Team Assignment
Reverse Engineering Thomson Reuters News Analytics
Reference: Fact Sheet (Cached)

We saw in class that Thomson Reuters News Analytics (TRNA) is one of the primary systems used and studied in the financial community (i.e., we listed the articles by Borovkova; another example; another.). (Note that TRNA now belongs to Refinitiv: Fact Sheet ; Refinitiv News Analytics (by Borovkova). Why Refinitiv News Analytics Programming Overview Video on Programming Interface
Recall that TRNA profiles news about a Company according to the following 19 attributes:

0. Date, Time, and Company Symbol
1. Item Type - Is the news an Alert, Article, Update, Correction, or ...
2. Item Genre - interview, exclusive, wrap-up [news jargon: see NewsML and genre for others]
3. Headline - The actual alert or headline text
4. Relevance - [between] 0 - 1.0 [a heuristic]
5. Prevailing Sentiment - 1, 0, -1 [a heuristic based on the previous classifications of the Company]
6. Current Sentiment: Positive, Neutral, Negative - Provides more detailed sentiment indication [heuristic; all between 0 and 1]
7. Location of 1st Mention - Sentence location of the first time the Company is mentioned
8. Total Sentences - Used for article length
9. Number of Companies - How many other companies are tagged to the news
10. Number of Words/Tokens - How many words/tokens are about the Company
11. Total Words/Tokens - Total words/tokens in the news
12. Broker Action - Denotes broker actions: upgrade, downgrade, maintain, undefined or whether it [the news] is [about] the broker itself
13. Price/Market Commentary - Used to flag items describing pricing/market commentary [is the news an opinion?]
14. Volume or News Item Count - How many news items have been published on the Company over different time periods
15. Novelty or Linked Count - Denotes level of repetition of this item's news about the Company from 12 hours to 7 days [is it really new news?]
16. Topic Codes - Describes what the story is about, i.e. RCH=Research; MRG=Mergers & Acquisitions. Examples: see news codes.
17. Other Companies - What are the other companies tagged to the article [provides a list of Symbols]
18. Other Metadata - Index IDs, linked references, story chains, etc. [citations to previous news, i.e., "see previous article."]

ASSIGNMENT
Divide the class into at 8 most teams corresponding to companies AMZN, KO, ORCL, CAT, GE, T, CSCO, DIS.
Each team will specify which news vendor it uses -- Yahoo, Google, Bloomberg, etc. Is the news source free?

Research Note that only attributes 4-6 are concerned with sentiment and judgement; the the other 16 attributes are either entered manually by the author or are automatically entered when the author submits the article to the news processing system. This "meta-data" can be easily extracted or easily computed by finding the correct XML tag or by counting words or tokens (words/tokens) or other lexical elements. The XML tags are standard (e.g. by NewsML) as is the type or genre of the article. (Computers are good at matching words.)

For this class: each team will provide their own formula (specified as formal heuristic rules) for the three attributes 4-6 in terms of the other attributes.
NOTE: It is easiest to use the other 18 TRNA attributes to derive a formula.
EXAMPLE 1 [for Attribute 4: Relevance]:

Relevance =(X-Y) / X
with X = Total Sentences [attribute 8]; Y = Location of First Mention [attribute 7].

You can also use specific keywords or locations of words that computers can easily compute.

EXAMPLE 2 [for Attribute 6:Positive Sentiment]: Define 3 lists of parsed keywords like "good", "increasing", "positive", "decreasing", etc. for determining 3 lists of "positive words": "neutral words", or "negative words". Use these for sentiment formula: match the words in the news articles with the words in your 3 lists to determine sentiment. For example:

Positive Sentiment =Y/ X
with Y = number of sentences in the article where a positive word is mentioned; X = Total number of sentences [attribute 8].

Make sure that your formulas conform to the data requirements (e.g. between 0 and 1; +1, 0, -1, etc.).

Deliverable Each team will review 3 recent news headline articles for your company.

A. Show the following:

1. Formalize and show your 3 lists (of positive words, negative words, and neutral words) for attribute 6 or find such a list on the web (provide the hyperlink).
2. For each article: Show how you compute your scores using your Team's Heuristic Formulas for attributes 4-6.
3. Which codes will you use for genre [attribute 2] or topic [attribute 16]? If you define your own codes, list them. If you find such a list on the web provide the hyperlink.
4. For each article: show a sample spreadsheet showing the 18 attributes.

B. Discuss in a few brief points:

1. Can you do better (speed, quality and quantity) than a computer at determining news sentiment, genre and topic?
2. For high frequency trading (multiple times per day): Should traders (humans or computers) read news profiles or should they read the original news items when making trading decisions?
3. For slow frequency trading (multiple times per year): Should traders (humans or computers) read news profiles or should they read the original news items when making trading decisions?
4. You work for RhenHao Bank as an equity analyst covering Microsoft Corporation. You receive a tweet from the Associated Press announcing that the US Government is initiating a lawsuit against Microsoft for being a monopoly in order to break up the company. What do you do?
5. Could a traditional financial analyst be replaced by a computer? What skills can be replaced? What skills cannot be replaced?

C. Heuristics vs. Regression

1. Set up a linear regression model to compute sentiment formulas for attrributes 4-6: (i) Show a simple example of the model function and weights; (ii) identify the inputs and outputs you would use. (iii) Can linear regression compute your heuristic formulas? Why? (or Why Not?)
2. Set up a nonlinear regression model ("neural network") to compute sentiment formulas for attrributes 4-6: (i) Show a simple example of the model function and weights; (ii) identify the inputs and outputs you would use. (iii) Can a nonlinear regression model compute your heuristic formulas? Why? (or Why Not?)

Document all results in an easily readable manner.
Prepare your answers (and citations and references) in a PowerPoint presentation. (See the course FAQ for citation and referencing.)

Be prepared to discuss your work next week.
