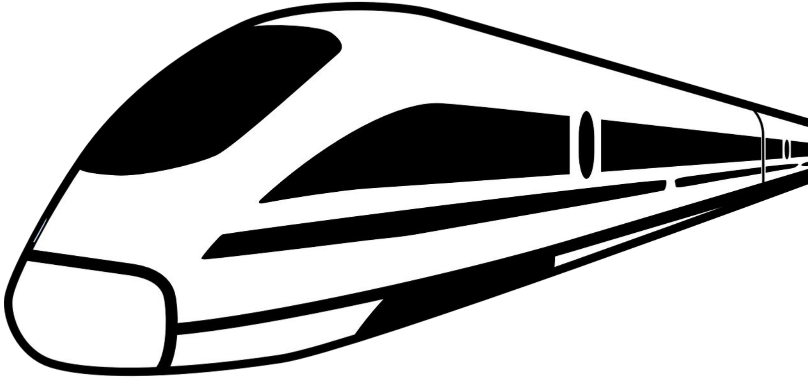
Google Cloud

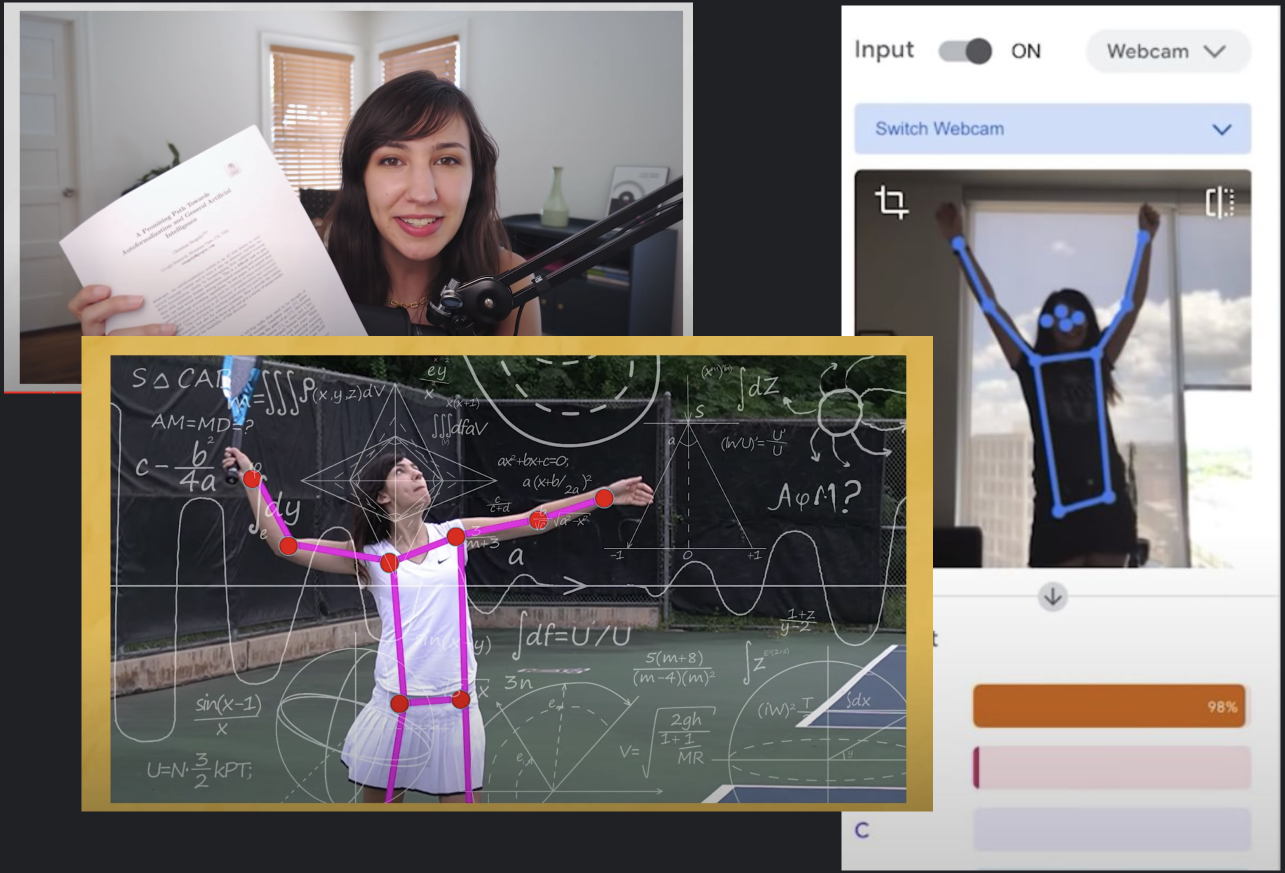
NLP In a Hurry

Using Pretrained Models to Train Language Apps Fast



Dale Markowitz, Applied AI Engineer
@dalequark

Who am I?



Let’s build a smart, AI-powered news article app

AI Article App Features

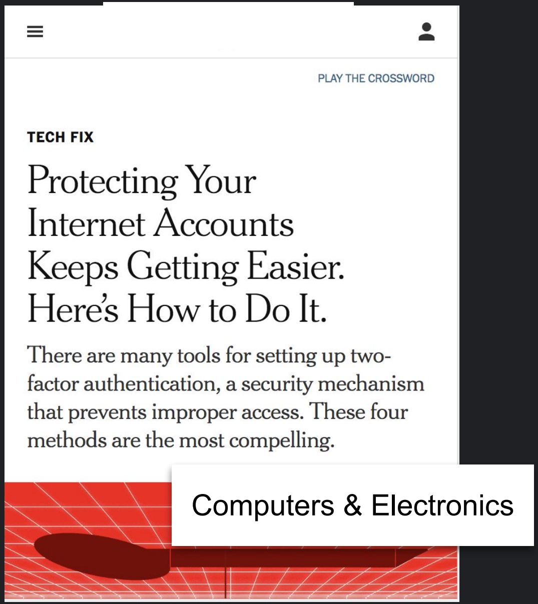
- Categorize articles by type
- Recognize important people, places, dates, products, etc
- Provide context
- Answer questions using news articles
- Recommend similar articles
- Smart semantic search

Using (almost) no labeled training data!

AI Article App Features

1. Transfer Learning
2. Pre-Trained General Purpose Models
3. Knowledge Graphs
4. Language Embeddings

Predicting article categories



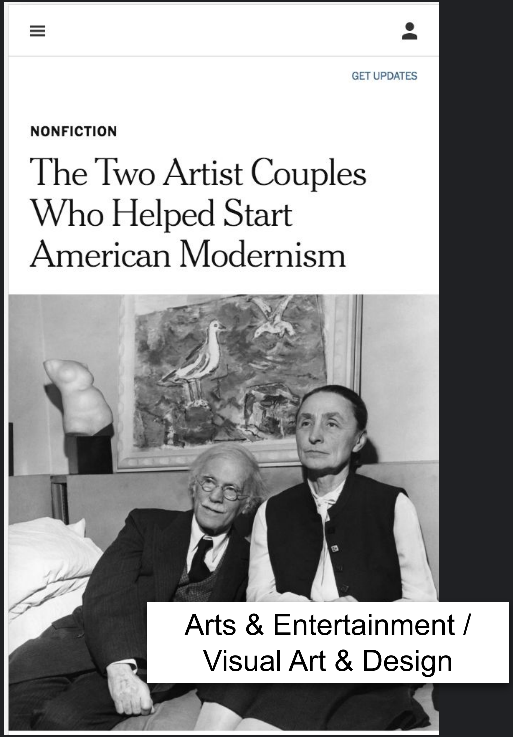
PLAY THE CROSSWORD

TECH FIX

Protecting Your Internet Accounts Keeps Getting Easier. Here's How to Do It.

There are many tools for setting up two-factor authentication, a security mechanism that prevents improper access. These four methods are the most compelling.

Computers & Electronics

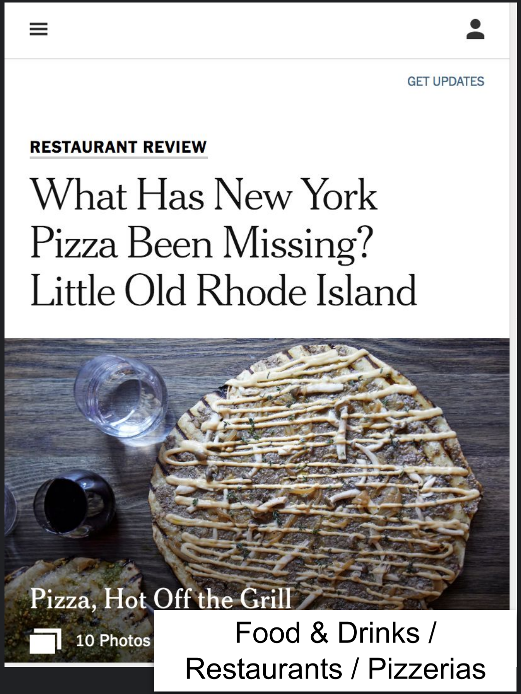


GET UPDATES

NONFICTION

The Two Artist Couples Who Helped Start American Modernism

Arts & Entertainment / Visual Art & Design



GET UPDATES

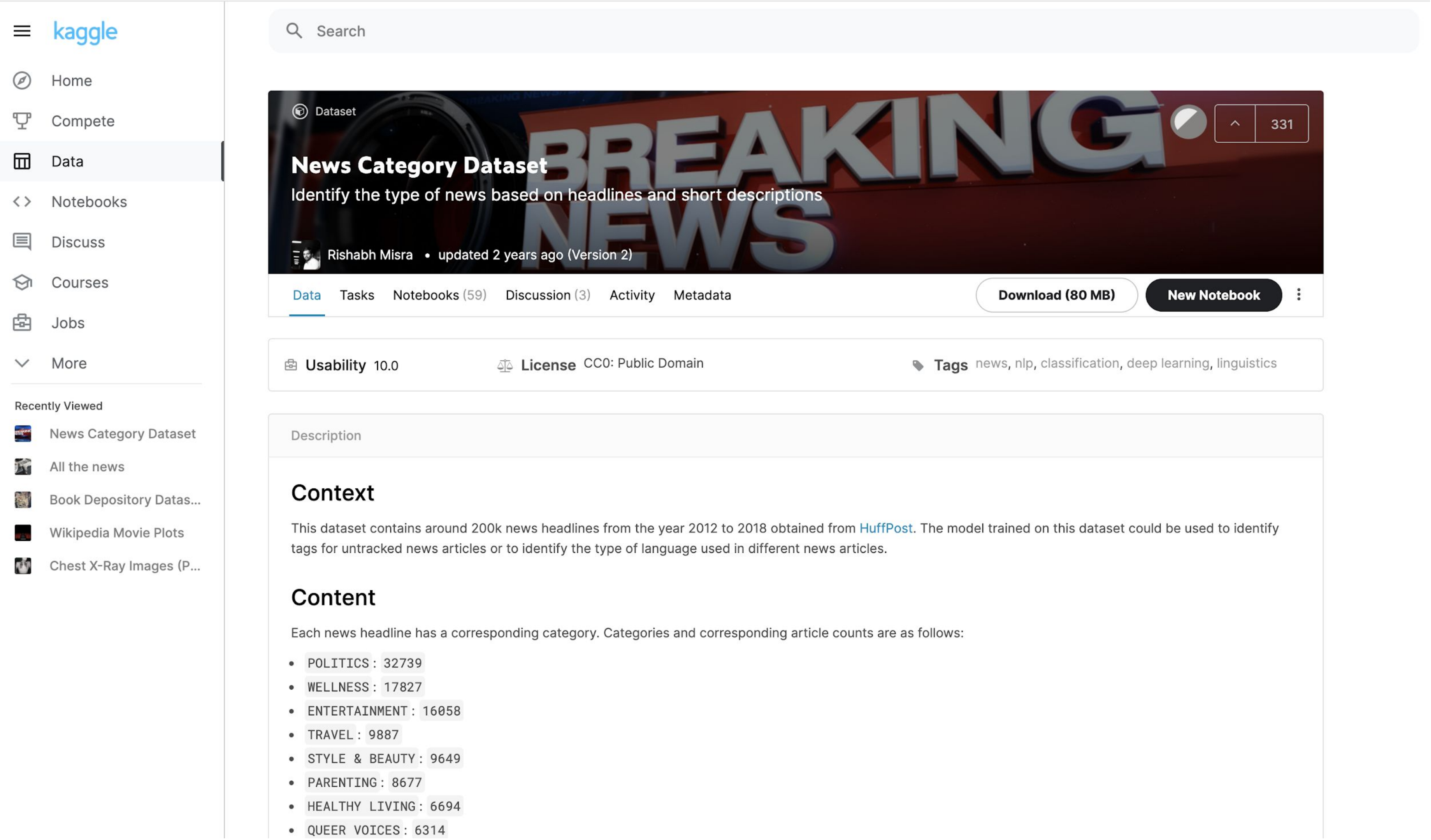
RESTAURANT REVIEW

What Has New York Pizza Been Missing? Little Old Rhode Island

Pizza, Hot Off the Grill

10 Photos

Food & Drinks / Restaurants / Pizzerias



News Category Dataset

Identify the type of news based on headlines and short descriptions

Rishabh Misra • updated 2 years ago (Version 2)

Usability 10.0 **License** CC0: Public Domain **Tags** news, nlp, classification, deep learning, linguistics

Context

This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. The model trained on this dataset could be used to identify tags for untracked news articles or to identify the type of language used in different news articles.

Content

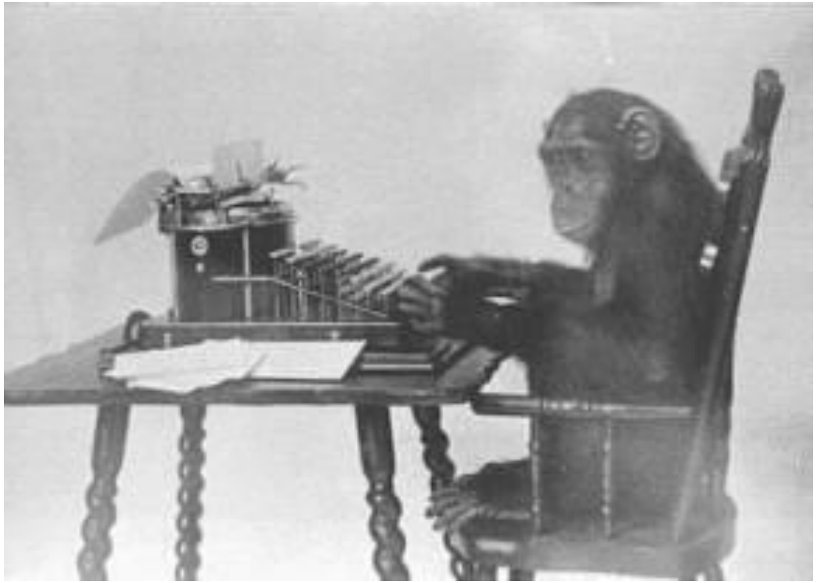
Each news headline has a corresponding category. Categories and corresponding article counts are as follows:

- `POLITICS`: `32739`
- `WELLNESS`: `17827`
- `ENTERTAINMENT`: `16058`
- `TRAVEL`: `9887`
- `STYLE & BEAUTY`: `9649`
- `PARENTING`: `8677`
- `HEALTHY LIVING`: `6694`
- `QUEER VOICES`: `6314`

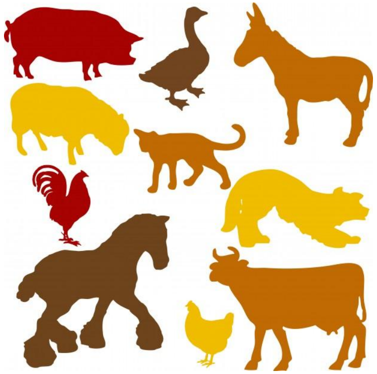
1000’s of human labels!

Headline	Description	Label
Jimmy Kimmel Knows Why Iran's Supreme Leader Watches 'Tom And Jerry' (HuffPo)	The host reimagined the cartoon after Ayatollah Ali Khamenei brought up the cat and mouse in defiance of U.S. threats.	COMEDY
Mystery 'Wolf-Like' Animal Reportedly Shot In Montana, Baffles Wildlife Officials (HuffPo)	“We have no idea what this was until we get a DNA report back."	WEIRD NEWS

1 Transfer Learning



Transfer Learning for Vision



Animal Tagging Model

+

Labeled Dog Breed Dataset

=

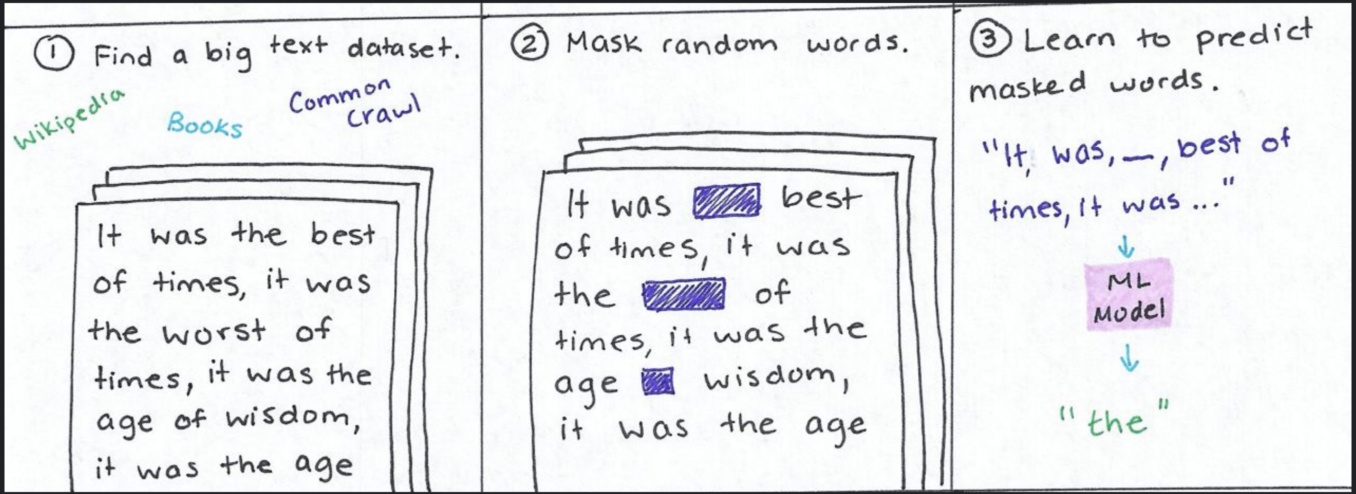
Dog Breed Model

BERT

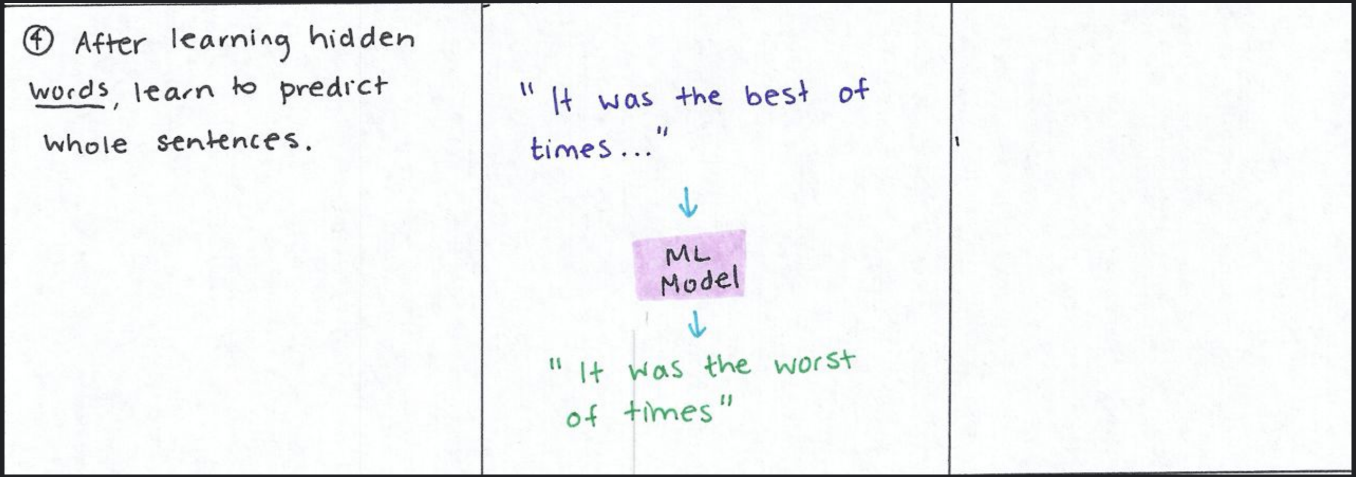
A big language model trained on **Wikipedia** and **BooksCorpus**, built by Google, that you can use as a base for your own NLP model.

Get it on TensorFlow Hub

How was BERT trained?


① Find a big text dataset.
Wikipedia
Books
Common Crawl
It was the best of times, it was the worst of times, it was the times, it was the age of wisdom, it was the age
② Mask random words.
It was ▨ best of times, it was the ▨ of times, it was the age ▨ wisdom, it was the age
③ Learn to predict masked words.
"It was, —, best of times, It was ..."
↓
ML Model
↓
"the"

How was BERT trained?



④ After learning hidden words, learn to predict whole sentences.

"Self-Supervised Learning"

Find a clever way to learn from data *without* having humans label it first.

BERT and other large NLP models demonstrate a general understanding of the structure of language.

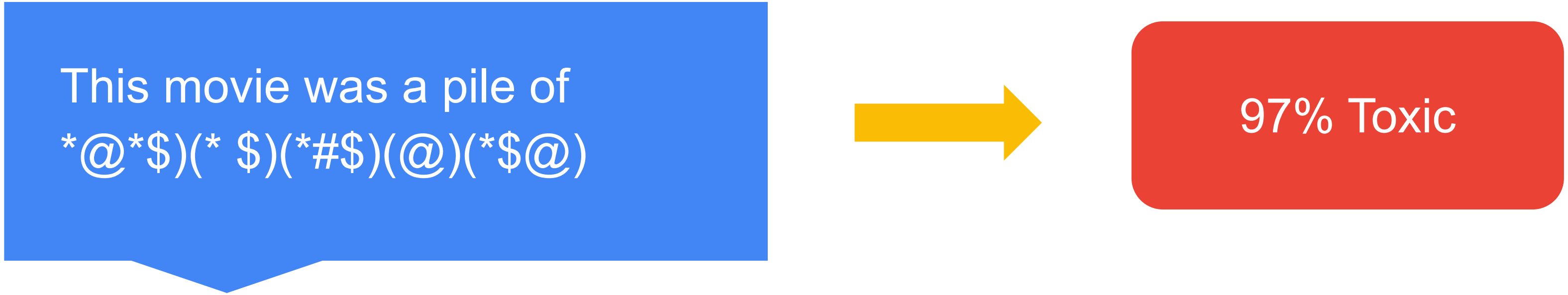
Using BERT the easy way with AutoML Natural Language

AI Article App Features

- **Categorize articles by type**
- Recognize important people, places, dates, products, etc
- Provide context
- Answer questions using news articles
- Recommend similar articles
- Smart semantic search

2 Pre-Trained, General-Use Models

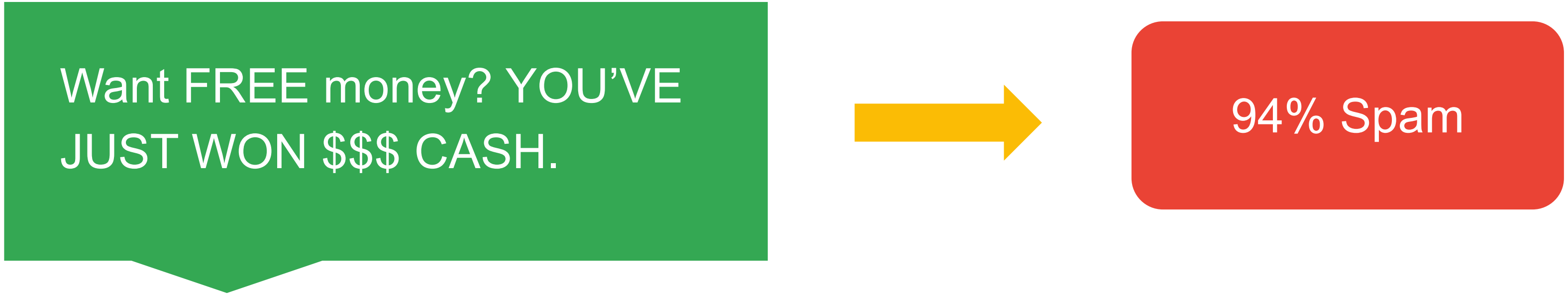
Toxic Speech Detection



This movie was a pile of *@*$)(* $)(*#$)(@)(*$@)

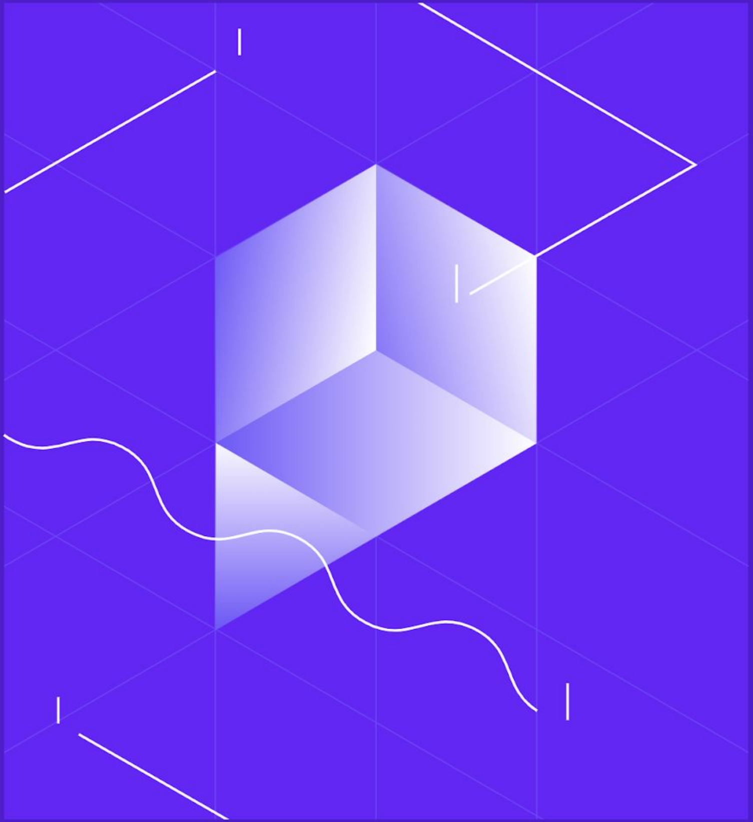
97% Toxic

Spam Detection



Want FREE money? YOU’VE JUST WON $$$ CASH.

94% Spam

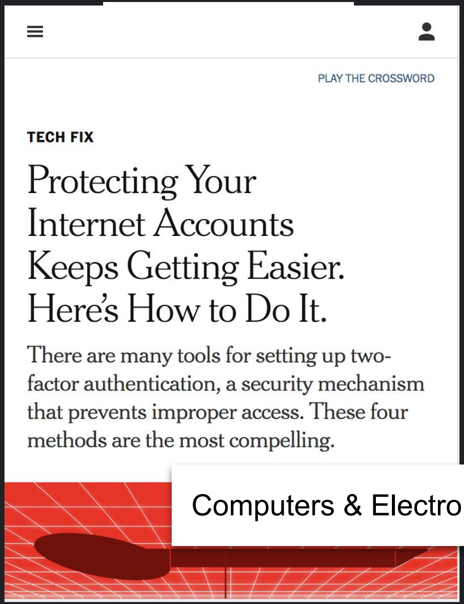


Perspective API

Perspective is an API that makes it easier to host better conversations. The API uses machine learning models to score the perceived impact a comment might have on a conversation. Developers and publishers can use this score to give realtime feedback to commenters or help moderators do their job, or allow readers to more easily find relevant information, as illustrated in two experiments below. Our first model identifies whether a comment could be perceived as “toxic” to a discussion.

perspectiveapi.com

Natural Language API
cloud.google.com/natural-language



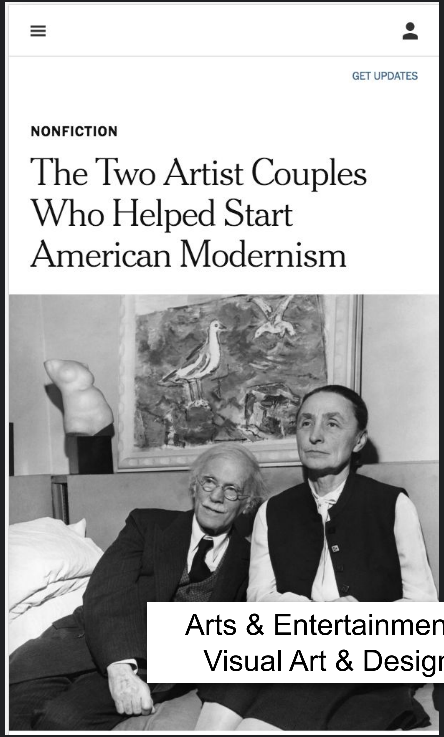
PLAY THE CROSSWORD

TECH FIX

Protecting Your Internet Accounts Keeps Getting Easier. Here's How to Do It.

There are many tools for setting up two-factor authentication, a security mechanism that prevents improper access. These four methods are the most compelling.

Computers & Electronics

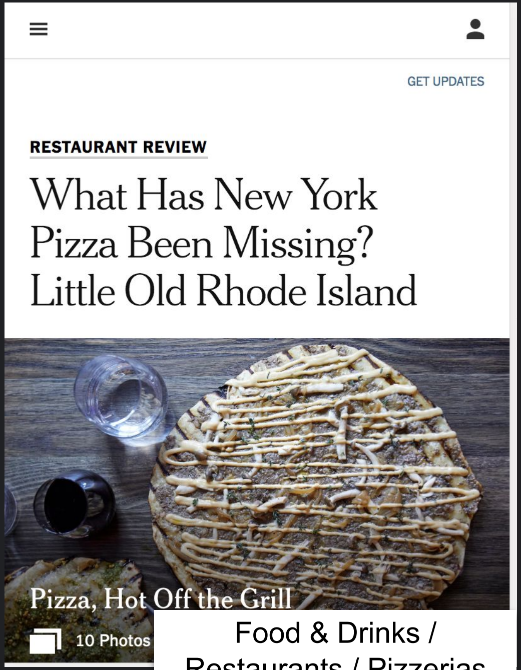


GET UPDATES

NONFICTION

The Two Artist Couples Who Helped Start American Modernism

Arts & Entertainment / Visual Art & Design



GET UPDATES

RESTAURANT REVIEW

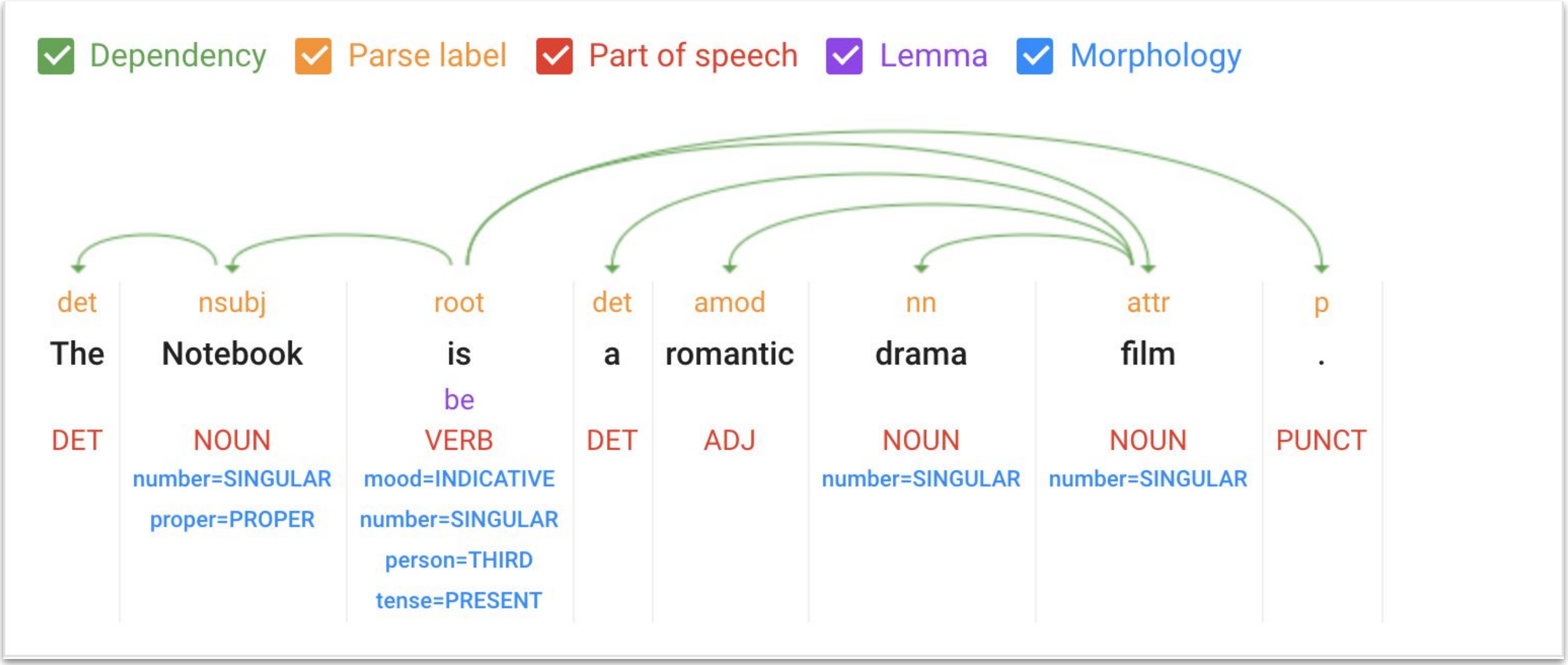
What Has New York Pizza Been Missing? Little Old Rhode Island

Pizza, Hot Off the Grill

10 Photos

Food & Drinks / Restaurants / Pizzerias

Part of Speech Tagging



☑ Dependency ☑ Parse label ☑ Part of speech ☑ Lemma ☑ Morphology

det	nsubj	root	det	amod	nn	attr	p
The	Notebook	is	a	romantic	drama	film	.
		be					
DET	NOUN	VERB	DET	ADJ	NOUN	NOUN	PUNCT
	number=SINGULAR proper=PROPER	mood=INDICATIVE number=SINGULAR person=THIRD tense=PRESENT			number=SINGULAR	number=SINGULAR	

Named Entity Recognition (NER)

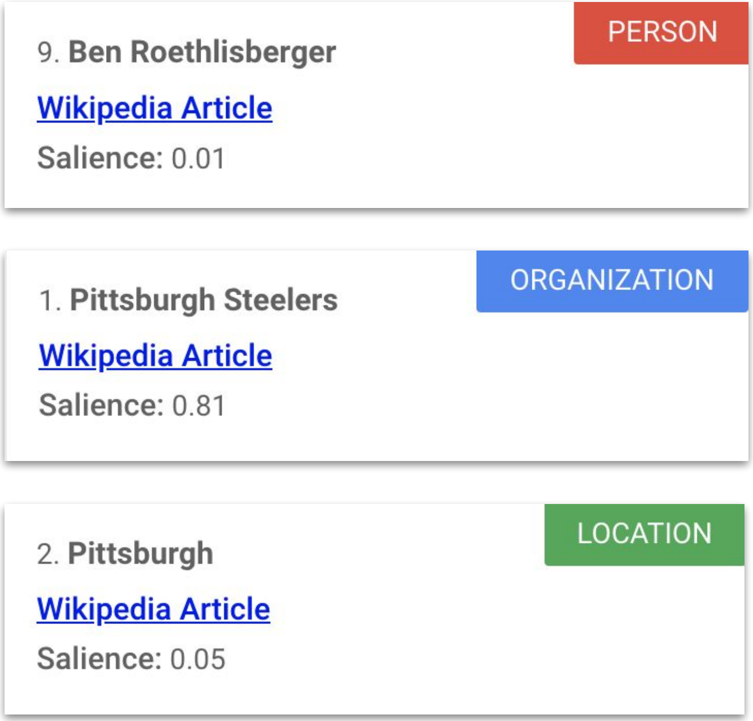
“The Pittsburgh Steelers are a professional American football team based in Pittsburgh. They compete in the NFL and are members of the AFC North division. Founded in 1933, the Steelers are the seventh-oldest franchise in the NFL. Ben Roethlisberger, Joshua Dobbs, and Mason Rudolph are their quarterbacks.”

Named Entity Recognition (NER)

“The Pittsburgh Steelers are a professional American football team based in Pittsburgh. They compete in the NFL and are members of the AFC North division. Founded in 1933, the Steelers are the seventh-oldest franchise in the NFL. Ben Roethlisberger, Joshua Dobbs, and Mason Rudolph are their quarterbacks.”

The ⟨Pittsburgh Steelers⟩$_1$ are a professional ⟨American football team⟩$_1$ based in ⟨Pittsburgh⟩$_2$. They compete in the ⟨NFL⟩$_5$ and are ⟨members⟩$_3$ of the ⟨AFC North⟩$_7$ ⟨division⟩$_4$. Founded in ⟨1933⟩$_{11}$ ⟨1933⟩$_{12}$, the ⟨Steelers⟩$_1$ are the seventh-oldest ⟨franchise⟩$_1$ in the ⟨NFL⟩$_5$. ⟨Ben Roethlisberger⟩$_9$, ⟨Joshua Dobbs⟩$_8$, and ⟨Mason Rudolph⟩$_{10}$ are their ⟨quarterbacks⟩$_6$.

Named Entity Recognition (NER)



9. **Ben Roethlisberger**

PERSON

Wikipedia Article

Salience: 0.01

1. **Pittsburgh Steelers**

ORGANIZATION

Wikipedia Article

Salience: 0.81

2. **Pittsburgh**

LOCATION

Wikipedia Article

Salience: 0.05

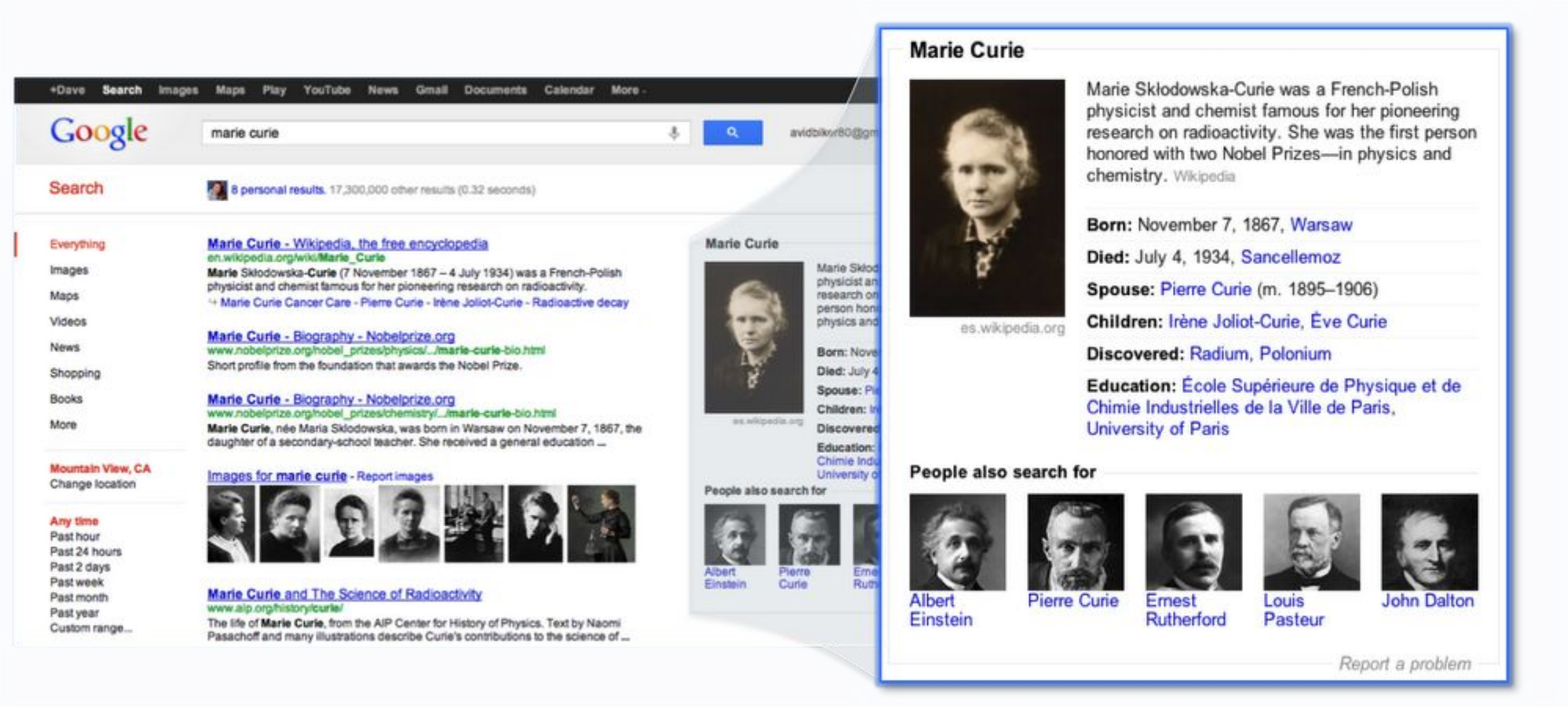
The ⟨Pittsburgh Steelers⟩$_1$ are a professional ⟨American football team⟩$_1$ based in ⟨Pittsburgh⟩$_2$. They compete in the ⟨NFL⟩$_5$ and are ⟨members⟩$_3$ of the ⟨AFC North⟩$_7$ ⟨division⟩$_4$. Founded in ⟨1933⟩$_{11}$ ⟨1933⟩$_{12}$, the ⟨Steelers⟩$_1$ are the seventh-oldest ⟨franchise⟩$_1$ in the ⟨NFL⟩$_5$. ⟨Ben Roethlisberger⟩$_9$, ⟨Joshua Dobbs⟩$_8$, and ⟨Mason Rudolph⟩$_{10}$ are their ⟨quarterbacks⟩$_6$.

AI Article App Features

- **Categorize articles by type**
- **Recognize important people, places, dates, products, etc**
- Provide context
- Answer questions using news articles
- Recommend similar articles
- Smart semantic search

We can add context to entities with knowledge graphs.

Knowledge Graphs



8 personal results. 17,300,000 other results (0.32 seconds)

Marie Curie - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Marie_Curie**
Marie Skłodowska-**Curie** (7 November 1867 – 4 July 1934) was a French-Polish physicist and chemist famous for her pioneering research on radioactivity.
↪ Marie Curie Cancer Care - Pierre Curie - Irène Joliot-Curie - Radioactive decay

Marie Curie - Biography - Nobelprize.org
www.nobelprize.org/nobel_prizes/physics/.../**marie-curie**-bio.html
Short profile from the foundation that awards the Nobel Prize.

Marie Curie - Biography - Nobelprize.org
www.nobelprize.org/nobel_prizes/chemistry/.../**marie-curie**-bio.html
Marie Curie, née Maria Sklodowska, was born in Warsaw on November 7, 1867, the daughter of a secondary-school teacher. She received a general education ...

Images for **marie curie** - Report images

Marie Curie and The Science of Radioactivity
www.aip.org/history/**curie**/
The life of **Marie Curie**, from the AIP Center for History of Physics. Text by Naomi Pasachoff and many illustrations describe Curie's contributions to the science of ...

Marie Curie

Marie Skłodowska-Curie was a French-Polish physicist and chemist famous for her pioneering research on radioactivity. She was the first person honored with two Nobel Prizes—in physics and chemistry. Wikipedia

es.wikipedia.org

Born: November 7, 1867, Warsaw

Died: July 4, 1934, Sancellemoz

Spouse: Pierre Curie (m. 1895–1906)

Children: Irène Joliot-Curie, Ève Curie

Discovered: Radium, Polonium

Education: École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris, University of Paris

People also search for

Albert Einstein · Pierre Curie · Ernest Rutherford · Louis Pasteur · John Dalton

Report a problem

https://developers.google.com/knowledge-graph

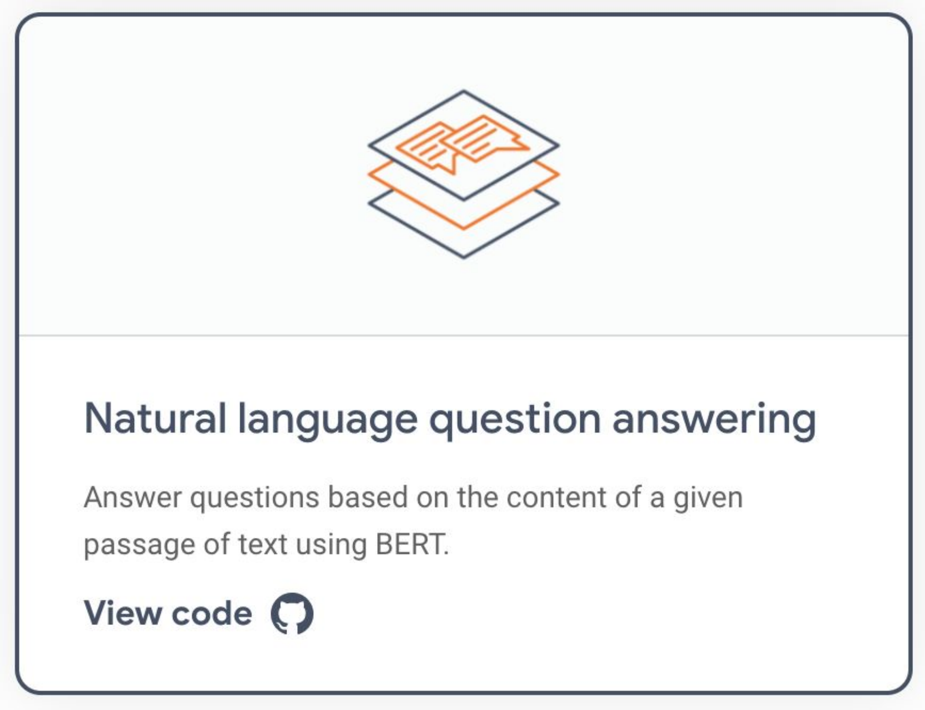
AI Article App Features

- **Categorize articles by type**
- **Recognize important people, places, dates, products, etc**
- **Provide context**
- Answer questions using news articles
- Recommend similar articles
- Smart semantic search

But what about specific questions answered in each article?

Talk to Books

https://books.google.com/talktobooks/



Natural language question answering

Answer questions based on the content of a given passage of text using BERT.

View code

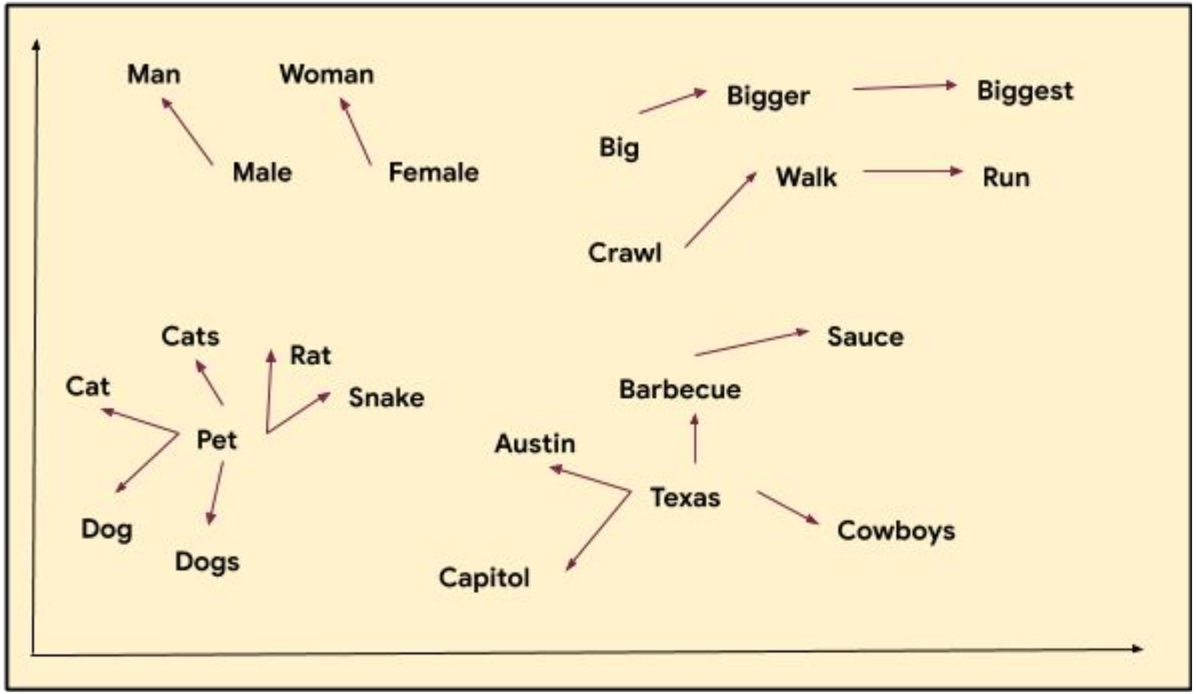
https://codepen.io/dalequark/pen/ZEWxOjN

AI Article App Features

- **Categorize articles by type**
- **Recognize important people, places, dates, products, etc**
- **Provide context**
- **Answer questions using news articles**
- Recommend similar articles
- Smart semantic search

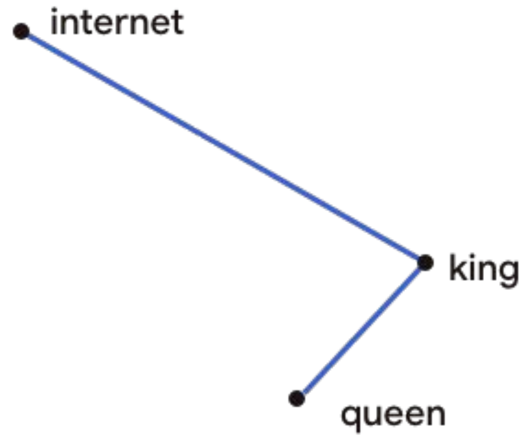
3 Embeddings

Word Vectors (“Word2Vec”)



Let’s explore word vectors!

The *distance* between word vectors tells you how similar they are.



deduplication

finding synonyms

Changing word tense, quantity, gender

Embeddings are VERY USEFUL!

clustering

topic modeling

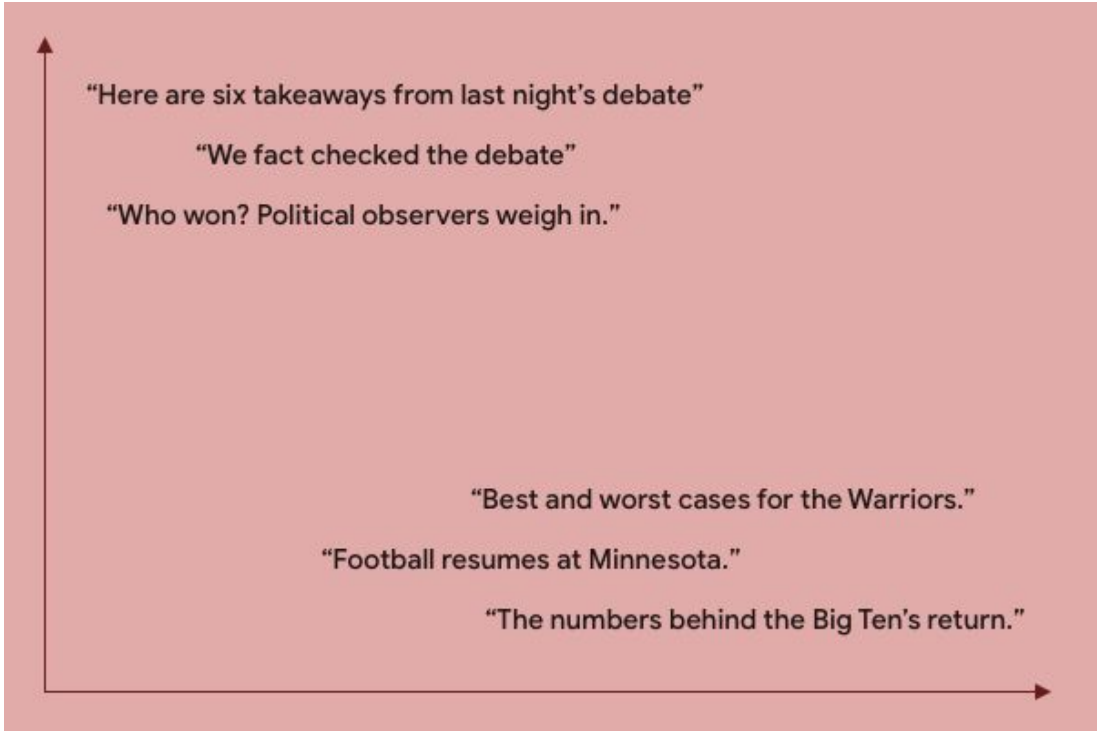
search

preprocessing for other NLP models

translation

Sentence Embeddings

Built on top of BERT!



“Here are six takeaways from last night’s debate”

“We fact checked the debate”

“Who won? Political observers weigh in.”

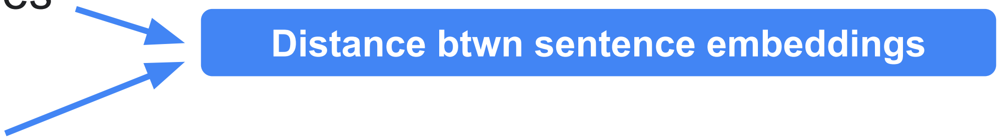
“Best and worst cases for the Warriors.”

“Football resumes at Minnesota.”

“The numbers behind the Big Ten’s return.”

AI Article App Features

- **Categorize articles by type**
- **Recognize important people, places, dates, products, etc**
- **Provide context**
- **Answer questions using news articles**
- Recommend similar articles
- Smart semantic search



Distance btwn sentence embeddings

Easily prototype language apps with Semantic Reactor

https://research.google.com/semanticexperiences/semantic-reactor.html

Thanks!

Some Useful Links

- daleonai.com (Dale's blog for all things ML)
- daleonai.com/semantic-ml
- https://www.tensorflow.org/js/models (embedding and question answering models)
- cloud.google.com/natural-language
- cloud.google.com/natural-language/automl/