Monitoring Social Media using Machine Learning

Summer 2019

Joseph Jinn, Professor Keith VanderLinden

Calvin University 1876

Introduction

Our research is an extension of prior work by CSIRO - Commonwealth Scientific and Industrial Research Organization, Australia's national research laboratory. Our focus is on utilizing Twitter data, Tweets, as a dataset by which we measure the SLO - Social License to Operate - of various mining, gas, and oil companies. SLO is defined as the acceptability of a company's business operations by its employees, stakeholders, and the general public. The primary purpose of the summer 2019 research project is to investigate and find a methodology by which we can effectively model the topics of all the Tweets in our dataset. Topic modeling is a way of defining abstract "topics" that are prevalent in a corpus of textual documents. It is statistical in nature and is essentially unsupervised machine learning by which we attempt to cluster the Twitter data to find similarities and patterns among groups of words.

To that end, we first utilized standard data science techniques to investigate the nature of our Twitter dataset. This involves the use of the Python programming language, the Pandas data analysis library, the Matplotlib data visualization library, and other processing and visualization software. Our discoveries and results are recorded in Jupyter Notebooks – an interactive web-based application that allows researchers to easily share code, equations, visualizations, and text. We also utilize Scikit-Learn, a machine learning software suite, and Gensim, a topic modeling software suite, along with various 3rd party libraries, to implement baseline topic models from which we can begin to investigate how to best extract relevant topics from the Tweet texts.

Objectives

1. Construct a dataset processor to extract/derive Tweet attribute fields we deem relevant from the raw JSON dataset file and convert/output to CSV file format.
2. Perform data analysis on the Twitter CSV dataset file to determine the nature of our data.
3. Use data analysis results to determine our approach to implementation of Natural Language Processing techniques on the Tweets in our dataset.
4. Pre-process, post-process, and tokenize Tweet text to prep for use as the input feature for baseline topic modeling algorithms.
5. Implement baseline topic modeling algorithms using our tokenized Twitter data to perform topic extraction on all 650k+ Tweets.
6. Analyze topic extraction results to infer any visible patterns among the top N words associated with each topic.
7. Visualize topic extraction results using pyldAVIS topic modeling visualization library.
8. Attempt to understand the statistical and mathematical construct behind each topic modeling algorithm to better infer which approach is best for our data.
9. Create a modified baseline topic modeling algorithm that generates more coherent topics and associated words.

Results - Topic Extraction

Latent Dirichlet Allocation

Topic 0: work court risk world group port native live title coal
Topic 1: want good coalmine claim profit finance clear gov voter poll
Topic 2: gas time coal know oppose thank industry narrabri think ask
Topic 3: project fund reef climate coal change wo barrier kill watch
Topic 4: fortescue protest cost local pm forest fine coal financial hand
Topic 5: support fight right join wrong campaign demand planet country happen
Topic 6: need farmer deal cut rule coal premier run royalty investment
Topic 7: water australian future land question open coal pollution morning concern
Topic 8: labor coal rail energy slo_cashn line win shorten election stop
Topic 9: stop billion public action destroy dollar naif sign national party
Topic 10: plan say business big council use end law year production
Topic 11: new loan govt coal bank price giant make support long
Topic 12: slo_mention tell help leave start environment talk massive away turn
Topic 13: job government turnbull lnp look 000 year way lose create
Topic 14: people queensland power let coal alp community thing stand close
Topic 15: coal ½i build ½i² ¼i galilee ahead basin approve china
Topic 16: company state iron indian ore news high lie face carbon
Topic 17: tax pay break approval federal ceo subsidy issue slo_cash miner
Topic 18: india money report come taxpayer point day barnaby joyce week
Topic 19: green vote minister share sell canavan matt promise buy read

Author-Topic

Label: 1 -> Words: leard go look maules work think well get have property
Label: 2 -> Words: project road beach news win rd downer ceo 's peter
Label: 3 -> Words: gas coal farmer seam water stop community people want protest
Label: 4 -> Words: creek write 's wa coal | project company timor plan
Label: 5 -> Words: gas coal forest petroleum tycoon energy price oil sale whc
Label: 6 -> Words: tax joyce barnaby slo_cashil pay chevron liberal rail coal origin
Label: 7 -> Words: narrabri water gas risk basin coal national artesian > farmer
Label: 8 -> Words: í ° ½í² beach field ¼í¶□ tour $ win home
Label: 9 -> Words: eis land coal water inland go seam want appliance fracking
Label: 10 -> Words: $ wpl sto video sire share field day gain result

Hierarchical Dirichlet Process

(0, "0.025*coal + 0.008*job + 0.007*'s + 0.006*project + 0.006*stop + 0.006*want + 0.006*labor + 0.005*water + 0.005*fund + 0.005*loan")
(1, '0.082*í + 0.050*° + 0.043*tax + 0.041*½í² + 0.023*¼í¶\x93 + 0.016*pay + 0.013*$ + 0.012*coal + 0.010*energy + 0.008*½í')
(2, '0.170*$ + 0.009*cba + 0.009*anz + 0.009*wbc + 0.009*nab + 0.007*coal + 0.006*price + 0.005*fmg + 0.005*bxb + 0.005*syd')
(3, "0.016*coal + 0.005*'s + 0.005*job + 0.005*$ + 0.004*stop + 0.004*project + 0.004*want + 0.004*í + 0.003*fund + 0.003*reef")
(4, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*project + 0.004*stop + 0.004*want + 0.003*fund + 0.003*gas + 0.003*labor")
(5, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*project + 0.004*stop + 0.004*want + 0.003*gas + 0.003*water + 0.003*fund")
(6, "0.015*coal + 0.005*$ + 0.005*job + 0.005*'s + 0.004*stop + 0.004*project + 0.004*want + 0.003*gas + 0.003*fund + 0.003*support")
(7, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*project + 0.004*stop + 0.004*want + 0.003*gas + 0.003*fund + 0.003*labor")
(8, "0.015*coal + 0.005*job + 0.004*'s + 0.004*$ + 0.004*stop + 0.004*project + 0.003*want + 0.003*gas + 0.003*fund + 0.003*support")
(9, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*project + 0.004*stop + 0.004*want + 0.003*water + 0.003*gas + 0.003*fund")
(10, "0.015*coal + 0.005*'s + 0.005*job + 0.004*$ + 0.004*stop + 0.004*project + 0.004*argyle + 0.004*want + 0.003*gas + 0.003*fund")
(11, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*project + 0.004*stop + 0.004*want + 0.003*í + 0.003*fund + 0.003*gas")
(12, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*project + 0.004*stop + 0.004*want + 0.003*gas + 0.003*fund + 0.003*labor")
(13, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*stop + 0.004*project + 0.004*want + 0.003*gas + 0.003*fund + 0.003*water")
(14, "0.015*coal + 0.005*job + 0.005*'s + 0.004*$ + 0.004*stop + 0.004*project + 0.004*want + 0.004*fund + 0.003*gas + 0.003*labor")
(15, "0.015*coal + 0.005*job + 0.005*'s + 0.004*project + 0.004*stop + 0.004*$ + 0.004*want + 0.004*gas + 0.003*water + 0.003*fund")

Non-Negative Matrix Factorization

Topic 0: coal build away indian massive environment india mean minister power
Topic 1: tell say full wrong way ask make question happen try issue
Topic 2: job 000 10 create claim lie thousand tourism renewable pm
Topic 3: australian company oil face govt financial use fail write set
Topic 4: project gas narrabri land forest farmer seam sign community field
Topic 5: stop tax pay profit million corporate haven office rate island
Topic 6: time thank good come late stand start leave week long
Topic 7: reef barrier fight help kill destroy risk protect let save
Topic 8: labor support green vote lnp win election shorten alp party
Topic 9: ½i beach ¼i read love ½i² slo_hash letter oh ¾i
Topic 10: want people know billion future dollar listen industry mega video
Topic 11: day action protest court group right join wrong native meet

Biterm

Topic 0 | Coherence=-146.13 | Top words= coal stop labor want climate support project tax new shorten
Topic 1 | Coherence=-145.24 | Top words= support stop green labor people climate change queensland government oppose
Topic 2 | Coherence=-134.33 | Top words= rail line basin water galilee land support project farmer gas
Topic 3 | Coherence=-183.76 | Top words= coal iron ore fortescue year price climate people need stop
Topic 4 | Coherence=-147.70 | Top words= coal power energy price new india renewable solar year plant
Topic 5 | Coherence=-155.60 | Top words= coal job reef project barrier gas 000 create time stop
Topic 6 | Coherence=-156.48 | Top words= gas coal farmer project want people water govt pipeline tell
Topic 7 | Coherence=-157.07 | Top words= coal water queensland loan point fund new time year line
Topic 8 | Coherence=-166.53 | Top words= coal gas group people seam want plan water climate north
Topic 9 | Coherence=-155.40 | Top words= coal india power company stop need year indian plan report
Topic 10 | Coherence=-196.49 | Top words= tax pay title native company loan energy slo_cash government island
Topic 11 | Coherence=-152.78 | Top words= job coal court people 000 labor 10 time reef queensland
Topic 12 | Coherence=-132.85 | Top words= barnaby joyce gas coal project taxpayer money rail new india
Topic 13 | Coherence=-161.65 | Top words= job fund want project bank people work tell good public

Processes

Twitter API

- Tweet Attributes
- JSON file format
- UTF-8 Encoding

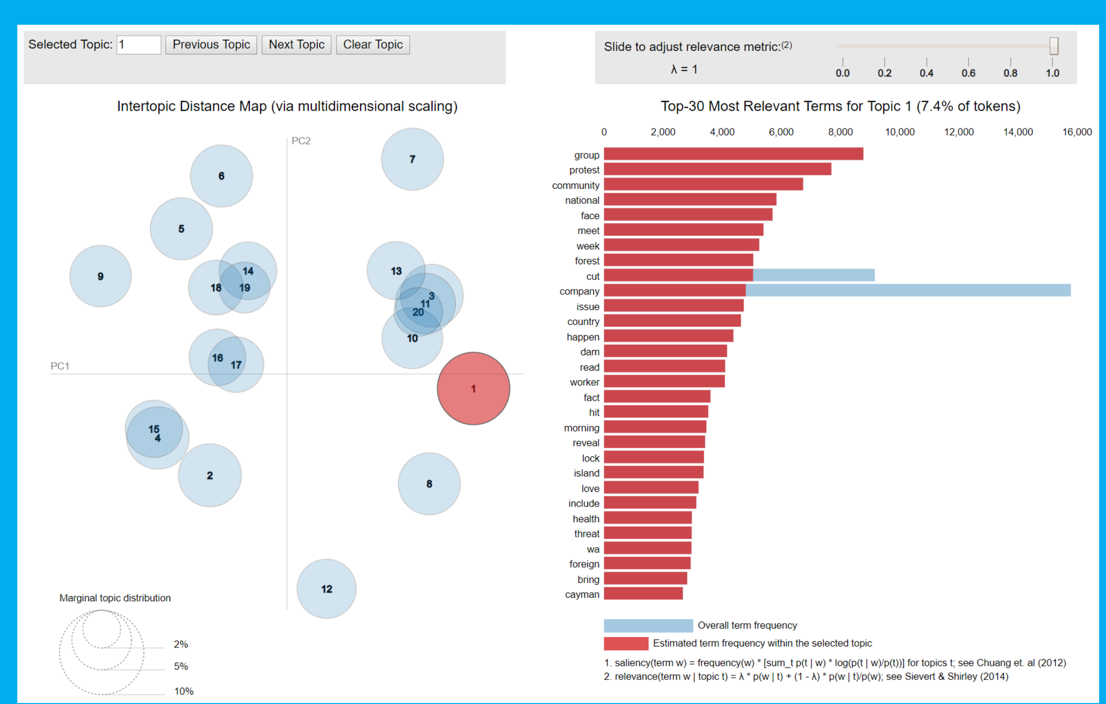
Raw Tweet Dataset

- Determine company
- Determine language
- Drop no-Company Tweets
- Drop non-English Tweets
- CSV file format

Dataset

- Preprocessing
- Postprocessing
- Tokenization
- Topic Extraction

Data Visualization



Conclusion

Baseline topic modeling algorithm libraries do not work well on our dataset. Latent Dirichlet Allocation and all derivatives of this algorithm work best on cor-pora containing documents that are long and written in a formal grammatical style. Tweets suffer from limited character length, grammatical inconsistency, and Twit-ter-specific linguistic elements, which makes it difficult to extract coherent topics.

Hyperparameter tuning may improve results to some extent but would be a time-consuming and ex-haustive process. Biterm execution runs take almost half a day per. Hierarchical LDA suffers from RAM overflow issues due to its recursive nature. Utilization of Calvin College's Borg Supercomputer could expe-dite matters but would require parallelization of our codebase and the construction of a Singularity contain-er.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (March 2003), 993-1022.

David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2003. Hierarchical topic models and the nested chinese restaurant process. In Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'03), S. Thrun, L. K. Saul, and B. Schölkopf (Eds.). MIT Press, Cambridge, MA, USA, 17-24.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI '04). AUAI Press, Arlington, Virginia, United States, 487-494.

R. Zhao and V. Y. F. Tan, "Online Nonnegative Matrix Factorization With Outliers," in IEEE Transactions on Signal Processing, vol. 65, no. 3, pp. 555-570, 1 Feb.1, 2017.
doi: 10.1109/TSP.2016.2620967

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). ACM, New York, NY, USA, 1445-1456. DOI: https://doi.org/10.1145/2488388.2488514

Yee Whye Teh, Michael I Jordan, Matthew J Beal & David M Blei (2006) Hierarchical Dirichlet Processes, Journal of the American Statistical Association, 101:476, 1566-1581, DOI: 10.1198/016214506000000302

Future Work

Our plans are to continue Objectives 8 and 9. We do not have a full grasp of the statistical and mathematical construct behind each topic modeling algorithm. This understanding will be essential to any attempt to create a modified baseline topic modeling algorithm that will hopefully improve topic extraction results on our Twitter dataset. It is our hope that we can minimize model perplexity while maximizing topic coherence metric values.

We also hope to obtain an updated Twitter dataset with more recent Tweets from CSIRO.