Exploring Commodity and Stock Volatility using Topic Modeling on Historical News Articles: Application to Crude Oil Prices

Rui Jiang (forestj@stanford.edu), Olufolake Ogunbanwo (folakeo@stanford.edu), and Mustafa Al Ibrahim (malibrah@stanford.edu)

Abstract

Thorough commodity or stock analysis is a time consuming process because it involves extracting and summarizing information from large amount of data. The majority of financial studies only focuses on numerical data because it is easier to automate using statistical methods and because they represent concrete unchanging values. However, focusing solely on numerical data ignores the context and circumstances of events that led to those values. This study presents a semi-automated workflow for studying stock and commodity prices that incorporates numerical and text based data. The workflow uses topic modeling (Latent Dirichlet Allocation algorithm) to extract important topics from historical news articles. Initial testing of the algorithm was done on 765 articles from PetroWiki. Topic interpretation is made by the user incorporating general knowledge and a simple investigation of representative articles. Time-dependent frequency of each topic is calculated and correlated with the price. The workflow is tested on 28,415 New York Times articles mentioning "oil prices" between 1970 and 2015. As expected, oil prices are controlled by a number of inter-connected factors related to politics and financial events. Ten topics were extracted: corporate finance, commodity and US currency, world economy, US Energy policy, emerging economies, Middle East conflict, OPEC, stock market, US elections and politics, and world-OAPEC relations. Time correlation plots show that, in the long run, some topics such as US currency and US Energy policy show a consistent correlation with oil price while others, such as OPEC, change with time. In addition, investigation into a larger number of randomly sampled articles revealed some interesting correlations such as the correlation between oil price and unemployment. Whether the correlation is a causational relationship is unknown. In general, the workflow is useful in identifying important factors that affect the price of a commodity or a stock, aiding the investigator in financial analysis.

Introduction

Studying the relationship between stock or commodity prices and external factors is an essential component of the investment decision making process. Knowledge of the critical external factors and how the relevance of each factor changes over time is important. For example, a company in the United States that utilizes locally made raw materials in its manufacturing plant will not be susceptible to political and economic changes in China. Relocation of its manufacturing plant to China will however, greatly increase its susceptibility to China. Unfortunately, studying this susceptibility is a time consuming process as it requires going through a large number of reports and financial data. This is especially true for obscure stocks as the investigator does not know in advance the general factors affect the price.

In this study, we present a workflow for exploring factors related to stock or commodity prices and study their correlation using machine learning techniques applied on historical newspaper articles. We use the volatility of crude oil price as a case study in this paper because it is of great interest to consumers, companies, and governments. Given the importance of the oil price to global landscape, the observations from this study may shed some lights on hidden factors which should be considered while making financial and political decisions related to the energy industry.

Topic Modeling and Latent Dirichlet Allocation

The most commonly used method for topic modeling, or topic discovery from a large number of documents, is Latent Dirichlet Allocation (LDA), first introduced by Blei et al. (2003). Simply speaking, the LDA algorithm assumes that each document, as a bag-of-words, is a mixture of different topics, and each topic is a mixture of different words.

More formally, the smoothed LDA we used assumes the generative process shown in Figure 1Figure 2. There are four steps in this generative process. (a) $\boldsymbol{\varphi}$, probability of words in K topics, is drawn from Dirichlet distribution with $\boldsymbol{\beta}$ as prior weight of words; (b) $\boldsymbol{\theta}$, distribution of topics in each document, is drawn from Dirichlet distribution with $\boldsymbol{\alpha}$; (c) $\boldsymbol{Z}$, identity of topic of each word in each document, is drawn from the categorical distribution with probability $\boldsymbol{\theta}$; (d) $\boldsymbol{W}$, identity of each word in each document, is drawn from the categorical distribution with probability $\boldsymbol{\varphi}$, given the topic identity specified by $\boldsymbol{Z}$.

Given $\boldsymbol{W}$, the collection of documents, LDA uses a variational EM algorithm to estimate the latent variables $\boldsymbol{\varphi}$ (word distribution for each topic), and $\boldsymbol{\theta}$ (topic distribution in each document), with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as hyperparameters. For example, we ran LDA for 10 topics on 765 articles from PetroWiki (PetroWiki, 2015). The results, including interpretation of 10 topics, word-topic relationship, and topic-document relationship, are shown in Figure 2. With the topics neatly interpreted, one can have a quick and high level overview of all the content inside PetroWiki.

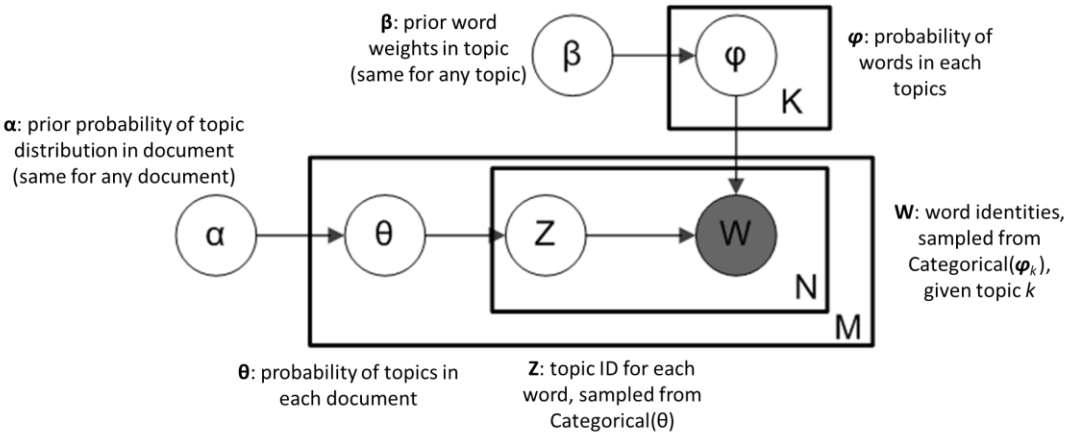

Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. Figure is modified from Blei et al. (2003).

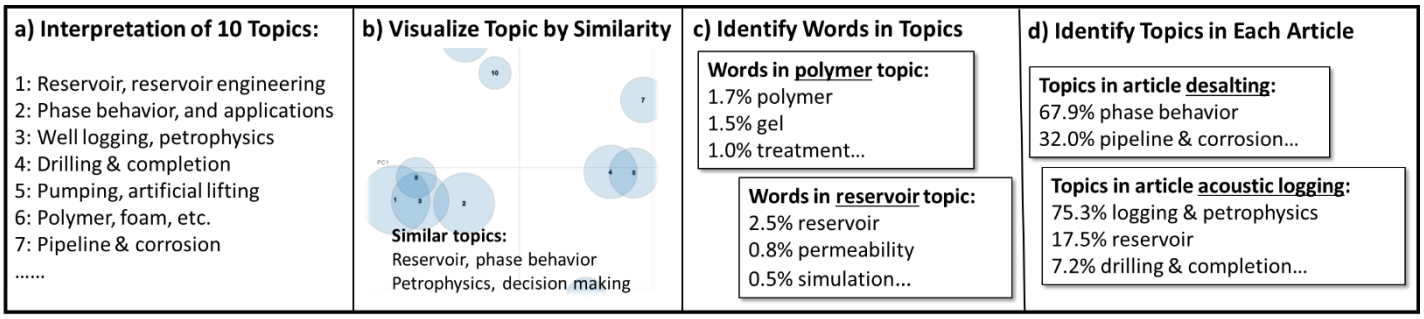

Figure 2: LDA result on PetroWiki articles: (a) interpretation of the topics, (b) visualize topics based on similarity, (c) sample topics with word distribution, (d) sample articles with topic distribution.

Workflow

Figure 3 shows the general workflow developed for exploring factors from a set of articles and relating them to prices. To create the dataset, a snippet of the first paragraph from articles on the news website of interest is scraped. Application programming interfaces (APIs) provided by news websites can be used to automate the scraping process. To obtain a dataset that is representative of our interest, a query term is specified to restrict the search to the newspaper articles that are related to the topic. The resulting dataset is the main input to topic modeling. Preprocessing the dataset includes removing all stop words and non-word character using regular expressions.

Topic modeling is then applied using the LDA algorithm in the Python package, *gensim* (Řehůřek and Sojka, 2010). In this workflow, users need to specify the number of topics according to the diversity of the corpus. A visualization tool, *LDAVis* (Sievert et al., 2014), which shows the inter-topic relationship and word distribution is used to enhance topic selection and interpretation. In *LDAVis* (Sievert et al., 2014), the inter-topic distance is calculated and projected using multi-dimensional scaling (MDS) with a Jensen-Shannon divergence distance. The weight of each topic is captured by the size of the circle used to depict the topic in the visualization. The bigger the circle depicting the topic, the greater the weight of the topic in the dataset.

The terms within the topics are used to interpret the topics. *LDAvis* defines the relevance, *r*, of a term in a topic with a parameter called lambda (λ) (Equation 1). λ accounts for the lift which can be defined as the ratio of a terms probability within a topic, $\boldsymbol{\varphi}_{kw}$, to its marginal probability across the corpus, $\boldsymbol{p}_w$. Modifying lambda helps to reduce the noise in terms and increases interpretability of the topics. λ can lie between 0 and 1 and is optimally found to have a value of about 0.6 (Sievert et al., 2014).

$$r(w,k|\lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\log\left(\frac{\phi_{kw}}{p_w}\right) \tag{1}$$

Once the topics have been interpreted, the topic distributions of all articles in each year are calculated and plotted with time (Ng, 2015). To study the time dependence of factors related to historical prices, linear correlation coefficients between topic trends and the price are computed using a temporal moving window. The correlation coefficients for the entire time period are also computed.

Finally, using the visualization, knowledge of the major historical events, and changes in the stock or commodity price with time, the results are analyzed to find interesting correlations between the weights of different topics and the price over time. The ability to navigate across topics, words, articles, and time, enables thoughtful analysis of the correlations with support from individual articles.

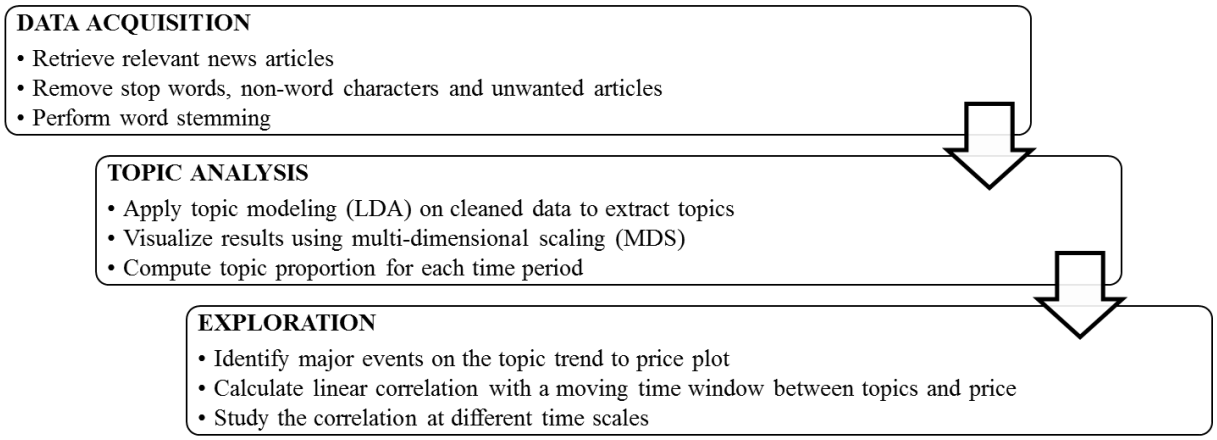

Figure 3: Workflow for exploring commodity and stock volatility using topic modeling on historical news articles.

Results on Articles Mentioning "Oil Prices"

We applied the workflow on a clean dataset from New York Times. We used "oil prices" as the query term, and obtained the headlines and the first paragraphs of 28,415 articles published between 1970 and 2015. Studying the topics extracted (Figure 4), we observed that some topics are more similar to each other when they are examined using multi-dimensional scaling (Figure 4, right). The correlations observed are reasonable. For example, the major words in topics 4 and 3 are "*energy, tax, gas, bill, congress*" and "*economic, world, president, Russia*" respectively. This might be indicative of the relationship between national and international policies related to the oil industry in the US. After testing the algorithm with different number of topics, we observed that ten topics are sufficient to capture the variability in the topics without over-segmenting. This is both a subjective and objective decision based on the topics extracted and the multi-dimensional scaling results.

#	Word Distribution for Topics	Topic Interpretation	Color
1	1.6% company, 0.9% quarter, 0.5% profits, 0.4% industry	Corporate Finance	
2	1.2% dollar, 0.9% futures, 0.8% trade, 0.4% commodity	Commodity and US Currency	
3	1.2% economic, 0.9% Russia, 0.6% Mexico, 0.5% debt	World Economy	
4	2.2% energy, 1.1% tax, 1.0% gas, 0.6% bill, 0.5% congress	US Energy Policy	
5	1.9% economy, 1.8% rates, 1.7% growth, 1.6% interest	Emerging Economies	
6	1.2% Iraq, 1.1% Saudi, 0.7% Iran, 0.6% war, 0.3% military	Middle East Conflict	
7	3.1% OPEC, 1.8% production, 1.8% crude, 1.0% output	OPEC Production	
8	2.5% stocks, 2.0% market, 1.8% dow, 1.2% shares	Stock Market	
9	0.6% president, 0.3% America, 0.2% public, 0.2% election	US Elections and Politics	
10	1.0% countries, 0.7% world, 0.5% arab, 0.5% OPEC	World-OAPEC Relations	

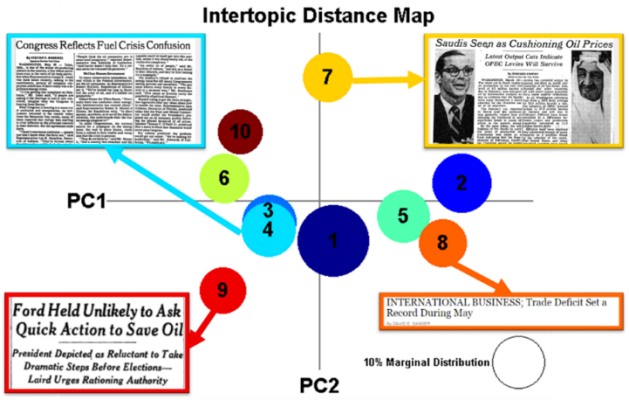

Figure 4. Left: Topics extracted from the "oil prices" dataset examined in this study. Examples of the most frequent words are given. The topics are interpreted based on the words and examination of representative articles. Right: Multi-dimensional scaling visualization of the first two principle components showing the relationship between the topics. Note that some topics (e.g. 3 and 4) are closer to each other. The colors match the colors used in the rest of the figures.

Comparing the results with historical events, we observed good correlations between topic trends and oil price (U.S. Energy Information Administration, 2015). The comparison is done on yearly basis to ensure a sufficient number of articles for each year. Figure 5 shows the trend of each topic over the years as well as the historical inflation-adjusted oil price. The area between the curves represents the proportion of different topics.

The signature of major world events can be easily identified. For example, the First Oil Crisis (1973-1974) occurred when the Organization of Arab Petroleum Countries (OAPEC) started an oil embargo on selling oil to the United States, which resulted in a substantial price hike. Topic modeling results show that the topic #10 (World-OAPEC relations, with words "*countries, nations, world, econ, arab, opec*") has a relatively high proportion in that period. Another example is related to the Middle East in general, and Iraq specifically in topic #6 (Middle East conflict, with words like "*iraq, united, saudi, states, iran*"). We observed a higher trend in this

topic related to the decreased oil output during the Iranian Revolution (1979), the Gulf War (1991), and the Iraq War (2003). Around 1995, we observed a huge increase in the importance of the topic #2 (commodity and US currency, with words like *"dollar, future, trade, commodity*"). This coincides with the Clinton presidency in the United States during which the dollar was very strong. In the same period, the oil price fell. This suggests an inverse relationship between the dollar value and the price of oil which studies have shown to be true. Finally, financial related topics dominate the signature in the last ten years including the 2008 Financial Crisis.

To obtain a more quantitative result regarding the relationship between oil prices and topics, correlation plots are constructed using a moving temporal window. Figure 5 (right) shows the result when a window of nine years is used. The relatively long time window is used to reveal any long-term correlation. Results show that topics' importance generally changes with time. For example, topic#7 which is related to the Organization of Petroleum Countries (OPEC) shows a varying correlation with time. This is because OPEC's stance changes with time. Currently, OPEC is driving the oil price down with overproduction which is visible on the plot by the downward trending average correlation. Other topics show generally constant correlation. For example, topic #2 (commodity and US currency) is generally negatively correlated with price. This constant correlation could reflect the fact that oil price has an inverse correlation with the US dollar. Another example is topic #3 (US energy policies) which shows a general positive correlation reflecting the fact that the majority of US energy policies are in favor of the oil and gas industry.

Sometimes, topics extracted by LDA are informative in an unexpected way. A previous LDA result suggested a topic including words: "*coffee, world, markets, cents, sugar, cotton, pound*". It implies an interesting fact that certain commodities are usually mentioned along with the oil prices. Individual article examinations showed several folds of relationship: coffee, sugar, cotton, along with crude oil, are main indicators of the economy, hence the tendency of being mentioned together. Oil prices will increase the cost of crops as harvesting, processing and transporting require fuel. Polyester made from petroleum derivatives is a major type of clothing material that can replace cotton, so we expected that cotton demand increased as oil price increased. Moreover, this investigation helped in locating an issue where the dataset inappropriately included "business digest" where the oil price was mentioned as one of many uncorrelated news entries.

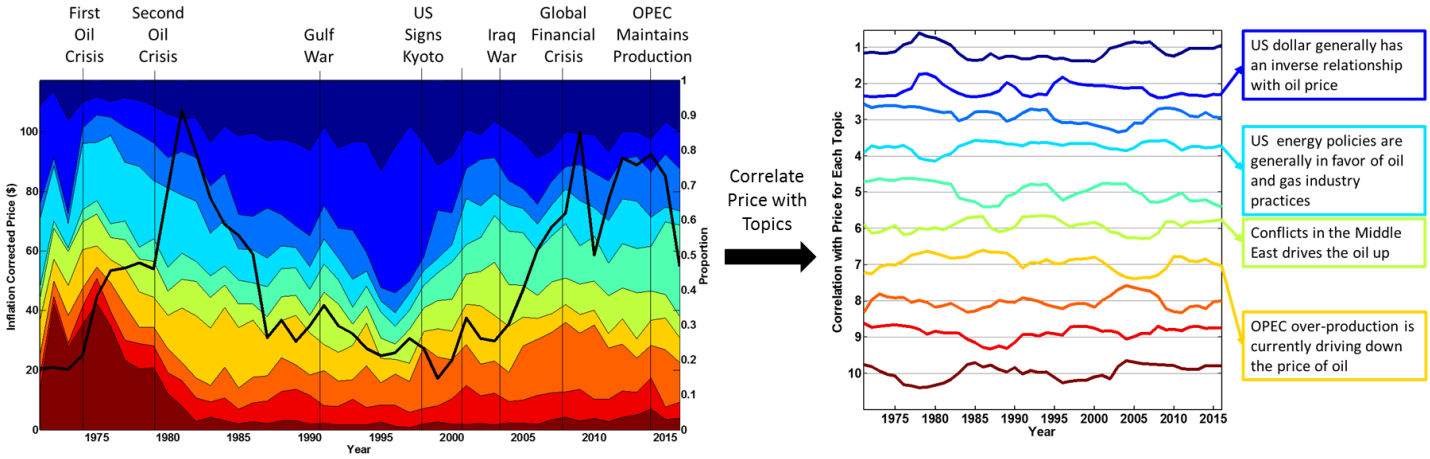

Figure 5. Left: Topics trends overlaid by the inflation corrected oil price and major events related to oil price. Right: Correlation between each topic and oil price using an average window of 9 years (i.e. long term correlation). Topic colors and titles are defined in Figure 4.

Results from General Articles

To explore more hidden factors related to the oil price, we applied the workflow to a collection of 338,828 New York Times articles randomly sampled daily from 1970 to 2015. In this case, given the vast variety in the data, 50 topics are required to adequately model the topics. Some of the topics clearly interpreted from the terms distribution include sports, war, crime, and law (Figure 6, left). A linear correlation of the topic distribution trend in time against the oil price showed correlations of various degrees. As expected, there was a strong positive correlation between the oil price and war-related topics (e.g. topic #40). We investigated topics which showed strong correlations to oil price. While trying to understand the strong correlation of crime (e.g. topic #36) with oil price, we discovered that crime rate is closely linked to unemployment. Intuitively, this is expected. The advantage in discovering this link with unemployment rate is that there is readily available information on the employment rate (U.S. Bureau of Labor, 2015) which can be plotted versus oil price to confirm if there is a correlation (Figure 6, right).

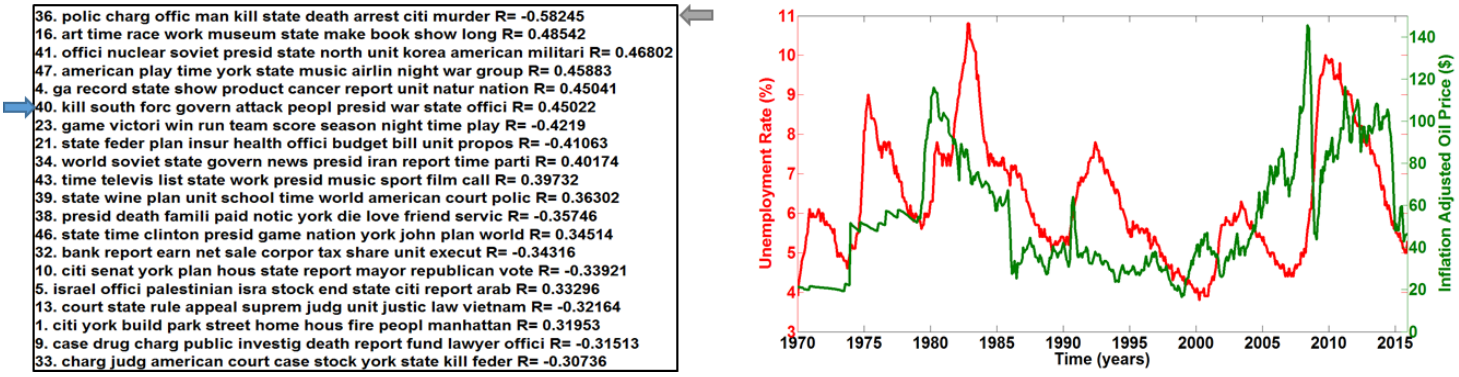

Figure 6. Left: Correlation coefficient of some topics from the general data set ordered by the absolute correlation coefficient. Topic #36 (crime) and topic #40 (war) are pointed. Right: Plot of unemployment rate and oil prices versus time which shows a relatively good correlation with some time lag.

Conclusions and Final Remarks

Topic modeling enables users to have a bird's-eye view of a large volume of documents, and to examine the topics, words and documents in dynamic ways, hence provides a ground for insightful investigations. This validates the workflow as a viable means of quickly exploring large volumes of texts to understand the factors affecting the stock and commodity prices. In this study, topics extracted related to oil price fit into reasonable defined groups and the trends identified agree with the major historical events. The study also showed that the topic's importance changes with time. In addition, the unexpected connection between the oil prices and unemployment rate was uncovered. It is important to note that the correlation observed does not necessarily mean causation in any direction. Further analysis is needed. The results does, however, provide a starting direction to look for the causation.

The study can be expanded in a number of directions. Hierarchal clustering and/or multi-dimensional scaling to group topics sequentially and attempt to estimate the optimum number of topics automatically. The results can be used to study the topics at a multi-scale level depending on the objectives and time constraints. In addition, predicting the stock or commodity price can be attempted by predicting the trend of the topics and using the correlation coefficient estimated.

References

Blei, D. M., Ng, A. Y., and Jordan, M., 2003: Latent Dirichlet Allocation: Journal of Machine Learning Research, v. 3, p. 993-1022

Ng, A., 2015, Automated Biography for a Nation: website, annalyzin.wordpress.com/2015/06/04/automated-biography/, retrieved on 12/11/2015.

PetroWiki, 2015: website, petrowiki.org, retrieved on 12/05/2015.

Řehůřek, R., and Sojka, P., 2010, Software Framework for Topic Modeling with Large Corpora: in Witte, R., Cunningham, H., Patrick, J., Beisswanger, E., Buyko, E., Hahn, U., Verspoor, K., and Coden, A. R., eds., Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, University of Malta, 5 p.

Sievert, C., and Shirley, K.E., 2014, *LDAvis*: A Method for Visualizing and Interpreting Topics: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, p. 63–70. Interactive Language Learning, Visualization, and Interfaces, p. 63–70.

U.S. Bureau of Labor, 2015, website, www.bls.gov/, retrieved on 12/04/2015.

U.S. Energy Information Administration, 2015, Spot Prices for Crude Oil and Petroleum Products, website, www.eia.gov, retrieved on 10/15/2015.