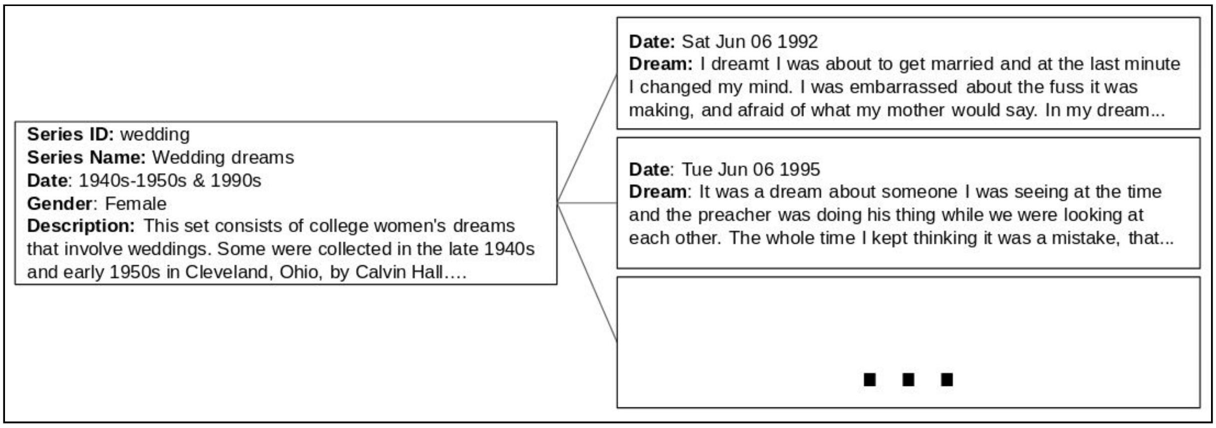
DreamBank Visualized
An Interactive Visualization of over 26,000 Dream Transcriptions

Jonathan Sauder (jonathan.sauder@student.hpi.de), Wael Barhoumi (wael.barhoumi@telecom-paristech.fr), Clement Landrin (clement.landrin@telecom-paristech.fr)

We present an interactive visualization of the DreamBank dataset [1, 2] which is comprised of over 33,000 dreams transcriptions, of which over 26,000 are in english. We use modern techniques from deep learning and natural language processing, most notably recent advances in dense representations of text, to enable the user of our system to explore the dataset by certain semantic attributes such as the topic or the sentiment. More specifically, we use Google's universal sentence encoder [3, 4] to assign each dream a dense representation in high-dimensional space and various dimensionality reduction techniques to project these high-dimensional vectors into two dimensions. This report highlights both the data manipulation techniques employed and the visual and interactive aspects of the system.

1. The DreamBank Dataset

The DreamBank dataset [1, 2] has been created and ever since maintained by Adam Schneider & G. William Domhoff of the Psychology Dept., UC Santa Cruz. As of June 2018, it counts over 33,000 dream transcriptions, of which over 26,000 are in English. These English dreams are categorized into 78 dream series, i.e. "who dreamt these dreams". Each dream series contains a description, the gender of the dreamers, and the years in which the dreams were collected. Next to the textual transcription of the dream, some dreams contain an exact date. Dream lengths range from a handful of words to many hundred words in length. Because DreamBank offers no option to download the entire dataset, the data was scraped [5] from the site. Some information might have been lost in the scraping process due to inconsistent HTML formatting.



Series ID: wedding
Series Name: Wedding dreams
Date: 1940s-1950s & 1990s
Gender: Female
Description: This set consists of college women's dreams that involve weddings. Some were collected in the late 1940s and early 1950s in Cleveland, Ohio, by Calvin Hall….

Date: Sat Jun 06 1992
Dream: I dreamt I was about to get married and at the last minute I changed my mind. I was embarrassed about the fuss it was making, and afraid of what my mother would say. In my dream...

Date: Tue Jun 06 1995
Dream: It was a dream about someone I was seeing at the time and the preacher was doing his thing while we were looking at each other. The whole time I kept thinking it was a mistake, that...

Figure 1: An example of the structure of the dataset. The dream series 'wedding' consists of many dreams

2. The System

The key to our visualization is to take each dream in its raw text form, and find a meaningful projection into two dimensions for it. Each dream is displayed as a small circle on the screen - clicking it will display the dream's content, date, and information about the dream series. The user can switch between different projections which highlight different aspects of the data. The user can select which dream series will be displayed, and can also perform a keyword search on all displayed dreams. It is also possible to select an area on the screen and get summarized information about all dreams whose current projection coordinates fall into that area to discover similarities such as the semantic meaning of that area.

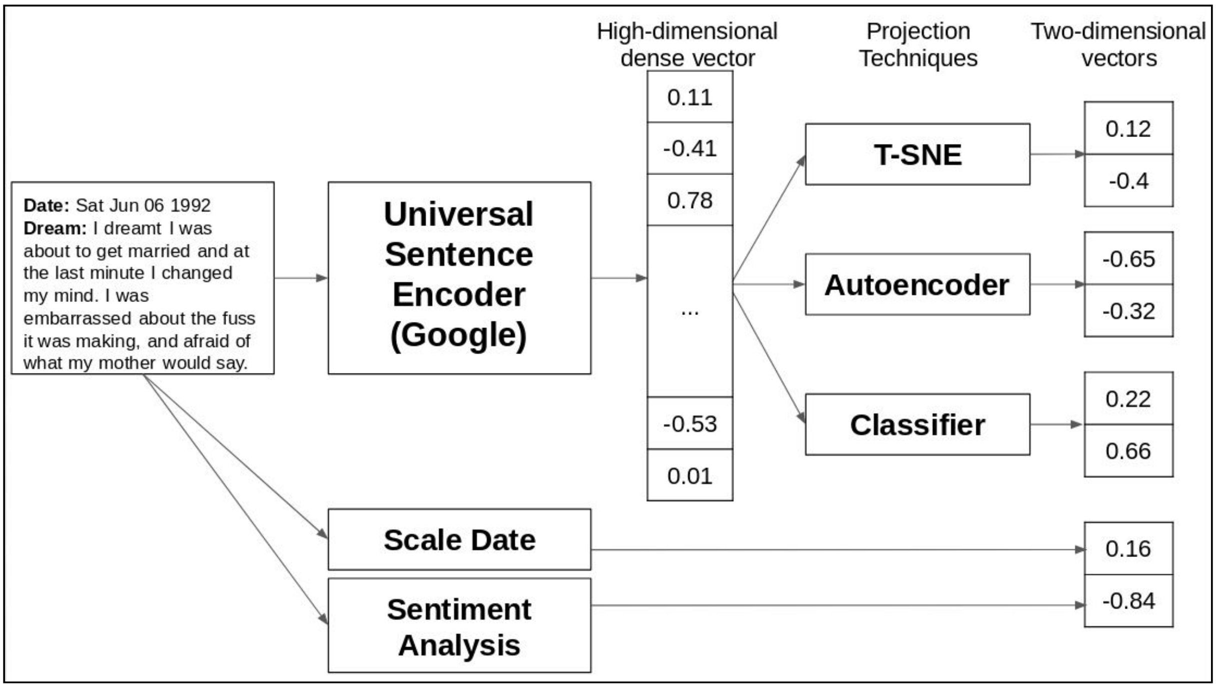

High-dimensional dense vector
Projection Techniques
Two-dimensional vectors
Date: Sat Jun 06 1992
Dream: I dreamt I was about to get married and at the last minute I changed my mind. I was embarrassed about the fuss it was making, and afraid of what my mother would say.
Universal Sentence Encoder (Google)
0.11
-0.41
0.78
...
-0.53
0.01
T-SNE
Autoencoder
Classifier
0.12
-0.4
-0.65
-0.32
0.22
0.66
Scale Date
Sentiment Analysis
0.16
-0.84

Figure 2: An overview of the system which assigns two-dimensional coordinates to dreams (textual data).

2.1 Dense Representations

Natural language processing was revolutionized within the past five years by word embeddings, most notably [6], a technique to assign dense vectors to words, encoding both their semantic and syntactic meaning. Since then, the top performing on models on almost all natural language processing tasks (such as machine translation, text classification, sentiment analysis, etc.) make use of dense representations of text. In this work we use the Universal Sentence Encoder [3, 4], a state-of-the-art method for encoding not only words but entire sentences. We use the Universal Sentence Encoder to assign a high-dimensional vector to each dream (i.e. 512 dimensions), containing semantic information about the dream.

2.2 Sentiment Analysis

A common tool for analyzing text data is sentiment analysis. This holds true for dreams, as both good dreams and nightmares are of special interest to all types of users. We used Google's sentiment analysis API [7] to assign a sentiment score to each dream.

2.3. Projections

The user can select between multiple different projections of the textual dream data into two dimensions.

2.3.1. T-Distributed Stochastic Neighbor Embedding (T-SNE)

T-Distributed Stochastic Neighbor Embedding [8] is a technique which attempts to keep the closest neighbors of each high-dimensional point the same in the low-dimensional projection by using pairwise distances as conditional probabilities of a point neighboring another point. T-SNE is chosen by default when starting the system, as it does a very good job of grouping dreams by semantic similarity.

2.3.2 Autoencoder

Autoencoders are a popular technique in machine learning in which a neural network is used to learn a lower-dimensional representation of high-dimensional data. Autoencoders learn to encode data from the input layer into an intermediate layer which has fewer dimensions than the input layer by simultaneously training a decoder which attempts to reconstruct the original input from the intermediate layer. Due to the intermediate layer being of lower dimensions, autoencoders need to reduce the dimensionality of the data while preserving the most important information. Autoencoders are trained by minimizing the reconstruction error.

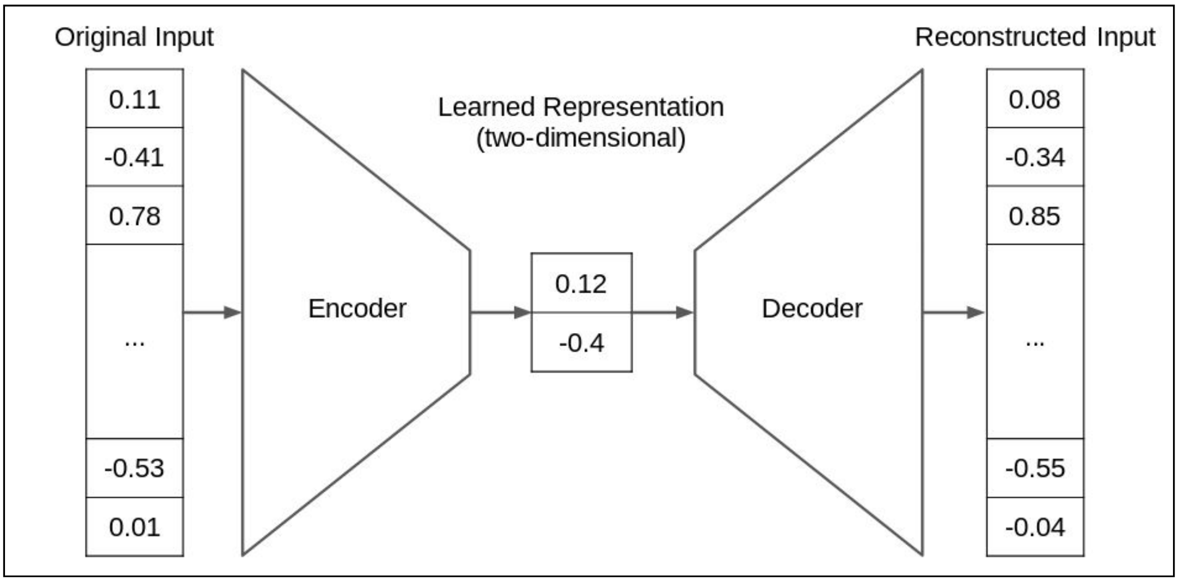


Figure 3: A simplified illustration of an autoencoder. It is trained by minimizing the reconstruction error

2.3.3 Binary Neural Network Classifier (Gender)

Binary classifiers are trained in a supervised fashion, on a labeled dataset of two distinct classes. They are maximized to correctly assign the right labels to the training training data. Neural network most commonly consist of multiple layers, the last layer being of two dimensions in the binary case. The output from the last layer is then passed through the softmax function to go from arbitrary numeric values to probabilities. In our specific case, we trained a neural network classifier using the gender of the dream series as a label. This network can give answers to the question “Do men and women dream differently?”, with the last layer’s two values containing information on how many “male” elements and how many “female”elements were found in a dream. Note that the neural network in our work strongly overfits on the training data, and therefore may lead erroneous or misleading dream classification - this problem would be mitigated with a larger dataset and a more sophisticated way of training the network, for example by performing stratified splits and evaluating the performance on a validation set rather than the training set itself. This is merely an example of how we can use binary neural network classifiers to give a meaningful two-dimensional projection of the data.

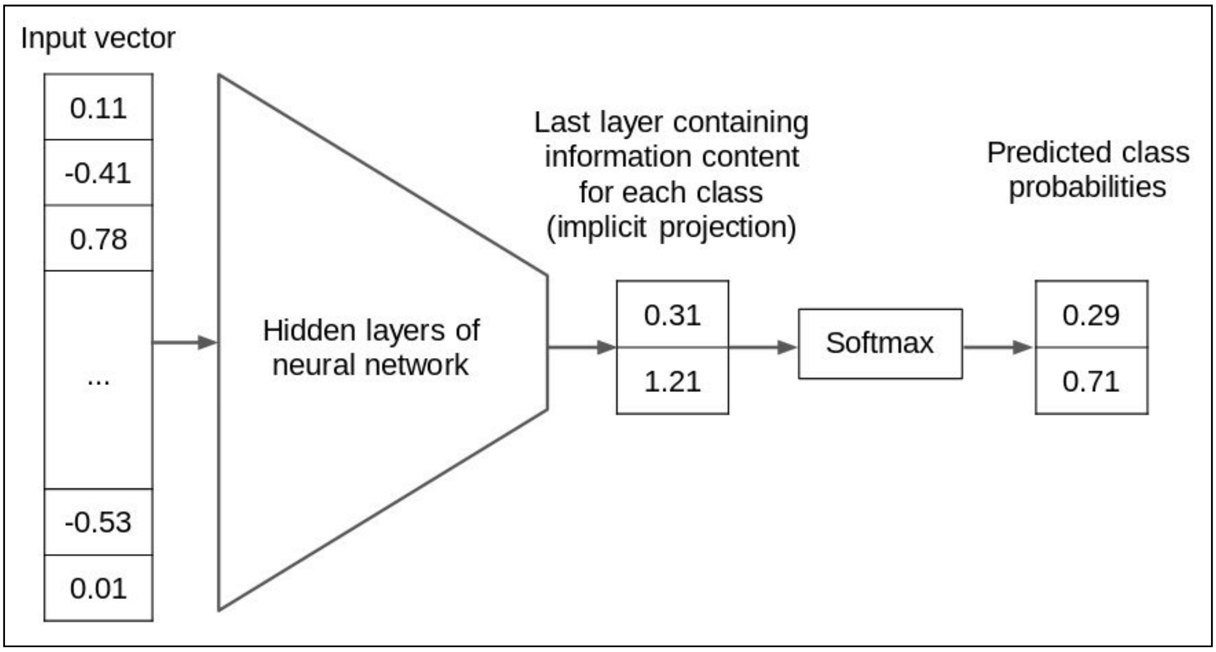


Figure 4: A simplified illustration of a binary neural network classifier and how to use it as a projection method

2.3.4 Date and Sentiment

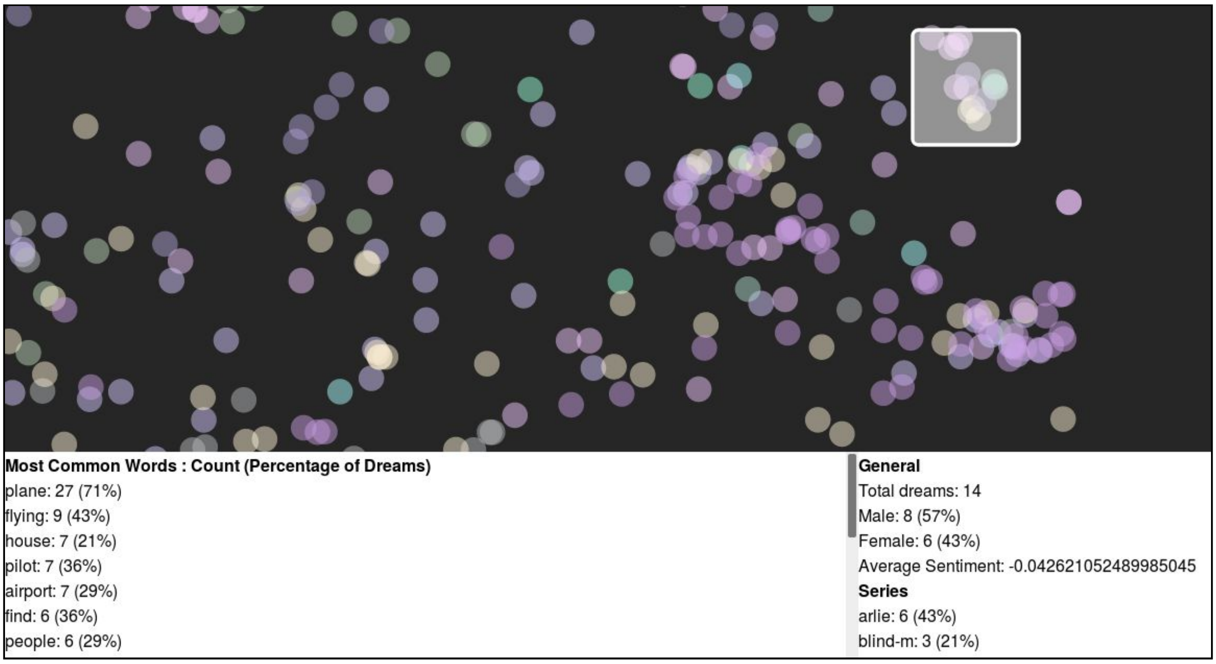
Another chosen projection in this work is the projection of date on the x-axis and the output of a sentiment analysis model on the y-axis. This projection lets the user explore dreams in their original chronological order, while highlighting exceptionally good and bad dreams. Combining this projection with the interactions described in section 2.4, this projection enables the user to also analyze shifts in dream themes over time and find common themes in particularly good and bad dreams.

2.4 Visualization and Interaction

When clicking on a dream, the user will see the dream's content and information about the dream series. However, this is a very low-level view as the user can only see information about one dream at a time. Our system thus makes use of color and spatial projections to help the user explore dreams and identify themes in clusters without having to read every single dream in a cluster.

2.4.1 Area Summary

Our system supports spatial slicing in order to summarize an area. The user can select a rectangular area on the screen to see a summary of all dreams in this area. The 100 most common words, along with their count and the percentage of selected dreams they were in will be displayed. The system also displays more general summarizing information such as the percentage of dreams by each gender in that area, the average sentiment, and what percentages of dreams originate which specific dream series. This feature helps the user evaluate why certain dreams appear close to each other in the projection and therefore helps explore dreams by their semantic information. This feature can also be used to analyze entire dream series at once, i.e. the user can display only a single dream series and draw a rectangle around all dreams to highlight the topics and themes in this dream series. When using the projection by date and sentiment, the area summary feature can be used to analyze how general themes shift over time and what generally determines positively or negatively classified dreams.



Most Common Words : Count (Percentage of Dreams)
plane: 27 (71%)
flying: 9 (43%)
house: 7 (21%)
pilot: 7 (36%)
airport: 7 (29%)
find: 6 (36%)
people: 6 (29%)

General
Total dreams: 14
Male: 8 (57%)
Female: 6 (43%)
Average Sentiment: -0.042621052489985045
Series
arlie: 6 (43%)
blind-m: 3 (21%)

Figure 5: An area on the screen is selected in the T-SNE projection, revealing the shared theme of flying

2.4.2 Color

By default, each dream is colored according to its dream series. However, this is not the only information that is interesting to the user. Dreams, no matter from which dream series, can be colored differently according to certain attributes. The user can switch to color by gender mode. This allows the user to explore the commonness of a gender dreaming about a certain topic. Another way to use color is in keyword search. The user enters a search query, and all dreams matching the query are colored in green whereas those not matching are colored in grey. Whereas the area selection described above answers the question “What semantic meaning does this area represent?”, using the search option gives an answer to “In what area can I find dreams about this topic?”.

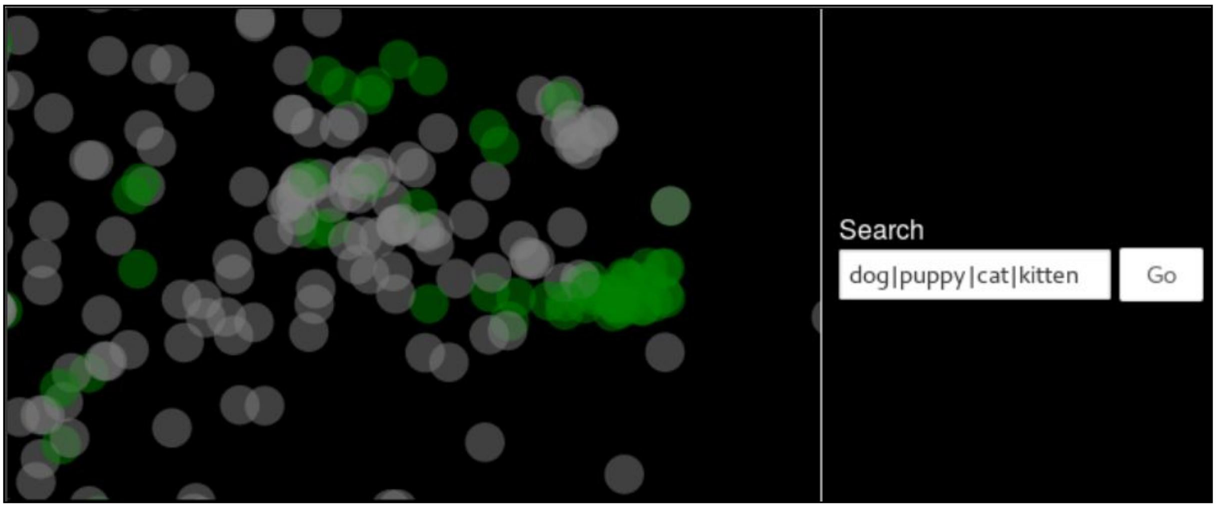


Figure 6: A keyword search for either “dog”, “puppy”, “cat”, or “kitten” reveals a group of dreams about pets

2.4.3 Nearest Neighbor Links

To facilitate the exploration of the dataset, we included nearest neighbor links in the original high-dimensional dream embedding, which are displayed when a dream is selected. The system displays links to up to ten nearest neighbors, but only to those which are already displayed by the current dream series selection. The closest neighbor is displayed in white with an opacity of one, the following neighbors are colored in with decreasing opacity. This enables the user to explore the dreams by semantic similarity because the euclidean distance between dreams in the high-dimensional embedding is an approximation to the semantic similarity of dreams. It also allows users to evaluate the quality of projection techniques.

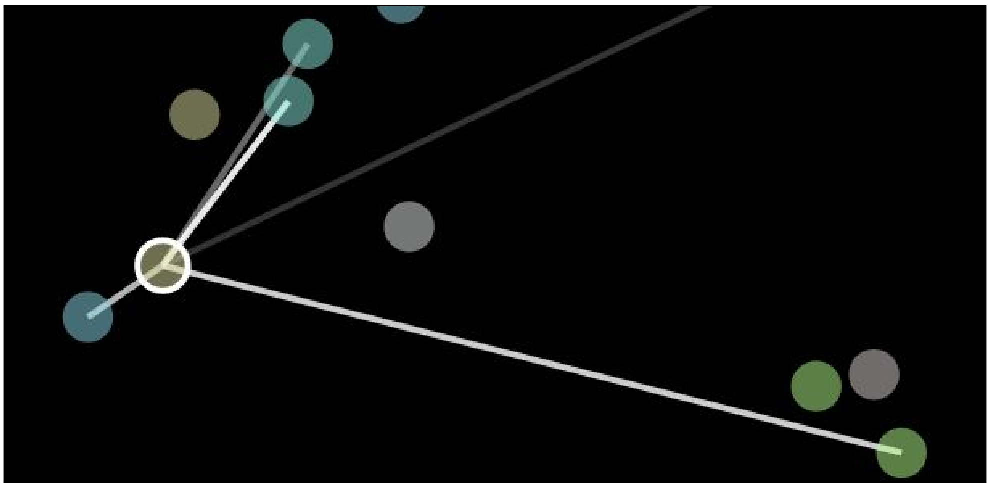


Figure 7: Nearest neighbor links displayed across multiple dream series. Closer neighbors are more opaque

3. Target Users and Representative Tasks

The target users of our system can be grouped into three major categories: psychologists with an interest in studying dreams, computer scientists or data scientists with an interest in the natural language processing and dimensionality reduction techniques used, and just regular users who have no specific interest or question in mind.

3.1 Psychologists and Dream Researchers

Dream research has been of interest to many psychologists all the way from ancient history, popularized through Sigmund Freud, and up into the modern day. Some popular theories reason that the subconscious nature of dreams reveals an honest insight into the dreamer's thoughts, desires, and emotions. In this project we treat dreams from the perspective of natural language processing and computer science, rather than from a psychological perspective. However, the system developed is designed to also help psychologists answer various questions about dreams. For example, a psychologist might be interested in the question if the symptoms of PTSD in the Vietnam veteran, whose dreams are contained in the DreamBank, diminish over time. Using the projection by date and sentiment together with the keyword search, psychologists can quickly identify nightmares. The area summary tool could also be used to give insights into a question like "How do the themes that teenages dream about differ from those of college students?". Dreams are often classified by psychologists using the Hall/Van de Castle Norms [9], which relies on identifying themes (e.g. violence, sexuality), certain keywords (e.g. words of action), and recurring characters. This process is painstaking and takes a large amount of time and resources - a system like our visualization could facilitate the classification process using tools like keyword search and area analysis.

3.2 Data Scientists

This project is an excellent demonstration of how powerful the modern natural language processing techniques employed are. Dense representations, especially contextualized sentence vectors have led to significant improvement in almost any natural language task. However, instead of reading about another benchmark being shattered by these new techniques, this system provides an interactive demonstration of their effectiveness. Indeed it is truly remarkable that we can take a sequence of words, encode them in a high-dimensional vector, and then remove all but two dimensions by projection and still keep such a considerable amount of semantic information. When our system is being used by data scientists, exploring the effectiveness of these techniques will be their main focus. Questions they will ask themselves are "How close is the true closest semantic neighbor of this dream in the projection into two dimension?", "What semantic, structural or topological information are preserved best by which projections?" or "Why was this dream assigned a positive sentiment?". To give answers to the first two questions, the user can make extensive use of the area summary tool and the nearest neighbor links displayed. To give an answer to the last question, the user could use the area summary tool on all dreams above a certain sentiment threshold and find common themes and words. In fact, data scientists could use a system like this to test the performance of their own sentiment classifier, their own sentence embeddings, or their own novel dimensionality reduction techniques in an interactive way. A numerical accuracy or performance on a validation set only says so much, whereas it would be very easy to test new machine learning models on the DreamBank dataset and evaluate the quality of the models in an interactive and qualitative rather than quantitative way.

3.3 Regular Users

The system was also designed with users in mind, that have no specific research purposes in mind and that are not particularly interested in the methods used in the system. These users simply want to enjoy the deep dive into other people's dreams and be immersed by the system - they want to read dreams and find those that are interesting to them. The system is well-suited for these kind of users because the tools given to explore the data (keyword search, choice of projection techniques, nearest neighbor links, and area summary) do not require any particular knowledge about the underlying technical/mathematical concepts, nor any particular knowledge about dreams. Another requirement for regular users is the ease of use, which is why the system is available as an online website, with an intuitive responsive design. Because users don't need to install any software, make any specific downloads, and the website runs on the vast majority of modern devices, there are practically no hurdles and the project can reach a large audience. We have also facilitated this by including share buttons for social media.

4. Discussion

In this section we briefly discuss the strengths and limitations of this system.

The dataset consists of over 26,000 english dreams, however common web browsers on average machines can only handle up to 10,000 dreams displayed at a time without showing signs of a bad user experience (i.e. stuttering while zooming and panning). This is a limitation inherent to the rendering speed of d3.js.

Another limitation of the system is the performance of the sentiment classifier. While many dreams are assigned a reasonable sentiment, some are not. For example this dream from the *vietnam_vet* series is assigned a positive sentiment: “I am with Jamie H. on a trail in Vietnam. We are in a hostile environment, though at the same time on vacation or holiday. The terrain is wooded and tropical. I begin to tell him about war, about Chicom's and combat and death. We are both armed. I am fearful of going out on ambush. The next moment I begin to weep and sob deeply. Recollection ends here.” Obviously this is wrong - the misclassification of sentiments could be mitigated by training a sentiment analysis system specifically on the DreamBank dataset, instead of using a publicly available system. Seeing the rapid improvements in sentiment analysis systems in general though, one can expect to see significant improvements in publicly available sentiment analysis systems’ accuracies within the coming years or even just months.

The system could also be improved with some manual work on data cleaning. There are 3406 dates missing. Some dreams contain dates within their dream content, which we were not able to parse without significant effort. Some dream series descriptions also contain broken references to other parts of the original DreamBank page.

5. Future Work

The quality of the projections in this system are all made by projections which are commonly used in the field of machine learning. A larger dream dataset would lead to an even more interesting visualization with higher quality projections. Because the data follows a very common structure - a large collection of short texts divided into a relatively small numbers of series - our system could be reused on other data without any major modifications. An example could be to collect thousands of tweets from a small selected group of accounts (e.g. United States senators and house representatives) and simply rerun the entire code with the new data. In many other text corpora, additional information that could make for insightful projections can be found. Examples would be a projection by geospatial coordinates to discover themes specific to certain areas of the globe, or a projection by age of the author to discover patterns and links between the age of the author and the contents of the texts.

6. Acknowledgements

A special thank you to the Adam Schneider & G. William Domhoff of the Psychology Dept., UC Santa Cruz, for their outstanding effort to collect and maintain the DreamBank dataset. We also want to thank Google for making the Universal Sentence Encoder, a very powerful tool, publicly available. We also want to thank the wonderful Freya Behrens for taking the time to thoroughly test our system and giving us thorough feedback on the user experience.

[1] Schneider, A., & Domhoff, G. W. (2018). The Quantitative Study of Dreams. Retrieved June 6, 2018 from http://www.dreamresearch.net/

[2] Schneider, A., & Domhoff, G. W. (2018). DreamBank. Retrieved June 6, 2018 from http://www.dreambank.net/

[3] Google via Tensorflow Hub. Universal Sentence Encoder. Retrieved June 6, 2018 from https://www.tensorflow.org/hub/modules/google/universal-sentence-encoder/1

[4] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Universal Sentence Encoder. arXiv:1803.11175, 2018.

[5] Matt Bierner. DreamScrape. Retrieved June 6, 2018 from https://github.com/mattbierner/DreamScrape

[6] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

[7] Google. Cloud Natural Language. Retrieved June 6, 2018 from https://cloud.google.com/natural-language/

[8] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov (2008): 2579-2605.

[9] Hall, Calvin S., and Robert L. Van de Castle. "The content analysis of dreams." (1966).