Report

Topic Modeling of TED Corpus

Designing a Recommendation Engine to identify Similar Topics

Introduction:

I consider TED transcripts as knowledge inputs for training a topic model to be used as a source for recommendation engine for TED.com. In this process I will be dealing with a considerably sizeable count of transcripts.

The criteria I would be interested to ensure the quality of inputs to the recommendation engine are

A. The interpretability of the topics to get a good view of the topic.
B. The generalizability of the topics to easily group or cluster the topics.

The thought behind the above criteria is to ensure the topics obtained from model are easy to process and optimize. For this I would be using LDA model for training topics for the following reasons

a. The volume of data to be trained is high and the data is unsupervised.
b. The output requires multiple human validations in retraining them for better interpretability and generalizability.
c. LDA is highly efficient in processing data at a faster rate which aids for the constraints mentioned above.

Topic Modeling:

A topic model has been trained using Mallet via Gensim Library on JupyterHub and it is included as

For an unsupervised learning that involves the use of natural language, it is important to carry out certain pre-processing steps. The pre-processing steps used in the code are

a. Removal of unwanted space in text to avoid multiple iterations of code while processing the transcripts.
b. Tokenizing the words to identify, sort and group the words in a better way. It is also done to.
c. Tokenizing also helps in matching words to their synonyms for identifying semantic structure between different topics thus adds to ease of topic generalization. This is done by matching tokens to dictionary.
d. Stop words in English are identified and removed as they occur frequently, have negligible impact in interpretation of topics and increases the processing time.
e. Scanning through the sample ted videos, the words ‘laughter’ and ‘applause’ seems to common in all the transcripts and they have been removed.

A number of iterations were carried out across topic sizes between 10 to 70 with an interval of 15. After all these trails, I felt the transcripts can be split across 45 diverse topics which could be interpreted and generalized easily without losing its meaning. When the number of topics spans above 45, in few topics the core theme of the topic is getting compromised to some extent.

The coherence score plot in Image 1. throws a clear view on the semantic coherence across a range of topics taken into consideration. It can be clearly seen that the coherence score for topic range 45 is better when compared with the other values.

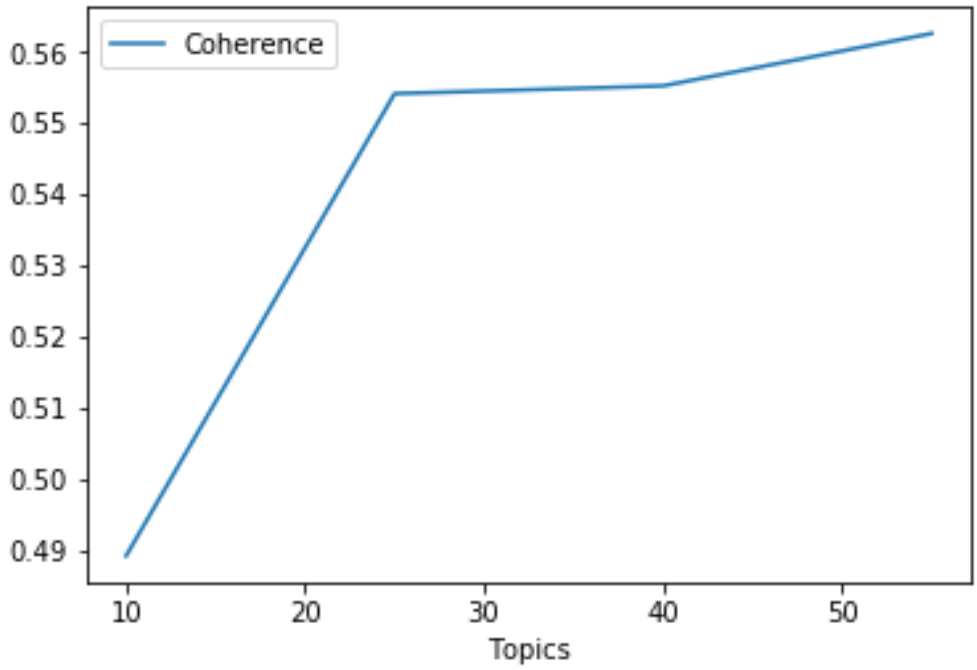

Image 1.

Recommendation Engine:

The three probable recommendations from TED.com transcript corpus for example videos of Kriti Sharma and Timothy Bartik are

Kriti Sharma:

Document Index: 2472
Similarity Score 0.44443643
I'm here to offer you a new way to think about my field, artificial intelligence. I think the purpose of AI is to empower humans with machine intelligence. And as machines get smarter, we get smarter. I call this "humanistic AI" — artificial intelligence designed to meet human needs by collaborating and augmenting people. Now, today I'm happy to see that the idea of an intelligent assistant is mainstream. It's the well-accepted metaphor for the interface between humans and AI. And the one I help ...

Document Index: 1148
Similarity Score 0.4228189
I work on helping computers communicate about the world around us. There are a lot of ways to do this, and I like to focus on helping computers to talk about what they see and understand. Given a scene like this, a modern computer-vision algorithm can tell

you that there's a woman and there's a dog. It can tell you that the woman is smiling. It might even be able to tell you that the dog is incredibly cute. I work on this problem thinking about how humans understand and process the world. The th ...

Document Index: 2479
Similarity Score 0.42236295
I'm going to talk about how AI and mankind can coexist, but first, we have to rethink about our human values. So let me first make a confession about my errors in my values. It was 11 o'clock, December 16, 1991. I was about to become a father for the first time. My wife, Shen-Ling, lay in the hospital bed going through a very difficult 12-hour labor. I sat by her bedside but looked anxiously at my watch, and I knew something that she didn't. I knew that if in one hour, our child didn't come, I ...

All the three topics are relevant and they speak about the short coming of AI when it comes to biased decisions when AI starts making the decisions which humans make. The key though in all three topics complement each other.

Timothy Bartik:

Document Index: 2855
Similarity Score 0.29231533
I'm an historian. And what I love about being an historian is it gives you perspective. Today, I'd like to bring that perspective to education in the United States. About the only thing people can agree on is that the most strategic time for a child to start learning is early. Over 50 years ago, there was a watershed moment in early education in the US called "Head Start." Now, historians love watersheds because it makes it so easy to talk about what came before and what's happened since. Befor ...

Document Index: 726
Similarity Score 0.26803243
The world is changing with really remarkable speed. If you look at the chart at the top here, you'll see that in 2025, these Goldman Sachs projections suggest that the Chinese economy will be almost the same size as the American economy. And if you look at the chart for 2050, it's projected that the Chinese economy will be twice the size of the American economy, and the Indian economy will be almost the same size as the American economy. And we should bear in mind here that these projections wer ...

Document Index: 3402
Similarity Score 0.26561737
Well, this is about state budgets. This is probably the most boring topic of the whole morning. But I want to tell you, I think it's an important topic that we need to care about. State budgets are big, big money — I'll show you the numbers — and they get very little scrutiny. The understanding is very low. Many of the people involved have special interests or short-term interests that get them not thinking about what the implications of the trends are. And these budgets are the key for our futu ...

All the three topics are partially relevant. The first topic speaks about the importance of early education. The second topic speaks on political part with the view of China citing

its unity and compares it with America where state migrations leads to partiality in providing facilities. The third topic matches on the budgets considering the long term aspects instead of short term aspects.

So the algorithm, though doesn't always recommend the exactly similar topics, it matches on some key parts of the topics that are relevant to be learned next to know better on the interested topics.

Further refinement of the tokenization by investing more time in analysing and removing some irrelevant tokens will surely be an improvement on this.

Appendix

[(0,
  '0.022*"hours" + 0.019*"minutes" + 0.016*"night" + 0.016*"sleep" + 0.013*"home" +
0.013*"place" + 0.013*"days" + 0.013*"morning" + 0.012*"moment" + 0.012*"start" +
0.010*"10" + 0.009*"hour" + 0.009*"end" + 0.009*"months" + 0.009*"week" +
0.009*"call" + 0.009*"room" + 0.008*"weeks" + 0.008*"half" + 0.008*"30"'),
 (1,
  '0.059*"data" + 0.023*"problem" + 0.018*"information" + 0.011*"understand" +
0.011*"simple" + 0.010*"big" + 0.010*"model" + 0.009*"start" + 0.008*"important" +
0.008*"show" + 0.008*"bit" + 0.007*"pretty" + 0.007*"process" + 0.007*"problems" +
0.007*"turns" + 0.006*"car" + 0.006*"test" + 0.006*"looked" + 0.006*"based" +
0.006*"step"'),
 (2,
  '0.034*"community" + 0.014*"communities" + 0.012*"local" + 0.012*"africa" +
0.011*"poor" + 0.009*"poverty" + 0.009*"started" + 0.009*"working" + 0.009*"access"
+ 0.008*"lives" + 0.007*"development" + 0.007*"places" + 0.007*"health" +
0.007*"village" + 0.007*"support" + 0.007*"country" + 0.006*"home" +
0.006*"resources" + 0.006*"public" + 0.006*"live"'),
 (3,
  '0.063*"school" + 0.047*"kids" + 0.036*"students" + 0.029*"education" +
0.019*"learning" + 0.016*"learn" + 0.015*"high" + 0.014*"schools" + 0.014*"teachers"
+ 0.014*"class" + 0.013*"teacher" + 0.013*"teach" + 0.013*"college" + 0.012*"student"
+ 0.010*"teaching" + 0.010*"young" + 0.010*"children" + 0.009*"university" +
0.007*"classroom" + 0.007*"math"'),
 (4,
  '0.041*"money" + 0.030*"dollars" + 0.019*"business" + 0.016*"companies" +
0.015*"market" + 0.014*"buy" + 0.014*"company" + 0.012*"pay" + 0.009*"percent" +
0.009*"industry" + 0.008*"cost" + 0.007*"product" + 0.007*"financial" + 0.007*"sell" +
0.007*"price" + 0.007*"000" + 0.006*"bank" + 0.006*"million" + 0.006*"paid" +
0.006*"economy"'),
 (5,
  '0.013*"business" + 0.012*"job" + 0.012*"team" + 0.012*"success" + 0.011*"working"
+ 0.010*"company" + 0.008*"ideas" + 0.008*"hard" + 0.007*"career" +
0.007*"important" + 0.007*"leadership" + 0.006*"leaders" + 0.006*"successful" +
0.006*"idea" + 0.006*"jobs" + 0.006*"change" + 0.006*"failure" + 0.006*"organization"
+ 0.006*"start" + 0.006*"worked"'),
 (6,
  '0.075*"music" + 0.036*"sound" + 0.020*"play" + 0.018*"hear" + 0.016*"sounds" +
0.015*"song" + 0.013*"voice" + 0.009*"singing" + 0.009*"playing" + 0.008*"listening"
+ 0.008*"noise" + 0.008*"piece" + 0.008*"listen" + 0.008*"dance" + 0.007*"la" +
0.007*"hearing" + 0.007*"audience" + 0.007*"sing" + 0.006*"musical" +
0.006*"yeah"'),
 (7,
  '0.020*"media" + 0.018*"information" + 0.013*"news" + 0.009*"internet" +
0.009*"facebook" + 0.008*"twitter" + 0.008*"online" + 0.007*"security" +
0.007*"government" + 0.007*"public" + 0.007*"found" + 0.006*"privacy" +

0.006*"happened" + 0.005*"message" + 0.005*"tv" + 0.005*"report" + 0.005*"social" +
0.005*"post" + 0.005*"email" + 0.005*"journalists"'),
 (8,
  '0.036*"technology" + 0.031*"computer" + 0.024*"machine" + 0.023*"human" +
0.015*"robot" + 0.013*"computers" + 0.013*"machines" + 0.013*"robots" +
0.012*"build" + 0.010*"future" + 0.010*"system" + 0.009*"intelligence" + 0.009*"real"
+ 0.008*"lab" + 0.007*"device" + 0.007*"humans" + 0.007*"artificial" + 0.006*"ai" +
0.006*"learning" + 0.006*"working"'),
 (9,
  '0.028*"ocean" + 0.025*"water" + 0.022*"fish" + 0.022*"sea" + 0.010*"animals" +
0.009*"deep" + 0.008*"oceans" + 0.007*"ice" + 0.007*"species" + 0.006*"coral" +
0.006*"boat" + 0.006*"fishing" + 0.006*"surface" + 0.006*"place" + 0.006*"places" +
0.006*"marine" + 0.006*"sharks" + 0.005*"coast" + 0.005*"high" + 0.005*"bottom"'),
 (10,
  '0.092*"brain" + 0.015*"brains" + 0.013*"memory" + 0.011*"neurons" +
0.008*"activity" + 0.007*"information" + 0.006*"signals" + 0.006*"understand" +
0.006*"visual" + 0.005*"human" + 0.005*"cells" + 0.005*"control" + 0.005*"body" +
0.005*"normal" + 0.005*"mind" + 0.005*"electrical" + 0.005*"areas" + 0.005*"cortex"
+ 0.005*"ability" + 0.005*"area"'),
 (11,
  '0.058*"city" + 0.029*"cities" + 0.022*"car" + 0.016*"cars" + 0.013*"york" +
0.012*"street" + 0.011*"place" + 0.011*"public" + 0.010*"urban" + 0.010*"live" +
0.009*"places" + 0.008*"space" + 0.008*"park" + 0.008*"road" +
0.008*"neighborhood" + 0.007*"map" + 0.007*"streets" + 0.007*"built" +
0.007*"home" + 0.007*"building"'),
 (12,
  '0.055*"story" + 0.038*"stories" + 0.016*"feel" + 0.012*"change" + 0.011*"told" +
0.010*"truth" + 0.010*"share" + 0.010*"felt" + 0.009*"voice" + 0.009*"feeling" +
0.008*"hope" + 0.008*"remember" + 0.008*"identity" + 0.008*"moment" +
0.008*"talk" + 0.008*"live" + 0.007*"telling" + 0.007*"hear" + 0.006*"mind" +
0.006*"speak"'),
 (13,
  '0.036*"war" + 0.018*"africa" + 0.010*"african" + 0.010*"military" + 0.010*"country"
+ 0.009*"peace" + 0.008*"conflict" + 0.008*"international" + 0.008*"east" +
0.007*"afghanistan" + 0.007*"refugees" + 0.007*"countries" + 0.006*"security" +
0.006*"middle" + 0.006*"soldiers" + 0.006*"iraq" + 0.006*"west" + 0.006*"weapons" +
0.005*"killed" + 0.005*"europe"'),
 (14,
  '0.035*"game" + 0.034*"play" + 0.031*"number" + 0.016*"games" + 0.016*"2" +
0.014*"1" + 0.014*"playing" + 0.013*"numbers" + 0.011*"dog" + 0.011*"3" +
0.010*"video" + 0.009*"answer" + 0.009*"win" + 0.009*"times" + 0.008*"room" +
0.008*"head" + 0.006*"10" + 0.006*"players" + 0.006*"dogs" + 0.006*"set"'),
 (15,
  '0.030*"art" + 0.015*"images" + 0.015*"image" + 0.015*"show" + 0.015*"film" +
0.011*"project" + 0.010*"artist" + 0.009*"video" + 0.009*"picture" + 0.008*"camera" +
0.008*"pictures" + 0.008*"movie" + 0.008*"artists" + 0.007*"painting" +

0.007*"museum" + 0.007*"visual" + 0.007*"create" + 0.006*"started" + 0.006*"sort" +
0.006*"piece"'),
 (16,
  '0.037*"body" + 0.012*"move" + 0.010*"legs" + 0.009*"hand" + 0.009*"arm" +
0.008*"feel" + 0.008*"feet" + 0.007*"foot" + 0.007*"leg" + 0.006*"head" +
0.006*"movement" + 0.006*"physical" + 0.006*"walk" + 0.006*"bodies" +
0.006*"control" + 0.006*"muscles" + 0.005*"moving" + 0.005*"blind" + 0.005*"left" +
0.005*"walking"'),
 (17,
  '0.017*"paper" + 0.013*"hand" + 0.012*"box" + 0.012*"show" + 0.009*"wall" +
0.008*"bit" + 0.007*"hands" + 0.007*"hold" + 0.007*"yeah" + 0.007*"big" +
0.007*"top" + 0.007*"piece" + 0.006*"cut" + 0.006*"audience" + 0.006*"magic" +
0.006*"end" + 0.006*"inside" + 0.005*"thought" + 0.005*"stick" + 0.005*"line"'),
 (18,
  '0.021*"internet" + 0.019*"technology" + 0.017*"phone" + 0.013*"digital" +
0.012*"online" + 0.012*"web" + 0.012*"video" + 0.011*"network" +
0.011*"information" + 0.010*"google" + 0.009*"open" + 0.008*"mobile" +
0.007*"content" + 0.007*"project" + 0.007*"software" + 0.007*"share" +
0.006*"phones" + 0.006*"tools" + 0.006*"networks" + 0.006*"data"'),
 (19,
  '0.046*"earth" + 0.027*"space" + 0.024*"planet" + 0.013*"universe" + 0.013*"stars" +
0.012*"mars" + 0.012*"moon" + 0.011*"sun" + 0.010*"planets" + 0.009*"black" +
0.009*"system" + 0.009*"star" + 0.008*"sky" + 0.008*"hole" + 0.007*"surface" +
0.007*"solar" + 0.007*"galaxy" + 0.006*"light" + 0.006*"atmosphere" +
0.006*"galaxies"'),
 (20,
  '0.028*"human" + 0.017*"question" + 0.014*"wrong" + 0.011*"reality" +
0.011*"answer" + 0.010*"idea" + 0.010*"reason" + 0.009*"true" + 0.009*"sense" +
0.008*"knowledge" + 0.007*"moral" + 0.007*"questions" + 0.007*"beings" +
0.007*"view" + 0.007*"nature" + 0.006*"problem" + 0.006*"choice" + 0.006*"ideas" +
0.005*"thought" + 0.005*"point"'),
 (21,
  '0.034*"cancer" + 0.030*"cells" + 0.024*"disease" + 0.016*"blood" + 0.014*"body" +
0.010*"cell" + 0.010*"drug" + 0.010*"diseases" + 0.009*"drugs" + 0.009*"hiv" +
0.009*"virus" + 0.007*"tissue" + 0.006*"immune" + 0.006*"stem" + 0.006*"heart" +
0.006*"risk" + 0.006*"tumor" + 0.005*"malaria" + 0.005*"breast" + 0.005*"medicine"'),
 (22,
  '0.021*"death" + 0.015*"fear" + 0.011*"die" + 0.011*"lost" + 0.010*"dead" +
0.009*"died" + 0.008*"lives" + 0.008*"end" + 0.008*"face" + 0.008*"days" +
0.007*"alive" + 0.006*"human" + 0.006*"kill" + 0.006*"live" + 0.006*"dying" +
0.006*"pain" + 0.005*"suffering" + 0.005*"chance" + 0.005*"home" + 0.005*"killed"'),
 (23,
  '0.026*"talk" + 0.022*"stuff" + 0.021*"sort" + 0.018*"bit" + 0.018*"idea" +
0.016*"thought" + 0.015*"ca" + 0.014*"big" + 0.014*"basically" + 0.014*"guy" +
0.013*"interesting" + 0.013*"started" + 0.012*"pretty" + 0.012*"yeah" + 0.011*"guys" +
0.011*"thinking" + 0.010*"happened" + 0.010*"couple" + 0.010*"talking" +
0.010*"ted"'),

(24,
  '0.019*"law" + 0.019*"police" + 0.012*"prison" + 0.011*"violence" + 0.011*"justice"
+ 0.010*"system" + 0.010*"crime" + 0.010*"case" + 0.009*"legal" + 0.008*"court" +
0.007*"state" + 0.007*"states" + 0.007*"jail" + 0.006*"gun" + 0.006*"rights" +
0.006*"laws" + 0.006*"criminal" + 0.006*"society" + 0.005*"drug" + 0.005*"judge"'),
(25,
  '0.021*"study" + 0.017*"social" + 0.015*"research" + 0.012*"attention" +
0.011*"found" + 0.010*"behavior" + 0.010*"group" + 0.008*"stress" + 0.008*"studies"
+ 0.008*"mind" + 0.007*"positive" + 0.006*"feel" + 0.006*"happiness" +
0.006*"negative" + 0.006*"experience" + 0.006*"effect" + 0.006*"asked" +
0.006*"experiment" + 0.005*"science" + 0.005*"emotions"'),
(26,
  '0.022*"dna" + 0.021*"human" + 0.015*"species" + 0.013*"humans" +
0.013*"bacteria" + 0.011*"genes" + 0.011*"evolution" + 0.010*"genetic" + 0.010*"cell"
+ 0.009*"gene" + 0.009*"biology" + 0.008*"genome" + 0.007*"molecules" +
0.007*"animals" + 0.007*"living" + 0.007*"organisms" + 0.006*"biological" +
0.006*"microbes" + 0.006*"animal" + 0.005*"ago"'),
(27,
  '0.031*"countries" + 0.023*"china" + 0.018*"country" + 0.017*"india" +
0.014*"growth" + 0.014*"economic" + 0.012*"global" + 0.011*"economy" +
0.011*"chinese" + 0.011*"united" + 0.011*"states" + 0.010*"percent" + 0.009*"income"
+ 0.008*"population" + 0.007*"europe" + 0.007*"wealth" + 0.007*"state" +
0.006*"jobs" + 0.006*"rich" + 0.006*"progress"'),
(28,
  '0.014*"species" + 0.014*"trees" + 0.013*"animals" + 0.013*"tree" + 0.013*"forest" +
0.010*"nature" + 0.009*"plants" + 0.009*"birds" + 0.008*"river" + 0.008*"plant" +
0.007*"land" + 0.007*"insects" + 0.007*"bees" + 0.006*"forests" + 0.005*"ants" +
0.005*"bird" + 0.005*"natural" + 0.005*"animal" + 0.005*"desert" + 0.005*"sand"'),
(29,
  '0.026*"science" + 0.017*"universe" + 0.015*"theory" + 0.012*"physics" +
0.011*"space" + 0.011*"matter" + 0.009*"particles" + 0.009*"quantum" + 0.008*"idea"
+ 0.008*"scientists" + 0.007*"mathematics" + 0.007*"speed" + 0.007*"energy" +
0.007*"force" + 0.007*"atoms" + 0.006*"einstein" + 0.006*"field" +
0.006*"mathematical" + 0.006*"shape" + 0.006*"structure"'),
(30,
  '0.107*"women" + 0.051*"men" + 0.031*"woman" + 0.026*"girls" + 0.023*"sex" +
0.015*"female" + 0.015*"male" + 0.014*"gender" + 0.014*"girl" + 0.012*"sexual" +
0.011*"man" + 0.011*"young" + 0.009*"boys" + 0.009*"talk" + 0.008*"gay" +
0.005*"violence" + 0.005*"culture" + 0.005*"marriage" + 0.005*"talking" +
0.004*"society"'),
(31,
  '0.042*"health" + 0.028*"care" + 0.024*"medical" + 0.023*"patients" +
0.018*"patient" + 0.017*"heart" + 0.015*"doctor" + 0.015*"doctors" + 0.014*"hospital"
+ 0.011*"treatment" + 0.010*"surgery" + 0.009*"medicine" + 0.008*"pain" +
0.007*"blood" + 0.007*"disease" + 0.007*"depression" + 0.006*"system" +
0.005*"high" + 0.005*"hospitals" + 0.005*"healthy"'),
(32,

'0.021*"wanted" + 0.017*"started" + 0.017*"thought" + 0.014*"told" + 0.013*"man" +
0.013*"knew" + 0.012*"father" + 0.011*"asked" + 0.010*"looked" + 0.010*"home" +
0.009*"decided" + 0.009*"found" + 0.009*"mother" + 0.009*"house" + 0.009*"family"
+ 0.008*"learned" + 0.008*"felt" + 0.008*"met" + 0.007*"gave" + 0.007*"dad"'),
(33,
'0.049*"percent" + 0.025*"000" + 0.025*"change" + 0.024*"million" +
0.018*"climate" + 0.017*"10" + 0.015*"billion" + 0.011*"number" + 0.011*"100" +
0.011*"global" + 0.011*"30" + 0.010*"20" + 0.010*"ago" + 0.010*"50" + 0.010*"1" +
0.009*"times" + 0.009*"half" + 0.008*"carbon" + 0.008*"problem" + 0.007*"future"'),
(34,
'0.051*"food" + 0.043*"water" + 0.016*"eat" + 0.010*"waste" + 0.008*"plastic" +
0.008*"plant" + 0.008*"farmers" + 0.007*"grow" + 0.006*"eating" + 0.006*"produce" +
0.006*"feed" + 0.006*"agriculture" + 0.006*"growing" + 0.005*"farm" + 0.005*"meat"
+ 0.005*"sugar" + 0.005*"bread" + 0.005*"plants" + 0.005*"oil" + 0.005*"system"'),
(35,
'0.046*"design" + 0.029*"building" + 0.014*"build" + 0.012*"material" +
0.012*"materials" + 0.011*"architecture" + 0.011*"space" + 0.011*"process" +
0.011*"built" + 0.010*"create" + 0.010*"designed" + 0.010*"project" + 0.009*"idea" +
0.009*"buildings" + 0.008*"structure" + 0.007*"working" + 0.007*"form" +
0.007*"designers" + 0.006*"place" + 0.006*"making"'),
(36,
'0.034*"black" + 0.024*"god" + 0.024*"white" + 0.018*"american" + 0.013*"america"
+ 0.012*"religion" + 0.010*"religious" + 0.009*"race" + 0.009*"compassion" +
0.008*"man" + 0.008*"faith" + 0.007*"african" + 0.007*"history" + 0.007*"community"
+ 0.006*"culture" + 0.006*"americans" + 0.006*"church" + 0.005*"color" +
0.005*"free" + 0.005*"young"'),
(37,
'0.052*"love" + 0.032*"feel" + 0.018*"person" + 0.013*"talk" + 0.009*"talking" +
0.009*"friends" + 0.008*"relationship" + 0.008*"experience" + 0.008*"happy" +
0.008*"feeling" + 0.008*"makes" + 0.008*"thinking" + 0.007*"bad" + 0.007*"real" +
0.006*"job" + 0.006*"conversation" + 0.006*"friend" + 0.006*"hate" + 0.005*"joy" +
0.005*"means"'),
(38,
'0.065*"light" + 0.020*"color" + 0.018*"blue" + 0.017*"red" + 0.014*"skin" +
0.014*"eyes" + 0.012*"eye" + 0.012*"green" + 0.010*"turn" + 0.009*"colors" +
0.008*"lights" + 0.008*"dark" + 0.007*"invisible" + 0.006*"white" + 0.006*"inside" +
0.006*"smell" + 0.006*"tiny" + 0.006*"hair" + 0.005*"waves" + 0.005*"high"'),
(39,
'0.069*"children" + 0.042*"child" + 0.036*"family" + 0.022*"parents" +
0.022*"mother" + 0.020*"baby" + 0.019*"age" + 0.018*"kids" + 0.017*"young" +
0.015*"born" + 0.013*"families" + 0.012*"babies" + 0.012*"home" + 0.011*"care" +
0.011*"older" + 0.011*"lives" + 0.011*"birth" + 0.007*"mothers" + 0.006*"adults" +
0.006*"autism"'),
(40,
'0.020*"history" + 0.017*"century" + 0.009*"ancient" + 0.008*"modern" +
0.007*"ago" + 0.006*"age" + 0.006*"story" + 0.006*"king" + 0.005*"man" +
0.005*"began" + 0.005*"centuries" + 0.005*"famous" + 0.005*"early" + 0.005*"000" +

0.004*"cultures" + 0.004*"20th" + 0.004*"stone" + 0.004*"cultural" + 0.004*"gold" + 0.004*"past"'),
 (41,
  '0.044*"energy" + 0.023*"air" + 0.014*"power" + 0.012*"nuclear" + 0.012*"solar" + 0.010*"wind" + 0.010*"water" + 0.010*"oil" + 0.010*"heat" + 0.009*"fly" + 0.009*"electricity" + 0.009*"ice" + 0.009*"gas" + 0.009*"fuel" + 0.008*"temperature" + 0.007*"flying" + 0.007*"high" + 0.006*"miles" + 0.006*"coal" + 0.005*"degrees"'),
 (42,
  '0.022*"political" + 0.017*"government" + 0.017*"country" + 0.013*"democracy" + 0.011*"states" + 0.010*"politics" + 0.010*"president" + 0.010*"rights" + 0.009*"citizens" + 0.009*"change" + 0.008*"power" + 0.008*"public" + 0.008*"united" + 0.007*"vote" + 0.007*"state" + 0.007*"movement" + 0.007*"election" + 0.006*"national" + 0.006*"party" + 0.005*"values"'),
 (43,
  '0.020*"system" + 0.020*"power" + 0.018*"future" + 0.013*"change" + 0.012*"systems" + 0.010*"ways" + 0.010*"problems" + 0.010*"social" + 0.010*"create" + 0.008*"powerful" + 0.008*"understand" + 0.007*"term" + 0.007*"rules" + 0.007*"society" + 0.007*"technology" + 0.006*"century" + 0.006*"thinking" + 0.006*"problem" + 0.006*"order" + 0.006*"global"'),
 (44,
  '0.037*"language" + 0.035*"book" + 0.032*"words" + 0.027*"read" + 0.024*"word" + 0.020*"write" + 0.020*"books" + 0.019*"english" + 0.016*"writing" + 0.013*"reading" + 0.011*"wrote" + 0.011*"written" + 0.008*"learn" + 0.008*"languages" + 0.007*"letters" + 0.007*"speak" + 0.007*"letter" + 0.007*"library" + 0.006*"found" + 0.006*"poetry"')]

Works Cited

Bisgin, Halil, et al. “Mining FDA Drug Labels Using an Unsupervised Learning Technique - Topic Modeling.” BMC Bioinformatics, Eighth Annual MCBIOS Conference https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-12-S10-S11

Kriti Sharma, “How to keep human bias out of AI”, TEDxWarwick https://www.ted.com/talks/kriti_sharma_how_to_keep_human_bias_out_of_ai

Timothy Bartik, “The economic case for preschool”, TEDxMiamiUniversity https://www.ted.com/talks/timothy_bartik_the_economic_case_for_preschool