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

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

Stanford
SCHOOL OF EARTH, ENERGY
& ENVIRONMENTAL SCIENCES
Energy Resources Engineering

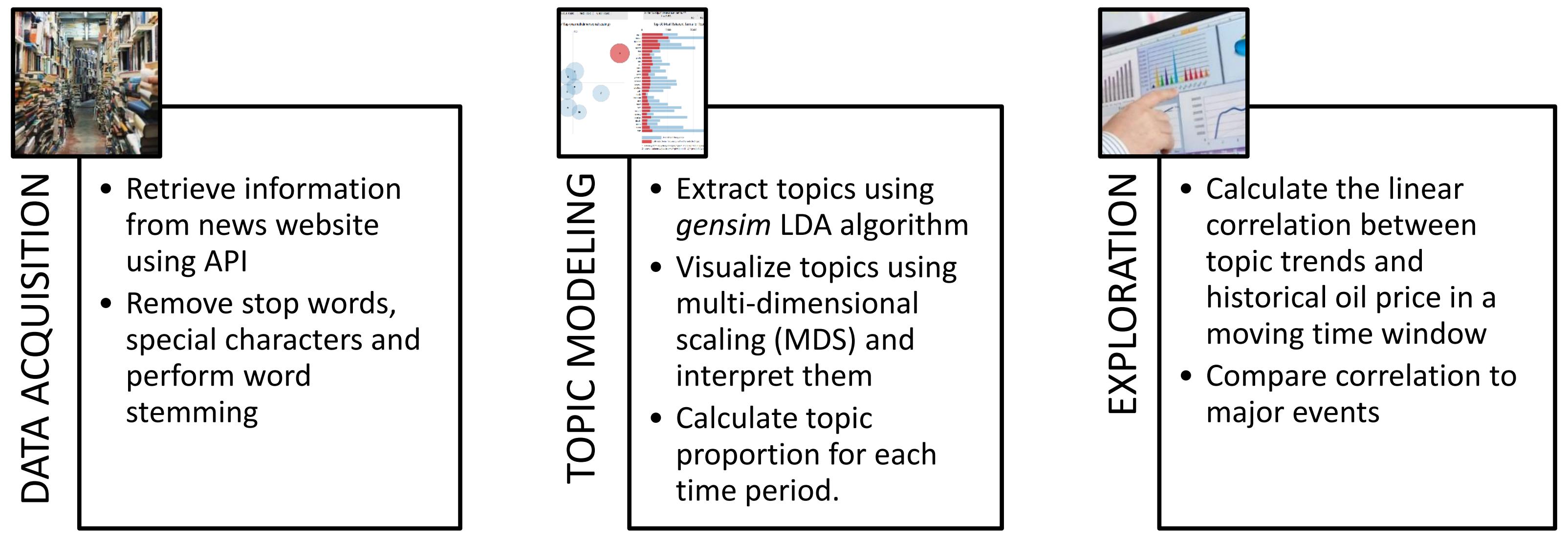
1. MOTIVATION & OBJECTIVES

Analyzing historical commodity and stock prices is a pre-requisite to investment. The procedure is time consuming and requires knowledges of many related factors, which may be hidden within a large volume of texts. This study develops a workflow to help users:

1. Summarize topics from large amount of news article and identify the relationships among topics, word occurrences and articles.
2. Explore various factors related to historical price volatility by visualizing topic trends over time with correlation analysis.

We use this workflow to reveal factors related to the crude oil price, and to show how their importance change with time.

2. GENERAL WORKFLOW



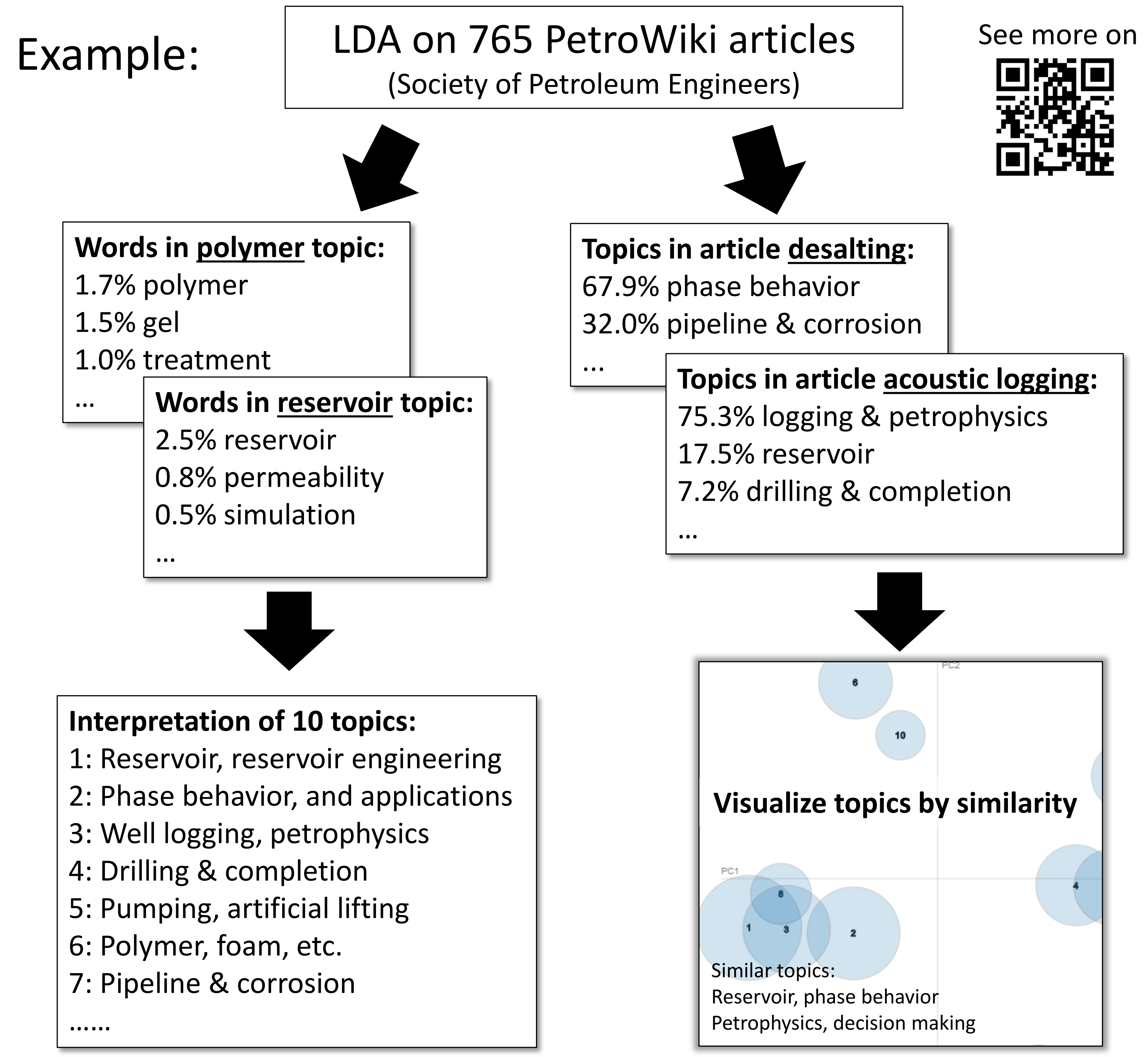
DATA ACQUISITION
- Retrieve information from news website using API
- Remove stop words, special characters and perform word stemming

TOPIC MODELING
- Extract topics using *gensim* LDA algorithm
- Visualize topics using multi-dimensional scaling (MDS) and interpret them
- Calculate topic proportion for each time period.

EXPLORATION
- Calculate the linear correlation between topic trends and historical oil price in a moving time window
- Compare correlation to major events

3. TOPIC MODELING: LATENT DIRICHLET ALLOCATION

Assumptions: Each document (bag-of-words) is a mixture of latent topics; each topic is a mixture of words. LDA uses EM algorithm to estimate the following hidden variables from many documents: **word distribution for each topic**, and **topic distribution in each document**.



Example:

LDA on 765 PetroWiki articles
(Society of Petroleum Engineers)

See more on

Words in polymer topic:
1.7% polymer
1.5% gel
1.0% treatment
…

Words in reservoir topic:
2.5% reservoir
0.8% permeability
0.5% simulation
…

Topics in article desalting:
67.9% phase behavior
32.0% pipeline & corrosion
…

Topics in article acoustic logging:
75.3% logging & petrophysics
17.5% reservoir
7.2% drilling & completion
…

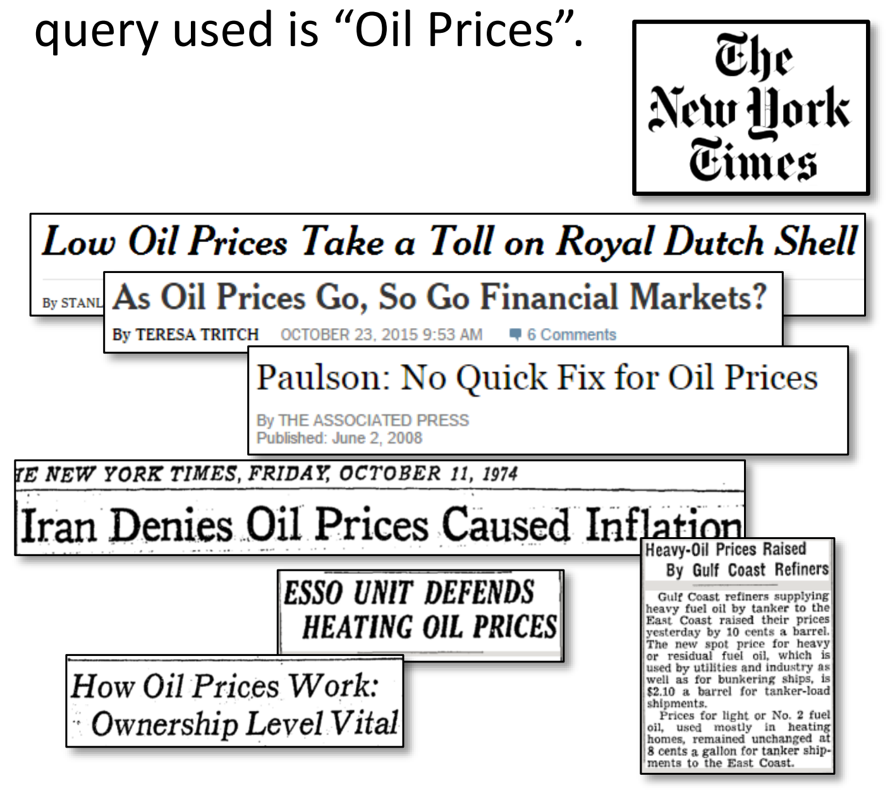
Interpretation of 10 topics:
1: Reservoir, reservoir engineering
2: Phase behavior, and applications
3: Well logging, petrophysics
4: Drilling & completion
5: Pumping, artificial lifting
6: Polymer, foam, etc.
7: Pipeline & corrosion
……

Visualize topics by similarity

Similar topics:
Reservoir, phase behavior
Petrophysics, decision making

4. ANALYSIS OF ARTICLES MENTIONING "OIL PRICES"

28,415 articles are extracted from the New York Times using application programming interface. The search query used is "Oil Prices".



The New York Times

Low Oil Prices Take a Toll on Royal Dutch Shell
By STANL

As Oil Prices Go, So Go Financial Markets?
By TERESA TRITCH OCTOBER 23, 2015 9:53 AM 6 Comments

Paulson: No Quick Fix for Oil Prices
By THE ASSOCIATED PRESS
Published: June 2, 2008

HE NEW YORK TIMES, FRIDAY, OCTOBER 11, 1974
Iran Denies Oil Prices Caused Inflation

ESSO UNIT DEFENDS HEATING OIL PRICES

How Oil Prices Work: Ownership Level Vital

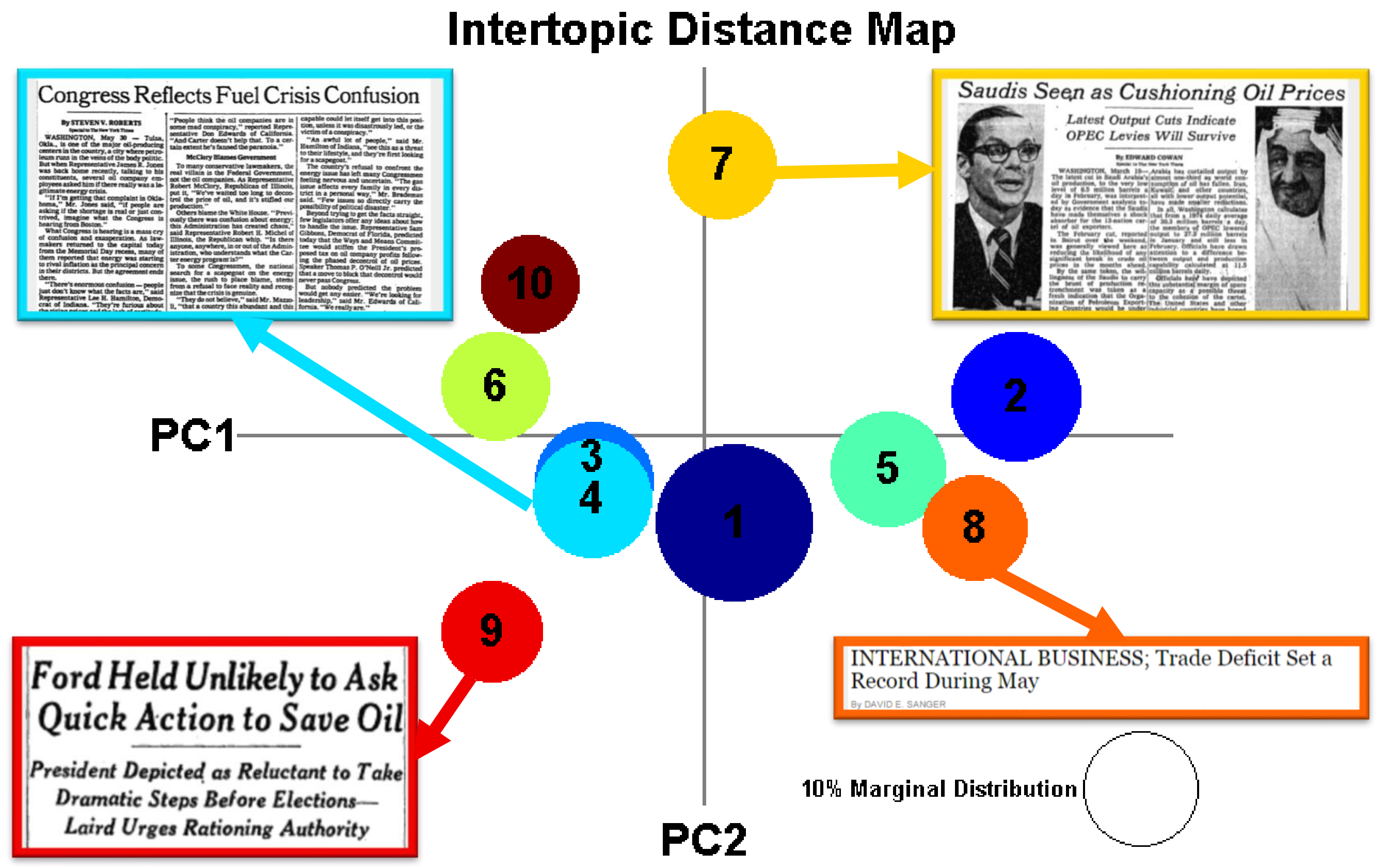
Heavy-Oil Prices Raised By Gulf Coast Refiners

Gulf Coast refiners supplying heavy fuel oil by tanker to the East Coast raised their prices yesterday by 10 cents a barrel. The new spot price for heavy or residual fuel oil, which is used by utilities and industry as well as for bunkering ships, is $2.10 a barrel for tanker-load shipments.
Prices for light or No. 2 fuel oil, used mostly in heating homes, remained unchanged at 8 cents a gallon for tanker shipments to the East Coast.

Topic Modeling

Topics are extracted using LDA. MDS is used to study relationships between topics. Topics are interpreted by studying the representative terms.

#	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-OLD OPEC Relations**	🟫



Intertopic Distance Map

Congress Reflects Fuel Crisis Confusion

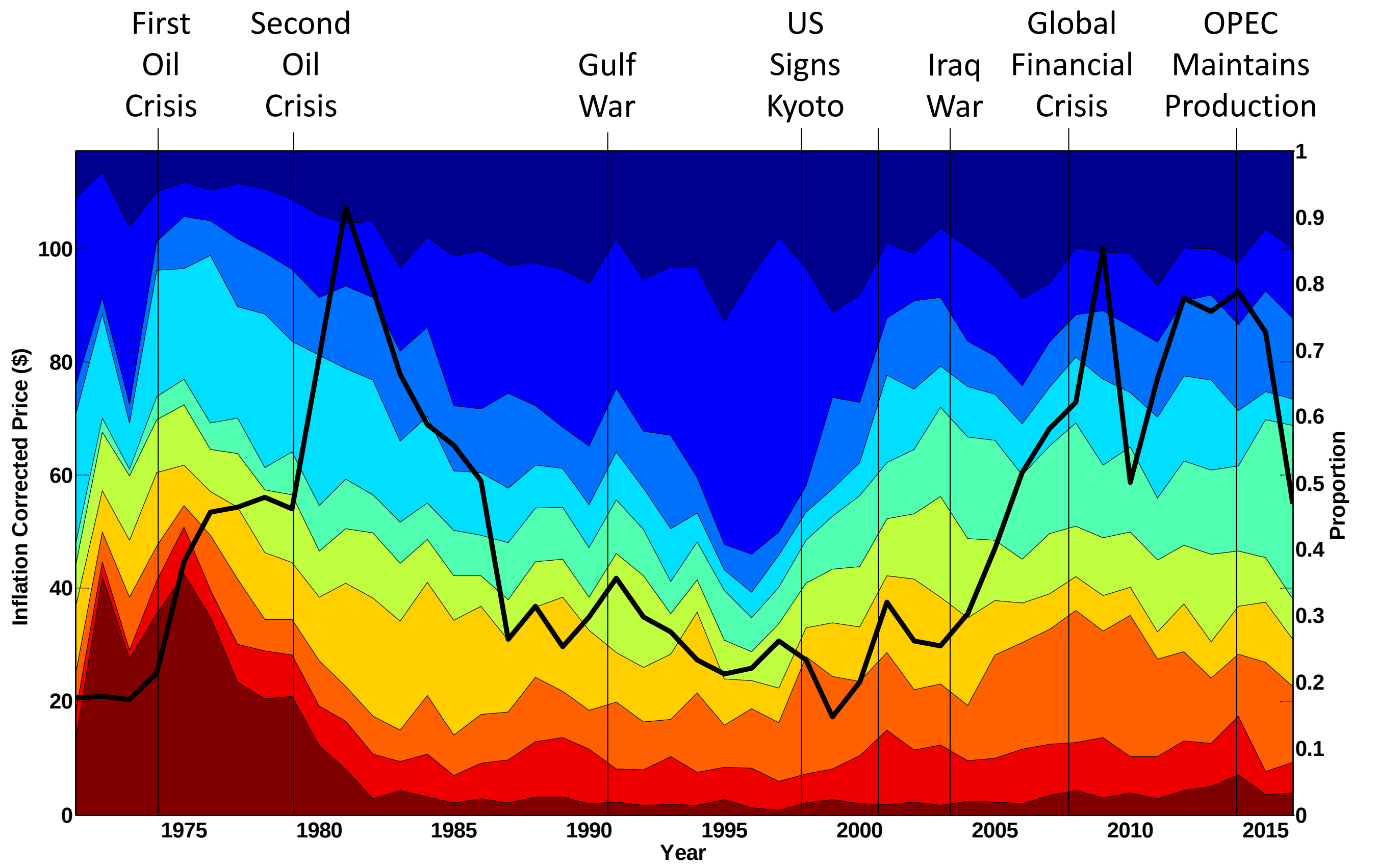
Saudis Seen as Cushioning Oil Prices
Latest Output Cuts Indicate OPEC Levies Will Survive

Ford Held Unlikely to Ask Quick Action to Save Oil
President Depicted as Reluctant to Take Dramatic Steps Before Elections— Laird Urges Rationing Authority

INTERNATIONAL BUSINESS; Trade Deficit Set a Record During May
By DAVID E. SANGER

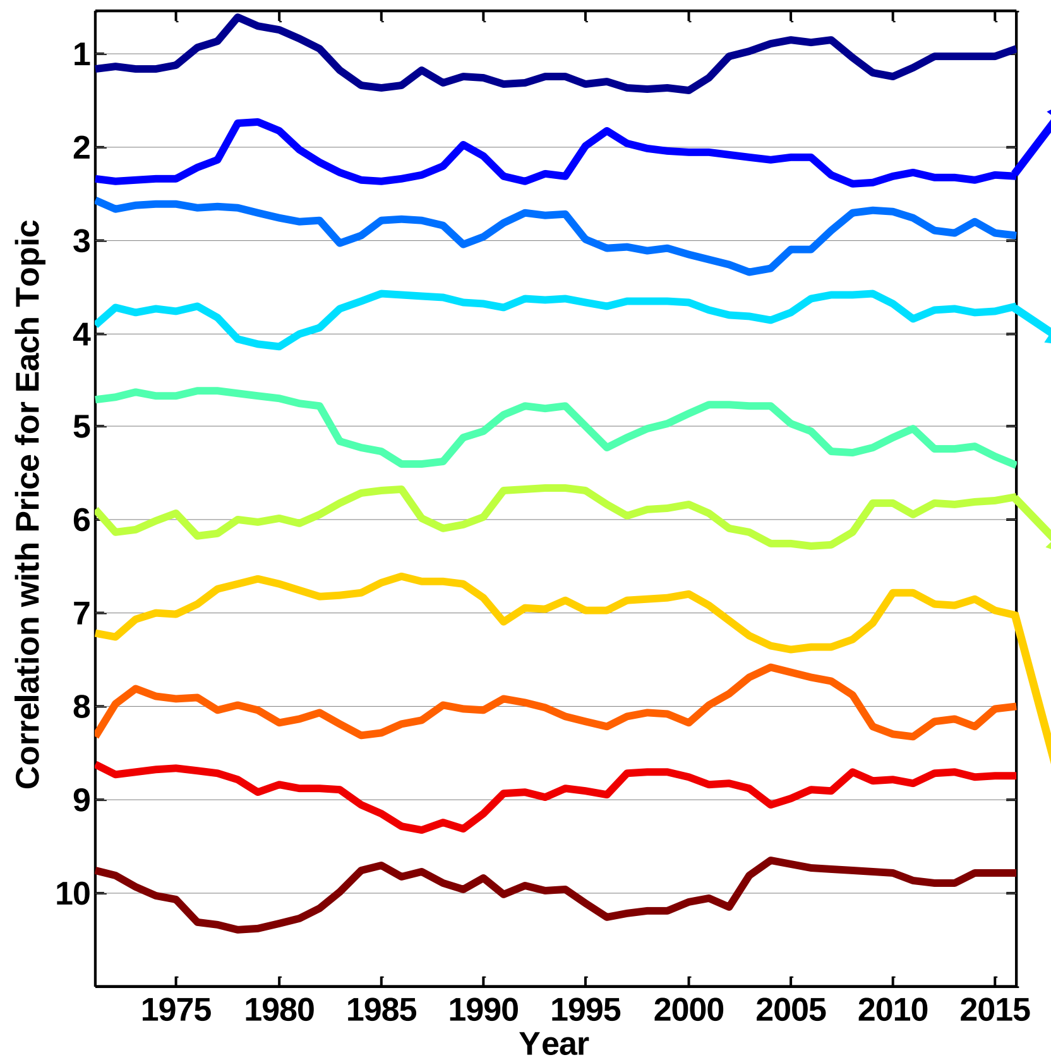
Topic trends are extracted on a yearly or monthly bases. Major events and the price are overlaid.

Some trends observed are consistent with known global events. For example, topic #6 (Middle East Conflicts) sees a relative increase in proportion during the Gulf War (1990) and the Iraq War (2003).



Topic trends are correlated with price

using a moving temporal window (9 years)



US Dollar generally has an inverse relationship with oil price

US Policies are generally in favor of oil and gas industry practices

Conflicts in the Middle East drives the oil up

OPEC over-production is currently driving down the price of oil

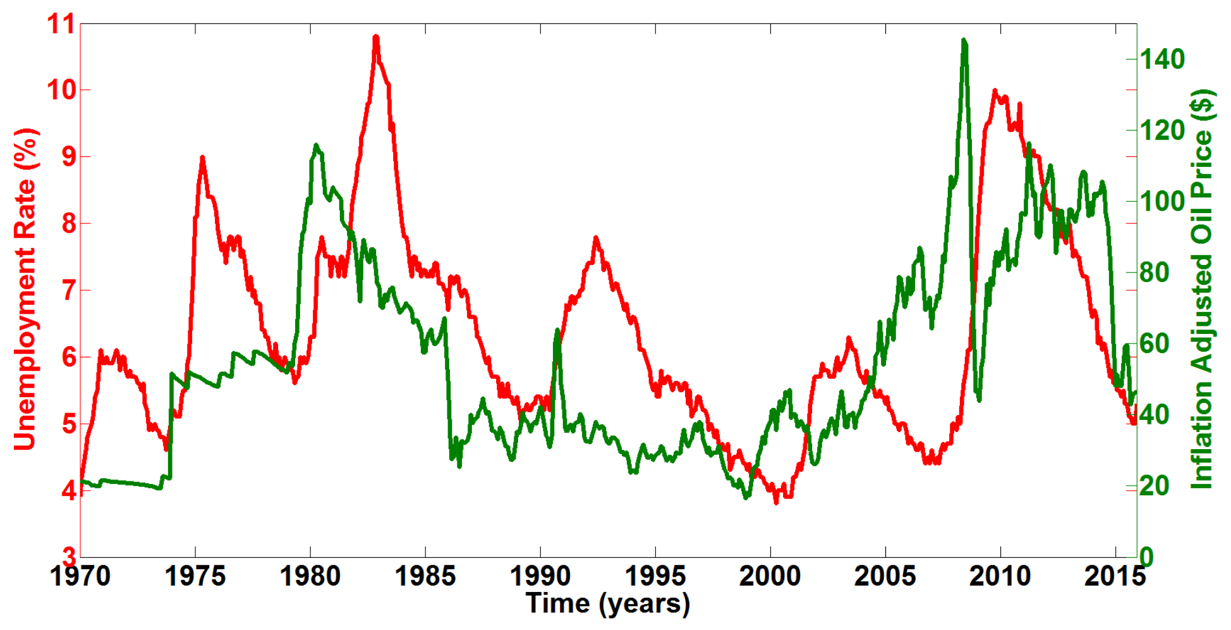
5. FACTORS FROM GENERAL ARTICLES

The workflow is applied to a collection of 338,828 NYTimes articles, randomly sampled daily from 1970 to 2015. Oil price is correlated with the trends of 50 topics.

Interestingly, we see a negative correlation between crime-related topics (36, 9, 33) and oil price, and a positive correlation for entertainment-related topics (16, 47 43, 39).

Topics sorted by absolute correlation coefficient

36. polic charg offic man kill state death arrest citi murder R= -0.58245
16. art time race work museum state make book show long R= 0.48542
41. offici nuclear soviet presid state north unit korea american militari R= 0.46802
47. american play time york state music airlin night war group R= 0.45883
4. ga record state show product cancer report unit natur nation R= 0.45041
40. kill south forc govern attack peopl presid war state offici R= 0.45022
23. game victori win run team score season night time play R= -0.4219
21. state feder plan insur health offici budget bill unit propos R= -0.41063
34. world soviet state govern news presid iran report time parti R= 0.40174
43. time televis list state work presid music sport film call R= 0.39732
39. state wine plan unit school time world american court polic R= 0.36302
38. presid death famili paid notic york die love friend servic R= -0.35746
46. state time clinton presid game nation york john plan world R= 0.34514
32. bank report earn net sale corpor tax share unit execut R= -0.34316
10. citi senat york plan hous state report mayor republican vote R= -0.33921
5. israel offici palestinian isra stock end state citi report arab R= 0.33296
13. court state rule appeal suprem judg unit justic law vietnam R= -0.32164
1. citi york build park street home hous fire peopl manhattan R= 0.31953
9. case drug charg public investig death report fund lawyer offici R= -0.31513
33. charg judg american court case stock york state kill feder R= -0.30736



The correlation between the crime rate and oil prices was investigated. Crime-related articles suggested a deep connection to unemployment rate. A plot of unemployment rate versus the oil price proved there is a trend.

With LDA and topic trend correlation with price, it is easy to delve into topics to discover uncommon interactions

6. CONCLUSION & FINAL REMARKS

Results show that the workflow is a viable means of exploring large text corpus to understand the factors affecting the oil price.

- By comparing topic trends with the oil price, known historical events associated with changes in the oil price in time were captured
- The unexpected connection between the oil prices and unemployment rate was uncovered.
- Correlation observed does not mean causation in any direction. Further analysis is needed. The results does provide a starting point to where to look for the causation.

Future directions for the study include:

- Using hierarchal clustering and/or PCA to group topics sequentially and obtain insights. This can be used to obtain the optimum number of topics automatically.
- Predicting the stock or commodity price by predicting the trend of the topics and using correlation factors observed.

7. REFERENCES

1. Blei, D. M., Ng, A. Y., and Jordan, M., 2003: Latent Dirichlet Allocation: Journal of Machine Learning Research, v. 3, p. 993-1022
2. Chen, E., 2011, Introduction to Latent Dirichlet Allocation: website http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/, retrieved on 10/16/2015.
3. Řehůřek, R., and Sojka, P., 2010, Software framework for topic modelling 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.
4. 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.
5. PetroWiki, 2015. http://http://petrowiki.org/
6. Energy Information Administration
7. U.S. Bureau of Labor, www.bls.gov