DHUM25A43 - 4 Investigating with AI

NLP

Feb 18th, 2025

What we saw last time

- Classic NLP tasks
  - classification (Sentiment, toxicity, spam, …)
  - POS (adj, nouns, verbs, …)
  - stopwords
  - NER : persons, location, GPEs, ORGs, …
- Embeddings
  - text to vectors
  - similarity score

Today: Hands on

- Demo: Wikipedia AI page (30mn)
  - NER
  - Spacy
  - Embeddings with transformers
  - corpus normalization
  - live notebook
- Your turn (1h)
  - similar analysis using wikipedia but related to your subject
- Demo Next (30mn)
  - Next step in data analysis using the NYT API

Normalization

Text normalization for frequency based analysis on words

- remove stopwords
- lowercase
- lemmatize
    - gaming, to game, gamer, games => game
    -

Topic modeling

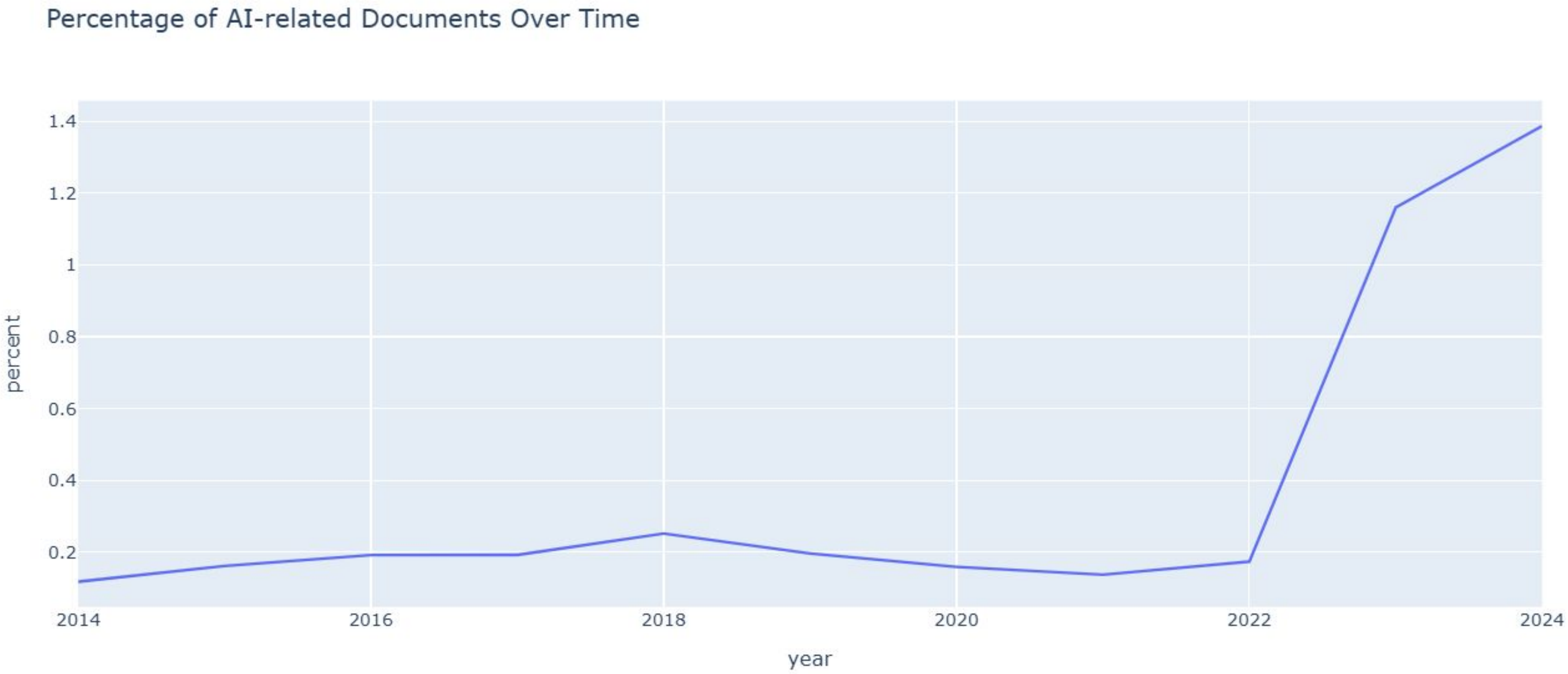
Automatically detecting the topics in a corpus based on the relative frequency of the different words

- requires normalization
- crystal ball
- difficult to find the right number of topics beforehand

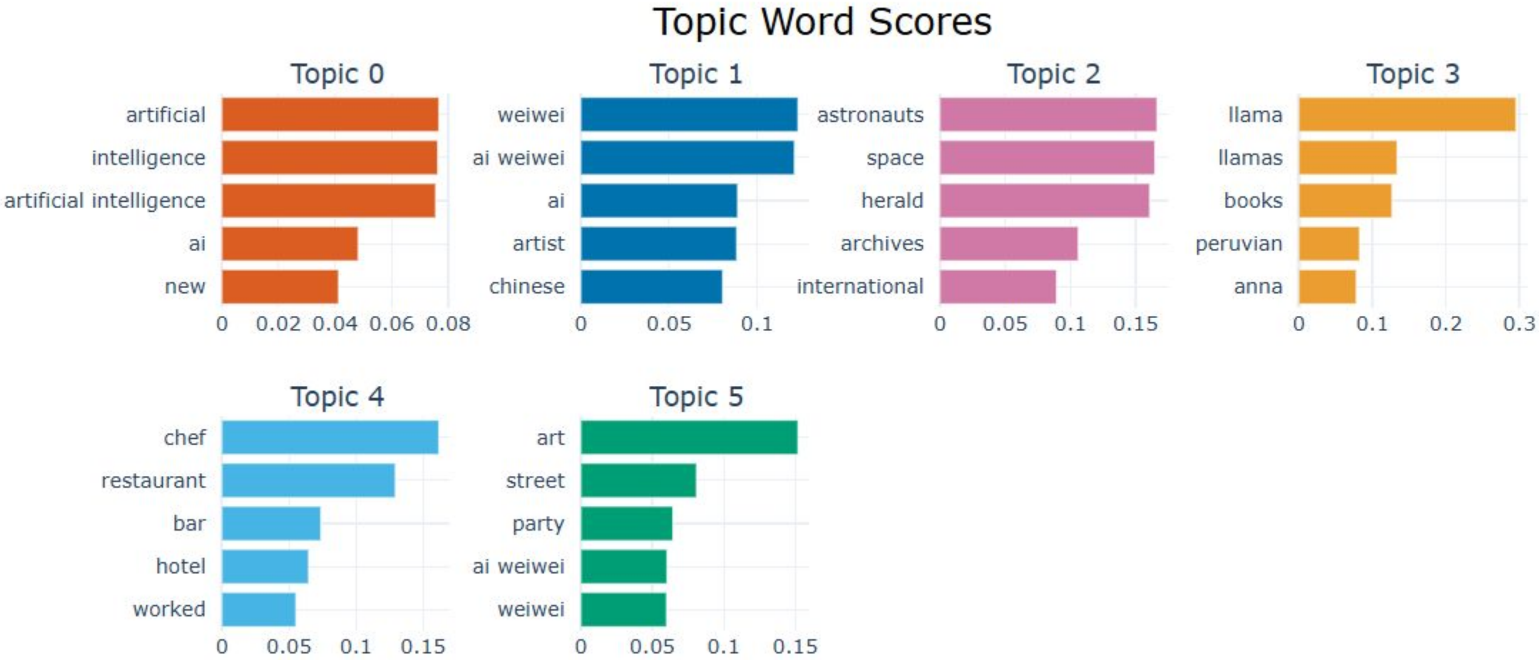
Exploratory textual analysis using a **New York Times (NYT) dataset.** The analysis includes:

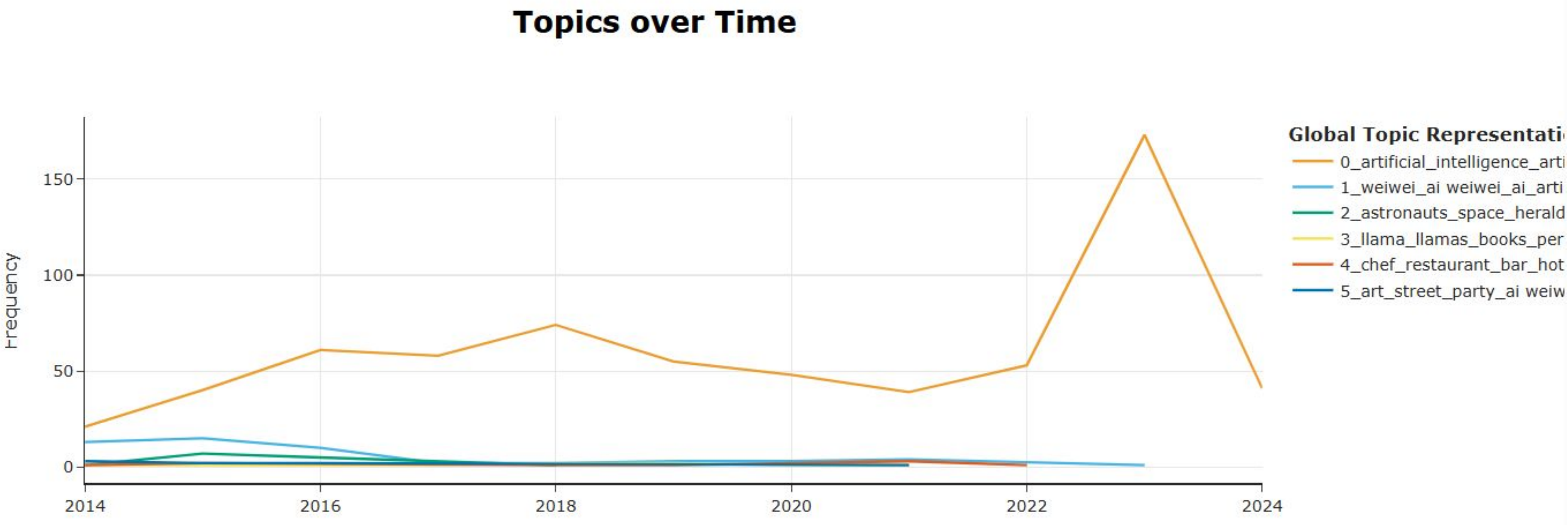
- **Data preprocessing**: Cleaning and extracting a subset of relevant data
- **AI-related publication detection**: Using keyword-based filtering
- **Descriptive statistics**: Time trends and document counts
- **Topic modeling**: Identifying main themes in the dataset
- **Document embeddings & interactive 2D mapping**
- **Time series analysis**: Tracking topic evolution over time

https://drive.google.com/file/d/1GFR_SvUU8w5CO81KJsDR5SNNNtmDC3WK/view?usp=sharing

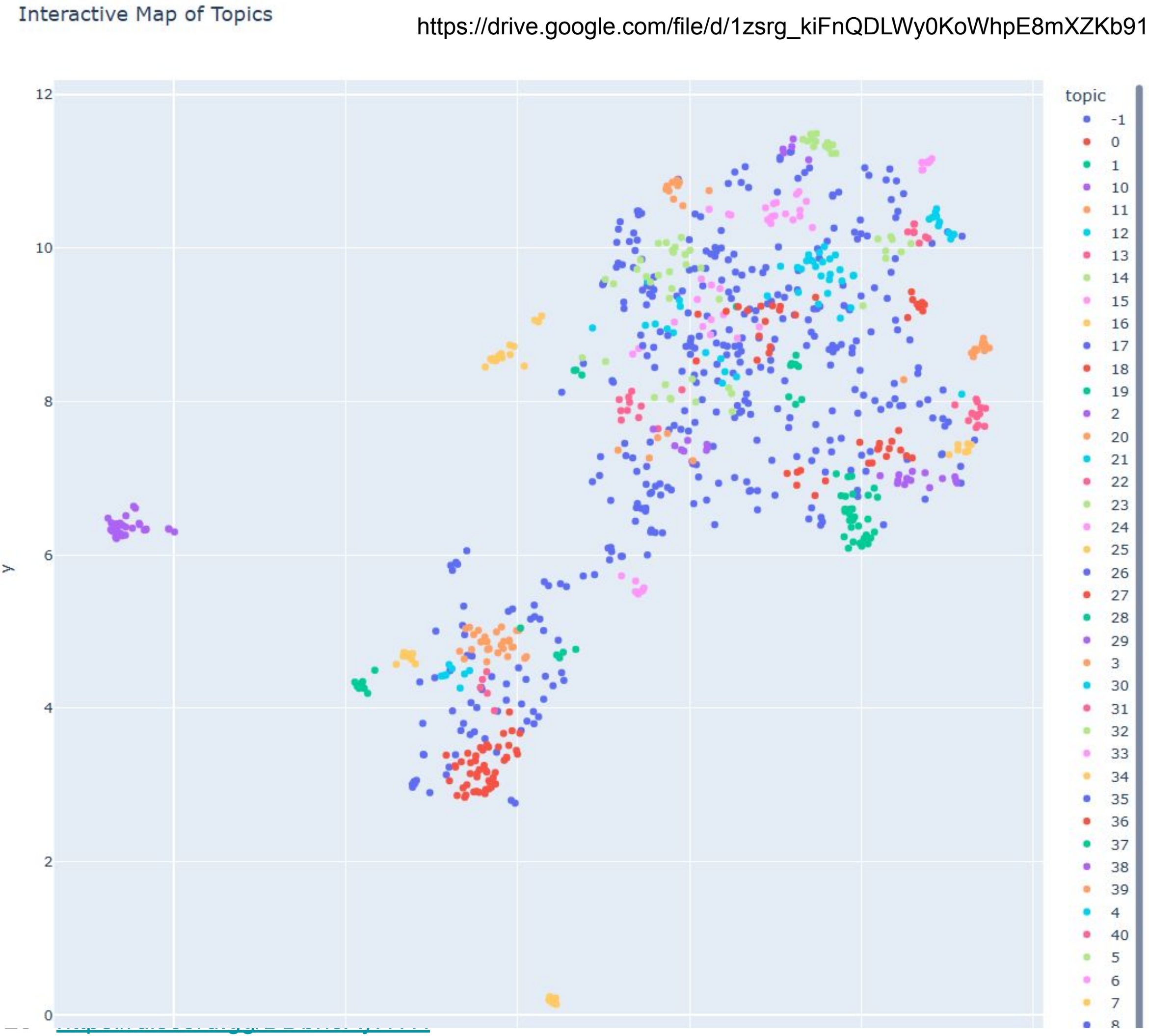


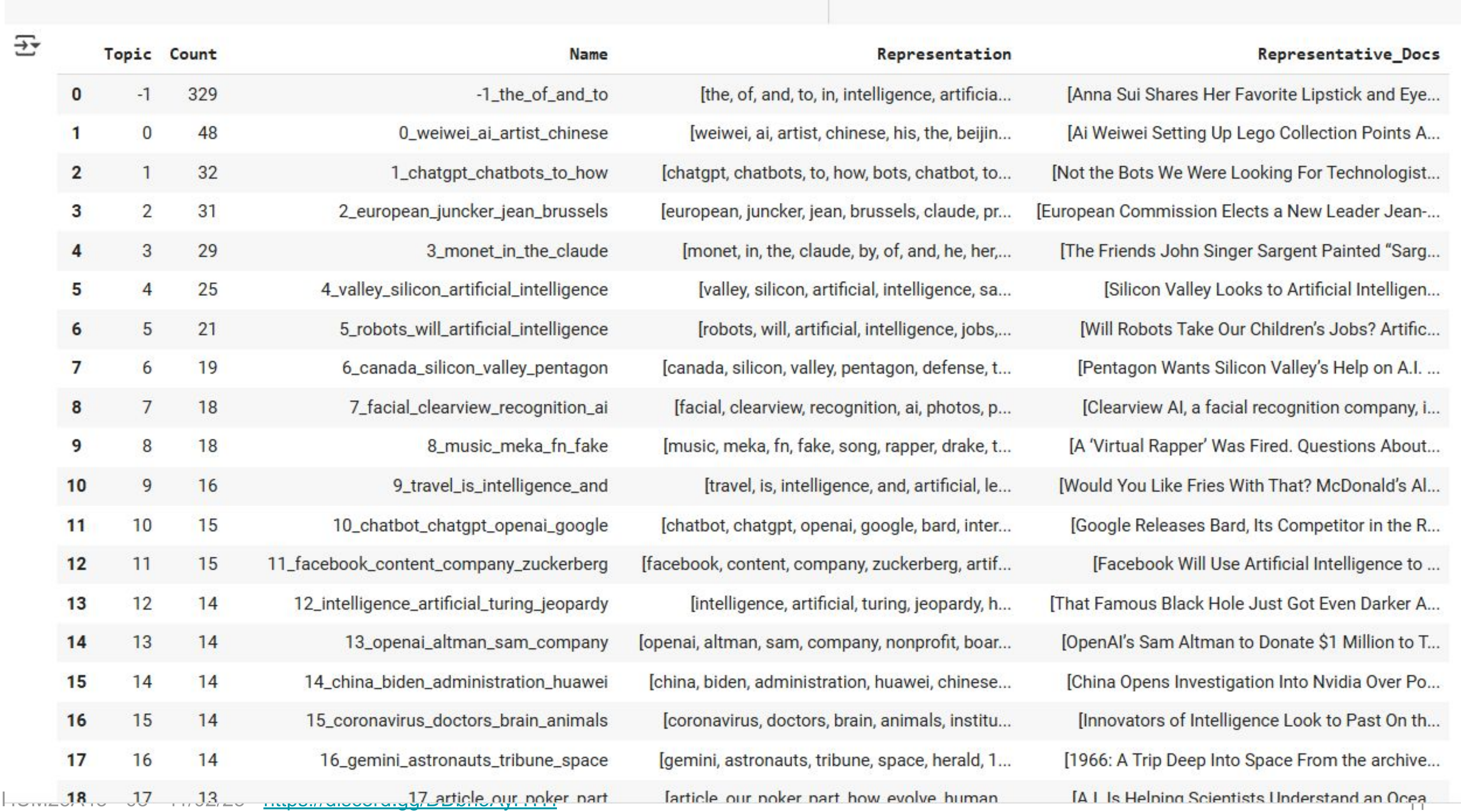
Percentage of AI-related Documents Over Time


Topic Word Scores
Topic 0
artificial
intelligence
artificial intelligence
ai
new
0
0.02
0.04
0.06
0.08
Topic 1
weiwei
ai weiwei
ai
artist
chinese
0
0.05
0.1
Topic 2
astronauts
space
herald
archives
international
0
0.05
0.1
0.15
Topic 3
llama
llamas
books
peruvian
anna
0
0.1
0.2
0.3
Topic 4
chef
restaurant
bar
hotel
worked
0
0.05
0.1
0.15
Topic 5
art
street
party
ai weiwei
weiwei
0
0.05
0.1
0.15


Topics over Time
Frequency
0
50
100
150
2014
2016
2018
2020
2022
2024
Global Topic Representati
0_artificial_intelligence_arti
1_weiwei_ai weiwei_ai_arti
2_astronauts_space_herald
3_llama_llamas_books_per
4_chef_restaurant_bar_hot
5_art_street_party_ai weiw

https://drive.google.com/file/d/1zsrg_kiFnQDLWy0KoWhpE8mXZKb91zkb/view?usp=sharing





	Topic	Count	Name	Representation	Representative_Docs
0	-1	329	-1_the_of_and_to	[the, of, and, to, in, intelligence, artificia...	[Anna Sui Shares Her Favorite Lipstick and Eye...
1	0	48	0_weiwei_ai_artist_chinese	[weiwei, ai, artist, chinese, his, the, beijin...	[Ai Weiwei Setting Up Lego Collection Points A...
2	1	32	1_chatgpt_chatbots_to_how	[chatgpt, chatbots, to, how, bots, chatbot, to...	[Not the Bots We Were Looking For Technologist...
3	2	31	2_european_juncker_jean_brussels	[european, juncker, jean, brussels, claude, pr...	[European Commission Elects a New Leader Jean-...
4	3	29	3_monet_in_the_claude	[monet, in, the, claude, by, of, and, he, her,...	[The Friends John Singer Sargent Painted “Sarg...
5	4	25	4_valley_silicon_artificial_intelligence	[valley, silicon, artificial, intelligence, sa...	[Silicon Valley Looks to Artificial Intelligen...
6	5	21	5_robots_will_artificial_intelligence	[robots, will, artificial, intelligence, jobs,...	[Will Robots Take Our Children’s Jobs? Artific...
7	6	19	6_canada_silicon_valley_pentagon	[canada, silicon, valley, pentagon, defense, t...	[Pentagon Wants Silicon Valley’s Help on A.I. ...
8	7	18	7_facial_clearview_recognition_ai	[facial, clearview, recognition, ai, photos, p...	[Clearview AI, a facial recognition company, i...
9	8	18	8_music_meka_fn_fake	[music, meka, fn, fake, song, rapper, drake, t...	[A ‘Virtual Rapper’ Was Fired. Questions About...
10	9	16	9_travel_is_intelligence_and	[travel, is, intelligence, and, artificial, le...	[Would You Like Fries With That? McDonald’s Al...
11	10	15	10_chatbot_chatgpt_openai_google	[chatbot, chatgpt, openai, google, bard, inter...	[Google Releases Bard, Its Competitor in the R...
12	11	15	11_facebook_content_company_zuckerberg	[facebook, content, company, zuckerberg, artif...	[Facebook Will Use Artificial Intelligence to ...
13	12	14	12_intelligence_artificial_turing_jeopardy	[intelligence, artificial, turing, jeopardy, h...	[That Famous Black Hole Just Got Even Darker A...
14	13	14	13_openai_altman_sam_company	[openai, altman, sam, company, nonprofit, boar...	[OpenAI’s Sam Altman to Donate $1 Million to T...
15	14	14	14_china_biden_administration_huawei	[china, biden, administration, huawei, chinese...	[China Opens Investigation Into Nvidia Over Po...
16	15	14	15_coronavirus_doctors_brain_animals	[coronavirus, doctors, brain, animals, institu...	[Innovators of Intelligence Look to Past On th...
17	16	14	16_gemini_astronauts_tribune_space	[gemini, astronauts, tribune, space, herald, 1...	[1966: A Trip Deep Into Space From the archive...
18	17	13	17_article_our_poker_part	[article, our, poker, part, how, evolve, human	[A.I. Is Helping Scientists Understand an Ocea