

Project Check In

In Your Dreams: Coherent Topic Identification from Dream Journals

Ben Attix, Jay Cordes, Kari Ross
W266: Natural Language Processing with Deep Learning
UC Berkeley School of Information
{battix, jcordes, kari.ross}@ischool.berkeley.edu
October 12th, 2017

Abstract

Our goal is to apply NLP techniques to a new domain: dream analysis. Coherent topics will be identified from dream journal entries in order to expand the variety of categorizations available for analysis and reduce the variation, bias, and time requirement of human classification. The most predictive topics will be highlighted as the most potentially useful for dream analysis, since they will be the dream topics that most distinguish a subject from others.

Introduction

The main question dream analysts are typically interested in is knowing how unusual someone's dreams are, which is done by classifying the people, actions, and objects from their dream journals into categories which can then be compared with the general population.


The most successful and reliable categorization thus far is the **Hall/Van de Castle system**.^[1] This system has been replicated several times, thus instilling confidence that categories and analysis can show variances from the norms.


The benefits of using NLP

Historically, human judges are almost always involved in the categorization of dreams and the analysis has been limited to broad categories that can be easily and consistently coded.

The topics arrived at from an unsupervised analysis of dream journal entries will not be as readily understandable as the ones carefully designed by dream researchers, but the most important thing has never been the categories

themselves; it's been the differences between dreamers and the ability to identify what is most surprising or unusual about an individual's dreams.

We will extract **"topics"** from the dreams, **which will simply be groups of words that often occur together**. There will initially be no assurance that these topics will be sensible, and therefore useful for analysis, so we will filter out the topics that fail to meet a certain threshold of coherency. 

It's important to note that our purpose is not to use NLP techniques to simply predict the dreamer with the highest accuracy; the goal is to find coherent dream topics that can be used to predict dreamers. For example, a word that dreamer misspells in their journals would be useful in terms of identifying the dreamer, but would not provide any useful information for dream analysis. In our case, coherence is mandatory, even though some prediction accuracy will be sacrificed. 

Data: Dream Bank www.dreambank.net

The data to be analyzed is a collection of 26,000 dreams from various individuals and groups. All of it has been labeled as male or female, and some of them manually coded with other categories that may also be used for training. It's unlikely that we will be able to draw meaningful general conclusions from the data, as the dream journal entries are obviously not drawn from the population at random. For example, the number of female dreams outnumbers the number of male dreams 18,187 to 7,813, so the dreams are far from a representative sample of all English speakers.

There were 40 different individual dreamers who provided sets of dream journal entries, while the rest of the collections were from groups of dreamers such as West Coast teenage girls or Peruvian women.

```
Alta: a detailed dreamer (422 dreams)
ID: alta
type: series
sex: F
age: A
time: 1985-1997
sample dream:
  number: 1
  date: 1957
  report: The one at the Meads's house, where i
  obblestone street and a Pied-Piper sort of man wi
```

Fig 1. Sample dream series in the dataset

We won't initially consider the age of the dreamers, but if we do go that direction, the data is already coded with the categories of Adult (10,067 dreams), Young Adults (3,116), College-Aged (6,453), Teenagers (857), Young Teenagers (5,126), and Children (381).

Dataset Preparation

For ease of use, we converted the data to a single denormalized list of dreams with a couple of new fields: A unique, surrogate ID for each of the dreams, and an ID for each of the 40 individual dreamers we are interested in studying.

We sometimes had to combine different collections of dreams under a single ID when they belonged to the same dreamer. The reason a single person would have multiple IDs is that sometimes their dreams were split up if the dreams came from different time periods of their life. For example, someone named Phil originally had 3 different IDs for each of the following life stages: teens, late 20s, and retirement.

For the first set of experiments, we will assume that dreams are relatively consistent over people's lifetimes and group them together. However, if time permits, we may break these back apart later to examine the question of whether or not dream topics change significantly between different periods of people's lives.

Methods

For our baseline, we implemented a **bag-of-words** (BoW) model for each of our 40 dreamers. The models are **logistic regressions**, which are modeling whether a given dream is from that dreamer or from any of the other dreamers (one-vs-all).

We used BoW initially because it was the easiest to conceptualize and implement. Counting the frequency of words in a dream and seeing if those words show up consistently enough to predict the dreamer seemed like a logical first step. We also tried using TF-IDF, but **in a limited sample, the TF-IDF results were worse than BoW.**

In addition, the reason we chose to run a separate model for each dreamer was because **we wanted to know which words were the most predictive for each given dreamer.** In other words, find out which words best separated someone from all other dreamers. This is in line with the expected use case for this kind of analysis if it's found to be useful. A subject would provide a series of dreams, and our best predictive model would be trained to extract coherent topics that best differentiate their dreams from the others in the DreamBank. The output of the model would be the most predictive topics (typical topics that are most unique to their dreams).


By running 40 different one-vs-all models, we can extract the most relevant words in each of those 40 people's dreams. We experimented with random forest classifiers and ridge classifiers, but found that logistic regression performed the best.

Results and Discussion


One of our concerns for the bag-of-words approach was that **the most predictive words might be words that a dreamer consistently misspelled or names that appeared frequently.** While misspellings and names might be **predictive, they aren't particularly informative** about the content of the dream and aren't helpful for analysis or future topic-modeling. Since our topic extraction will skip misspellings and names (they're unlikely to be grouped with other words in any sensible way), we felt like


the bag-of-words had a bit of an unfair advantage over our topic extraction approach. In other words, the fact that we're limiting ourselves to coherent topics means that our predictive success probably won't be as high as even a simple model with no such restrictions.



We were pleasantly surprised when we studied the top 5 most predictive words for each of the 40 dreamers and found only one occurrence of a misspelling or name. This means that the accuracy of our topic-based approach should be closer to the bag-of-words benchmark than we originally anticipated.

Next, we used our logistic regression models for prediction to see if we could identify who a dream came from given a dream that wasn't in the training set. To accomplish this,  we scored 5,200 dreams from our test dataset against the 40 logistic regression models. Whichever model produced the highest predicted probability of a dreamer, then we chose the dreamer that corresponded with that model for our prediction. Using this method, we were able to successfully identify the correct dreamer 82% of the time.

Next Steps

Next, we plan on implementing topic modeling using Latent Dirichlet Allocation (LDA) and measuring the topic coherence with word intrusion and observed coherence. We will look  into using Mallet software for the topic extraction. Chang et al.^[11] first introduced the method of word intrusion. The idea is that a person or system will try to detect which word in a topic (a set of related words) was a random word that "intruded" into the topic. The easier it is to identify the random word, the more interpretable the original topic must have been. While word intrusion was first performed by humans, Lau et al. showed that the interpretability of topics using word intrusion can be fully automated^[4] and thus we can avoid involving human judges. We can also measure topic coherence using observed coherence, a method that was introduced by Newman et al.^[10] Observed coherence originally involved asking humans to rate the topics on a 3-point scale based off how coherent they thought the topics were, however, today this can be automated as well and does not require human judges^[4].

We will remove topics that don't meet a minimum threshold of coherency because the closer they are to a collection of random words, the less valuable they would be to dream analysts.  Even if they are highly predictive, topics with low coherency would almost certainly not be fruitful for providing insights into the dreamer's mind (similar to the words we identified with the bag-of-words technique as being the most predictive).

Martin and Johnson^[3] have shown that lemmatization and reducing our corpus to nouns-only  will simplify and speed up topic extraction without any downside in terms of coherency. During our exploratory analysis, we already stripped the dream journal entries down to nouns-only and performed lemmatization in preparation for the next steps. However the corpus reduction and lemmatization was not yet used in our analysis. Going forward, we will tie it into the analysis. 

We will experiment with different parameter values (such as the number of topics to extract from each set of dreams) and find the approach that maximizes the ability to predict dreamers from previously unseen dreams, while preserving coherency. We believe that this would be the most useful for dream analysts: not only providing common and coherent dream topics for a particular dreamer, but identifying which of those topics are most noteworthy.

We are also looking into different models, including Support Vector Machine (SVM)^[9] and Naive Bayes (NB), and we are investigating our evaluation metrics. We're considering using ROC and Precision-Recall Curves to compare our results against our baseline (BoW) to the topic-based representation. F1 score and accuracy percentage are also being considered as a performance metric.

References

1. Domhoff, G. W. (1999). New directions in the study of dream content using the Hall/Van de Castle coding system. *Dreaming*, 9, pages 115–137.
www2.ucsc.edu/dreams/Library/domhoff_1999a.html
2. Domhoff, G. W. (2000). Methods and measures for the study of dream content. In M. Kryger, T. Roth, & W. Dement (Eds.), *Principles and Practices of Sleep Medicine: Vol. 3* (pages 463–471). Philadelphia: W. B. Saunders.
www2.ucsc.edu/dreams/Library/domhoff_2000a.html
3. Fiona Martin and Mark Johnson. 2015. More Efficient Topic Modelling Through a Noun Only Approach . In *Proceedings of Australasian Language Technology Association Workshop*, pages 111–115.
www.aclweb.org/anthology/U15-1013
4. Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)* (pp. 530–539). Gothenburg, Sweden: Association for Computational Linguistics.
www.aclweb.org/anthology/E14-1056
5. Chandler May et. al (August 2016) .Analysis of Morphology in Topic Modeling.
[arXiv:1608.03995v1](https://arxiv.org/abs/1608.03995v1)
6. Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology–Volume 1*, (pp. 173–180). The Association for Computational Linguistics.
<https://nlp.stanford.edu/pubs/tagging.pdf>
7. Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
<http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf>
8. Blei, D. M., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022
<http://ai.stanford.edu/~ang/papers/nips01-lda.pdf>
9. Wang S. Manning, C. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification
https://nlp.stanford.edu/pubs/sidaw12_simple_sentiment.pdf
10. Newman, J.H. Lau, K. Grieser, and T. Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 100–108, Los Angeles, USA.
<https://mimno.infosci.cornell.edu/info6150/readings/N10-1012.pdf>
11. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 21 (NIPS-09)*, pages 288–296, Vancouver, Canada.
<https://www.umi.acs.umd.edu/~jbg/docs/nips2009-rtl.pdf>

Appendix: Bag-of-Words Most Predictive Words per Dreamer

ID	Most Predictive Words				
1	rather	somewhere	somebody	here	re
2	children	quarter	campus	meet	pancakes
3	poison	hometown	picture	husband	model
4	helpful	nightmares	nightmare	neat	scary
5	stands	leading	red	ghost	world
6	street	bunk	note	stranger	reach
7	surprise	standing	real	burglar	few
8	looked	toilet	red	bathroom	period
9	definitely	stack	abstract	buick	insured
10	mixed	preaching	teaching	miss	hats
11	dream	mood	both	clearly	looks
12	laid	medicine	chest	radio	pill
13	sheets	witnesses	repaired	email	waking
14	visit	toes	painting	shift	frank
15	beside	seven	joint	suddenly	army
16	ex	slammed	huge	hospital	fight
17	mom	video	promptly	diary	god
18	sound	laugh	tape	woke	kind
19	sat	thought	physics	remember	scores
20	sets	ago	gift	plastic	sad
21	grandpa	girlfriend	may	fellatio	missing
22	rubbing	dreamt	cowboys	boyfriend	observatory
23	supposedly	cheeseburger	marketed	chicken	belt
24	tests	comic	john	lost	tooth
25	accident	end	hugged	grade	mike
26	mumbling	freezer	drunk	restaurant	art
27	spotting	giraffes	pond	sort	husband
28	thru	although	grain	recall	wife
29	drunk	24	hell	team	softball
30	match	museum	though	fragment	hometown
31	patient	printing	certain	ward	sister
32	ft	thru	trouble	saw	winner
33	agency	educational	wife	apparently	assassinated
34	stood	earnestly	woke	much	morning
35	police	after	started	set	away
36	created	circus	dances	occaisionaly	travelling
37	lights	everyone	now	re	wake
38	alarm	should	because	ever	talking
39	everyplace	long	white	scary	corridor
40	life	ends	begin	wake	recollection