Submission for
Exercise task 4
of UTU course TKO_8964-3006
Textual Data Analysis
by Botond Ortutay

Instructions:

Duplicate detection is one of the applications of embeddings. Let's try!

Grab the first 2000 examples of this dataset: https://huggingface.co/datasets/sentence-transformers/quora-duplicates (the "pairs" version) which contains examples of duplicate questions in Quora. Each example has an *anchor* and a *positive* and they form a duplicate question pair.

Embed both anchors and positives with some embedding model. You can get away with a small, simple model like `all-MiniLM-L6-v2` which allows you to run this on CPU. Index the *positive* embeddings in FAISS (`IndexFlatL2` is quite enough!). Then, query the index with the *anchors* and evaluate how often the correct hit (i.e. the corresponding *positive* to the query *anchor*) is in top 1 and how often it is in top 5 (say). In other words, you are evaluating the accuracy of the retrieval.

Solutions:

Importing libraries & environment setup:

NOTE: we use the CPU only version of FAISS

In [1]:
```
import datasets                 # For downloading datasets off of huggingface
import sentence_transformers    # For generating embeddings
import faiss
import numpy as np              # For using with faiss
import random                   # For getting random samples
```

Dataset setup:

In [2]:
```
# Downloading
quoraData = datasets.load_dataset("sentence-transformers/quora-duplicates", "pair", split="train")
# Limiting dataset to first 2000 members as per the exercise instructions
quoraData = quoraData[:2000]
"""
Data now in the following format:
quoraData               dict    2 keys:
quoraData["anchor"]     list    length: 2000
quoraData["positive"]   list    length: 2000
"""
print("")    # This is here so that running this module wouldn't result in printing artifacts
```

Generating embeddings:

In [3]:
```
# Download model and configure to run on CPU and calculate cosine similarity
model = sentence_transformers.SentenceTransformer(model_name_or_path = "sentence-transformers/all-MiniLM-L6-v2"
```

In [4]:
```
# Calculating embeddings
anchorEmbeddings = model.encode(quoraData["anchor"])
positiveEmbeddings = model.encode(quoraData["positive"])
```

Indexing `positiveEmbeddings` using FAISS:

So basically: we take the embeddings we have generated above and put them into a searchable index, so that they can be queried.

In [5]:
```
# Creating new FAISS index
EMBEDDING_DIMENSIONS = len(positiveEmbeddings[0])     # How many dimensions do our embeddings have?
posIndex = faiss.IndexFlatL2(EMBEDDING_DIMENSIONS)    # Generating n dimensional search index where n matches o

"""
Note for self: different FAISS index options listed at:
https://github.com/facebookresearch/faiss/wiki/Faiss-indexes
Useful website!
"""

# adding our embeddings to the index. Note the use of numpy here!
posIndex.add(np.array(positiveEmbeddings, dtype=np.float32))
```

Functions for querying:

In [6]:
```
"""
Searches a given index using pre-calculated embeddings as a query
---
In:
embeddingQuery          numpy.ndarray                pre-calculated embedding to be used as a query    NOTE:
index                   faiss.swigfaiss.IndexFlatL2  index to be searched                              NOTE:
k                       int                          amount of search results to be returned
---
Out:
distanceIndexTuple      tuple
distanceIndexTuple[0]   numpy.ndarray, shape: (1,k)  contains the cosine distance of the search results to t
distanceIndexTuple[1]   numpy.ndarray, shape: (1,k)  contains the positions of the search results within the
"""
def embeddingBasedSearch(embeddingQuery, index, k):
    print("Performing embedding based search!")
    print("")
    print("---")
    print("")

    # Transforming input (regular python array) to numpy array and performing search
    return index.search(np.array([embeddingQuery], dtype=np.float32), k)

"""
Searches a given index using a text query
---
In:
textQuery               str                                                          query
embeddingGenerator      sentence_transformers.SentenceTransformer.SentenceTransformer  model we use to gener
index                   faiss.swigfaiss.IndexFlatL2                                  index to be searched
k                       int                                                          amount of search resu
---
Out:
distanceIndexTuple      tuple
distanceIndexTuple[0]   numpy.ndarray, shape: (1,k)                                  contains the cosine d
distanceIndexTuple[1]   numpy.ndarray, shape: (1,k)                                  contains the position
"""
def textBasedSearch(textQuery, embeddingGenerator, index, k):
    print("Searching for: \"" + textQuery + "\"")
    print("")
    print("---")
    print("")

    # Calculating embeddings for input (string), converting calculated embeddings to numpy array and performing
    return index.search(np.array([embeddingGenerator.encode(textQuery)], dtype=np.float32), k)

"""
Function used to print search results from functions embeddingBasedSearch() and textBasedSearch()
---
In:
distanceIndexTuple      tuple
distanceIndexTuple[0]   numpy.ndarray, shape: (1,k)  contains the cosine distance of the search results to t
distanceIndexTuple[1]   numpy.ndarray, shape: (1,k)  contains the positions of the search results within the
---
Out:
---
"""
def printSearchResults(distanceIndexTuple, documents):
    print("Search results:")

    for i in range(len(distanceIndexTuple[0][0])):    # Where len(distanceIndexTuple[0][0]) is k
        print(str(i+1) + ". " + documents[distanceIndexTuple[1][0][i]] + "    Distance: " + "{:.2f}".format(dis

    print("")
    print("")
    print("")
    print("")
```

Analysis:

So now that we have done embedding calculations, created an index and even written some querying functions, it should be super easy to search the index and determine how "good" it is and how "well" it works. In order to provide some demonstration, but then also keep my sanity, I'm going to write a script that takes 20 random samples from our data, uses the `embeddingBasedSearch` function and `k = 5` and searches the index. My hypotheses is that the correct positive will be the #1 result at least 90% of the time (so at least 18 out of 20 times) and in the top 5 100% of the time. Let's write and run the code!

In [7]:
```
for i in range(20):
    # Taking random sample
    currentSample = random.randint(0,1999)
    print("Randomly selected anchor question: " + quoraData["anchor"][currentSample])

    # Getting search results
    results = embeddingBasedSearch(anchorEmbeddings[currentSample], posIndex, 5)

    # Is the correct answer the #1 result?
    if (results[1][0][0] == currentSample):
```

```
        print("#1 MATCH!!!")

    # If the correct answer is not the #1 result: Is the correct answer in the top 5?
    else:
        for j in range(1,5):
            if (results[1][0][j] == currentSample):
                print("#" + str(j) + " MATCH!!!")

    # Print the search results
    printSearchResults(results,quoraData["positive"])
```

```
Randomly selected anchor question: U.S. Presidential Elections: Would Trump beat Sanders if they were the nomin
ees for President?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Could Trump beat Sanders in a general election?    Distance: 0.32
2. What would happen if the presidential nominee died before the November election?    Distance: 1.00
3. Is there a big chance that Trump will win the election?    Distance: 1.00
4. Does Donald Trump have any chance of winning the forthcoming election?    Distance: 1.02
5. Who is the better candidate for being the President of the United States of America: Hillary Clinton or Dona
ld Trump?    Distance: 1.02



Randomly selected anchor question: Does global warming exist?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Is it possible that global warming is a hoax?    Distance: 0.42
2. Is it possible that global warming is a hoax?    Distance: 0.42
3. Is the global warming climate change things for real or a hoax?    Distance: 0.54
4. Do we have any good ways to stop global warming?    Distance: 0.66
5. Can the global climate change be reversed or halted?    Distance: 0.68



Randomly selected anchor question: Quora: How do you post a question on Quora?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. How do I post something in Quora?    Distance: 0.28
2. What are the best ways to ask a question on Quora?    Distance: 0.49
3. How do you delete a question on Quora?    Distance: 0.69
4. How do I delete a question from Quora?    Distance: 0.72
5. How do you find good questions on Quora?    Distance: 0.77



Randomly selected anchor question: What are part time jobs that can work from home?
Performing embedding based search!

---

#2 MATCH!!!
Search results:
1. What are some of the best paid part time jobs that can be done from home?    Distance: 0.15
2. What are some of the best paid part time jobs that can be done from home?    Distance: 0.15
3. What are some of the best paid part time jobs that can be done from home?    Distance: 0.15
4. What are the best ways to earn money from home?    Distance: 0.57
5. How can I make money online for job?    Distance: 0.81



Randomly selected anchor question: What are the daily life examples of shear stress?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What is shear stress? Also give any real life example of it    Distance: 0.41
2. Why is pressure an intensive property?    Distance: 1.16
3. What are the most common examples of solid matter?    Distance: 1.25
4. How do earthquake resistant buildings work?    Distance: 1.26
5. What are some human effects on the water cycle?    Distance: 1.26
```

```
Randomly selected anchor question: What countries are socialist?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Are there any socialist countries left?    Distance: 0.36
2. What is the difference between socialism, marxism, and communism?    Distance: 0.88
3. What are some political affiliations?    Distance: 1.09
4. Who is the richest country in the world?    Distance: 1.11
5. What's the best European country to start a business in (as a U.S. citizen)?    Distance: 1.17



Randomly selected anchor question: What is the difference between transgender man and transgender woman?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What is the difference between a transgender man and a transgender woman?    Distance: 0.02
2. If society reversed gender roles would transgender people be the same people or different people?    Distanc
e: 0.65
3. Is sex more pleasurable for men or for women?    Distance: 1.19
4. Who are better drivers woman or man?    Distance: 1.25
5. What is the difference between humans and the other animals?    Distance: 1.28



Randomly selected anchor question: Will Modi win in 2019?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Can Narendra Modi become Prime Minister of India in 2019?    Distance: 0.58
2. Does Rahul Gandhi have chances to become next Pm of India after Modi?    Distance: 0.74
3. Will India win on Kashmir issue.?    Distance: 0.87
4. What will be the next move by PM Modi to improve India?    Distance: 0.87
5. Who will be next CM of Gujarat?    Distance: 0.89



Randomly selected anchor question: How Can I impress a girl who hate me?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. How can impress a girl who hate me?    Distance: 0.03
2. How do I flirt with any girl?    Distance: 1.02
3. How do I impress my mother in law?    Distance: 1.08
4. How do I know if this girl likes me?    Distance: 1.17
5. My ex-girlfriend is in my class and I am unable to face her. I feel emotionally tortured because I still hav
e the same love for her. How do I face her?    Distance: 1.17



Randomly selected anchor question: What do Pakistani people think about the Uri attack on 18th September 2016?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What do Pakistani people think about the Uri attack on 18th September 2016?    Distance: 0.00
2. What do Pakistani residents think of Uri attack?    Distance: 0.18
3. What is the best option to respond to Pakistan after Uri attack?    Distance: 0.56
4. How India can respond to the Uri terror attack?    Distance: 0.73
5. What steps should be taken immediately by our country against Pakistan for 29 th November attack?    Distanc
e: 0.75



Randomly selected anchor question: How will Trump's presidency affect Indian students who are planning to do a
PhD in the US?
```

```
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. How would Trump presidency affect Indian students in the US?    Distance: 0.19
2. How will a Trump presidency affect the students presently in US or planning to study in US?    Distance: 0.4
6
3. How is Trump becoming the president affect the Indians applying for an MS in the US (Mech)?    Distance: 0.4
7
4. How will Donald Trump benefit India?    Distance: 0.57
5. How might Trump affect the status of foreign students at top universities in the US?    Distance: 0.62



Randomly selected anchor question: Is mechanical keyboard really helpful for touch typing?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Is mechanical keyboard helpful for Touch Typing?    Distance: 0.01
2. How can I create a typing effect on my website?    Distance: 1.16
3. How can I have good handwriting?    Distance: 1.33
4. How can I have good handwriting?    Distance: 1.33
5. Is there a way I could learn to play the piano?    Distance: 1.39



Randomly selected anchor question: Why did Mahatma Gandhi not get Bharat Ratna or the Nobel Peace Prize?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Why Mahatma Gandhi didn't get Bharat Ratna?    Distance: 0.26
2. Was giving Nobel Prize to Malala a complete joke?    Distance: 0.98
3. Why has an Indian born person never received the Fields medal?    Distance: 0.99
4. Did Mahabharata really happen or only an allegory according to M.K.Gandhi?    Distance: 1.03
5. Does Rahul Gandhi have chances to become next Pm of India after Modi?    Distance: 1.07



Randomly selected anchor question: What is the best method of losing weight?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are the best way of loose the weight?    Distance: 0.30
2. What are the best things to do when working on losing weight?    Distance: 0.32
3. How can I efficiently lose weight?    Distance: 0.42
4. How do i lose weight?    Distance: 0.44
5. What is the easiest way to lose weight faster?    Distance: 0.46



Randomly selected anchor question: Is it possible to root an iOS into an Android phone?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Is it possible to install iOS to an Android phone?    Distance: 0.33
2. Is it possible to run iOS apps on an Android phone?    Distance: 0.45
3. How do I root htc desire 826?    Distance: 1.21
4. How much would it cost to build your own iPhone?    Distance: 1.23
5. What is the main purpose of jailbreaking an iPhone? How is it done?    Distance: 1.26



Randomly selected anchor question: What are the qualities of a good leader?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are some good qualities of a leader?    Distance: 0.04
```

```
2. What personality traits do all successful salespeople have in common?    Distance: 1.04
3. Who is your role model and why?    Distance: 1.05
4. Why will Donald Trump be a good president?    Distance: 1.08
5. What are the best strengths of Indian Army?    Distance: 1.09



Randomly selected anchor question: What is the simplest way to forget someone?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What is the quickest and easiest way to forget someone?    Distance: 0.11
2. How can you completely get over someone?    Distance: 0.76
3. How does a person get over a broken heart?    Distance: 1.04
4. How do you get over a broken heart?    Distance: 1.04
5. How do you turn someone down?    Distance: 1.06



Randomly selected anchor question: Can height increase after 25?
Performing embedding based search!

---

#2 MATCH!!!
Search results:
1. Can height increase after 25?    Distance: 0.00
2. Can Height be increased after 18 or 19 years of age?    Distance: 0.31
3. Can someone increase their height naturally after 19?    Distance: 0.34
4. How do I increase our height after 21?    Distance: 0.51
5. How can I increase in height after 20 years?    Distance: 0.53



Randomly selected anchor question: What is a black hole? How can we understand it?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What is called a black hole?    Distance: 0.41
2. Do black holes exist?    Distance: 0.78
3. Would a black hole be the exit of this universe?    Distance: 0.98
4. Does a black hole have a finite mass?    Distance: 0.98
5. What happened to the event horizons of the two black holes that merged sendig to us gravitational waves?
Distance: 1.10



Randomly selected anchor question: What is the best App for downloading films for free?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are the best apps for downloading films for free?    Distance: 0.04
2. What is the best app for video editing?    Distance: 0.91
3. What is a good free animation software?    Distance: 1.02
4. What are the best educational apps?    Distance: 1.05
5. How do you promote your app for free?    Distance: 1.06
```

So because these results are based on random samples, the output of the above script will always be different (or not "always" but you get the point). Therefore whoever runs it after me will get a different result. And I will also get a different result, when I turn off the kernel and turn it back on and run all the cells to test if everything works correctly before submitting the exercise. Therefore: I saved the example output I will use in my analysis. I'll paste it below.

```
Randomly selected anchor question: Why does Dubai Police drive fast car?
Performing embedding based search!

---

#1 MATCH!!!
```

```
Search results:
1. Why do the Dubai Police have super cars?    Distance: 0.32
2. Why do people on Bay Area highways drive so slowly in the left lane?    Distance: 0.95
3. Is it legal for a traffic police to stand at the middle of the road to stop vehicles?
Distance: 1.19
4. Why are police lights red and/or blue?    Distance: 1.22
5. How much does Uber driver earn in India?    Distance: 1.28



Randomly selected anchor question: What are some best examples of Presence of mind?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are some of the examples of presence of mind?    Distance: 0.03
2. What are some of the examples of presence of mind?    Distance: 0.03
3. How can I increase my presence of mind?    Distance: 0.61
4. How can I maintain my peace of mind?    Distance: 1.08
5. What are some books that expand our mind?    Distance: 1.11



Randomly selected anchor question: Is there any proof which can be given for the existence of
the GOD? If yes, what are those?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Is there any proof that there is no god?    Distance: 0.38
2. Can math prove the existence of God?    Distance: 0.42
3. Is there really the existence of Aliens and is there any proof available realted to them?
Distance: 0.91
4. Who created the "GOD"?    Distance: 0.98
5. Do Greek gods exist? Why or why not?    Distance: 1.03



Randomly selected anchor question: What are some of the best life tips?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are some of your best life coaching tips?    Distance: 0.44
2. What is the best advice you ever received?    Distance: 0.88
3. What are the best school life hacks?    Distance: 0.97
4. What is the most important lesson you have learned from life?    Distance: 0.97
5. What is the best thing we learned from our life?    Distance: 1.02



Randomly selected anchor question: What are some Cyanide and Happiness comics on countries?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are the best Cyanide & Happiness comics?    Distance: 0.33
2. What are the best logos ever created?    Distance: 1.19
3. What are the few things that make Indians happy?    Distance: 1.20
4. What are some cool python scripts?    Distance: 1.23
5. What are some baby shower games that are actually fun?    Distance: 1.28



Randomly selected anchor question: How do I improve my drawing skills and techniques?
Performing embedding based search!
```

```
---

#1 MATCH!!!
Search results:
1. How can I improve my drawing skills?    Distance: 0.04
2. How do you improve your drawing?    Distance: 0.21
3. My skill is drawing. How Can I make money out of it?    Distance: 0.67
4. How can I have good handwriting?    Distance: 0.83
5. How can I have good handwriting?    Distance: 0.83



Randomly selected anchor question: Which are the best movies of 2016?
Performing embedding based search!

---

#2 MATCH!!!
Search results:
1. What is the best film of 2016?    Distance: 0.12
2. Which was the best film of 2016?    Distance: 0.17
3. What is your best 2016 movie?    Distance: 0.17
4. What are some best horror movies of 2016?    Distance: 0.36
5. What are some of the best movies of 2014?    Distance: 0.59



Randomly selected anchor question: Can anyone become good at mathematics?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Can everyone become good at math?    Distance: 0.17
2. How do you learn algebra 1 fast?    Distance: 0.90
3. Can math prove the existence of God?    Distance: 0.93
4. Is math an art or a science?    Distance: 1.02
5. How do I score good marks in mathematics (9 cbse)?    Distance: 1.02



Randomly selected anchor question: How close are we to World War Three, and how bad would it
be?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. How close is a World War III?    Distance: 0.30
2. Are we heading toward World War 3?    Distance: 0.37
3. Do you think we are on the verge of World War III?    Distance: 0.45
4. Is World War 3 more imminent than expected?    Distance: 0.49
5. Is World War 3 more imminent than expected?    Distance: 0.49



Randomly selected anchor question: How can I wake up early in the morning?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. How can I get up early in the morning?    Distance: 0.18
2. How can I efficiently learn while sleeping?    Distance: 1.03
3. How does one sleep Less but not feel tired?    Distance: 1.04
4. Do you need to wake up in the middle of REM sleep in order to remember you dreams?
Distance: 1.08
5. What's the one thing you think about when you wake up?    Distance: 1.12
```

Randomly selected anchor question: Does the female body undergo changes after losing virginity? If not, why would people think it does?
Performing embedding based search!

#1 MATCH!!!
Search results:
1. What might be the reasons why people sometimes believe a woman's body changes after she loses her virginity? Distance: 0.34
2. What's the right age to lose virginity? Distance: 1.10
3. What did it feel like when you first had sex? Distance: 1.15
4. What is maturity? Is it only the physical change? Distance: 1.17
5. Does DNA change when growing up from baby to adult? Distance: 1.22

Randomly selected anchor question: What are some symptoms of eccentric and concentric contractions?
Performing embedding based search!

#1 MATCH!!!
Search results:
1. How do concentric and eccentric contraction compare and contrast? Distance: 0.51
2. How do concentric and eccentric contraction compare and contrast? Distance: 0.51
3. What are the early and common signs of pregnancy? Distance: 1.12
4. What are the causes of a yellow jelly discharge? Distance: 1.33
5. What are some causes that make your period come early? Distance: 1.36

Randomly selected anchor question: How do I prepare for software interviews?
Performing embedding based search!

#1 MATCH!!!
Search results:
1. What are the best ways to prepare for software interviews? Distance: 0.05
2. How can I prepare for interview? Distance: 0.38
3. How do I prepare for KVPY sa interview? Distance: 0.76
4. What is to be done to be a good software developer? Distance: 0.94
5. How should I start preparing for UPSC(IAS) exams? Distance: 0.99

Randomly selected anchor question: Can I recover my email if I forgot the password?
Performing embedding based search!

#1 MATCH!!!
Search results:
1. What should I do if I forgot my email password? Distance: 0.28
2. I can't remember my Gmail password or my recovery email. How can I recover my e-mail? Distance: 0.51
3. What should I do if I forgot my iCloud email and password? Distance: 0.53
4. I do not remember my password to my Gmail account, how can I recover my account? Distance: 0.55
5. How can you recover your Gmail password? Distance: 0.58

Randomly selected anchor question: What do you think about the Bermuda Triangle?
Performing embedding based search!

#1 MATCH!!!
Search results:
1. What are your theories about Bermuda Triangle? Distance: 0.31
2. What do you think about the movie Interstellar? Distance: 0.97

```
3. What happened to MH370?    Distance: 1.20
4. What is your opinion on brexit?    Distance: 1.21
5. What is your view/opinion about Brexit?    Distance: 1.27



Randomly selected anchor question: Do atheists who celebrate Christmas call it something
different?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. Do atheists call Christmas something different?    Distance: 0.08
2. Why do people say "bless you" whenever someone sneezes?    Distance: 1.35
3. Why do we say god bless you when we sneeze?    Distance: 1.41
4. What is the origin of saying "bless you" when someone sneezes?    Distance: 1.42
5. Are Indians so obsessed with the notion of religion and caste?    Distance: 1.43



Randomly selected anchor question: How can I know if my boyfriend is using dating apps?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. How can I find out whether my partner is using dating sites?    Distance: 0.45
2. How can I know if my spouse is cheating?    Distance: 0.96
3. Do dating apps and sites really work?    Distance: 0.98
4. How do I tell if a girl has a boyfriend?    Distance: 1.00
5. How do you find out whether a hot guy is gay?    Distance: 1.06



Randomly selected anchor question: What are some of the high salary income jobs in the field
of biotechnology?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are some high paying jobs for a fresher with an M.Tech in biotechnology?    Distance:
0.29
2. How can medical doctors move into biotech?    Distance: 0.90
3. What is the salary of a doctor in India?    Distance: 0.95
4. What is the salary for engineer?    Distance: 1.07
5. What are the best part time jobs we can do in Bangalore?    Distance: 1.13



Randomly selected anchor question: What are the best car technology gadgets?
Performing embedding based search!

---

#1 MATCH!!!
Search results:
1. What are the best car gadgets and tools?    Distance: 0.23
2. What are the best available technology gadgets?    Distance: 0.46
3. What are some of the best smartphones technology gadgets?    Distance: 0.64
4. Which is your best gadget?    Distance: 0.64
5. What are some mind blowing technology gadgets that most people don't know?    Distance:
0.72



Randomly selected anchor question: Is mechanical keyboard really helpful for touch typing?
Performing embedding based search!

---
```

```
#1 MATCH!!!
Search results:
1. Is mechanical keyboard helpful for Touch Typing?    Distance: 0.01
2. How can I create a typing effect on my website?    Distance: 1.16
3. How can I have good handwriting?    Distance: 1.33
4. How can I have good handwriting?    Distance: 1.33
5. Is there a way I could learn to play the piano?    Distance: 1.39
```

So with this particular output 19 of the 20 tests produced a #1 match! And the remaining 1 question was a #2 match. There the question was "Which are the best movies of 2016?", the #1 match was "What is the best film of 2016?" and the correct answer (#2 match) was "Which was the best film of 2016?". So basically the system not finding the "correct" duplicate was due to several duplicates existing within the data; We can safely say that if the system's goal is to find duplicates, than it did its job perfectly despite not flagging the "correct" duplicate as the most likely match.

Based on this one run (not enough for a scientific analysis, but good enough for this demonstration), I can conclude: Yes, the system indeed appears to be working. Although there was one run when I was writing the script where the system did actually find one query where the correct answer wasn't in the top 5, which is curious. However I did not save that run, and I'm not willing to go comb the 2000 datapoints I have to find that one outlier again, that is not the point of this exercise.

Because I was curious enough to want to search these 2000 quora questions using custom text queries I wrote a function which can help me perform just that. I now also want to run that function just for fun and to show that it works & stuff. Since the very first item in the dataset is about astrology, I'm gonna use the question "Should I believe in astrology?" as my query.

In [8]:
```
printSearchResults(textBasedSearch("Should I believe in astrology?",model,posIndex,5),quoraData["positive"])
```

```
Searching for: "Should I believe in astrology?"

---

Search results:
1. Do you believe in horoscope?    Distance: 0.56
2. Do you believe in horoscopes?    Distance: 0.59
3. Are there any good free online astrologers?    Distance: 1.06
4. Do you believe that everything happens for a reason?    Distance: 1.20
5. How can I learn astronomy?    Distance: 1.21
```

As you can see even with a custom query, all the search results (at least in the top 5) are relevant. This gives even more credibility to the system working as intended. Anyone can run the `textBasedSearch` function with any query, and the system will deliver relevant results, as long as there are relevant results in the index.