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.

Processes

Twitter API

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

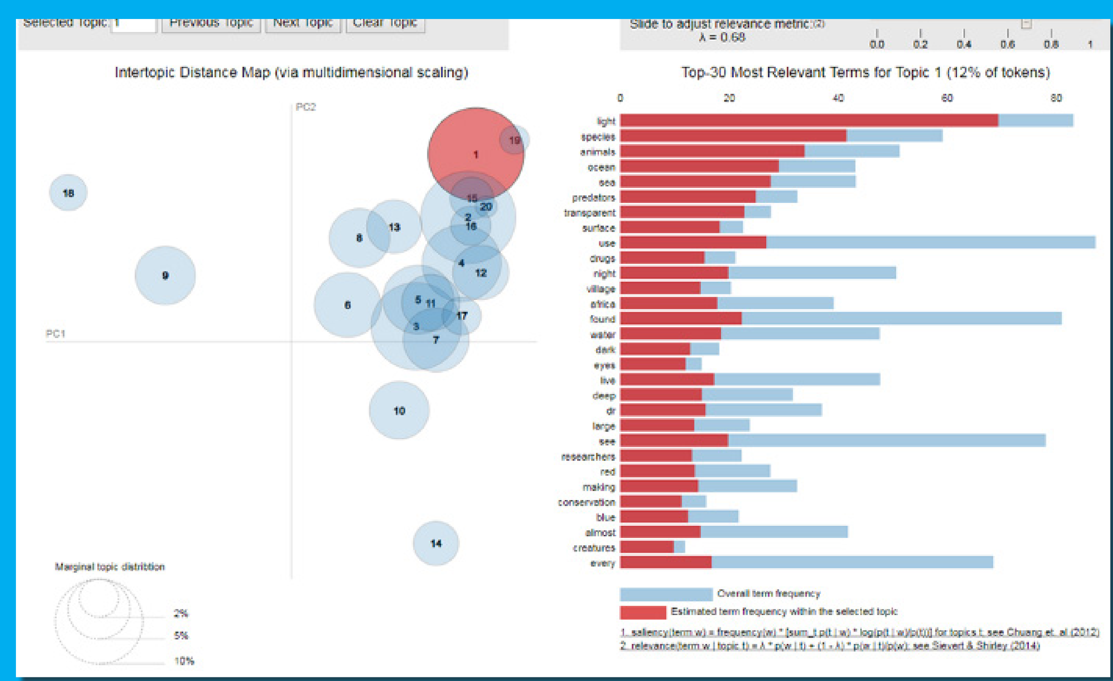
Raw Tweet Dataset

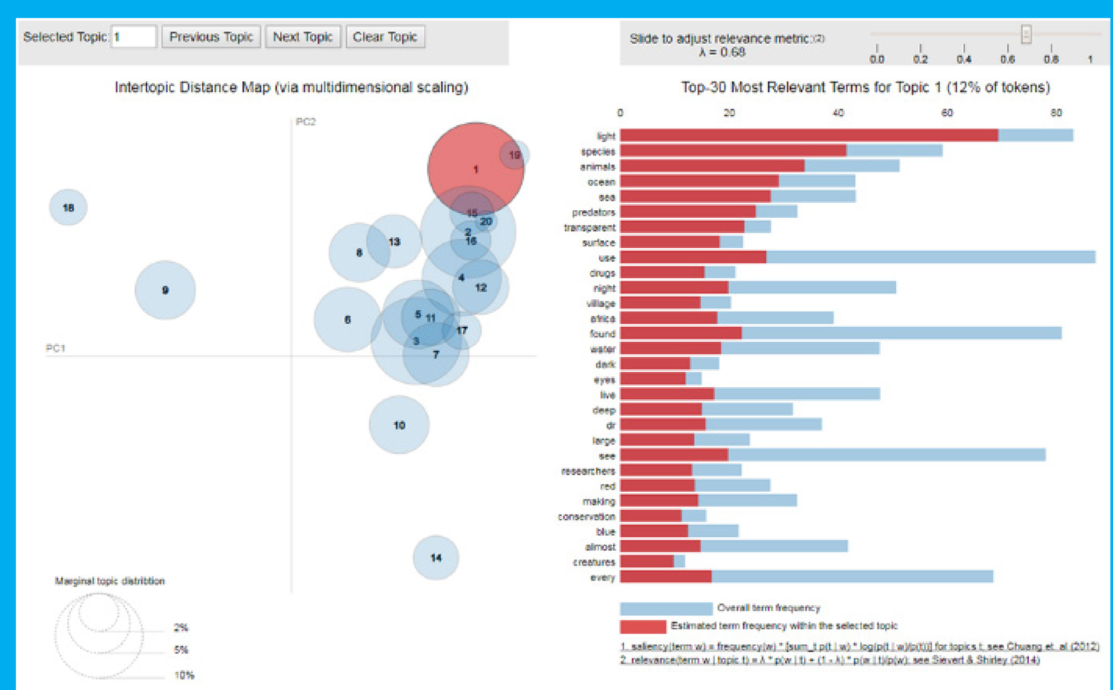
- Determine company
- Determine language
- Drop no-Company Tweets
- Drop non-English Tweets
- CSV file format

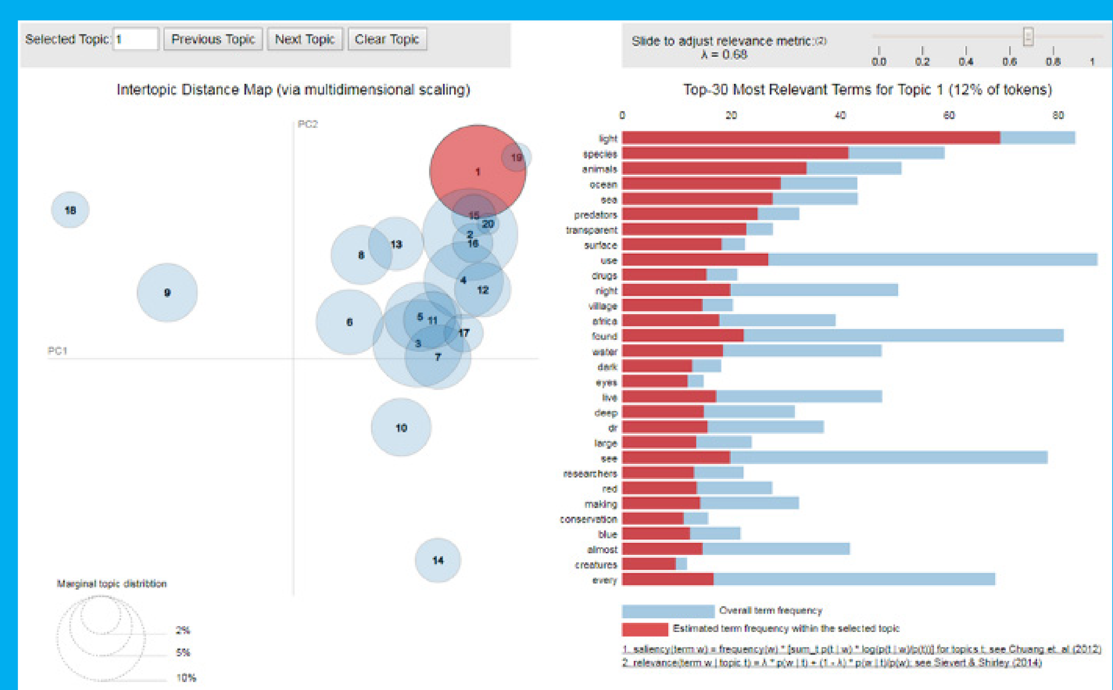
Dataset

- Preprocessing
- Postprocessing
- Tokenization

Data Visualizations







Results - Topic Extraction

Latent Dirichlet Allocation

Topic 0: adani coal reef great point loan barrier abbot fund turnbull
Topic 1: adani council environmental india company townsville corruption question know need
Topic 2: santos woodside bhp whitehaven beach oil dam year day iluka
Topic 3: adani labor support qld election stop vote loan shorten want
Topic 4: adani coal stop court people whitehaven action protest native carmichael
Topic 5: santos gas nsw water pilliga project narrabri coal csg seam
Topic 6: adani coal carmichael fund bank project power india new australian
Topic 7: bhp rio tinto billiton fortescue iron ore price share cut
Topic 8: climate coal change australia future energy need clean time world
Topic 9: job adani 000 coal create canavan 10 matt think lie
Topic 10: tax ½í pay adani ½í² bhp ¼í company slo_cash australia
Topic 11: adani water coal basin qld galilee great queensland rail free

Time taken to process dataset: 757.5309031009674 seconds, 12.62551505168279 minutes, 0.21042525086137984 hours.

Author-Topic

Label: 1
Words: project written woodside santos eis north downer group coal .slo_mention
Label: 2
Words: tax joyce barnaby inland woodside slo_cashil pay property chevron go
Label: 3
Words: gas coal seam water rd stop farmer people time pipeline
Label: 4
Words: woodside coal new 's energy | coleman wa project cfs
Label: 5
Words: santos nsw gas great australia water narrabri pilliga basin pipeline
Label: 6
Words: beach road win 1 tour day > video morning park
Label: 7
Words: santos route have party breed be right well v like
Label: 8
Words: woodside whitehaven lng coal tycoon rio oil price gas sale
Label: 9
Words: í whitehaven $ ° ½í² ¼í¶□ sto james whc ltd
Label: 10
Words: santos forest coal gas narrabri water petroleum csg creek field

Time taken to process dataset: 9772.976831436157 seconds, 162.8829471906026 minutes, 2.7147157865100433 hours.

Biterm

Topic coherence ..
Topic 0 | Coherence=-176.49 | Top words= adani coal bhp qld santos job australian labor rio gas
Topic 1 | Coherence=-146.07 | Top words= adani coal need climate australia job queensland new labor build
Topic 2 | Coherence=-124.21 | Top words= adani labor qld coal stop want fund support project need
Topic 3 | Coherence=-135.82 | Top words= adani coal loan point abbot government new slo_cashn turnbull project
Topic 4 | Coherence=-145.06 | Top words= adani australian job want oppose project loan new people govt
Topic 5 | Coherence=-126.36 | Top words= adani reef coal great barrier australia want need australian project
Topic 6 | Coherence=-129.98 | Top words= adani coal stop project australia carmichael labor need fund want
Topic 7 | Coherence=-150.89 | Top words= adani coal carmichael court title native federal queensland fund new
Topic 8 | Coherence=-160.12 | Top words= adani santos water coal stop people land whitehaven want farmer
Topic 9 | Coherence=-165.47 | Top words= santos bhp rio new day tinto nsw water gas project
Topic 10 | Coherence=-124.98 | Top words= santos gas coal nsw project pilliga seam narrabri csg farmer
Topic 11 | Coherence=-145.88 | Top words= coal water qld reef climate project money stop loan public
Topic 12 | Coherence=-142.86 | Top words= adani coal loan fund want rail australia qld line veto
Topic 13 | Coherence=-139.23 | Top words= bhp rio tinto tax australia australian ore iron billiton fortescue
Topic 14 | Coherence=-122.33 | Top words= water adani santos basin great artesian nsw risk support australia
Topic 15 | Coherence=-147.24 | Top words= bhp adani tax pay coal company billiton australia year cut
Topic 16 | Coherence=-150.63 | Top words= adani coal job 000 reef create 10 australian queensland kill
Topic 17 | Coherence=-128.56 | Top words= adani joyce barnaby india money coal taxpayer think rail spend
Topic 18 | Coherence=-176.93 | Top words= adani australia santos tax energy pay year woodside people action
Topic 19 | Coherence=-119.81 | Top words= gas field land barnaby narrabri propose nsw inland near joyce

Time taken to process dataset: 40567.238966464996 seconds, 676.1206494410833 minutes, 11.268677490684722 hours.

Hierarchical Dirichlet Process

(0, "0.044*adani + 0.022*coal + 0.008*santos + 0.007*job + 0.007*'s + 0.006*bhp + 0.006*project + 0.006*stop + 0.005*australia + 0.005*want")
(1, '0.076*í + 0.048*° + 0.042*tax + 0.040*½í² + 0.022*¼í¶\x93 + 0.021*adani + 0.016*pay + 0.013*$ + 0.011*coal + 0.010*energy')
(2, '0.144*$ + 0.027*bhp + 0.013*rio + 0.012*adani + 0.008*cba + 0.008*anz + 0.007*wbc + 0.007*nab + 0.006*coal + 0.006*price')
(3, "0.028*adani + 0.013*coal + 0.009*santos + 0.007*bhp + 0.006*rio + 0.005*australia + 0.005*tinto + 0.004*'s + 0.004*job + 0.004*$")
(4, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*job + 0.004*'s + 0.004*australia + 0.004*rio + 0.004*$ + 0.004*stop")
(5, "0.029*adani + 0.014*coal + 0.009*santos + 0.007*bhp + 0.004*australia + 0.004*'s + 0.004*job + 0.004*rio + 0.004*project + 0.004*$")
(6, "0.026*adani + 0.013*coal + 0.011*bhp + 0.007*santos + 0.007*rio + 0.006*tinto + 0.004*$ + 0.004*australia + 0.004*job + 0.004*'s")
(7, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*'s + 0.004*job + 0.004*australia + 0.004*stop + 0.004*rio + 0.004*$")
(8, "0.028*adani + 0.014*coal + 0.008*santos + 0.008*bhp + 0.004*rio + 0.004*'s + 0.004*australia + 0.004*job + 0.004*$ + 0.004*project")
(9, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*'s + 0.004*job + 0.004*australia + 0.004*rio + 0.004*$ + 0.004*stop")
(10, "0.029*adani + 0.014*coal + 0.008*santos + 0.008*bhp + 0.004*$ + 0.004*'s + 0.004*rio + 0.004*australia + 0.004*job + 0.004*project")
(11, "0.029*adani + 0.014*coal + 0.009*santos + 0.007*bhp + 0.004*'s + 0.004*job + 0.004*australia + 0.004*rio + 0.004*stop + 0.004*project")
(12, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*'s + 0.004*job + 0.004*australia + 0.004*rio + 0.004*$ + 0.004*project")
(13, "0.029*adani + 0.014*coal + 0.008*santos + 0.008*bhp + 0.004*rio + 0.004*'s + 0.004*australia + 0.004*job + 0.004*$ + 0.004*project")
(14, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*'s + 0.004*australia + 0.004*job + 0.004*rio + 0.004*$ + 0.004*project")
(15, "0.028*adani + 0.014*coal + 0.008*santos + 0.008*bhp + 0.007*$ + 0.005*í + 0.004*rio + 0.004*australia + 0.004*job + 0.004*'s")
(16, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*$ + 0.004*'s + 0.004*job + 0.004*australia + 0.004*rio + 0.004*project")
(17, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*'s + 0.004*job + 0.004*australia + 0.004*rio + 0.004*$ + 0.004*project")
(18, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.004*'s + 0.004*job + 0.004*australia + 0.004*rio + 0.004*$ + 0.004*stop")
(19, "0.029*adani + 0.014*coal + 0.008*santos + 0.007*bhp + 0.005*'s + 0.004*job + 0.004*australia + 0.004*rio + 0.004*$ + 0.004*project")

Time taken to process dataset: 1077.2826988697052 seconds, 17.95471164782842 minutes, 0.2992451941304737 hours.

Non-Negative Matrix Factorization

Topics using generalized Kullback-Leibler divergence:
Topic 0:
adani loan government minister turnbull india canavan face ahead question
Topic 1:
rio tinto iron ore bhp business close new mining fall
Topic 2:
santos gas nsw pilliga narrabri csg barnaby forest farmer water
Topic 3:
bhp billiton ceo dam disaster loss cut brazil boss asx
Topic 4:
coal new india build power open port dirty solar giant
Topic 5:
job 000 destroy create 10 lie real pm tourism claim
Topic 6:
great reef barrier help fight right world join kill basin
Topic 7:
stop labor election lnp vote greens win alp qld shorten
Topic 8:
australia energy future big clean industry thing demand planet fossil
Topic 9:
adani carmichael court coalmine approval land native challenge title owner
Topic 10:
pay tax company money billion cut dollar indian million billionaire

Time taken to process dataset: 306.614928483963 seconds, 5.11024880806605 minutes, 0.0851708134677675 hours.

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 corpora containing documents that are long and written in a formal grammatical style. Tweets suffer from limited character length, grammatical inconsistency, and Twitter 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 exhaustive 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 expedite matters but would require parallelization of our codebase and the construction of a Singularity container.

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.