbell et al., 2014) | Topic modeling  
LDA  
Dictionary-based  
Search engine based |
|---|---|---|---|
| 10 | Naïve Bayes | The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. (Feng Li, 2010)  
Accounting variables, deception, and a bag of words: assessing the tools of fraud detection. (Purda and Skillicorn, 2015) | Bayesian Probability  
Naïve Bayes Algorithm |
| 11 | Neural Network | Make your own Neural Network E-Book (Rashid, 2016) | Neural networks Part 1  
Neural network basics |
| 12 | Neural Network | Make your own Neural Network E-Book (Rashid, 2016) | Neural networks Part 2  
Convolutional networks  
Recursive networks  
Sentence classification  
Tensor Flow |
| 13 | Sentiment Analysis | Evaluating sentiment in financial news articles. (Schumaker et al., 2012)  
Processing fluency and investors’ reactions to disclosure readability. (Rennekamp, 2012). | Word2Vec  
Doc2Vec  
Dictionaries |
| 14 | Student Research Presentations |  |