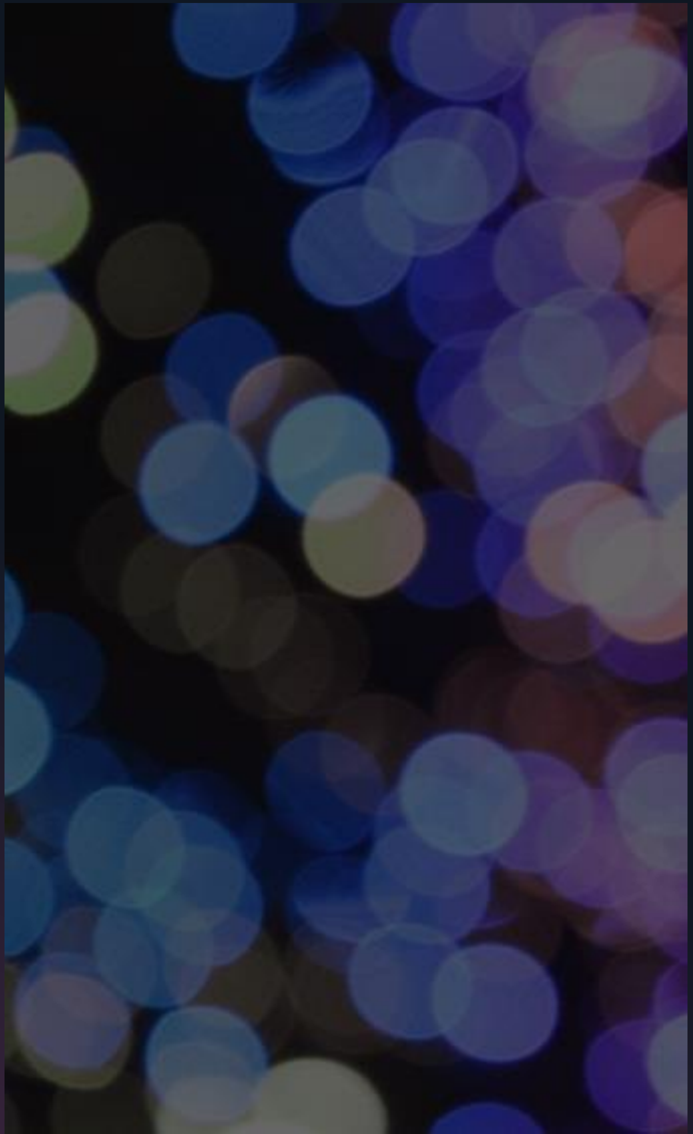
Finding similar Questions using SBERT

Prototype with ELI5 Questions

Ozan Yilmaz



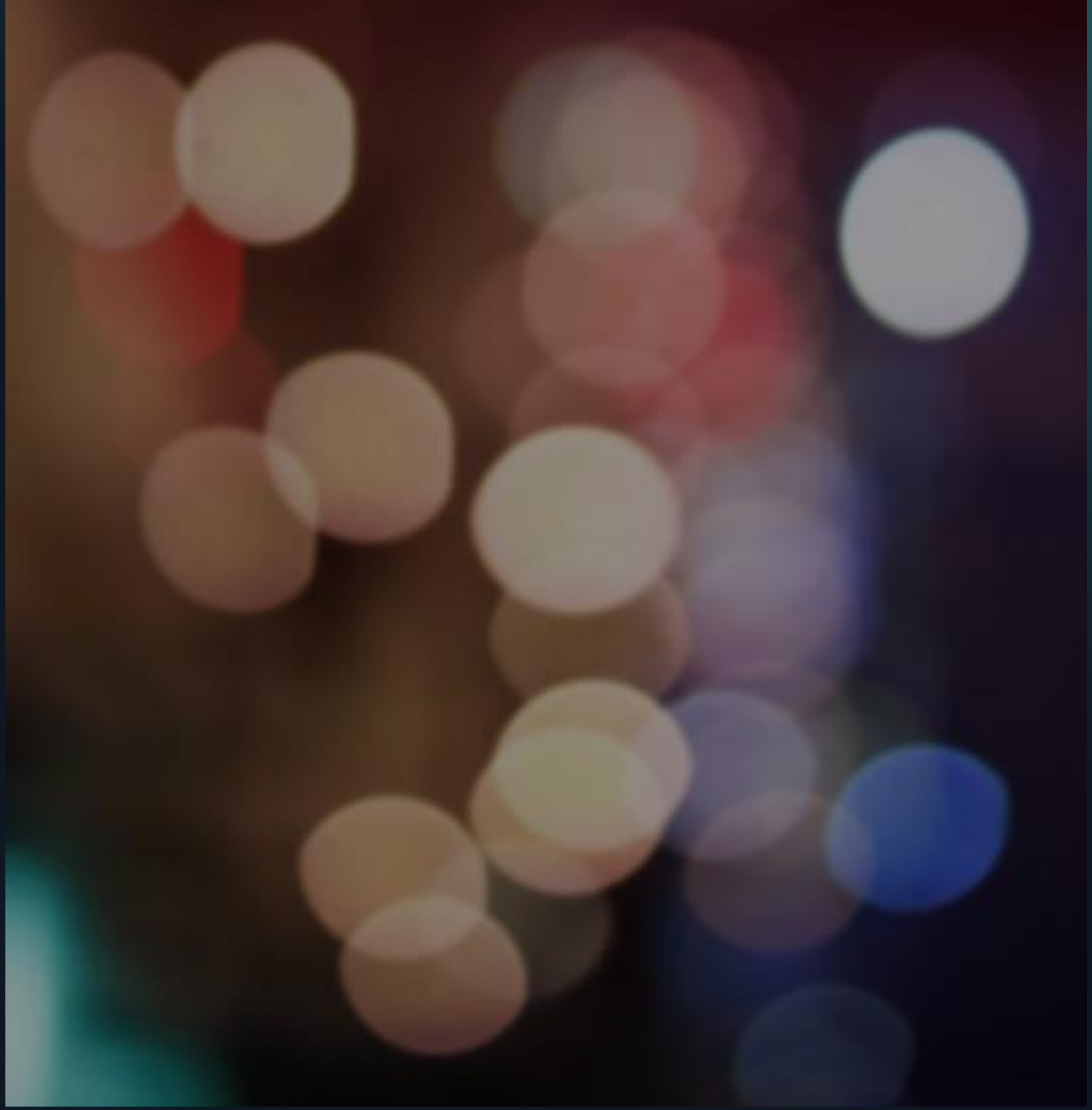
PROBLEM

Customer Support for Big Companies try to automate redundant support services

- Chatbots don’t work that well yet
- FAQ’s can’t answer all relevant questions, often overwhelming
- Forum’s provide a good platform for inter-customer dialogue and open access to similar problems and solutions -> problem with duplicates and poor phrasing of questions

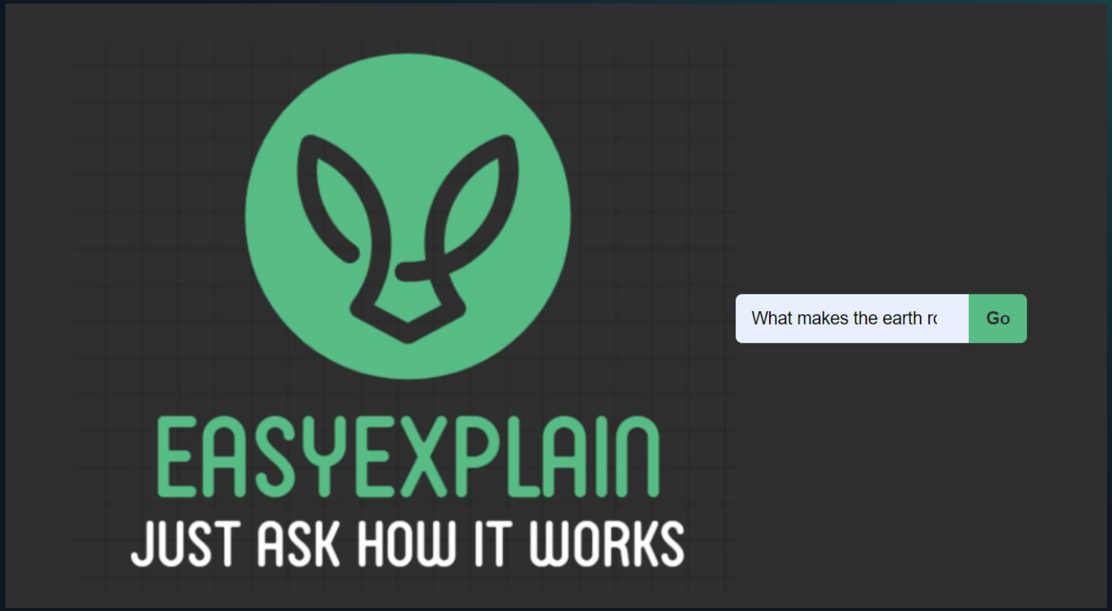
SOLUTION

A search platform to find similar asked forum questions with links and top-rated answers!



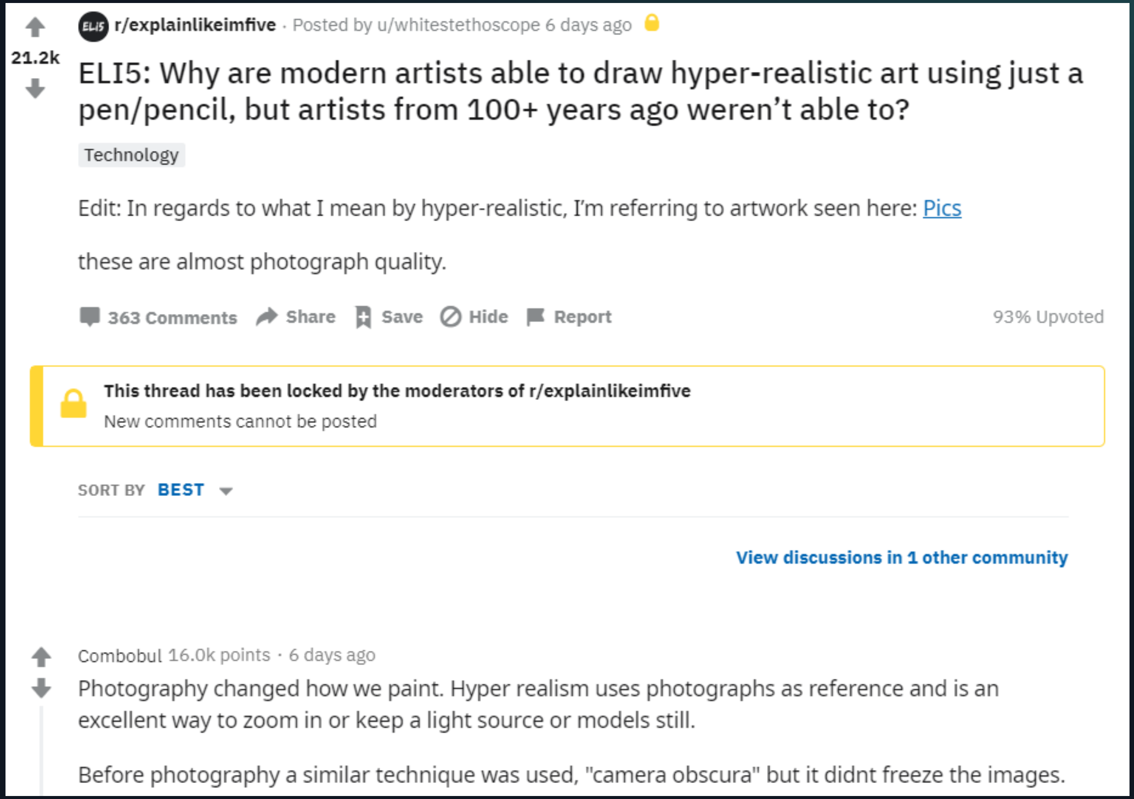
EasyExplain

PoC Semantic Search Project to find 4 most similar questions on Reddit’s famous subreddit ELI5



Explain like I’m Five

Users can ask questions about anything – The community tries to provide layman-friendly answers



r/explainlikeimfive · Posted by u/whitestethoscope 6 days ago

21.2k

ELI5: Why are modern artists able to draw hyper-realistic art using just a pen/pencil, but artists from 100+ years ago weren’t able to?

Technology

Edit: In regards to what I mean by hyper-realistic, I’m referring to artwork seen here: Pics

these are almost photograph quality.

This thread has been locked by the moderators of r/explainlikeimfive

New comments cannot be posted

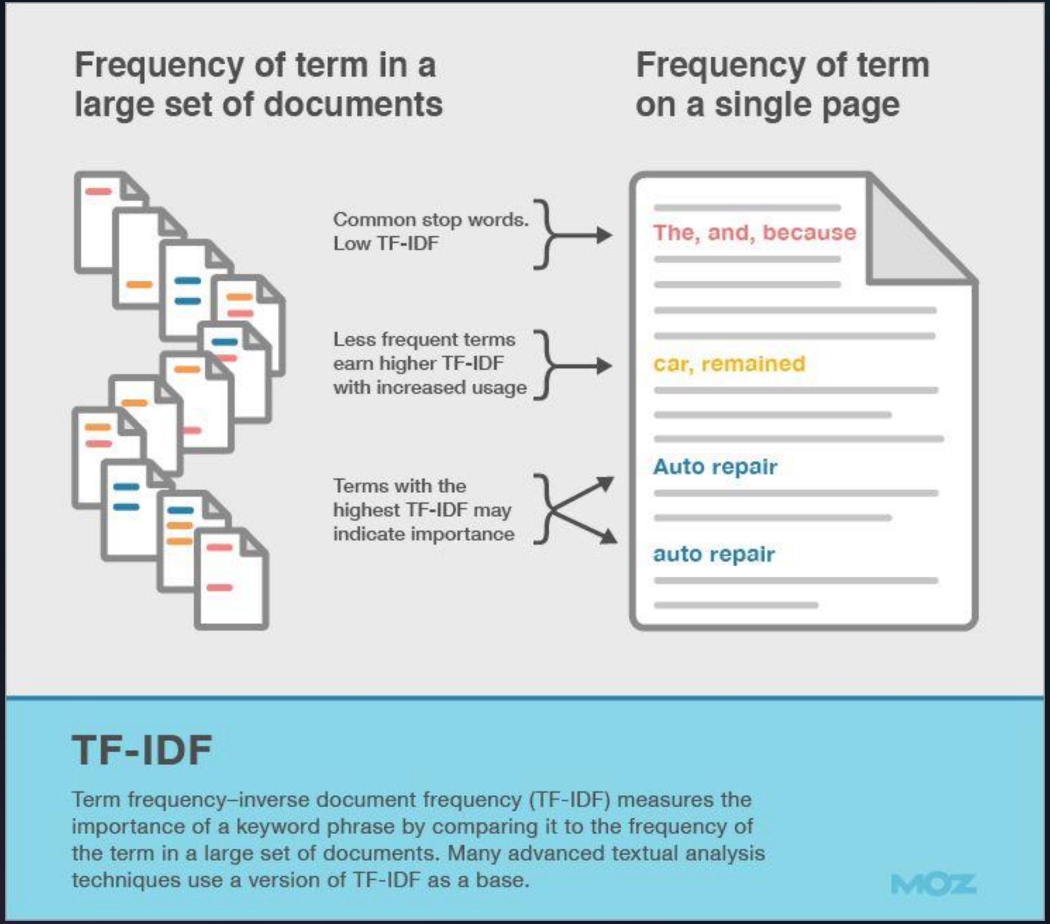
Combobul 16.0k points · 6 days ago

Photography changed how we paint. Hyper realism uses photographs as reference and is an excellent way to zoom in or keep a light source or models still.

Before photography a similar technique was used, "camera obscura" but it didnt freeze the images.

Backend Technology

TF-IDF



Frequency of term in a large set of documents

Frequency of term on a single page

Common stop words. Low TF-IDF

Less frequent terms earn higher TF-IDF with increased usage

Terms with the highest TF-IDF may indicate importance

TF-IDF

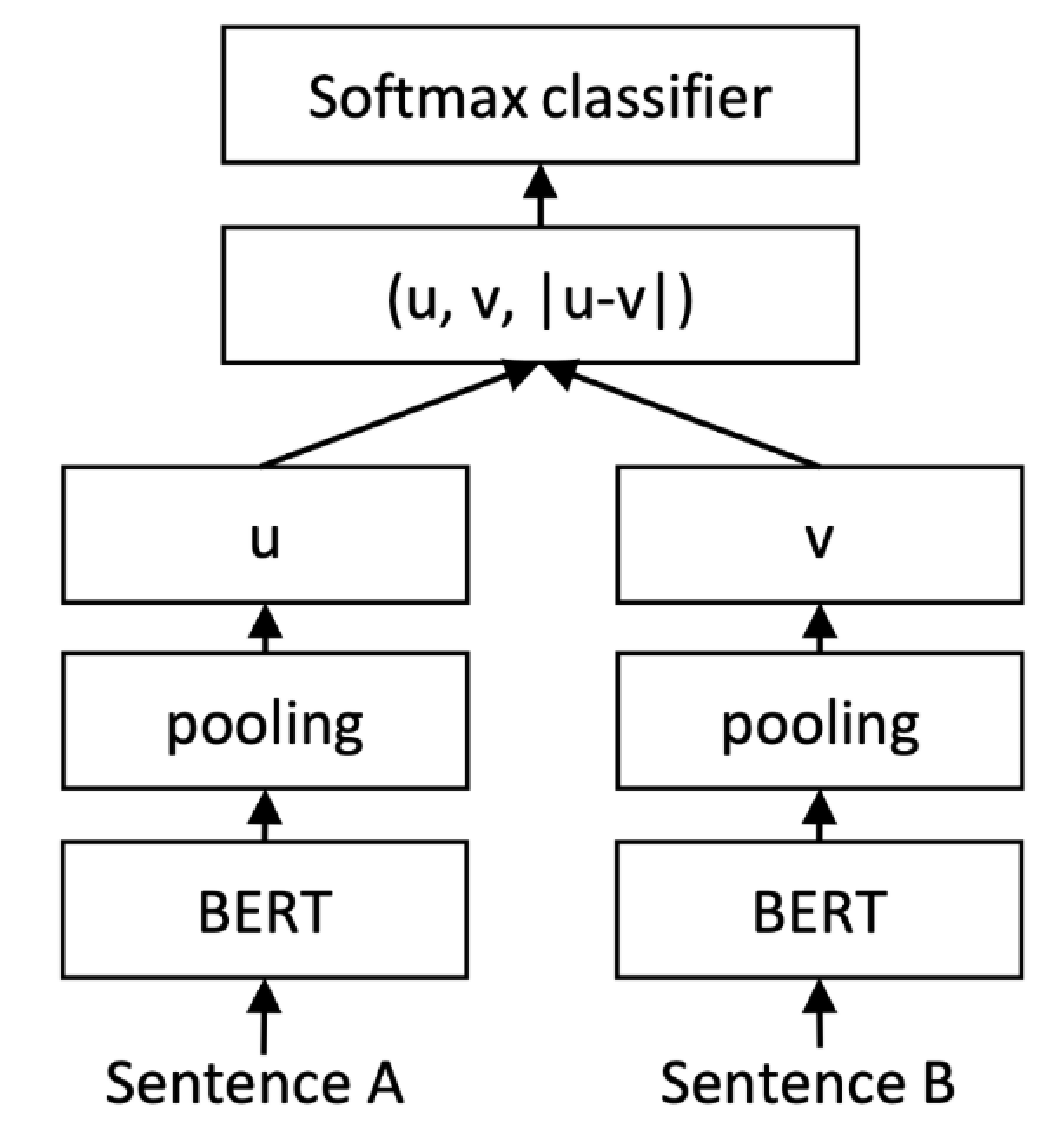
Term frequency–inverse document frequency (TF-IDF) measures the importance of a keyword phrase by comparing it to the frequency of the term in a large set of documents. Many advanced textual analysis techniques use a version of TF-IDF as a base.

Sentence-BERT

- Modification of BERT to derive meaningful sentence embeddings
- Utilize siamese networks to pool sentence inputs and finetune them with SNLI and the Multi-Genre NLI(5) corpus (Natural Language Inference)
- Creates vector space where Cosine Similarity works
- State-of-the-Art for many STS (Semantic Textual Similarity) Tasks

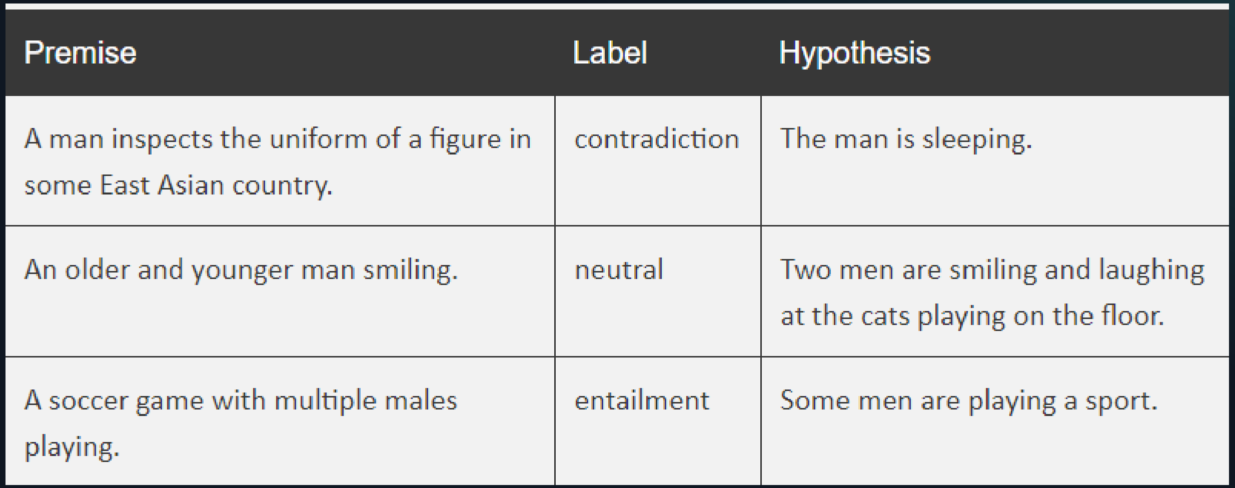
SBERT

Pool input sequences A and B to u and v, concatenate with pointwise difference |u-v| and solve NLI Task for finetuning Bert



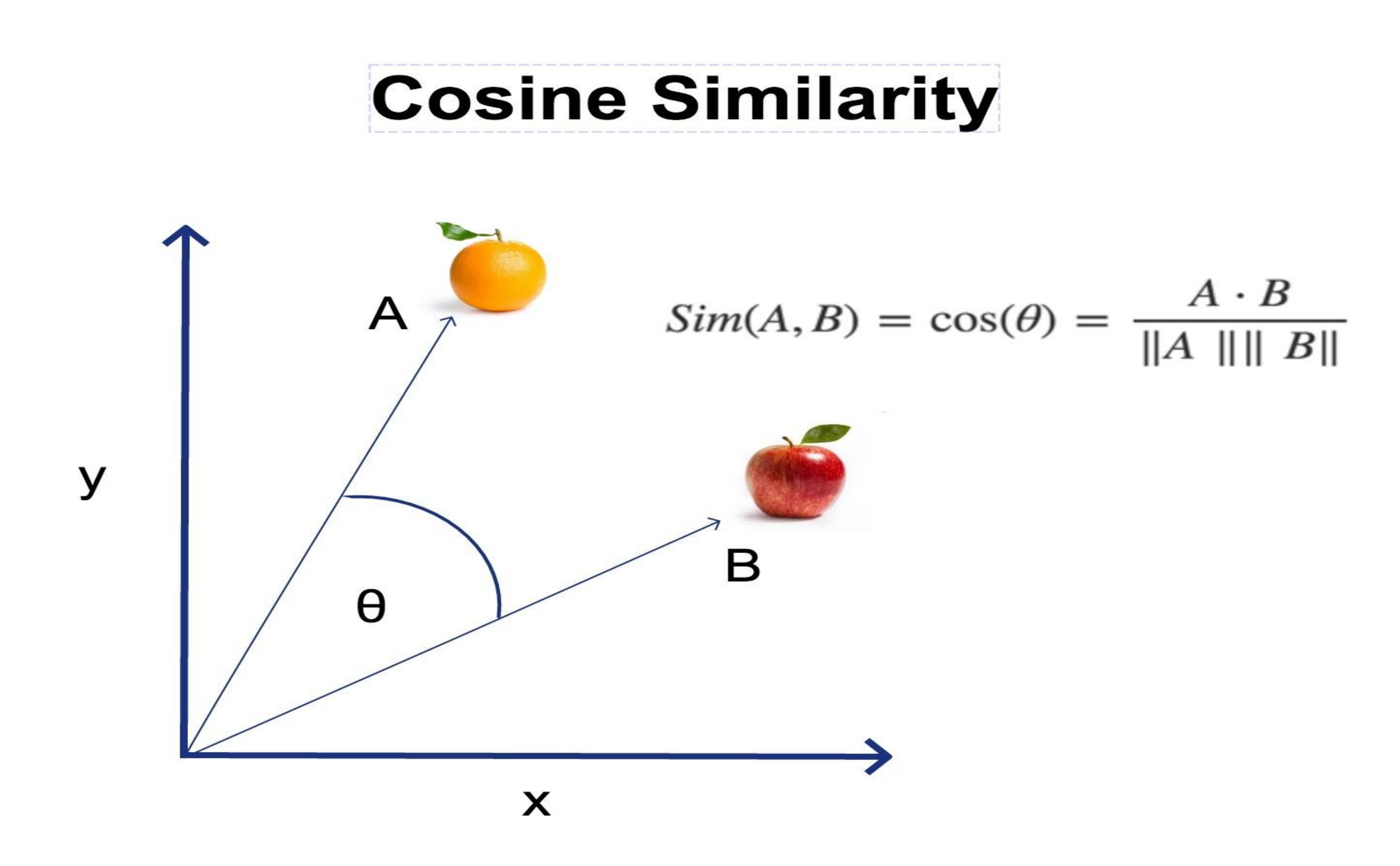
NLI Task

Natural language inference is the task of determining whether a “hypothesis” is true (entailment), false (contradiction), or undetermined (neutral) given a “premise”.



Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

Cosine Similarity



Cosine Similarity

$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Comparing TF-IDF & SBERT

TF-IDF

+ Statistically determines importance of terms
+ Works very good for longer documents
+ Not much preprocessing necessary (only removing stop-words)
+ Fast
- Doesn't capture semantics of sentences
- Only fitted for current dataset (closed domain)

SBERT

+ Captures semantics of arbitrarily long sentences in a fixed length vector
+ Works good for short/ middle-length sentences
+ Faster than other BERT-based Sentence encoding methods
+ Creates a feature-space where sentences can be compared using cosine similarity
- Slower than TF-IDF -> need to pre-encode document vectors
- Could be necessary that SBERT needs to be fine-tuned for specialized domains

Data Creation

Explain Like I'm Five Dataset Creation Pipeline

Download ELI5 Posts

Use script from Facebook's ELI5 project to download 250k Posts from ELI5 subreddit between 2011 and 2018

Extract Questions & Links

Extract Questions and Links from downloaded data in a separate file

Preprocessing

Use TF-IDF to create vectors of all questions for the baseline system after removing stopwords.

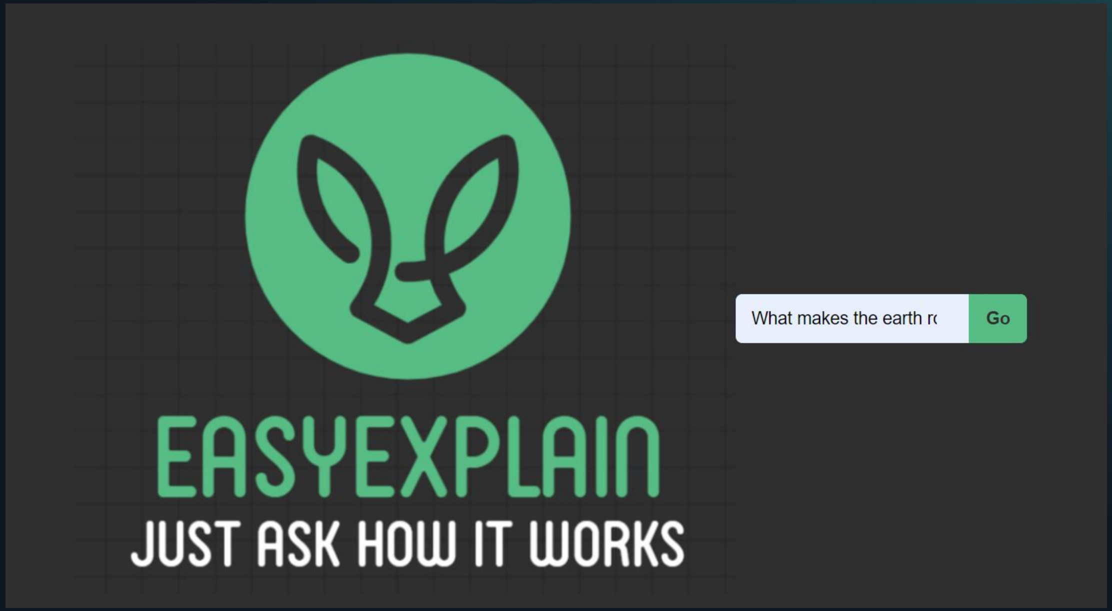
Use SBERT model to encode all questions and save them as npy file.

Extract Top Answers

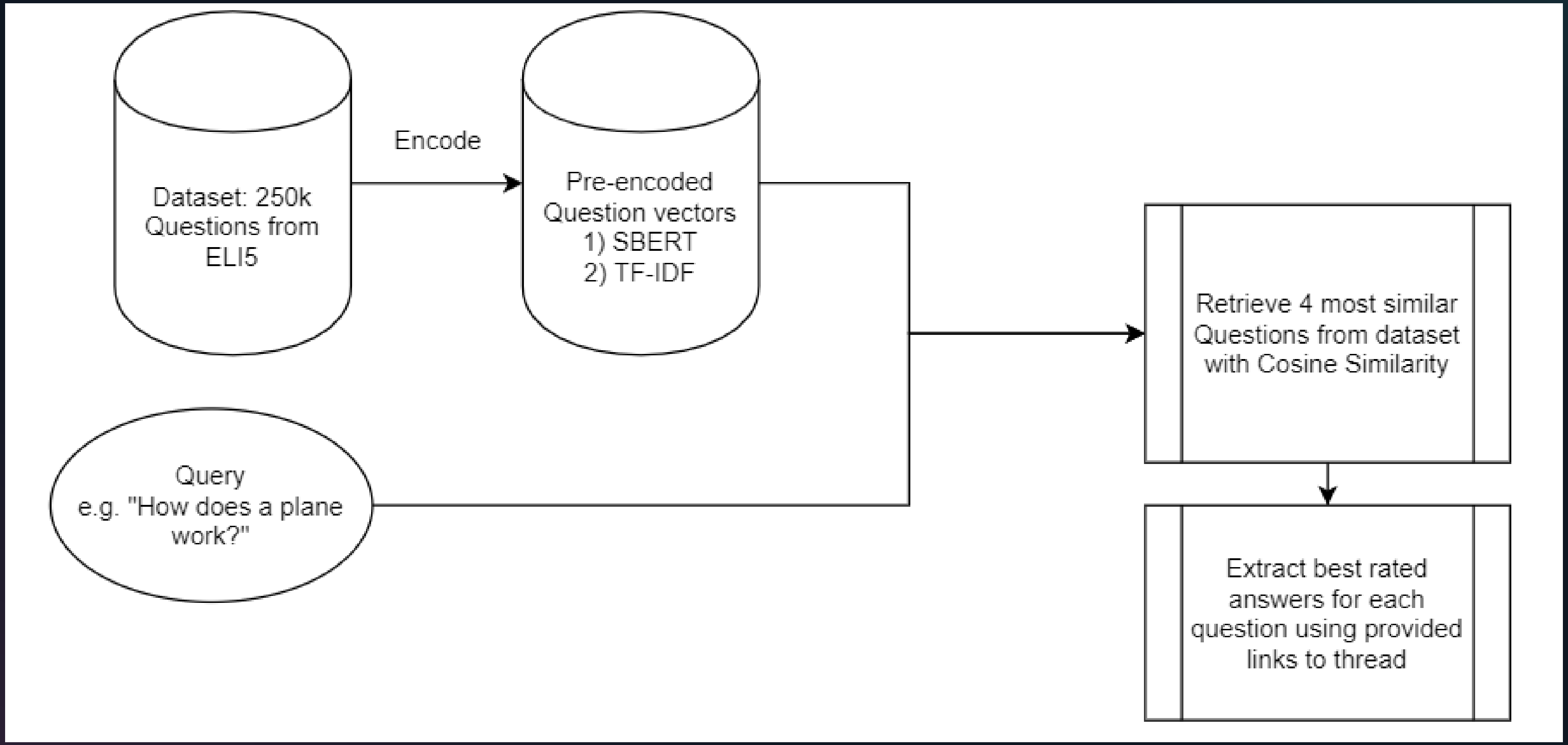
Use reddit API PRAW to extract top comments for each of the four most similar questions

Deploying Website

This webapp is built using Flask and runs on a pythonanywhere server



Pipeline



Dataset: 250k Questions from ELI5

Encode

Pre-encoded Question vectors
1) SBERT
2) TF-IDF

Query
e.g. "How does a plane work?"

Retrieve 4 most similar Questions from dataset with Cosine Similarity

Extract best rated answers for each question using provided links to thread

Tests

Example Questions

3 Example Questions to see the difference between TF-IDF and SBERT

’Is there racism in animals?’

TF-IDF

1. ELI5 : Why is there racism ?
2. ELI5 : Why is there racism ?
3. ELI5 : why is racism still a thing ?
4. ELI5 : Why is racism so common ?

SBERT

1. ELI5 : Do animals express discrimination / racism based on the colour of fur / skin ?
2. ELI5 : Do animals get embarrassed or feel shame ?
3. ELI5 : Why is incest breeding bad for animals ?
4. ELI5 : Are there murderers , psychopaths or other behavioral deviations commonly associated with human beings among animals ?

’Why didn’t we fly to Mars already?’

TF-IDF

1. ELI5 : Could we see ” someone ” on Mars ?
2. ELI5 : Why not mars ?
3. ELI5 How high can a fly fly ?
4. ELI5 : Why does it sometimes cost more to fly from A -> B than it does to fly from A -> B -> C ?

SBERT

1. ELI5 : why haven’t we been able to land on Mars ?
2. ELI5 : Why haven’t we put a man on Mars yet ?
3. ELI5 : Why are we making our expedition to mars a one way trip ?
4. ELI5 : Why do we need to go to Mars ?

’Why do we need a social circle?’

TF-IDF

1. ELI5 : If you were to draw the universe as a circle on a piece of paper , what would be outside the circle ?
2. ELI5 : A circle is 360 . Is that arbitrary ? Could we divide a circle by 100 ?
3. ELI5 : Can someone describe why a circle only has one dimension ?
4. ELI5 : Why do humans need social interaction ?

SBERT

1. ELI5 : Can someone please explain to me what a social imaginary is ?
2. ELI5 : Why do humans need social interaction ?
3. ELI5 : What is a Caucus and why do we have them ?
4. ELI5 : Why are humans wired for social connection ?

Observations

- Both systems work well enough for straight forward queries
- SBERT ‘understands’ semantically complicated queries and provides satisfying results - ! given enough data !
- TF-IDF model could frustrate customers with ‘unintelligent’ search results
- Empirical tests need to be conducted with human evaluators for more insight

Perfect End User vs Reality

How do cameras work?

Perfect End User Angela

gravity

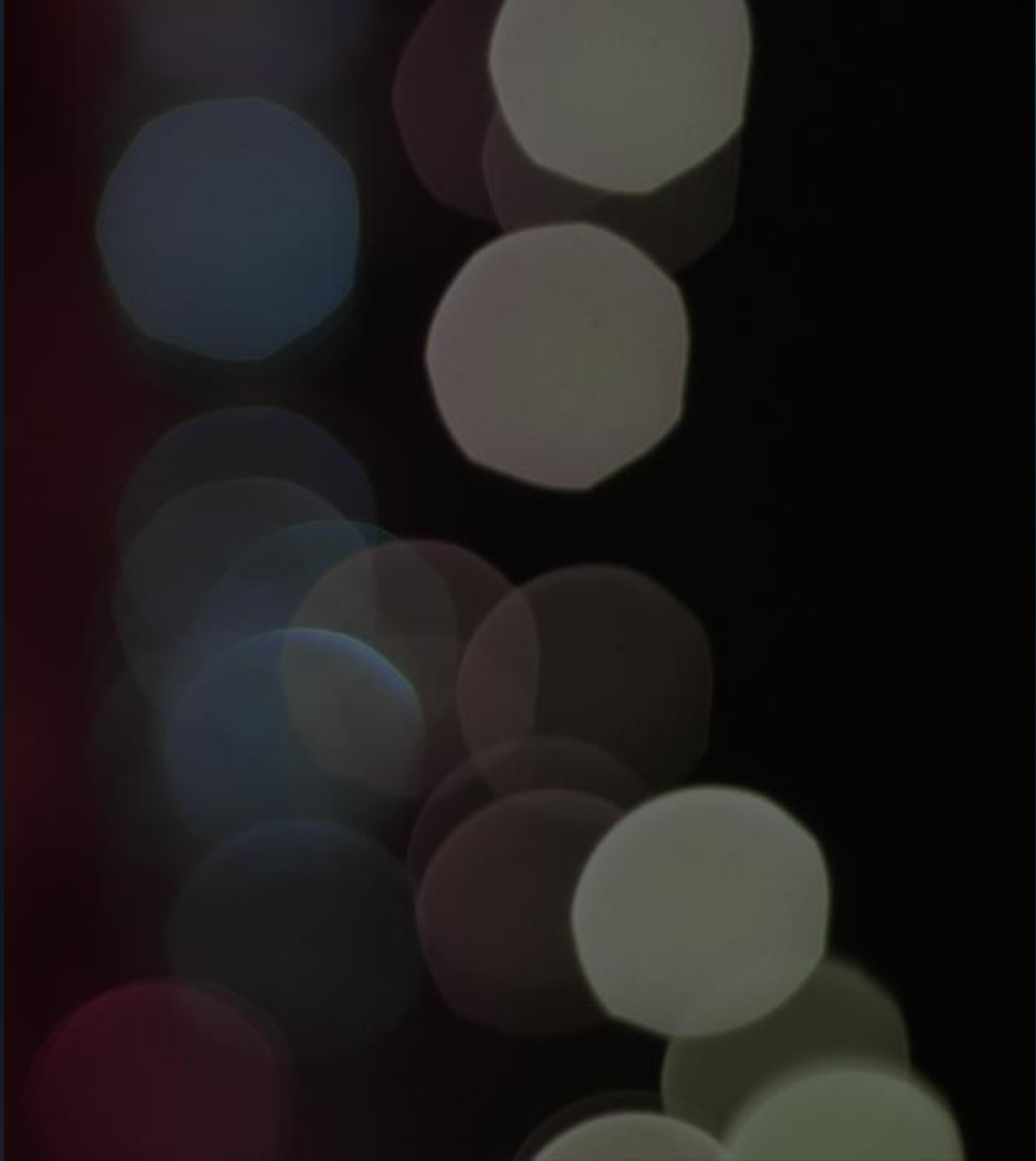
Average End User Simon

boobs

‘That’ End User Donald

Live Demo

THANK YOU!



References

See Github

Feedback from Reddit

typed in boobs, got what i was looking for. thanks op.

metaljerky