

THE QUANTIFICATION OF UNSTRUCTURED INFORMATION AND ITS USE IN PREDICTIVE MODELING

Prae Dumrong
Jared Gould
Greg Lee
Logan Nicholson
Kelly Gao
Peter Beling
beling@virginia.edu

Matthias Blume
Jeff Robinson
matthiasblume@fairisaac.com

Fair, Isaac & Co.
200 Smith Ranch Road
San Rafael, CA 94903 USA

Department of Systems and Information Engineering
University of Virginia

ABSTRACT

Managing text-based information is crucial when trying to extract valuable information from documents. Assigning a numerical value to the text-based (unstructured) information is one of the ways to extract value. This research studied the quantification of unstructured text and its forecasting power.

In order to examine unstructured information that related to predictive models, the Beige Books were utilized to investigate and predict changes in the U.S. economy. The Beige Books describe current economic conditions and discuss fluctuations in real Gross Domestic Product (GDP).

To quantify the text-based unstructured information, the Direct Scoring Algorithm (DSA) was proposed. It utilized the keywords in the document and their subjectively-determined numerical weights to score individual sentence. Statistical analyses were then conducted to verify which sections of the Beige Books contributed the most significant information to the prediction of GDP. Utilizing the significant sections, a linear regression model was constructed to predict future GDP growth.

The adjusted- R^2 values of the DSA model were compared to the scoring of the same documents by an economic expert. The comparison demonstrated that the DSA model using the Beige Book significantly contributed to the prediction of GDP, and it explained similar amounts of variance compared to the scores created by an economic expert. Also, a comparison between a structured predictive model and the DSA model was conducted to again prove the significance of text-based information.

research in this area and that the ideas have future potential in the quantification of unstructured information.

1 INTRODUCTION

To investigate techniques to quantify the unstructured information, the Federal Reserve's Beige Book was examined as a source. By extracting the meaningful information from the unstructured information using the software that applied the Direct Scoring Algorithm (DSA), the book's predictive power for real Gross Domestic Product (GDP) was addressed. The DSA method scores text based on a list keywords and their numerical weights. To determine the validity of the method, the statistical analysis was conducted to compare the historical GDP and DSA results. The analysis showed the DSA has the future predictability to forecast the GDP using the Beige Book.

1.1 Unstructured Text

Structured data are raw facts that are easily quantifiable and usable in predictive models. Unlike structured data, unstructured data are difficult to quantify, and it is often extremely complicated to capture the entire meaning of a text document with simple values. Since unstructured data may capture knowledge that cannot be derived from structured data, it is crucial to extract this information and transform unstructured data into a usable form by assigning numerical values.

Once unstructured information is quantified, it can be used as additional input to a predictive model. In general, a predictive model uses only structured data to estimate the value of a future outcome. Since structured data may not contain all the important information needed for prediction,

adding transformed unstructured data may increase the accuracy of the prediction.

1.2 Beige Books

Around 1970, then chairman of the Federal Reserve Board of the United States, Arthur Burns, “decided it would be a more efficient use of the FOMC’s [the Federal Open Market Committee makes decisions affecting short-term interest rates] time to have the reports on district conditions prepared in advance and compiled for the Committee’s edification” (Fettig, 1999). This was the beginning of the Beige Book, a compilation of these reports. Burns believed that this qualitative information could be useful when changing monetary policies. “Chairman Burns observed that one purpose of the new report, as he understood it, was to permit the Committee to more effectively draw on the knowledge of Reserve Bank’s directors. The emphasis was placed on qualitative information, such as “opinions and judgments...” (Fettig). The Beige Book was barely recognized by the public until the stock market crash of 1987, when the Fed was thrust into the spotlight. Even through 1998, the Beige Book was still viewed as a “crystal ball” that could offer a glimpse into the future of the economy. However, it is currently the focus of many analyses that attempt to connect the overall interpretation of the beige book to the future growth of the economy.

2 QUANTIFICATION OF TEXT

Before an unstructured document can be used in a predictive model, it must be broken down and quantified in a logical manner. This process was broken down into three main parts: Lexical Analysis, Syntactic Analysis, and Grammatical Analysis.

2.1 Lexical Analysis

Lexical Analysis is the first and most logical step in the quantification of a text document. It involves two steps, which are the breakdown of sentences and keyword detection.

The sentence breakdown process simply breaks up the document so that it contains one sentence per line. It involves parsing through the document and searching for periods. Certain exceptions were made for when a period did not indicate the end of a sentence (i.e. “St. Louis”). The document was then placed into a text file with each line containing only one sentence. All of this was done using an originally coded software tool, which was developed with Microsoft Visual Basic.

The keyword detection process was much more subjective than the sentence breakdown. After extensive research into the field of keyword detection and natural language processing, it became apparent that no automatic

methods would be adequate. Most computer-based methods rely on removing words that occur with great frequency, which is based on the ground-breaking research of H.P. Luhn where he “recognized that high-frequency words tended to be the common, or non-information bearing one” (Meadow, 1992). However, in testing it was discovered that this method removes very meaningful words from the document. Therefore, a more subjective approach was adopted, which was based on a method called Keyword-in-Context (KWIC) that was designed by Gerard Salton in 1968 (Salton, 1968).

The KWIC method for detecting keywords involved taking a small, random sample from the overall group of documents being analyzed. From this sample, all the relatively meaningful phrases were pulled out and then placed in a separate document. By leaving the keywords in their original context, it attempts to capture the meaning of the keywords based on the subject matter.

An identical process was conducted with the Beige Books, and a file containing all the subjectively-determined relevant keywords was created. The sample of Beige Books spanned all years (1983 – Present), and the readers were also students in economics, which added to the significance of the key-phrase detection process.

Also, while the use of an online dictionary and thesaurus was too difficult to implement, it was still necessary to capture all the synonyms and tenses of the keywords. Therefore, the keyword list was expanded by hand, using a dictionary and thesaurus, to include all relevant synonyms for the key adjectives and the different tenses of all the verbs.

2.2 Syntactic Analysis

Syntactic Analysis includes the relevant weighting of the keywords that were previously determined and the identification of parts of speech in the keyword list.

Again, the weighting of the keywords was a subjective process. Separate readers, who were familiar with economics and the Beige Books, read the keyword list and ‘scored’ the words based on their importance to the economy. A scale of [-3, 3] was chosen, and all words were assigned a positive or negative score. Since the keywords had been left in their original phrases, their original context could be determined. Negative adjectives were given negative scores, and vice versa. Nouns that had negative contexts (i.e. bankruptcies or cancellations) were assigned negative scores, and vice versa. And, nouns or verbs that did not give any indication of the condition of the economy were given a value of one, which would render them neutral during the scoring process (scoring uses multiplication).

Next, the different parts of speech had to be identified to prepare for the grammatical analysis process. Again, while the use of an online dictionary was not possible, it

was feasible to mimic the results of using one. Therefore, readers simply assigned the relevant parts of speech to each word in the keyword list. However, after this process was completed, it became apparent that the delineation between adjectives and verbs would not be necessary because both describe the economy (its changes or its state). Consequently, they were lumped together into the same part-of-speech category (denoted by 'A'). Also, it became necessary to differentiate between which nouns contained inherent meaning and which did not. Therefore, nouns with inherent meaning were put in one category (denoted by 'N') and non-meaningful nouns were placed in another category (denoted by 'O').

By the end of the Syntactic Analysis process, the documents had been examined, and a keyword list had been created. This list had then been broken down into individual words, and then those words were assigned relevant weights. The words were then categorized based on what part of speech they were, and the result was a spreadsheet with three columns: keywords, weights, and corresponding categories.

2.3 Grammatical Analysis

In order to develop the grammatical rules for scoring unstructured data, four Beige Books from year 1984, 1987, 1993, and 1998 were selected randomly to be analyzed. Exhaustive examinations yielded the following rules, which were derived from the most common sentence structures in the Beige Books. The scoring rules create key word groups and calculate the score for each group. For example, an adjective is grouped with the following noun that the adjective describes. The rules for grouping key words were based on the grammatical rules in *Link Grammar*, a syntactical parser software developed by John Lafferty, Daniel Sleator, and Davy Temperley from the School of Computer Science at Carnegie Mellon University. For each sentence, *Link Grammar* assigns a syntactic structure by tagging a part of speech to each word and grouping a pair of words.

Using the *Link Grammar* concepts, the grammatical rules for grouping key words were developed. Each sentence may have more than one group. Therefore, the sentence score was simply the sum of the group scores. If the sentence did not contain any key word group, that sentence was removed. Finally, the section score is calculated by averaging the total sentence scores. The remainder of this section below illustrates different scoring rules. The complete description of the scoring rules can be found in Appendix 1.

Rule 1 groups the non meaningful noun with the following adjective. If this noun stands alone, the noun is not included since it becomes meaningful only when grouped with the adjective. Since the noun's score is either 1 or -1, the magnitude of the group score takes on the adjective

score. The group score is calculated by multiplying the non meaningful noun's score with the adjective's score.

Rule 2 groups an adjective with the following noun. The score of the group is dependent on the signs of the adjective and the meaningful noun. For example, the score for the group, "slight decrease", is equal to the score of "decrease" (-2) minus the score of "slight" (-1) which is calculated to be -1. However, using rule 2, the group score of "strong decrease" is -4, which is computed by adding the score of "strong" (-2) to the score of "decrease" (-2). Since "slight decrease" has less negative effect on the state of the economy than the word "strong decrease", rule 2 was created to reflect the different levels of economic conditions.

In addition, rule 2 correctly calculates the score for the group containing the adjectives "no" and "not" in which case the group score is calculated by multiplying the noun's score with the adjective's score. For example, the score for the group "no deterioration" is equal to 2 which is obtained by multiplying the score of "no" (-1) with the score of "deterioration" (-2). Therefore, multiplying the scores for this case is appropriate since it yields a correct sign for the group score.

In order to capture the meaning of a group of words containing many adjectives and a noun, rule 3 was created. First, the average adjective score is computed prior to combining this score with the noun score. Similar to rule 2, the sign of the group score for rule 3 is dependent on the signs of both the adjective's score and the noun's score. In order to maintain the same scale of group scores as those calculated by other grammatical rules, the magnitude of the group score is equivalent to the average adjective score.

Rule 4 groups a meaningful noun with the following adjective unless this adjective is followed by a non meaningful noun, in which case the meaningful noun stands alone and rule 2 is in effect; the adjective is grouped with the following non meaningful noun. If the meaningful noun is grouped with the following adjective, the group score is equivalent to the plus or minus adjective score. Again the sign of the group score depends on the sign of each word in the group. The magnitude is also equal to the adjective's score to make the scales of magnitude of the group similar to those created by other rules.

All of these Rules were implemented in MS Visual Basic. And, from a random sampling of 50 sentences, the software scored the sentences based on these four main grammatical rules in an identical fashion to how a human reader would have interpreted the sentences 88% of the time.

3 DIRECT SCORING ALGORITHM

Using the Lexical, Syntactic, and Grammatical Analysis, the Direct Scoring Algorithm (DSA) was implemented in MS Visual Basic. It automatically completed all of the

steps previously defined and came up with a score for each sentence. The output from the DSA was a spreadsheet containing each sentence along with its corresponding score. These scores were then averaged for each section of the Beige Book (i.e. Manufacturing, Consumer Spending, etc.) and then treated as independent random variables in the regression analysis.

3.1 Estimating the Beige Book Scores using DSA

First order linear regression and stepwise regression were used to find the best model for estimating the Beige Book scores from their section scores. Figure 1 below illustrates the first order linear model used to find the beta parameters (β), the weight that each section contributes to the book score. GDP growth is calculated to be the percentage change between the current quarter (t) to the same quarter of the previous year (t-4). In order to test the lagged quality of Beige Books to GDP growth prediction, time lags of 0, 2, 4, and 6 were applied to this first order full model. A lag n model indicated that the two Beige Books in a current quarter could predict the GDP growth n quarters later. For a lag 2 model, the first quarter Beige Book score was modeled to fit the 3rd quarter GDP growth.

$\text{GDP Growth}_{t+h} = \beta_0 + \beta_1 \text{Overview}_t + \beta_2 \text{Manufacture}_t + \beta_3 \text{Consumer Spending}_t + \beta_4 \text{Agriculture}_t + \beta_5 \text{Real Estate}_t + \beta_6 \text{Finance}_t + \beta_7 \text{Energy}_t + \beta_8 \text{WIPE}_t$
<p>where $\text{GDP Growth}_t = (\text{Chained GDP dollars}_t - \text{Chained GDP dollars}_{t-4}) / \text{Chained GDP dollars}_{t-4}$</p>
<p>h (lag) = 0, 2, 4, 6</p>
<p>WIPE_t = Wage, Inflation, Price, and Employment</p>
<p>Estimation Period = 1983 (quarter 2) - 2002 (quarter 4)</p>

Figure 1: The First Order Full Model for the Beige Book Scores Using All Section Variables

Adjusted R-square was used to compare the significance of each of the lagged models. For this project, the significance level was .05. Table 1 below shows that the lag 2 model explained the largest portion of variation in the GDP growth due to its highest adjusted R² value of .359.

After comparing these models, the lag 2 model was chosen as the best fit. In order to possibly reduce the complexity of the model, a stepwise regression was applied. This could eliminate unneeded terms and keep those that

Table 1: Adjusted R² Values for Lag h Models

Lag h Models		
Model	Adjusted R ²	Observations
Lag 0	0.338	79
Lag 2	0.359	77
Lag 4	0.189	75
Lag 6	0.034	73

are the true driving force behind the model. Figure 2 below shows the reduced model after the stepwise regression technique was applied to the first order full model. The adjusted R² of this reduced model was .364 which was better than that of the lag 2 full model.

$\text{GDP Growth}_{t+h} = \beta_0 + \beta_1 \text{Overview}_t + \beta_2 \text{Manufacture}_t + \beta_3 \text{Consumer Spending}_t + \beta_4 \text{Agriculture}_t + \beta_5 \text{Real Estate}_t + \beta_6 \text{Finance}_t$
--

Figure 2: The First Order Reduced Model for the Beige Book Scores

The beta parameters of the reduced model in Figure 2 are shown below in Table 2.

Table 2: Parameter Values of the Reduced Model

Parameters	Values
β_0	1.591
β_1	0.545
β_2	0.479
β_3	0.439
β_4	-0.283
β_5	0.299
β_6	0.404

Since the Beige Books describe the state of each major economic sector, they should have a positive relationship with the GDP growth. For example, an increase in the agriculture market should increase GDP growth. However, the parameter pertaining to the agriculture section does not follow this positive relationship. According to the grammatical rules, "high price," received a negative score since it weakens the consumer spending section. However, "high price" improves the agriculture market since farmers received higher income. Because of this reverse interpretation of the agriculture section, the agriculture parameter has a negative relationship to the GDP growth.

3.2 The DSA Model vs. Payne's Model

Using the reduced model found in Figure 2 and the parameters in Table 2, the Beige Book scores were computed as a linear combination of the Beige Book section scores. These Beige Book scores can be treated as predicted GDP growths. Using the Beige Book scores as a predictor variable, and the GDP growth as a response variable, the adjusted R² value was estimated in Table 3 below.

Table 3: Adjusted R² Values for the Beige Book Model and the Payne's Model, Estimation Period: 1983 quarter 2 - 2000 quarter 2

Model	Adjusted R ²	Observations
Beige Book	0.356	69
Payne	0.45	69

Although the adjusted R² of the Beige Book model is not very high (.356), it was comparable to the result of Payne's model (.450) done by the economic expert. In 2000, David Payne, an economist with the Department of Commerce, created a method for scoring the Beige Books by human readers. Each section of the Beige Books was scored based on different levels of economic growth. The section scores were then used to predict the actual GDP growths (Payne 11-13).

Figure 3 below shows the time series of the predicted GDP growth (Beige Book scores) and the actual GDP growth. With the exception of the two extreme years, 1984 and 1991, the predicted GDP growth tends to follow the trend of the actual GDP growth quite well.

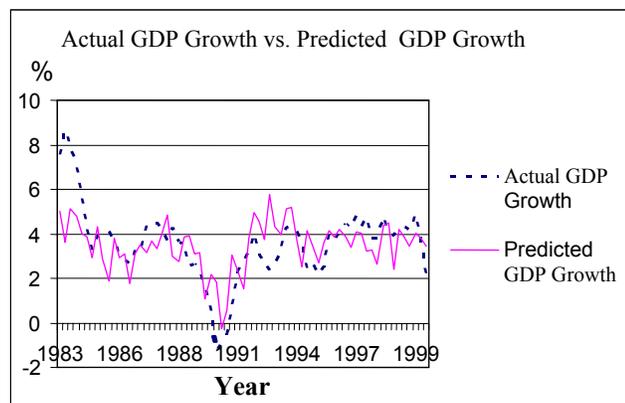


Figure 3: Actual GDP Growth vs. Predicted GDP Growth

3.3 Updating the Interest Rate Spread Model

In order to determine whether unstructured data added predictive power to an existing predictive model, linear regression of the combined model, the base predictive model

and the transformed unstructured data, was needed. For this project, an interest rate spread model was chosen as the base model, while the Beige Book scores were the transformed unstructured data.

In July 2000, Saito and Takeda developed the interest rate spread model which estimated the US GDP growth using the spread between 3 month Treasury bills and 10 year notes (Saito et al. 7). The linear model to predict GDP growth using the spread is shown in Figure 4 below.

$\text{GDP Growth}_{t+2} = 2.621 + .413 \text{ Spread}_t$ <p>where $\text{GDP Growth}_t = (\text{Chained GDP dollars}_t - \text{Chained GDP dollars}_{t-4}) / \text{Chained GDP dollars}_{t-4}$</p> <p>Estimation Period (t+2) = 1983 quarter 2 to 2002 quarter 4</p>
--

Figure 4: The Interest Rate Spread Model

Using the available data from 1983 to 2002, the spread model and the combined model were created. Table 4 below shows the adjusted R² and the p values of the spread model, the Beige Book model, and the combined model.

Table 4: Adjusted R² Values for the Spread, Beige Book, and the Combined Models

Model	Adjusted R ²	Observations
Spread	0.052	77
Beige Book	0.406	77
Combined	0.41	77

The adjusted R² of the Beige Book model was .406. Adding the Beige Book scores to the spread model, the adjusted R² of the combined model became .410 which was slightly higher than that of the Beige Book model. Since the spread model only explained 5.2% variation of the GDP growth, the spread variable was insignificant in predicting the GDP growth. Since the data only contained two economic cycles, it was not enough to capture important information in time series of the spread and the GDP growth. Thus, the spread did not contain good predictive power for the GDP growth. Therefore, combining the *Beige Book* scores with an insignificant variable such as spread did not truly improve the quality of the model because the added variable did not add a significant portion of variation it explained. In addition, when the stepwise regression was applied to the combined model, the spread variable was thrown out.

3.4 Clustering Scoring Algorithm (CSA)

The Clustering Scoring Algorithm (CSA) was a concept added to the linear regression model as a reengineering to

potential find more accurate scores. CSA was originally developed as an alternative algorithm to DSA, which would cluster books based on their similarity and compute scores based on those clusters. Because of limitations in implementation, the books' similarities could not be based on individual words or word phrases but rather on a single all-encompassing metric. The most obvious score here would be the predicted GDP score, making CSA an algorithm appended to DSA and not a stand-alone method. DSA finds book scores from the section scores, while CSA simply transforms those scores in hopes of getting an even better GDP approximation.

There are many available methods for computing similarity, and after analyzing the scale of the data, an equation was chosen. This similarity equation is shown in Figure 5:

$$S = e^{-D}$$

Figure 5: Similarity Equation

Here the distance, D, is the absolute value of the difference of scores. This equation ensures that all similarities are on a scale of 0 to 1. All of the similarities were then standardized to sum to 1, while excluding a particular book's similarity to itself. The CSA scores were then computed as the inner product of a book's similarity to all other books and corresponding actual GDP figures from the same quarter of each book.

3.5 CSA Results

The original method without the reengineering surely seems to be a more direct method than CSA and would produce better scores, but in its conception CSA did seem very intriguing. CSA would most likely improve scores that do not have a normal distribution by rescaling outliers. As previously mentioned, resource limitations changed the approach to CSA, forcing the similarity measures to be one-dimensional as opposed to multi-dimensional. Comparison to DSA did confirm that the new CSA is not an optimal method, but the original CSA, in theory, could offer better results. The results of CSA as implemented compared to actual GDP are as follows:

Table 5: CSA Statistical Results

Model	Adjusted R ²	Observations
CSA	.255	77

Because the CSA reengineering actually decreased model prediction, it will be left out of the remainder of the analysis. For future data sets, it could be an option, but DSA will be the only focus for the capturing of unstructured information.

4 CONCLUSIONS

This research project proved that it is possible to create an automated process to quantify unstructured data and that this quantified data does have predictive ability. The Beige Books were used as the unstructured data, or the predictor variables, that predicted the real GDP growth, the response variable. The project's recommended method for quantifying text-based documents is called the Direct Scoring Algorithm (DSA). Using the key word list and the grammatical rules, the DSA scored the sentences and the sections of the Beige Books. Based on linear regression and stepwise regression modeling, the best model for scoring the Beige Books was the lag 2 linear model with six predictor variables (overview, manufacturing, consumer spending, agriculture, real estate and finance). The final scoring model explained 35.6 % of the variation of the actual GDP growth which was less than the 45% variation explained by Payne's subjective scoring model.

Because of the lower percentage of variation, the final scoring model did not predict the actual GDP growths as well as Payne's model did. This result was expected since Payne's model relied on economic experts, whereas the final model used the computer program. Since the program cannot interpret human languages as well as humans do, the final model cannot capture all the significant meanings of sentences.

Although the variation of the GDP growth explained by the final model was not very high, it was comparable to that of the expert model. Therefore, the result showed that the DSA software, using the grammatical rules, was successful in interpreting unstructured data and transforming them to structured data.

Regarding the added value of the unstructured data to other predictive models, the conclusion was that the added value depends on the quality of the base model. First, if the base model does not have good predictive power, the unstructured data model may have a much better forecasting ability, and the result from the base model may be left out. This project reflected this case. The adjusted R² of the spread model was so low that when combining with the Beige Book scores, the quality of the combined model was only slightly better than that of the Beige Book model. Therefore, the prediction would still be good even when the unstructured data (the Beige Book scores) was included, and the result of the base model (the spread) was excluded. Second, if the base model has an excellent predictive ability, the interpreted unstructured data may not add any value. Since the base model already explains a high portion of variation, unstructured data may contain overlapped information with the existing base model. Lastly, the unstructured data may be useful in improving the predictive power of the base model if the base model has a medium forecasting ability. The unstructured data

may have useful information that the base model does not contain.

5 RECOMMENDATIONS

An important limitation of this project is the fact that the Beige Books are semi-structured while other types of unstructured data may not have simple key word phrases or sentence structure that the Beige Books do. From 156 Beige Books that were tested, the key words were consistent. Therefore, the keyword list was able to capture most of the important words. In addition, the sentence structures were simple and were similar to one another. The Beige Books are also divided so that each section contain only one topic. Finally, the Beige Books focus on one specific subject, which is the condition of the economy.

Since most text-based documents do not contain the semi-structured quality as the Beige Books do, it is critical to generalize the Direct Scoring Algorithm before applying it to other unstructured data. Depending on the type of the documents, it may be necessary to take the following steps prior to using the DSA:

- **Incorporate the Topic Detection Software**
This software should be able to determine whether the information from the whole document pertains to a specific subject. If the document focuses on more than one topic, it may be necessary to divide the documents up into sections so that each section contains one specific subject.
- **Incorporate a Thesaurus When Adding Words to the Key Word List**
After the key word list is created, a thesaurus can be used to add words with similar meanings to those that already exist in the list. Also, a thesaurus can be used to determine the part of speech of each key word which will be useful when scoring the sentences based on the key words and grammatical rules.
- **Modify the Grammatical Rules**
To obtain better scoring results, one need to develop more complex grammatical rules. For example, the rules may include compound sentences and different types of conjunctions.

Furthermore, additional research for the updating field is needed. In order to effectively use the information from unstructured data to improve the result of the predictive model, it is necessary to find updating algorithms that can update the end result of the base model with unstructured data instead of changing the base model by adding or deleting variables.

APPENDIX 1: Complete Grammatical Rules

Rule 1: An *O* is grouped with the following *A*.
If the *O* is not followed by an *A*, the *O* is thrown out.
Score: The group's score is equal to *O*'s score x *A*'s score.

Rule 2: An *A* is grouped with the following *O* or *N*
If the sign of the *A*'s score is negative, and an *N* is the following word.
Score:

- If the *N*'s score is negative, the group's score is equal to *N*'s score - *A*'s score.
- If the *N*'s score is positive, the group's score is equal to *N*'s score + *A*'s score.
- If the *A* is either "no" or "not", the group's score is equal to *N*'s score * *A*'s score.

If the sign of the *A*'s score is positive.

Score:

- If the *N*'s score is negative, the group's score is equal to -*A*'s score.
- If the *N*'s score is positive, the group's score is equal to *A*'s score

An *O* is the following word.

Score: The group's score is equal to *O*'s score * *A*'s score.

Rule 3: All *A*'s are grouped together until an *O* or a *N* encountered. Once *A*'s are followed by an *O* or a *N*, they are grouped with the *O* or the *N*. First the scores of all *A*'s are multiplied and divided by the number of *A*'s to obtain the average *A*'s score.

If *A*'s are followed by the *N*.

Score:

- If *A*'s score is positive, and *N*'s score is positive, the group's score is equal to average *A*'s score.
- If *A*'s score is positive, and *N*'s score is negative, the group's score is equal to -average *A*'s score.
- If *A*'s score is negative, and *N*'s score is positive, the group's score is equal to -average *A*'s score.
- If *A*'s score is negative, and *N*'s score is negative, the group's score is equal to average *A*'s score.

If A's are followed by an O.

Score: The group's score is equal to the average A's score * O's score.

Rule 4: A N is grouped with the following A unless the A is followed by an O, in which case the N stands alone and the A is grouped with the O.

The N is grouped with the following A

Score:

- If N's score is positive, and A's score is positive, the group's score is equal to A's score.
- If N's score is negative, and A's score is positive, the group's score is equal to -A's score.

The N is grouped with the following O

Score: The group's score is equal to N's score x O's score.

REFERENCES

Fettig, David. "The Federal Reserve's Beige Book: A better mirror than crystal ball." The Region. March, 1999.

Meadow, Charles T. Text Information Retrieval Systems. Academic Press Inc. San Diego, CA. 1992.

Payne, David R. 2000. Anticipating Monetary Policy With the Federal Reserve's Beige Book: Re-specifying the Taylor Rule. *Proceedings of the National Association for Business Economics Annual Meeting*, 1-24. Washington DC:

Saito, Yoshihito, Yoko Takeda. 2000. Predicting the US Real GDP Growth Using Yield Spread of Corporate Bonds. *International Department Working Paper Series 00-E-3*. 1-17.

Salton, Gerard. Automatic Information Organization and Retrieval. New York. McGraw-Hill. 1968.

AUTHOR BIOGRAPHIES

Prae Dumrong is a fourth year, Systems and Information Engineering major with a minor in Computer Science at the University of Virginia. She is originally from Bangkok, Thailand. Throughout her undergraduate curriculum, she

concentrated on computer science, data analysis, and economics. She also enjoys ballroom dancing and going to see the plays. In August 2003, Prae will be working as an Associate Consultant at Appian Corporation in Vienna VA. She can be reached at <prae@virginia.edu>.

Jared Gould is a fourth year Systems Engineering and Economics major at the University of Virginia. During his studies, Jared has focused on the data analysis and the financial aspect of economics. In July 2003, he will begin work at UBS Warburg as an analyst in their Investment Banking Division in New York, NY. Jared is also a member of Sigma Chi Fraternity and can be reached at <jjg4j@virginia.edu>.

Do-Hoon (Greg) Lee is a fourth year Systems and Information Engineering major at the University of Virginia. During his studies, Greg has focused on the computer science aspect of engineering. He can be reached at <dl7r@virginia.edu>.

Logan Nicholson is a fourth year Systems and Information Engineering major and Economics major at the University of Virginia. Throughout his undergraduate career, he has focused on the relationship between finance, economics, and engineering. He was a former member of the varsity football program and was involved in class council and fraternity council while at Virginia. After graduation, he will begin working for Goldman Sachs in the debt capital markets division in New York, NY. He can be reached at <ljn3m@virginia.edu>.