Final Project

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
December 19th, 2017

Abstract
Natural language processing (NLP) techniques were applied to a new domain: dream series analysis. Topics and most predictive words were automatically extracted from dream journal entries in order to expand the variety of categorizations and data available for analysis.

Topics were extracted from both individual and group dreams. The NLP methods used included: bag-of-words (BoW), Term Frequency - Inverse Document Frequency (TF-IDF), logistic regression, latent dirichlet allocation (LDA; Blei et. al[8]), and the author-topic model (Rosen-Zvi et al[16]). Cosine Distance Metrics were also used to quantitatively describe similarities between dreamers.

1 - Background
If eyes are the windows to the soul, then dreams are the windows to the mind. "The consistency and continuity found in past studies of dream series are based on a random spin through the dreamer's cognitive Rolodex."[12] In other words, dream journals are fertile with information about personal relationships, a dreamer's unprovoked thoughts, and insights into their minds and memories. This data can be helpful for psychotherapists seeking to discover important information about clients or for social scientists looking for consistent differences between demographic groups.

Dream content analysis typically centers around finding categories in a dream series with meaningful variation from norms found in the general population. The most successful and reliable categorization thus far is the Hall/Van de Castle system(HVdC)[1] , which is a manual system designed for consistency although it loses individualization of dreamers in an attempt to be broad and general.

2 - Introduction: Benefits of using NLP
By automatically extracting topics of interest from the dreams we are attempting to find a way to allow the dreams to "speak for themselves" and allowing for more nuanced topics that analysts may not have considered. This process can either supplement the HVdC categories by adding more categories to consider or possibly providing an alternative to the repetitive work of coding dreams manually.

Our main challenge was to find, without the benefit of human judges, topics that are "coherent" as well a method to distinguish a dreamer from the general population. The goal was to design a process that could create coherent topics out of any dream series that are also useful in distinguishing the dreamer.

This paper will discuss the DreamBank Dataset, data preparation, baseline models, and two LDA models, as well as the next steps for continued model improvement.

3.1 - Data: The DreamBank
The data utilized was a collection of 26,000 dreams from various individuals and groups. Since the dream journal entries are from people who voluntarily chose to record their dreams, data are clearly not drawn from the population at random and subject to selection bias.

There were 40 different individual dreamers who provided a series of dreams, while the rest of the collections were from groups of dreamers

such as West Coast teenage girls or Peruvian women.

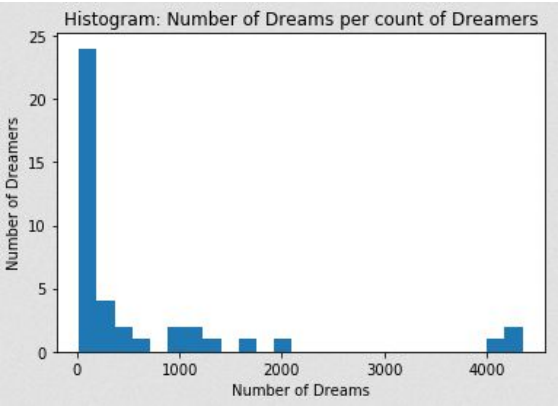

Figure 1

A majority of the dreamers have less than 1000 dreams.

3.2 - 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 were interested in studying. For our experiments, we assumed that dreams are relatively consistent over people's lifetimes and grouped them together. This allowed us to randomly draw training and test sets for each dreamer.

3.3 - Method of Data Cleaning

Martin and Johnson[3] have shown that lemmatization and reducing our corpus to nouns-only would simplify and speed up topic extraction without any downside in terms of coherency. During our topic analysis, we stripped the dream journal entries down to nouns-only and performed lemmatization in preparation for the next steps.

In summary, we pre-processed the data as follows: we (1) made all words lowercase, (2) tokenized the dream text, (3) tagged the tokens with part-of-speech, (4) reduced the text to noun-only, (5) removed numbers and words shorter than three characters, and (6) lemmatized the nouns. For example, tokenizing the sentence "I had these little cars in my hand, I was riding a bicycle" and extracting the nouns gives us "cars", "hand", and "bicycle." Lemmatizing the nouns switches "cars" to "car" and leaves the other words the same. This allowed the frequency counts to appropriately combine variations of words. Lemmatizing was used instead of stemming in order to take into consideration the morphological analysis of a word.

4.1 - Baseline Models

For the baseline, we implemented three models for each of our 40 dreamers. The models were each logistic regressions, which are predicting whether a given dream is from a given dreamer or from any of the other dreamers (one-vs-all). The predictive features were the word counts ("Bag-of-words", or BoW), term frequencies (TF), and term frequency / inverse document frequency (TF-IDF) for the three different models used.

We used BoW initially it is one of the simplest text models. BoW simply is a frequency count and enables prediction by seeing if those words show up consistently in dreams. Additionally, BoW is the vector representation utilized in several other models for NLP.

4.2 Baseline Results and Discussion

By running 40 different one-vs-all models, we extracted the most relevant words in each of those 40 people's dreams. One goal with the baseline models was to find out which words best separated someone from all other dreamers and the more sophisticated frequency-based models performed better than BoW.

Logistic regression models were used for prediction to see if we could identify who a dream came from. We had split the data as follows: 60% training data, 20% development data, and 20% test data. To make our prediction, we scored 5,200 dreams from our test dataset against the 40 logistic regression models.

The BoW word count model gave us 81% precision compared to 84% for the TF and TF-IDF models, respectively. Subjectively, the term frequency-based models appeared better suited for the task as well. For example, one of the words the TF approach found for Dreamer #4 was "wheelchair", which reflects how her

dreams about her limited mobility made her unique in the database.

It was noted that for the BoW models some of the most predictive features (defined as the words with the highest logistic regression coefficients) were words that only appeared once in the entire corpus. Since we were trying to determine which words set people apart from everybody else, valuing a word that shows up only once does not seem very useful, so it made intuitive sense that TF and TF-IDF had better accuracy.

We experimented with random forest classifiers but that yielded only 60% precision for BoW compared to 81% precision with logistic regression.

When we studied the top 5 most predictive words for each of the 40 dreamers (the TF-IDF approach had the highest F1 score), we were happy to find relatively few misspellings. Most of the important words were names or topics of clear particular interest to the dreamer (see Appendix 1).

5.1-Topic Extraction Introduction

In order to move to find nuanced topics and determine how similar or different dreamers are from each other, and how a dreamer might have excursions from their dream baseline, we implemented topic modeling using Latent Dirichlet Allocation (LDA). We also included coherency scores to find the topics that should be the most sensible to humans. We used the Gensim software package for the topic extraction.

The LDA model was chosen for exploration because it is a probability distribution of vocabulary over a collection of documents. Another well studied model is Latent Semantic Analysis/Indexing (LSA / LSI). We focused on LDA over LSI primarily because LDA is a generative model and enables the classification of an unseen document and is less susceptible to overfitting than the LSI model. The LSI model is similar to Primary Component Analysis (PCA) but for discrete data.[6]

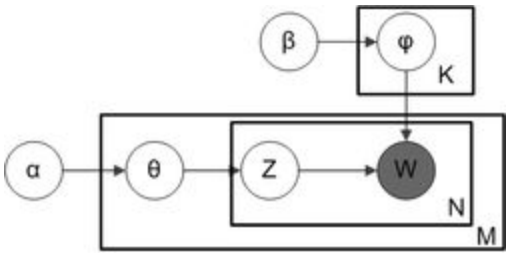

Figure 2 - LDA Model

Image credit:By Slxu.public
https://commons.wikimedia.org/wiki/File%3ASmoothed_LDA.png

In the LDA model, the topics, topics distributions, per-document per-word assignments are all hidden structures. In figure2, M = number of documents and N = Number of words in a document. The notion of hidden structures over a probability distribution is a good candidate for evaluating the hypothesis of dream data being drawn from a generative model, where the real generative model is the dreamer.

5.2 - LDA Base Model and Methods

To create the vocabulary, all of the individual dream data was used as the collection of documents. The preprocessing followed the methodology from Section 3.1 (no noun reduction) to end up with a Bag-of-Words.

From the entire dataset, we extracted 200 topics. Then, for each of the 40 series of dreams, we determined what their mix of those topics was. The topic coherence scores came from the UMass measure put forth by Mimno et al[12]. Their method improves coherence score while retaining the ability to identify bad topics.

5.3- LDA Base Model Results

The majority of topics extracted from the LDA base model didn't subjectively appear to have any sensible theme. The closer they are to a collection of random words, the less valuable they would be to content analysts without any additional quantitative metrics. Since we didn't feel that the topics were good enough to warrant continuing on with our original approach to find the most predictive ones, we found an alternate way in which the topics extracted could be put to use.

If someone subjectively looks down the list of discovered topics and finds ones that are sensible and interesting, then the DreamBank

could be analyzed to find the dreamers who match up best and worst with those topics. As a demonstration for this kind of approach, we came up with a "TopicScore" which quantifies how well a given topic matches with a dreamer's journal entries.

Take the LDA models topic #52, which subjectively appears both quite coherent and interesting, for example:

Word	Membership-in-Topic Score
Shot	0.107
Son	0.085
Kill	0.080
Gun	0.065
Shoot	0.055
Knife	0.049
Shooting	0.040
Killed	0.038
Jack	0.037

Table 1

The TopicScores were calculated as follows: (1) for each word in the topic, we found its frequency in the dreams for each of the 40 dreamers by counting up the instances of the words in their journal entries and dividing it by the total number of words in all of their dreams, (2) we then multiplied each of those frequencies (nine of them in this case) by the membership-in-topic score in the topic list, and (3) we totalled up those products up for each dreamer.

This method provides an unbiased metric with which dreamers can be compared in regards to this topic (see figures 3-6 below for the scoring of four interesting LDA-produced topics). Also, using this approach would allow researchers to discover things from dreams they didn't know to look for and wouldn't have tracked with any of the current categories used for dream coding.

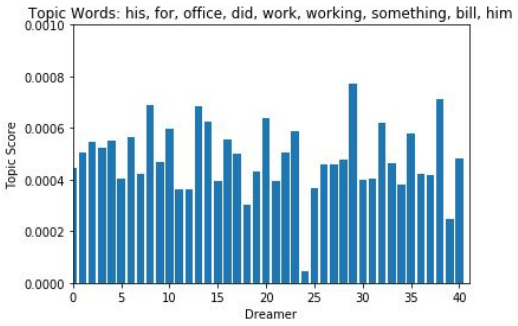

Figure 3. "young boy" (#24) almost never dreams about work, but "Nancy: Caring & headstrong" (#29) certainly does.

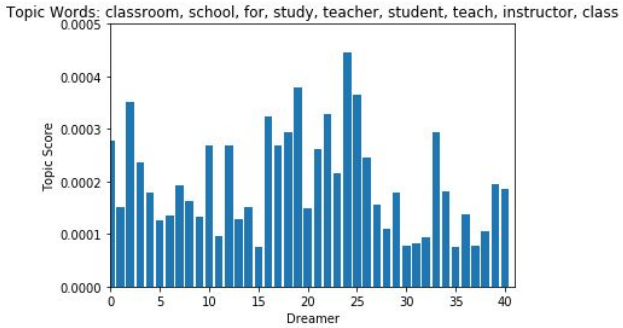

Figure 4. On the other hand, "young boy" (#24) is all about school.

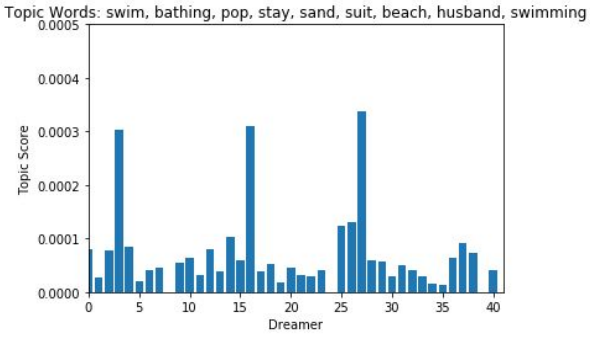

Figure 5. A few dreamers really seem to think about the beach more than the rest.

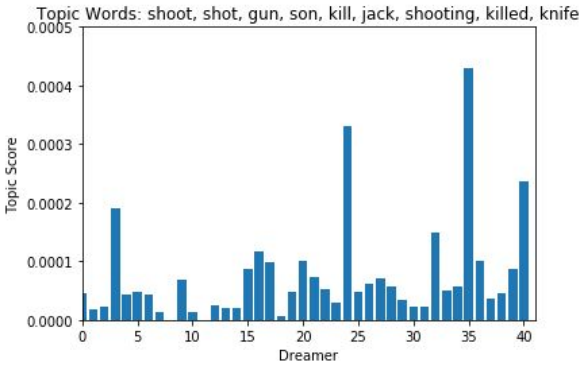

Figure 6. Should we be worried about #35 and #24? For comparison, #40 is a Vietnam vet with war dreams!

5.4 - LDA Author Model

In order to take into account the unbalanced dataset, with some dreamers producing far more dreams than others, the LDA Author Topic Model was explored, which is an extension of the base LDA model.[13]

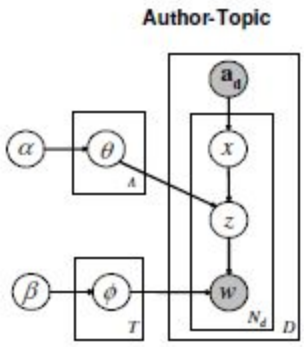

Figure 7 - Author-Topic Model [13]

As illustrated in Figure 7, the Author Topic trains is based on the distributions θ and ϕ in order to obtain information on both the author, α and the collection of documents, β in Figure 7. Use of this model enables a simplified use of the cosine distance metric to distances between dreamers.

The cosine distance metric was desired in order to have an unsupervised, less biased way of quantitatively grouping dreamers by their dream similarity. Additionally, the author-topic LDA model is an online algorithm:after a baseline model has been created, additional documents could be added to the model.

The LDA Author Model was trained on 50, 100, and 200 topics but the results were not significantly different, so the Model with 50 Topics was selected. The corpus was the same corpus as the Base LDA Model.

5.5 - LDA Author Model

Figures 8 and 9 show the results for 2 different dreamers and their cosine distances. Where this was deemed to be useful was in calculating which dreamers are similar. Also, Figure 9 was shown to be quite the anomaly: Dreamer 11 has very low cosine distances from other dreamers. This could be a method for seeing major deviations from a norm or baseline.

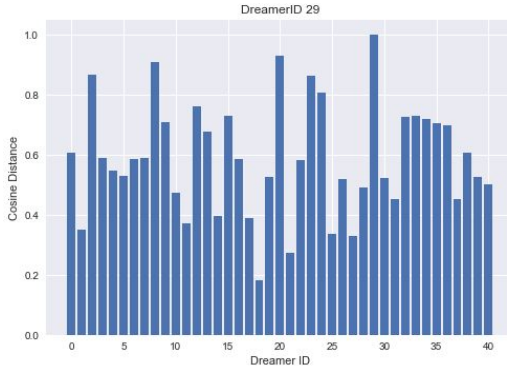

Figure 8 - Dreamer 29 shows similarity to 5 other dreamers (Cosine Distance > 0.8)

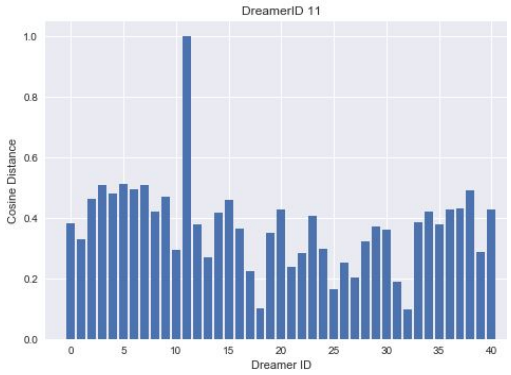

Figure 9 - Dreamer 11 shows low similarity with any other dreamers (Cosine Distance < 0.5)

Additionally, the author topic models could be used as a baseline or as a fingerprint, for what each individual author looks like. Again, this could be used as a prediction method or it could be used to see if a dreamer is deviating from their norm. See figures 10 and 11.

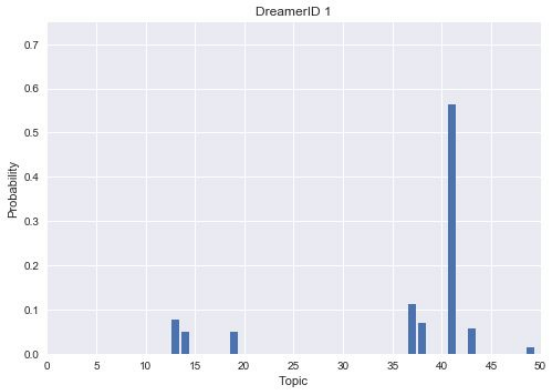

Figure 10 - Dream 1 Topic Distribution

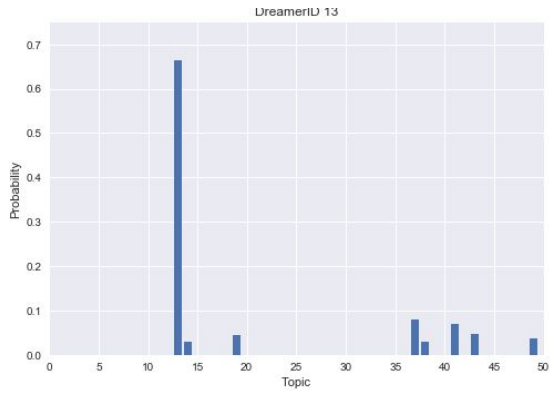

Figure 11 - Dreamer 13 Topic Distribution

Conclusion

While our study ended up requiring the intervention of a human judge, we believe that this is a promising start to a new approach for dream analysis. Our two main contributions that may be of value are: (1) an automated extraction of individual keywords from dreams that are considered important (they are useful in associating dreamers with previously unread dream entries), and (2) an automated process for extracting topics that may be of interest.

Overall, the topics were not typically as sensible as hoped for, but as can be seen from our examples, some topics clearly are. A content analyst can simply read down the list of topics produced and then "score" their dreamer of interest in respect to those topics and find out if they're notably different from others. In this way, we believe new and surprisingly meaningful topics may be discovered that analysts haven't considered before.

Lastly, the Author-Topic LDA model provided cosine distance metrics, which gives an unsupervised method for determining how similar or different a dreamer is from other dreamers. And in the online version of the model, new dreams could be scored against the baseline to see deviations for the established norm, as was deemed desirable.

Future Work

There are several opportunities identified which can improve on the work presented in this paper. A shortcoming in the LDA analysis as presented for the DreamBank data is that the quantitative evaluation and model checking is lacking, due to the abundance of incoherent topics.

If things went according to plan, the LDA analysis would have used the typical holdout dataset and then checked the accuracy of running the model against the unseen data. Perplexity could have also been added as a metric, although a study has shown that the desired low perplexity does not necessarily correlate with the ease human interpretation. [7]

Additionally, coherency could be automated, 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.

Another issue which was not addressed with the current LDA model is words which are unseen in the training corpus. This could be adjusted for with a modified Laplace smoothing or putting unseen words into a group of words with a probability in the long right tail of the corpus. In order to determine the usefulness of smoothing, the LDA model evaluation metrics would have to be determined.

The issue of sparsity in the DreamBank Data should also be addressed. The dreamer classes are highly unbalanced. Using TF-IDF in the baseline was a method for balancing the classes more, although that was done on a dream (record) basis and not an author basis. LSI / LSA could also be investigated , as it utilized a vector representation of TF-IDF into the model. The Author-Topic LDA model was another attempt to balance the classes. Other methods would include oversampling the smaller classes and/or undersampling the dreamers with more dreams.

Finally, the assumption in the data is that the dreamers have consistent dreams over time. As the data is time series panel data, this assumption could be explored and validated, on a dreamer by dreamer basis.

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
6. Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55(4), 77–84.* http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf
7. 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
8. Wang S. Manning, C. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification https://nlp.stanford.edu/pubs/sidaw12_simple_sentiment.pdf
9. 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
10. 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.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf
11. Domhoff, G. W., & Schneider, A. (2008). Studying dream content using the archive and search engine on DreamBank.net. Consciousness and Cognition, 17, 1238-1247. http://www2.ucsc.edu/dreams/Library/domhoff_2008c.html
12. Mimo, Wallach, Talley, Leenders,McCallum (2011) Optimizing Semantic Coherence in Topic Models https://mimno.infosci.cornell.edu/papers/mimno-semantic-emnlp.pdf
13. Rosen-Zvi et al (2004) The Author-Topic Model for Authors and Documents https://mimno.infosci.cornell.edu/info6150/readings/398.pdf
14. https://nbviewer.jupyter.org/github/rare-technologies/gensim/blob/develop/docs/notebooks/atmodel_tutorial.ipynb
15. https://markroxor.github.io/gensim/static/notebooks/lda_training_tips.html
16. www.dreambank.net

Appendix 1: Bag-of-Words Most Predictive Words per Dreamer

Dreamer	Most Predictive BoW (counts)	Most Predictive BoW (TF)	Most Predictive BoW (TF-IDF)
Alta: a detailed dreamer [ID=1]	['re', 'here', 'somebody', 'somewhere', 'rather']	['re', 'here', 'somebody', 'or', 'something']	['ve', 'somebody', 'maybe', 'nice', 'think']
Angie: age 18 & 20 [ID=2]	['pancakes', 'meet', 'campus', 'quarter', 'children']	['children', 'baby', 'boyfriend', 'then', 'people']	['boyfriend', 'pancakes', 'housemates', 'quarter', 'preschool']
Arlie: a middle-aged woman [ID=3]	['model', 'husband', 'picture', 'hometown', 'poison']	['town', 'husband', 'hometown', 'am', 'son']	['hometown', 'husband', 'town', 'local', 'picture']
Barb Sanders [ID=4]	['scary', 'neat', 'nightmare', 'nightmares', 'helpful']	['wheelchair', 'seated', 'lots', 'helpful', 'aware']	['bl', 'howard', 'ellie', 'bonnie', 'nate']
Robert Bosnak: a dream analyst [ID=5]	['world', 'ghost', 'red', 'leading', 'stands']	['very', 'world', 'night', 'life', 'that']	['world', 'pappie', 'dusty', 'freud', 'cellars']
Chris: a transvestite [ID=6]	['reach', 'stranger', 'note', 'bunk', 'street']	['note', 'and', 'street', 'reach', 'stranger']	['note', 'stranger', 'reach', 'continue', 'realize']
Chuck: a physical scientist [ID=7]	['few', 'burglar', 'real', 'standing', 'surprise']	['standing', 'myself', 'which', 'read', 'while']	['burglar', 'large', 'earth', 'read', 'standing']
Dahlia: concerns with appearance [ID=8]	['period', 'bathroom', 'red', 'toilet', 'looked']	['work', 'bathroom', 'dentures', 'period', 'legs']	['dentures', 'tattoo', 'period', 'chanel', 'sails']
David: teenage dreams [ID=9]	['insured', 'buick', 'abstract', 'stack', 'definitely']	['jimmy', 'brad', 'huge', 'life', 'all']	['laura', 'jimmy', 'brad', 'angela', 'wierd']
Dorothea: 53 years of dreams [ID=10]	['hats', 'miss', 'teaching', 'preaching', 'mixed']	['miss', 'chinese', 'must', '2nd', 'ask']	['miss', 'mrs', '2nd', 'asked', 'ask']
Ed: dreams of his late wife [ID=11]	['looks', 'clearly', 'both', 'mood', 'dream']	['dream', 'or', 'looks', 'are', 'when']	['mary', 'arms', 'adam', 'face', 'tells']
Edna: a blind woman [ID=12]	['pill', 'radio', 'chest', 'medicine', 'laid']	['all', 'yo', 'ladder', 'very', 'mother']	['ladder', 'dancing', 'mary', 'aged', 'pill']
Elizabeth: a woman in her 40s [ID=13]	['waking', 'email', 'repaired', 'sheets', 'bill']	['waking', 'tv', 'working', 'email', 'bill']	['matthew', 'waking', 'cas', 'realized', 'daddy']
Emma: 48 years of dreams [ID=14]	['frank', 'shift', 'painting', 'toes', 'visit']	['frank', 'shift', 'painting', 'purse', 'analyst']	['frank', 'andrew', 'pedro', 'shift', 'zena']
Emma's Husband [ID=15]	['army', 'suddenly', 'joint', 'seven', 'beside']	['suddenly', 'as', 'beside', 'or', 'comes']	['suddenly', 'comes', 'railroad', 'coffin', 'joint']
Esther: an adolescent girl [ID=16]	['fight', 'hospital', 'huge', 'slammed', 'ex']	['hospital', 'beach', 'path', 'carnival', 'fight']	['polly', 'wanda', 'carnival', 'bassist', 'heidi']
Izzy [ID=17]	['promptly', 'diary', 'god', 'video', 'vaguely']	['mom', 'video', 'saying', 'hell', 'god']	['ezra', 'eugene', 'nana', 'calvin', 'mom']
Jasmine [ID=18]	['kind', 'woke', 'tape', 'laugh', 'sound']	['and', 'kind', 'like', 'that', 'woke']	['like', 'kind', 'laugh', 'woke', 'going']
Jeff, a lucid dreamer [ID=19]	['scores', 'remember', 'physics', 'thought', 'sat']	['remember', 'then', 'think', 'really', 'if']	['remember', 'lucid', 'jamie', 'guinea', 'think']
Joan: a lesbian [ID=20]	['sad', 'plastic', 'gift', 'ago', 'sets']	['told', 'sad', 'me', 'yo', 'gift']	['eliza', 'gift', 'told', 'lori', 'brady']
Kenneth [ID=21]	['missing', 'fellatio', 'may', 'girlfriend', 'grandpa']	['may', 'girlfriend', 'fellatio', 'others', 'classmate']	['br', 'brimson', 'classmate', 'redding', 'stephen']

Madeline [ID=22]	['observatory', 'boyfriend', 'cowboys', 'dreamt', 'rubbing']	['boyfriend', 'somehow', 'sort', 'guess', 'maternal']	['jeremy', 'stuart', 'dreamt', 'maternal', 'sort']
Mack: a poor recaller [ID=23]	['belt', 'chicken', 'marketed', 'cheeseburger', 'supposedly']	['walked', 'ran', 'mark', 'john', 'kwon']	['tae', 'kwon', 'driveway', 'housemates', 'feds']
Mark: a young boy [ID=24]	['tooth', 'lost', 'john', 'comic', 'tests']	['john', 'killer', 'tooth', 'comic', 'lost']	['john', 'tooth', 'comic', 'killer', 'spelling']
Melissa: a young girl [ID=25]	['mike', 'grade', 'hugged', 'end', 'accident']	['so', 'all', 'end', 'big', 'said']	['squid', 'shirkan', 'said', 'rayna', 'big']
Merri: an artist [ID=26]	['art', 'restaurant', 'drunk', 'freezer', 'mumbling']	['the', 'not', 'hair', 'said', 'needed']	['dora', 'rudy', 'suppose', 'corinne', 'darkroom']
Melora (Melvin's wife) [ID=27]	['husband', 'sort', 'pond', 'giraffes', 'spotting']	['husband', 'sort', 'this', 'bob', 'going']	['sort', 'husband', 'going', 'bob', 'little']
Melvin (Melora's husband) [ID=28]	['wife', 'recall', 'grain', 'although', 'thr']	['recall', 'wife', 'some', 'thr', 'which']	['recall', 'wife', 'dream', 'sort', 'grain']
Nancy: Caring & headstrong [ID=29]	['softball', 'team', 'hell', '24', 'drunk']	['team', 'hell', 'decided', 'baby', 'softball']	['zack', 'carla', 'positive', 'softball', 'andrea']
The Natural Scientist [ID=30]	['hometown', 'fragment', 'though', 'museum', 'match']	['hometown', 'though', 'lady', 'which', 'probably']	['hometown', 'evidently', 'probably', 'chester', 'ave']
Norman: a child molester [ID=31]	['sister', 'ward', 'certain', 'printing', 'patient']	['patient', 'sister', 'ward', 'later', 'printing']	['later', 'patient', 'sister', 'printing', 'certain']
Pegasus: a factory worker [ID=32]	['winner', 'saw', 'thr', 'trouble', 'ft']	['thr', 'winner', 'fellow', 'heard', 'seemed']	['ann', 'saw', 'winner', 'fellow', 'ft']
Phil [ID=33]	['assassinated', 'apparently', 'wife', 'educational', 'agency']	['apparently', 'wife', 'daddy', 'assassinated', 'pretty']	['bonita', 'anita', 'apparently', 'sharon', 'daddy']
The Physiologist [ID=34]	['morning', 'much', 'woke', 'earnestly', 'stood']	['morning', 'stood', 'woke', 'suddenly', 'much']	['morning', 'stood', 'feature', 'newport', 'containing']
Ringo: from the 1960s [ID=35]	['away', 'set', 'started', 'after', 'police']	['out', 'away', 'friend', 'four', 'started']	['whale', 'hole', 'judy', 'truck', 'friend']
Samantha: in her 20s [ID=36]	['travelling', 'occaisionaly', 'dances', 'circus', 'created']	['really', 'guys', 'college', 'all', 'wave']	['tmm', 'walter', 'julia', 'diana', 'br']
Toby: a friendly party animal [ID=37]	['wake', 're', 'now', 'everyone', 'lights']	['wake', 're', 'and', 'but', 'girls']	['wake', 'highway', 'reason', 'starts', 'sudden']
Tom: an outgoing man [ID=38]	['talking', 'ever', 'because', 'should', 'alarm']	['am', 'talking', 'girls', 'girlfriend', 'walk']	['dog', 'girls', 'girlfriend', 'basically', 'talking']
Vickie: a 10-year-old girl [ID=39]	['corridor', 'scary', 'white', 'long', 'everyplace']	['go', 'this', 'said', 'white', 'yo']	['swords', 'nancy', 'wendy', 'morrison', 'valerie']
Vietnam Vet [ID=40]	['recollection', 'wake', 'begin', 'ends', 'life']	['wake', 'begin', 'recollection', 'no', 'ends']	['recollection', 'wake', 'begin', 'ends', 'resembles']