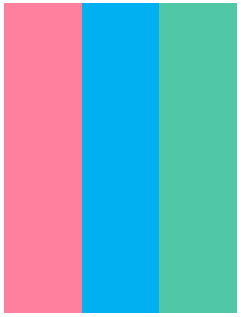


Latent Dirichlet Allocation Topic Modeling

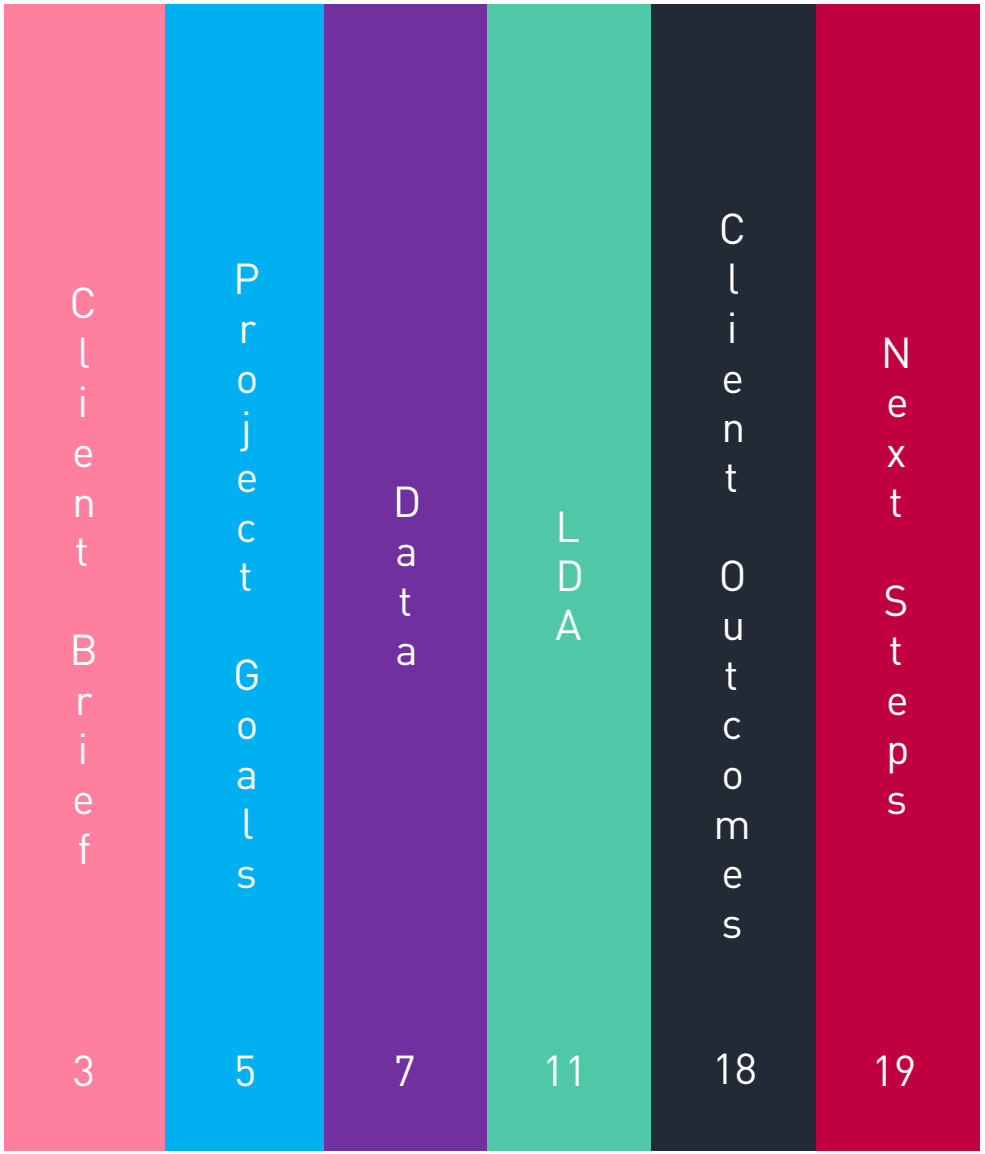
Leveraging an unsupervised learning technique to Identify relevant topics (i.e. clusters) of interest within the reddit parenting forum.

Client: [redacted]

Prepared by: Matthew B. Murrell

Date: 03.12.2020

Location: General Assembly, Boston



Topic Modeling with LDA

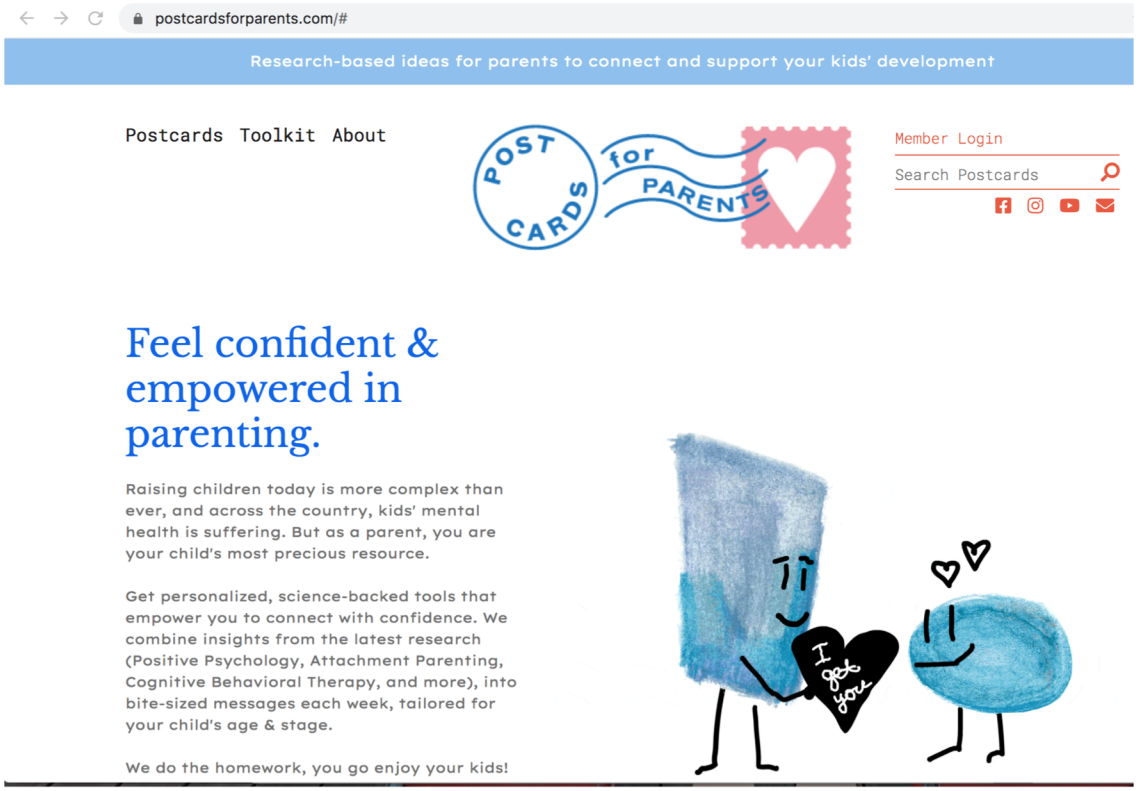
Overview

Client Brief

Introducing ...

Meet the Client

The client is a multi-channel, digital information platform for parents seeking clinically validated, holistic information about topics relating to the development of, and parenting issues about their school-aged children (roughly ages 5yrs to 18 yrs old).



Client Brief

If we only had an algorithm ...

Problem Statement

The client is interested in a way to classify an individual's "parenting style" in order to better align relevant content with the specific user. More specifically, the client is interested in exploring whether posts obtained from the subreddit, parenting, could be analyzed to identify distinct clusters of parenting "types" (according to the psychological concept of "attachment theory"), which could then be used as a basis to classify new users on the platform.

Project Goals

Value Generation ...

Insight and Automation

- Determine if it is possible to identify distinct parenting styles on the basis of their posts on www.reddit.com/r/parenting.
- If yes, extract the resulting clusters for use as the basis of a user classification algorithm.

Project Goals

It's good to have a plan, but ...

Reality Check

Short of having a purpose-built questionnaire designed to tease out aspects of an individual's parenting style attributes, it would not be possible to derive personality clusters from this particular body of documents without having trained a classification model in advance.

More specifically, there needs to be a baseline of responses to common questions, or responses to a common prompt and then labeled with a classifier in order to determine to which group a new user belongs.

A new, separate project that addresses the original goals is in the works, but for now ...

Adapting on the Fly

Given the realities and limitations of the available data, discussions with the client yielded an alternate exploration:

Could the reddit/parenting content be mined for topical information that might yield practical insights for the ongoing product development?

This new question gave rise to the following topic modeling (i.e. unsupervised clustering) analysis, for which the LDA algorithm was leveraged ...

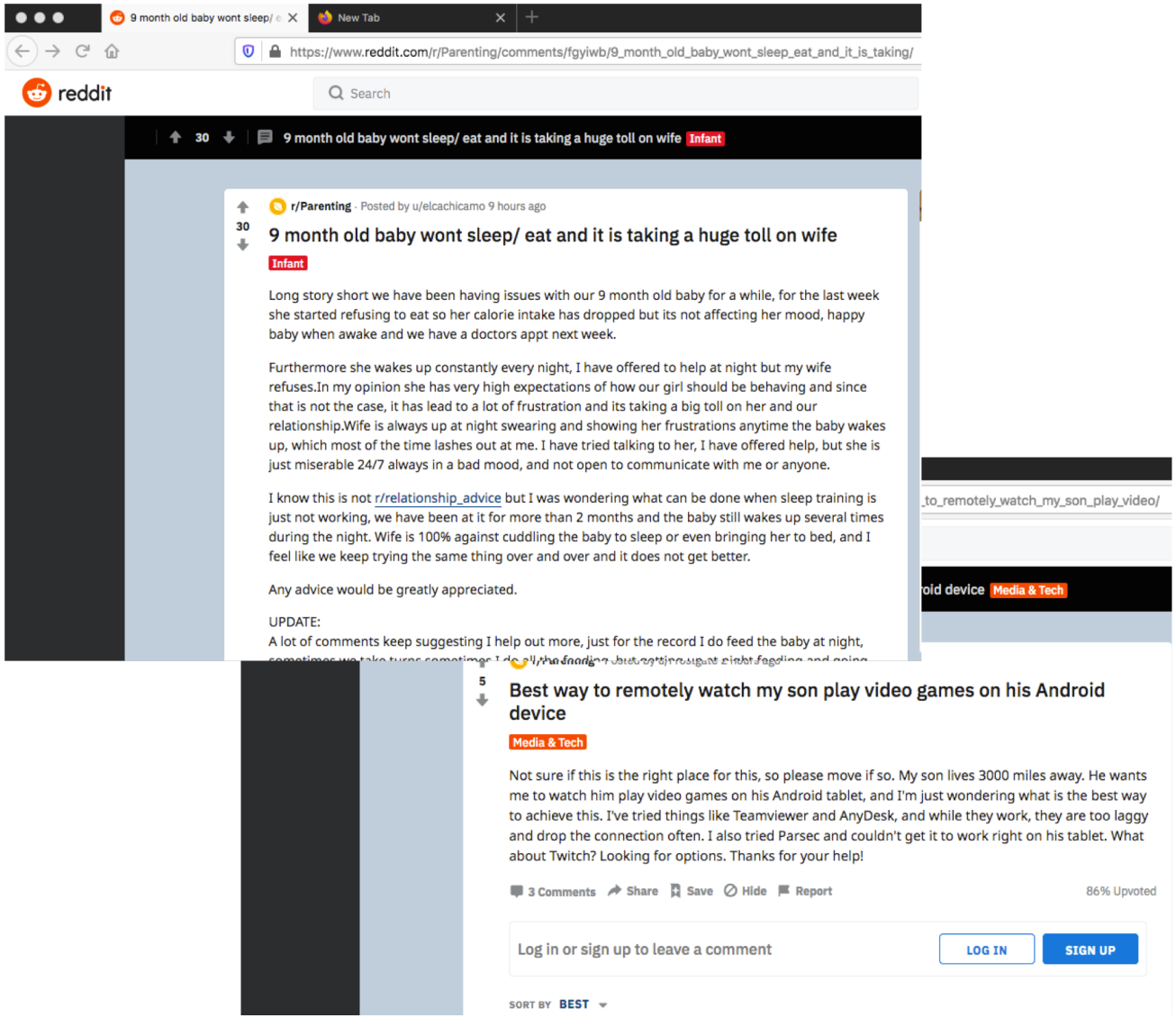
Data

What is this data anyway?

Source	www.reddit.com/r/parenting
Number of Observations in Sample	≈ 103K unique posts
Final Feature Count (i.e. model "vocabulary")	1,659 unique words in the lemmatized vocabulary 1,725 unique words in the stemmed vocabulary
Timeframe of Observations	Data spans posts submitted on reddit.com/r/parenting from 01 Jan 2015 through 31 Jan 2020
Parenting Group Size	≈ 1.9MM unique users (as 03.2020)

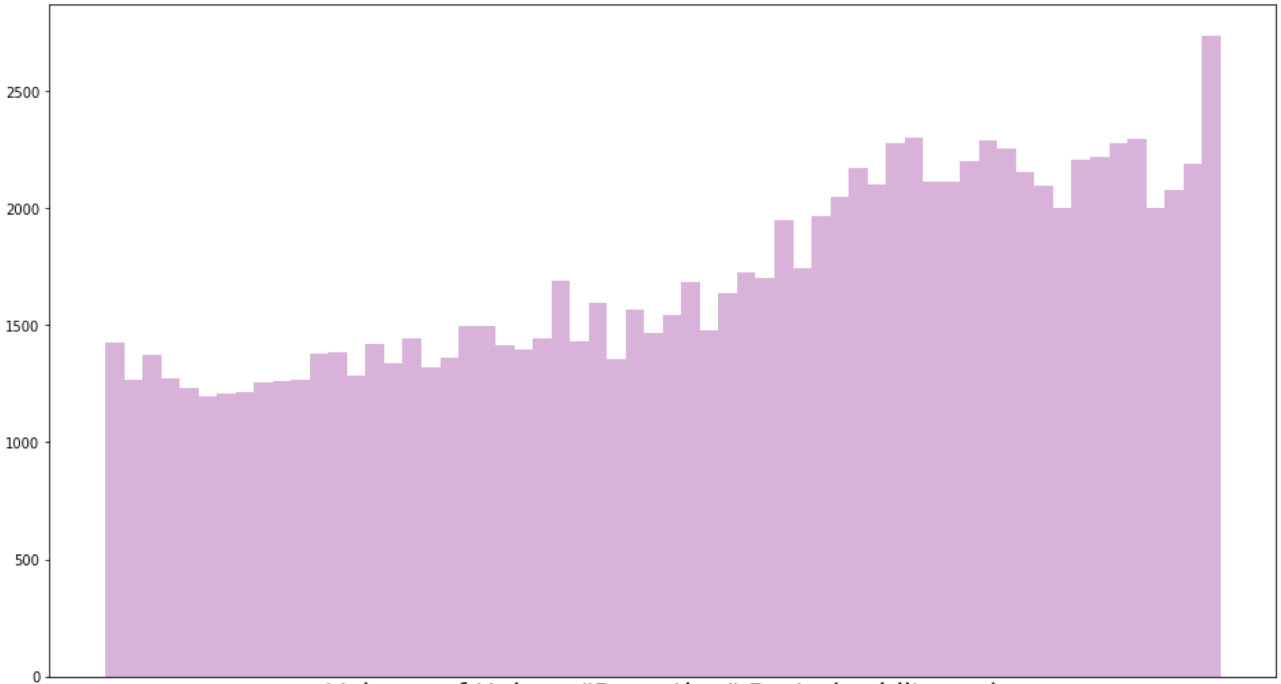
Data

Seeing it in the raw ...



Data

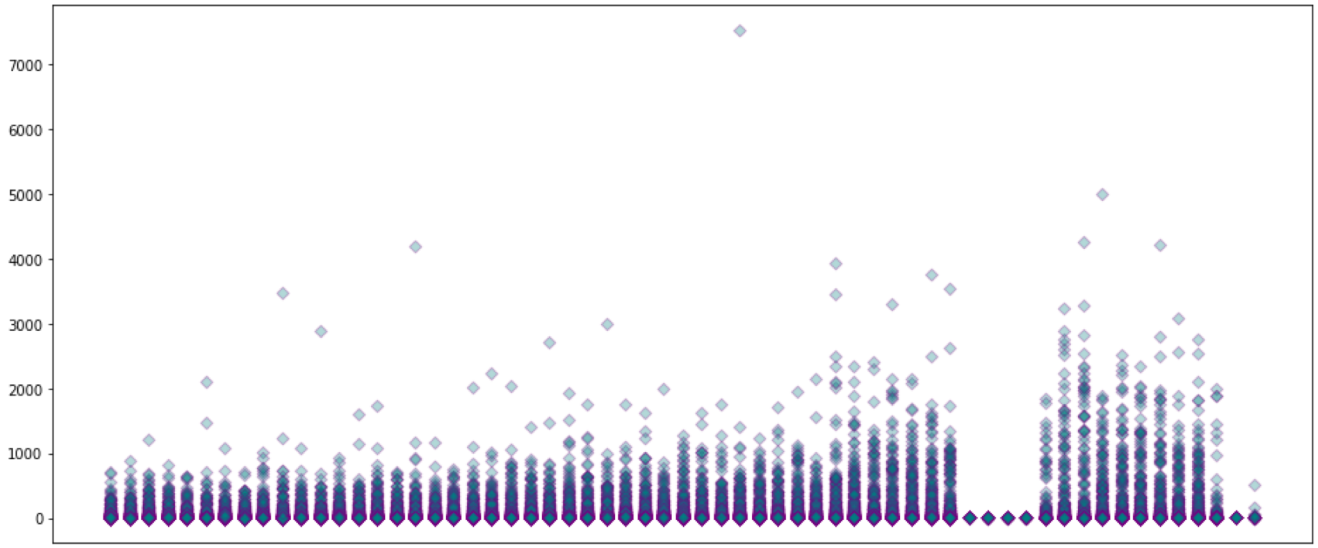
A validating perspective ...



- Over the roughly five-year period from which the raw data were pulled, there is a clear increase in the volume of parenting-related posts on this platform.
- The trend suggests that the appetite for help with parenting issues, broadly speaking, is on the rise.

Data

Validation reinforcement …



- This chart illustrates the frequency distribution of post scores (up-votes) by other members of the subreddit group.
- The trend similarly suggests an increase in appetite for parenting advice.
- Note the four-moth gap in the data set – after extensive review, it seems there was some technical issue with the data at the source data – realities of working with raw data.

LDA model

Clustering with LDA …

Some Context

- This is essentially an unsupervised clustering challenge
- Several potential models could be applied in this situation. A few of the most well known are listed below:
  - K means
  - DB Scan
  - Latent Dirichlet Allocation
  - Etc. …
- Various pros and cons to each approach, but LDA was ultimately chosen for this particular scenario because of its relative strength with Natural language Processing (NLP) applications.

LDA model

What is LDA?

So what, exactly, is LDA?

From the scikit learn LDA documentation:

> *Latent Dirichlet Allocation is a generative probabilistic model for collections of discrete datasets such as text corpora. It is also a topic model* [i.e. unsupervised clustering model] *that is used for discovering abstract topics from a collection of documents.*

Great. **But what does that actually mean???**

Source: https://scikit-learn.org/stable/modules/decomposition.html#latentdirichletallocation

LDA model

In other words ...

Conceptually, in NLP applications, LDA effectively treats the "data" as being comprised of two different, but related distributions ...

- A document is made up of a distribution of topics
- A topic is made up of a distribution of words
- Assumes K (prior) number of topics

A topic is represented as a weighted list of words, and the model also estimates the percentage each document talks about each topic.

Note that the K prior number of topics introduces the Bayesian aspect present in the LDA model.

LDA model

Key model parameters

Alpha parameter is the Dirichlet prior concentration parameter that represents ***document-topic density*** – with a higher alpha, documents are assumed to be made up of more topics and result in a more specific topic distribution per document.

Beta parameter is the same prior concentration parameter that represents ***topic-word density*** – with a high beta, topics are assumed to be made up of most of the words and result in a more specific word distribution per topic.

Source: Kapadia, "Topic Modeling in Python: Latent Dirichlet Allocation (LDA)"; https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

LDA model

Performance of competing specifications

Number of Topics	Coherence Score
N = 8	0.4266
N = 12	0.4531
N = 15	0.4227
N = 18	0.4139
N = 24	0.4011

1) It was unclear whether lemmatizing vs. stemming would have a meaningful impact on the ultimate model performance, so experiments were conducted with both methods for comparison. In the end, the differences appeared immaterial, so the above summarizes only lemmatized model results.

2) Note that this dual-vocabulary approach proved computationally expensive and time consuming, but given the initial lack of familiarity with the LDA algorithm it seemed a worthwhile experiment to explore.

LDA model

Audience
participation time!

Spin up the demo!

LDA model

Name that topic …

Time for some audience participation …

Record Demo Output Here:

Topic 1		Topic 7	
Topic 2		Topic 8	
Topic 3		Topic 9	
Topic 4		Topic 10	
Topic 5		Topic 11	
Topic 6		Topic 12	

Client Outcomes

Some unexpected insights ...

Despite the fact that it was necessary to revise the initial project goals because of the data limitations, the alternate exploration yielded valuable strategic insights for the client:

- Topic modeling of the corpus of reddit parenting posts from the last five years validated not only the range of topics covered by the client's own content strategy, but also one of their core concepts driving their broader product strategy (we obviously cannot be more specific because of proprietary concerns, but it was a meaningful insight).
- The quantitative aspects of the posting patterns also reinforced the client's belief that the market for this type of content has been growing, and continues to grow, further reinforcing the validation of one of the key assumptions driving the business model.

Limitations & Next Steps

Human interpretation and model limits.

- LDA performs better when there are both high N (number of docs) and when docs are longer (i.e. high word count/doc).[1]
  - However, this also drives two of the drawbacks to this approach
    1. Greedy for processing power
    2. Increasingly time consuming
- And of course in the converse situation of low-word-count docs (e.g. < 20 words/doc), LDA struggles.[2]
- Computers can't interpret the topics.

1. https://stackoverflow.com/questions/46326173/understanding-lda-topic-modelling-too-much-topic-overlap
2. Xiaohui Yan, et al, "A Biterm Topic Model for Short Texts"; http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.4032&rep=rep1&type=pdf

Limitations & Next Steps

Where do we go from here?

- Cluster topics in client's content modules (which we won't get into here for proprietary reasons)
- Compare topics across the two corpora
  - Determine if there are any gaps in relevant topic areas
- Explore optimization possibilities to improve the clustering model performance …
  - Additional removal of high frequency (non-stopword) terms
  - Additional hyperparameter tuning