A Clustering Tool for Petal

Brian A. Whiteaker[1]
University of California at San Diego, La Jolla, CA, 92093

Vikram Shyam[2]
NASA Glenn Research Center, Cleveland, OH, 44135

In this work we construct a topic modeling tool for the purpose of providing insights from biology to the engineer. We apply the machine learning text mining tools- latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF) with Kullback-Leibler divergence, to provide topic clusters to the user. Topic clusters are the underlying themes latent to a paper. For the text modeling problem, NMF-KL is the equivalent of probabilistic latent semantic analysis. Both LDA and NMF-KL are top performing modeling tools. We use these tools to identify biological specimens relevant to the user. Various organisms solve a particular survival problem in nature differently. The topic clusters allow the laymen to find these cross-topic themes in the body of documents and then branch out and examine papers whose target organisms solve a the engineers target problem.

Keywords – machine learning, lda, nmf, topic modeling, biomimicry, clustering

I. Introduction

This clustering tool is an addition to the Periodic Table of Life (PeTaL) developed at NASA Glenn Research Center. The goal is to provide tools for bio-inspired design, also called biomimetics or biomimicry. Human engineers are limited by the bounds of human imagination and experience. Nature is not similarly bound, and further, has had millions of years to explore solutions. This results in many highly efficient answers available when they can be discerned by a biologist.

Yet, how do we connect the engineer's search for inspired solutions to the domain knowledge of the biologists? The full body of knowledge from both fields is daunting. PeTaL aims to bridge that gap by providing a framework the engineer can approach to begin searching for useful instances. In support of this, the clustering tool is targeted on transferring findings held in the body of research documents, to the engineer.

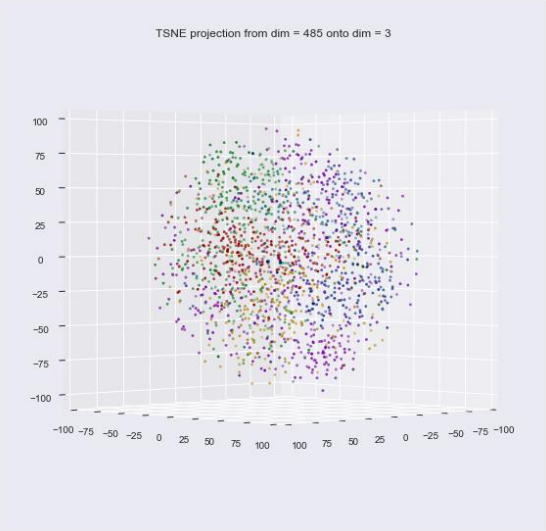

Figure 1. TSNE Topic Clusters projection. *Visualizing the documents as projected downward onto three dimensions.*

To transfer these findings we employ tools from machine learning called topic modelers. Topic modelers posit that any document is composed from a set of latent topics. These topics may be thought of as a set of words, or vocabulary, which describe a topic. The frequency of various words in a topic forms a distribution over the set of words. The sampling from these distributions generates the document.

Now consider a collection of thousands of documents, also called a corpus. How would one find the topics in this corpus? A topic modeler proceeds by first examining all of the documents in the corpus and forming a vocabulary.

[1] Intern, GVIZ, NASA Glenn Research Center, University of California at San Diego.

This vocabulary contains all words appearing across the entire corpus. A vocabulary can be many thousands of words. Next, documents are examined and word frequency counts are calculated for each term in the corpus appearing in any document. One would now be able to create a vector of counts for this document. This process would be repeated for each paper in the corpus with the resulting vectors then combined into a document-term matrix (dt).

Obtaining the dt-matrix gives us a mathematical representation on which we apply the topic modelers LDA and NMF. LDA is considered a soft clustering algorithm. This is due to the probabilistic nature of LDA. Words of the vocabulary may be shared across the different topics, with different probability. A word may strongly define a particular topic while playing a supporting role in another. In this way, LDA allows for some overlap among topics. A hard clusterer such as K-Means would look for a clear partition between the topic clusters occurring in a corpus.

Figure 2. Topics as distributions over words. *Topics are distributions over word counts of the vocabulary. Here we take plot the word features with feature selection p-val < 0.1*

For the field of topic modeling there are a few metrics that have been created to attempt to quantify the quality and usefulness of modeled topics. By and large the literature points to a high level of subjectivity. The goal of creating a general tool for unknown text datasets places certain constraints on design. We avoid biasing the model to the characteristics of a specific text corpus. However, we do offer a method for attempting to maximize the useful information achievable with the tools at hand.

II. Methods

The PeTaL project presents an interesting problem regarding generality. Many times, text mining is employed in a somewhat targeted fashion. There may be sentiment analysis required as in good or bad product reviews. A text mining team would form a specific vocabulary including slang terms to recognize positive or negative sentiment. In such a case, there is specific end goal.

For PeTal we require a general tool that will target a general question the laymen has, and bring them to specific results. The tool must be plug-and-play for any user provided corpus. To achieve this end we employ general preprocessing techniques, workhorse algorithms NMF with Kullback-Leibler loss, LDA, and recursive user interaction. The preprocessing acts while filtering text to its informative words simultaneously reducing the dimensionality of the document vector.

The topic modelers each have differences that produce qualitatively different topic clusters. By presenting topic

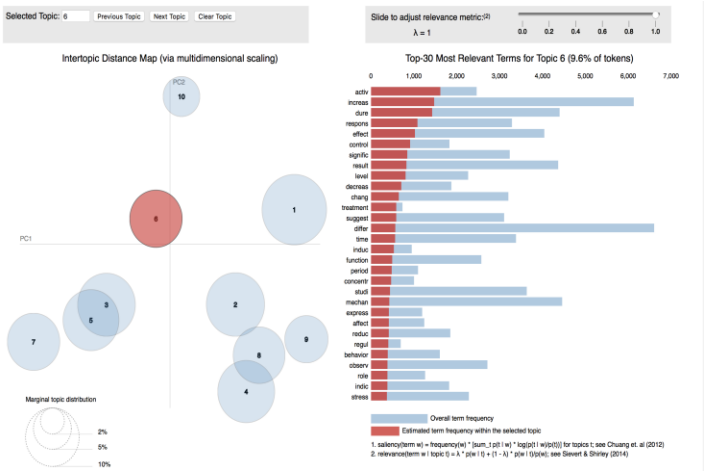

Figure 3. pyLDAvis topic choice 6. *Here is an initial launch of the cluster tool. The relevance slider is set to 1, presenting the top 30 relevance terms according to the relevance metric*[3]

clusters from both models we increase the possible selections leading to useful documents. At each level we present the topic clusters found. The user can then examine the topics and identify promising candidates. Then selections are made which are connected to the associated documents. The set union is taken to form a new sub-corpus for re-clustering upon. In this way we can recursively subset by topics of interest at each level and "zoom in" to the documents holding useful biological examples.

For the analysis of our clusters we focus our quality measures for topics the tool generate's on the results from LDA. The tools developed for use with LDA have been have proven the most valuable for examining topics. Currently there is a certain amount of subjectivity with topic models, more specificity would require solutions engineered towards a particular type dataset- built with domain knowledge. We are building a general purpose modeler and so the choice of topic modelers is for robustness, range of topics, and quality of topics.

A. Latent Dirichlet Allocation

LDA is what is known as a soft clustering topic modeler. The clusters it provides are the result of a generative statistical modeling process that treats each topic as a probability distribution. The documents of the corpus are each created by sampling from some mixture of the topics. It is an analogue to Gaussian mixture models. Here, Dirichlet priors as are the counterpart of Gaussians. The model attempts to reconstruct the likelihood of a set of words occurring together to form a topic via the assumption of Dirichlet priors. Various sampling methods are employed in the modeling process, for a deeper treatment see the original paper of Blei, Ng, and Jordan[1].

```
In [5]: topic_zoom.model()
Serving to http://127.0.0.1:8888/    [Ctrl-C to exit]
127.0.0.1 - - [07/Aug/2018 22:52:00] "GET / HTTP/1.1" 200 -

stopping Server...
the latent topics are:

1) structur protein mechan region form function observ type onli differ
2) prey light respons differ visual time movement signal behavior pattern
3) speci plant differ species growth resist correl relat trait high
4) muscl forc dure speed bodi energi mass power model increas
5) male femal size individu reproduct food predat behaviour time differ
6) activ increas dure respons effect control signific result level decreas
7) temperatur rate growth tradeoff popul increas thermal model differ adapt
8) stress mechan strength increas load length materi forc properti differ
9) water fish surfac flow insect pressur organ bodi dure speed
10) cell gene protein express size pattern growth select differ chang
11) onli expect data example explain requir rang possibl need observ
12) increas effect temperatur high differ higher exposur rate suggest environment
13) develop success development environment life adult affect studi growth ecolog
14) role express protein function provid mechan regul involv respons play
15) anim behavior visual allow predat field inform perform detect prey
16) dure result activ time studi suggest test record control muscl
17) energi signific rate cost relat mass bodi result metabol reduc
18) produc presenc form surfac normal region reach physic veri generat
19) support maintain period indic respons strong primari experi light result
20) speci species differ morpholog pattern correl relat result similar examin
Enter integer topic numbers one at a time


Enter an integer(type done when finished): 6

Enter an integer(type done when finished): done

Make a subset re-cluster(y/n)?y
Previous doc count: 1626
New subset doc count: 84
```

Figure 4. Presented topics for selection. *After ctrl-c to stop pyLDAvis we may select topics for subsetting.*

Let $\mathcal{D} \in \mathbb{R}^{n \times d}$ be a corpus whose rows are the document vectors $\vec{x}_j \ where \ 1 \leq j \leq d$, so, $\mathcal{D} = \{ \vec{x}_1 , \dots , \vec{x}_n \}$ for n row vectors. Let $z_l \ for \ 1 \leq l \leq k$ be a multinomial distribution over the vocabulary of words appearing in the corpus. Denote a word count w_j, and the distribution satisfies $\sum_{i=1}^{d} p(w_j | z_l) = 1$. The LDA algorithm then generates words from topics followed by topics from the documents. So, a distribution of words given a document δ is,

$$p(w_j \ | \delta) = \sum_{l=1}^{k} p(w_j | z_l) p(z_l | \delta)$$

The underlying assumption is the process by which the document is generated. Each topic is a word distribution which is sampled from. Each document derives from a topic distribution with assumed Dirichlet prior.

B. Non-Negative Matrix Factorization

NMF is a modeler that has its roots in singular value decomposition from linear algebra. The goal is to factor the matrix D into two matrices. One matrix is the documents by topic matrix and the other is the words by topic. The goal is to minimize the error after reconstruction of D. NMF has some interpretability advantages due to its non-negative nature. NMF simplifies classification when a new document is added to the existing corpus, since one may take the product of a document vector with a factored matrix to obtain a set of probabilities for topic membership. For further information see Lee and Seung[2].

We made use of NMF with Kullback-Leibler divergence as the second of our two topic clusterers. NMF solves the minimization problem,

$$\min_{W,H} \|D - WH^T\|_F^2$$
$$subject\ to:\ W \geq 0, H \geq 0$$

where $D \in \mathbb{R}^{n \times d}$ and $W \in \mathbb{R}^{n \times k}, H \in \mathbb{R}^{d \times k}$. When minimizing by use of the Kullback-Leibler divergence, this is equivalent to probabilistic latent semantic analysis (PLSA). PLSA finds strong employment for topic modeling. The KL-divergence is calculated as follows,

$$D_{KL}(P||Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)}.$$

Where P and Q are discrete probability distributions across the terms of the vocabulary. The KL divergence measures the expected value of the log difference of P and Q using the expectation of P. For a KL value of 1 the distribution Q diverges from P so much that given the expectation of P, the expectation approaches zero. If P and Q are highly similar, that is given P the expectation has no difference from Q, then KL-divergence approaches zero.

III. Summary

The choice of topic modelers has generated acceptable topics to biology domain experts with regard to the dataset tested here. For the user, pyLDAvis[3] tool provides a powerful set of tools for visualization and examination of topic feature words. The recursive subsetting is effective in reducing the number of documents for a user to peruse at a deeper level. Currently, the topics generated are able to provide various classification algorithms with topics that achieve precision-recall area under curve results at better than 80%. This preliminary result from future work indicates the topics partition reasonably well for general datasets. This was found even in the presence of "tradeoff" papers where dichotomies created some issue.g

III. Conclusion

For future work we feel it will be necessary to construct a classifier which works well with the initial topics of the clustering tool. The goal would be to connect the classified documents by topic in a probabilistic fashion. Then we may have multiple edges connect back into the PeTaL ontology graph.

It would also be useful to develop some type of "smart" sampling system for documents. As it stands now it is easy for the user to provide a biased dataset that does not classify well with the ontology. Currently a domain expert must review the initial topics and decide what labels in the ontology match the topics found by the modeler. This can be a problem if a topic/label pair is not well represented in the data. I would propose automating the sampling for a balance among the classes

The addition of some useful tool similar to pyLDAvis for analyzing the topics from NMF would be desirable. As of now there does not seem to be one and so topic quality is completely subjective.

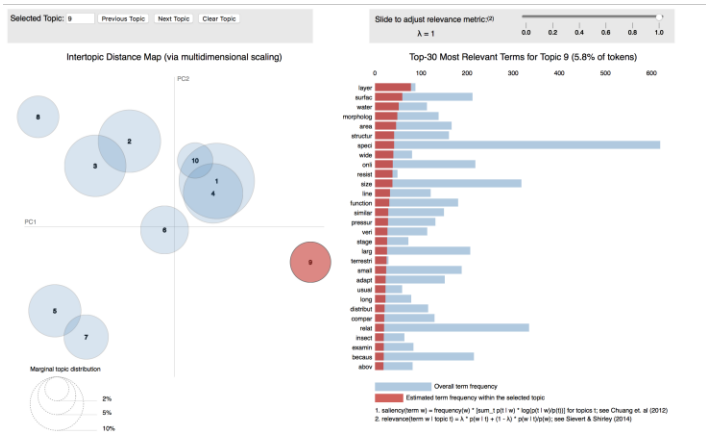

Figure 5. pyLDAvis topic choice 9. *The resulting clusters after topic 6 was chosen above. The cluster tool re-clusters on the documents of the selected topic(s) and presents the newly modeled clusters.*

Acknowledgments

We would like to acknowledge the guidance of his mentors Herb Schilling and Calvin Robinson at the GVIZ lab of NASA Glenn Research Center.

The opportunity to work on the PeTaL biomimicry project as created by principal investigator Vikram Shyam has been greatly appreciated.

Thank you to Ben Mabey for the port of LDAvis to his project pyLDAvis, making this outstanding tool available in Python.

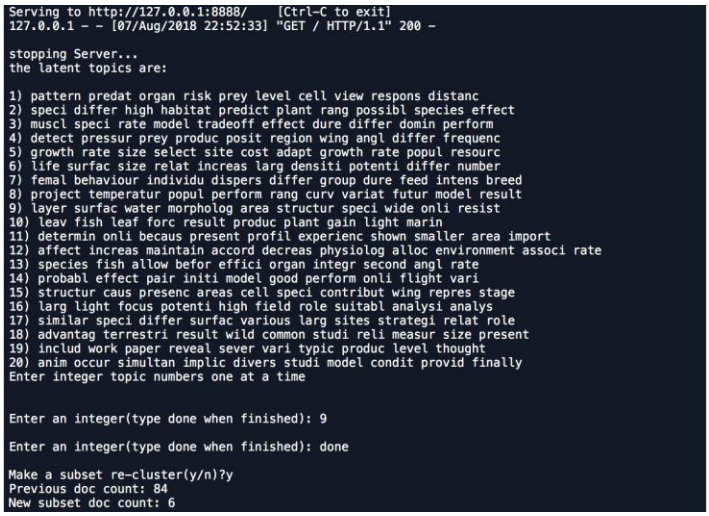

Figure 6. Presented topics for selection. *These are the resulting topics listed for perusal and selection*

References

Proceedings

[1]Blei, D, Ng, A. and Jordan, M., "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, No. 3, 2003, pp. 993-1022

[2]Lee, D. and Seung, H., "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems.* 2001, pp. 556-562

[3]Sievert, C. and Shirley, K.,"LDAvis: A method for visualizing and interpreting topics." *Proceedings of the workshop on interactive language learning, visualization, and interfaces.*,2014, pp.63-70

Books

[4]Aggarwal, C., *Machine Learning for Text*, 2nd ed., Springer International Publishing AG, Cham, Switzerland, 2018, Chaps. 1, 3, 4.

[5]Aggarwal, C. and Reddy, C., *Data Clustering: Algorithms and Applications*, CRC Press, New York, 2014, Chaps. 3, 4.

Electronic Publications

[6]B. Mabey, pyLDAvis, (2015), GitHub repository, https://github.com/bmabey/pyLDAvis