Financial Entity Identification and Information Integration (FEIII) 2017 Year Two Challenge

Introduction

The Financial Entity Identification and Information Integration (FEIII) challenge series aims to provide interesting datasets to researchers at the intersection of finance and big data. These datasets have been partially curated and have sufficient data for exploration, but they may be noisy and incomplete. Each year we provide a specific SCORED evaluation task. We also provide some ideas for UNSCORED tasks and we encourage participants to define tasks that match their research interests.

In Year One, we posed an IDENTIFIER ALIGNMENT challenge: Given four databases of financial entities from four different sources, participants needed to find the entities in common across the databases. See the report.

For Year Two, we are continuing with the theme of IDENTIFYING and understanding RELATIONSHIPS among financial entities and the roles that they play in financial contracts. The Year Two dataset was created from 10-K and 10-Q filings retrieved from the Securities and Exchange Commission (SEC) EDGAR website and additional resources from the National Information Center (NIC) of the Federal Financial Institution Examination Council. The task is a ranked classification task, to identify relevant and interesting sentences in the filings that provide evidence for a specific relationship, described by a role keyword, between the filing financial entity and a mentioned financial entity.

Dataset Summary

The dataset is drawn from the SEC 10-K and 10-Q filings of 27 holding companies with assets exceeding US$10Billion (*HC>$10B financial entities*) that have provided a resolution plan (Living Will). Using custom extractors developed by the University of Maryland team to leverage IBM SystemT tools, we have extracted the following **triples** from these filings:

- Financial Entity Mention: The text string containing the name of the **Mentioned** financial entity.
- Role keyword(s): The relationship between the **Mentioned** financial entity and the **Filing** financial entity, described by the role.
- The contextual text: Three sentences that surround the mentioned financial entity and the role keyword(s). The contextual text may provide evidence to support this relationship and it may also provide additional interesting financial information.

Scored Challenge Task

Participants will be provided with three datasets of triples: (1) A labeled Training dataset for training. (2) A Working dataset of additional unlabeled triples. (3) A Test dataset for evaluation of the scored task. [Details of the datasets are available in an Appendix of this document.]

Task Description

Participants will take as input the Test dataset (spreadsheet or csv) containing the triples and will return a file with an **additional column - score**. The score is the ranking score of that triple. Scores should be numbers between 0 and 1 inclusive, where 1 indicates the highest level of confidence. We interpret confidence broadly as follows: the sentences in the triple are relevant and interesting with respect to the filing financial entity, and/or the sentences provide evidence for the specific relationship, described by the role keyword, between the filing financial entity and the mentioned financial entity. [Details in a later section of this document.]

Only triples with scores will be evaluated. Participants should optionally group/sort the output by role / descending score.

Note: We will use the score to produce a ranked list for the evaluation of the task result; we request a score rather than a ranking for our convenience.

Scoring

We have asked domain experts to label each triple in the Test dataset according to whether it is highly relevant, relevant, neutral, or irrelevant. Details of the labeling for the Training and Test datasets are provided in an Appendix of this document. The result ranked list submitted by the participant, pRL, will be evaluated against the expert labels. We will compute the normalized discounted cumulative gain NDCG, a measure frequently used by search engines, as follows: The pRL gets one point of gain for each relevant triple and two points for each highly relevant triple, discounted by the logarithm of the position in pRL.

$$\mathrm{DCG}_p = \mathrm{rel}_1 + \sum_{i=2}^{p} \frac{\mathrm{rel}_i}{\log_2(i)}$$

Note: p is the length of pRL; rel_i is the relevance (gain) value at rank i.

This discounted cumulative gain is normalized by the DCG of the perfect ranking based on the expert labels, i.e., one that places all relevant and highly relevant triples before all neutral and irrelevant triples. This yields a score between 0 and 1.

We will divide the pRL by role keyword, order it by score, and compute the NDCG score for each role. These scores will be averaged to give the overall NDCG score for the pRL.

You might ask why we are not evaluating this task using the F-measure. This task is actually equivalent to a classification/ranking task where your system outputs a confidence for each item, rather than a single decision threshold. This allows scoring with more robust metrics including NDCG.

Additional Tasks

We can consider the dataset and the scored challenge tasks as a building block of a more sophisticated platform for financial big data analytics. We briefly describe two additional tasks as follows:

- Apply NLP and machine learning approaches to further embellish the role/relationship R between filing entity X and mentioned entity Y. Of particular interest are legal proceedings and financial settlements. The embellishment may include the directionality of the role/relationship R, e.g., the filing entity X is the insurer for mentioned entity Y, or further details about the relationship, e.g., the mentioned entity Y is the insurer for a specific asset or asset class.

- Create a (labeled) (directed) (weighted) graph between filing entity X and all mentioned entities. Determine the centrality of financial entity X, reciprocity of the graph, the exposure of filing entity X to some mentioned entity Y, etc. We will provide a file that will group mentioned financial entities that appear to be affiliates or subsidiaries of a (potential) parent financial entity. [To be released January 10, 2017.]

Timeline

Task announced:	December 2, 2016
Task details released:	December 16, 2016
Labeled Training dataset released:	December 23, 2016
Abstract for scored task:	February 24, 2017
Paper for non-scored-tasks:	February 24, 2017
Ranked results to be submitted by participants:	March 13-17, 2017
Scores released to participants:	April 2, 2017
Camera-ready papers for all tasks:	April 14, 2017
DSMM Workshop:	May 14, 2017

Advisory Committee

Doug Burdick, IBM
H.V. Jagadish, University of Michigan
Raghav Madhavan, UBS

Organizing Committee

Mark Flood, Office of Financial Research
John Grant, University of Maryland
Joe Langsam, University of Maryland and MIT
Louiqa Raschid, University of Maryland
Ian Soboroff, National Institutes of Standards and Technology
Elena Zotkina, University of Maryland

Domain Experts

Don Berndt
Mohamed Boraie
Mark Flood
Eric Fu
Joe Langsam
Chetan Saran Mehra
Joseph Proctor
Andrew Staedeli
Suzanne Stahl

Acknowledgements

This project was partially supported by NSF award CNS 1305368, the Office of Financial Research / NIST Award 70NANB15H194, and the Smith School of Business.

Appendix 1: Description of the Challenge Data

Organization of the Challenge Data

- Folder with the **Training** dataset. To be released December 23, 2016.
- File with the **Working** dataset. Released December 16, 2016.
- File with the **Test** dataset. To be released March 13-17, 2017.
- **Mentioned Financial Entities:** A file that groups together extracted financial entities that appear to be affiliates or subsidiaries of a (potential) parent financial entity. To be released January 9, 2017. Note: This will be relevant to a visual analytics task on the working dataset but it is not directly relevant to the scored ranked classification task.

The following files were released on December 16, 2016.

- **Role Keywords**: A file with the list of ten keywords that were used to represent relationships and to label triples. It also includes additional keywords that may occur in the sentences.
- **Seed Entities:** A file with data about financial entities from NIC. It contains the NIC name, RSSD ID identifier, and SEC CIK identifier for the 27 seed entities.
- **NIC Financial Entities**: Additional information from NIC for a set of financial entities related to the seed financial entities.
- **NIC Parent Offspring Relationships:** A file representing the parent financial entity for each financial entity.

Statistics of the datasets

We obtained 162 10-K filings and 486 10-Q filings from 27 seed entities filed between 2011 and 2016. A Financial_Entity/Role/3_Sentence extractor created 10000+ triples. Some of the filings produced empty results and some triples were discarded when the mentioned financial entity was incorrectly extracted.

The triples were randomly split into three groups: Training, Working and Test.

Training dataset: A collection of triples that were labeled by a group of domain experts. We provide the rating and comments of each expert. In some cases, where triples were rated by multiple experts, there was no consensus among experts. We intentionally provide these multiple expert ratings so that participants may use them in different ways, e.g., use the highest rating or average across the experts. To be released December 23, 2016.

Working dataset: Unlabeled triples (including triples from the Training dataset) that participants can use for additional tasks such as network / visual analytics. Participants may also wish to label additional triples to enlarge the training set. Released December 16, 2016.

Test dataset: A set of triples that were labeled by a similar group of domain experts. This set will be used to evaluate the results for the scored task. The test set will not include triples that were rated inconsistently by multiple experts. To be released March 13-17, 2017; participants will have to collect the dataset and return results within 24 hours.

A count of filings and triples in each group are as follows:

Dataset	Count of filings	Count of triples
Training	25	975
Test	26	1011
Working	536 + 25	8622 + 975
Total	587	10608

Format of the datasets

Each of the three datasets (Training, Test, Working) has the following format:

- DOCUMENT_TYPE: The type of document (10-K or 10-Q).
- FILER_NAME: The name of the financial entity that made this filing.
- FILER_CIK: The CIK for the financial entity.
- FILING_INTERVAL: The fiscal year corresponding to the filing.
- FILING_DATE: The date of the filing.
- MENTIONED_FINANCIAL_ENTITY: The string, identified by a Named Entity Recognition (NER) extractor, that represents the mention of a financial entity.
- PP_RSSD_ID: The RSSD_ID value from NIC of the potential parent financial entity of the mentioned financial entity. These values to be released January 10, 2017.
- ROLE: A keyword describing the relationship between the filing financial entity and the mentioned financial entity. The financial entity mention is associated with a Role keyword by a Role_Participant extractor.
- THREE_SENTENCES: Three sentences surrounding the Role keyword and the mentioned financial entity.

The Training dataset is made available as separate files, one for each filing, for ease of display. It has the following additional columns for one or more experts:

- RATING: the rating from expert(s).
- COMMENT: comment from expert(s).

Note: There appear to be duplicate triples in the datasets; they are a result of a financial entity being mentioned MULTIPLE times in the proximity of a role keyword. These duplicates can be ignored for the scored ranking task. The duplicates are relevant to any additional tasks related to text extraction, etc.

Appendix 2: Role Keywords, Relationships and Labeling

Role Keywords

The following TEN role keywords and definitions are adapted from the NYS Society of CPA glossary and from Investopedia.

Affiliate Agent Counterparty Guarantor Insurer
Issuer Seller Servicer Trustee Underwriter

Note: The term **PERSON** includes **LEGAL PERSON** which can be a **COMPANY**.

AFFILIATE(D) COMPANY: Company, or other organization related through common ownership, common control of management or owners, or through some other control mechanism, such as a long-term lease.
SUBSIDIARY COMPANY: A company that satisfies the criteria that more than 50% of the voting shares are owned by another company; the latter company is the **PARENT COMPANY**.

AGENT: A person that is designated by another person (the principal) to act on behalf of the principal in specified activities.
COUNTERPARTY: Given a person or party of interest (in a trade), the party faces a counterparty; there can be more than one counterparty.
GUARANTOR: A legal arrangement involving a promise by a person (guarantor) to perform the obligations of a second person (or many persons), in the event that the latter person fails to meet their obligations.
INSURER: A person who, through a contractual agreement, undertakes to compensate specified losses, liability, or damages incurred by the person of interest.

ISSUER: This term refers to an issuer of securities which are (1) registered under Section 12 of the Securities Exchange Act of 1934, or (2) required to file reports under Section 15(d) of that Act, or (3) has filed a registration statement with the SEC.

SELLER: Exchanges a good or service in exchange for a payment.
SERVICER: Typically a person that collects payments from one party and makes payments to another party.

TRUST: An (ancient) legal practice where one person (**GRANTOR**) transfers the legal title to an **ASSET**, also called the principal or corpus or property, to another person (**TRUSTEE**), with specific instructions about how the **ASSET** is to be managed and disposed.
TRUSTEE: Person who is given legal title to, and management authority over, the property placed in a **TRUST**.

UNDERWRITER: A person that assumed the risk of purchasing securities from the issuing entity and reselling them to the public, either directly or through dealers.

The role terms are often modified as follows:
administrative agent
calculation agent
co-documentation agents
co-syndication agents
life insurance agent
placement agent
remarketing agent
transfer agent
bankruptcy trustee
debenture trustee
guarantee trustee
property trustee
securitization trustee
successor trustee

When deciding whether a role is being described, we can ask these following questions:

- TRUSTEE:
  - Who is the Person (LEGAL Person) playing the role of TRUSTEE?
  - What is the associated TRUST?
  - What are the ASSETs and instructions of the associated TRUST?
  - What action is being taken by the TRUSTEE or why is the trustee being mentioned in this statement?

- AFFILIATE COMPANY:
  - Who is the PARENT COMPANY?
  - What is the business purpose of the AFFILIATE COMPANY?
  - Why is the AFFILIATE COMPANY being mentioned in this statement?

- ISSUER:
  - What was issued?
  - Why is the ISSUER being mentioned?

- GUARANTOR:
  - What is being guaranteed?
  - What are the terms of the guarantee?
  - Who is the GUARANTOR?

Expert Ratings

Highly Relevant Sentences: One type of highly relevant sentences will identify potential sources of significant (large) expenses and/or significant business opportunities. Examples of the source of the expenses or opportunities include litigation, spin-offs, acquisitions, etc. Most of these sentences describe a change from the status quo or current situation. Another type of highly relevant sentence will identify corporate character, e.g., the compensation of senior executives or commentary about business activities.

Relevant Sentences: One type of relevant sentences will identify existing assets, liabilities, revenues, or expenses. They may be very specific, e.g., interest rate expenses. Another type of relevant sentences will also identify the size and nature of current business activities, e.g., retail division, underwriting, investment banking, etc.

Note:It may be difficult to differentiate highly relevant and relevant sentences.

Neutral Sentences: These sentences may describe the type of business activity, the location of some business entity or activity. They are informative sentences but convey less information value compared to the highly relevant or relevant sentences.

Irrelevant: This is boilerplate text that is not informative. In some cases, the extracted sentences may be irrelevant to the filing financial entity or the mentioned entity or the role.

Guide for labeling:

- Provide a rating of highly relevant, relevant, neutral or irrelevant in the RATING column.
- If you cannot provide a rating for the set of three sentences then provide a rating for each sentence or make a note in COMMENT.
- It may be difficult to differentiate between highly relevant and relevant; make a note in COMMENT.
- The sentences may include additional information about the role relationship, e.g., “transfer agent” rather than “agent” or “trustee for some specific asse”. Make a note in COMMENT.
- The sentence may include the directionality of the role/relationship R, e.g., the filing entity X is the insurer for mentioned entity Y.
- The sentences may be unrelated to the role or mentioned financial entity; this may be due to an error of our extraction process or for other reasons. Make a note in COMMENT.