1 Search Engine

1.1 Dataset

The data was acquired from crawling simple.wikipedia.com pages using wikipedia_crawler_2.py script.
The crawling script was run for around 5 hours and collected **89 535 articles** about various topics.
The starting point for crawling was page "World War 2" with max depth set to 4.
Many of the articles are related to history, but due to the depth of the search there are lots of articles which have zero connection to history (e.g. the dataset contains "(You Drive Me) Crazy" article, which is a song by Britney Spears).

1.2 Data processing

1.2.1 Raw articles processing

After collecting articles, the raw text was preprocessed using ArticleProcessor.py script, which stripped the sections at end of the article (references and other websites section), then it used ntlk library to remove stop words and punctuation and them, used Porter stemming algorithm to stem the words. The processed articles were saved into new folder to process further.
This was done with multi threading to speed the process up.

1.2.2 Dictionary creation

The processed articles were used by DictionaryCreator.py script, which collected every word used in articles into a set, it also keep the count of each word occurrence. It also removed every word which had length lower than 3, and words which occurred less than 15 times.
This was done with multi threading to speed the process up.
The final size of the dictionary was **35 030 words**

1.2.3 Article lookup map creation

The processed articles were used by ArticleLookupMapCreator.py, which assigned an index to each article. This map will be used later by the engine to map the resulting index to a specific article name. This is done by inserting the article name into an array.

1.2.4 Creating term-by-document matrix

The ArticleVectorizer.py script utilizes the processed articles and the dictionary. It creates a MxN matrix, where M is the number of articles and N is the number of words in dictionary. Each row $\mathrm{R_i}$ represents how many occurrences of the dictionary words are in a article whose index is i. For example cell $\mathrm{C_{i,j}}$ represents how many times word with index j occured in article with index i. The created matrix is sparse, so we handle it using scipy.sparse methods. After creating the matrix, each column is multiplied by inverse document frequency (IDF) value, which is calculated for each word:

$$IDF(w) = log\frac{M}{n_{\mathrm{w}}} \tag{1.1}$$

Where M is the count of articles, and n_w is the count of articles contacting at least one word w. We do that to reduce common word significance (word occurring in only one document will have very high IDF value, while word occurring in every document will have IDF equal to 0)
After calculating, the matrix is normalized by $\mathrm{L_2}$ norm and then saved on the disk to be loaded later.

1.2.5 Low rank approximation

The created term-by-document matrix is loaded by ArticleVectorSVD.py script, and then it uses sklearn.decomposition.TruncatedSVD class to calculate matrices for low rank approximation of our original matrix.
My computer could not handle computing the full matrix approximation(U*S*Vh), so the script returns two matrices (U*S) and Vh. I tried some normalization techniques, and i managed to get very good results by using normalized US matrix by rows and normalized Vh matrix by columns. The matrices are then saved into a file to be loaded later.
I calculated approximations of the matrix using values of k: 500,1000,2500,5000

1.3 Querying the data

The data can be queried using the class inside search_engine.py file. The SearchEngine loads the computed dictionary, article lookup map and search matrices(without low rank approximation and with low rank approximation). To query the data, search method is called. It requires user string input, which is sanitized the same way as in 1.2.1 section, then the query vector is created. The query vector has size N, and each cell $\mathrm{C_i}$ represent how many times word with index i occurred in the query, then the vector is normalized by $\mathrm{L_2}$ norm. The search method then performs a search matrix multiplication with created query vector, the resulting vector has size M, and each cell C_i represent how similar the article with index i, is to the provided query, the similarity is based on cosine similarity.

2 Frontend application

The frotend layout is served by Flask server, and the frontend files were created using React.

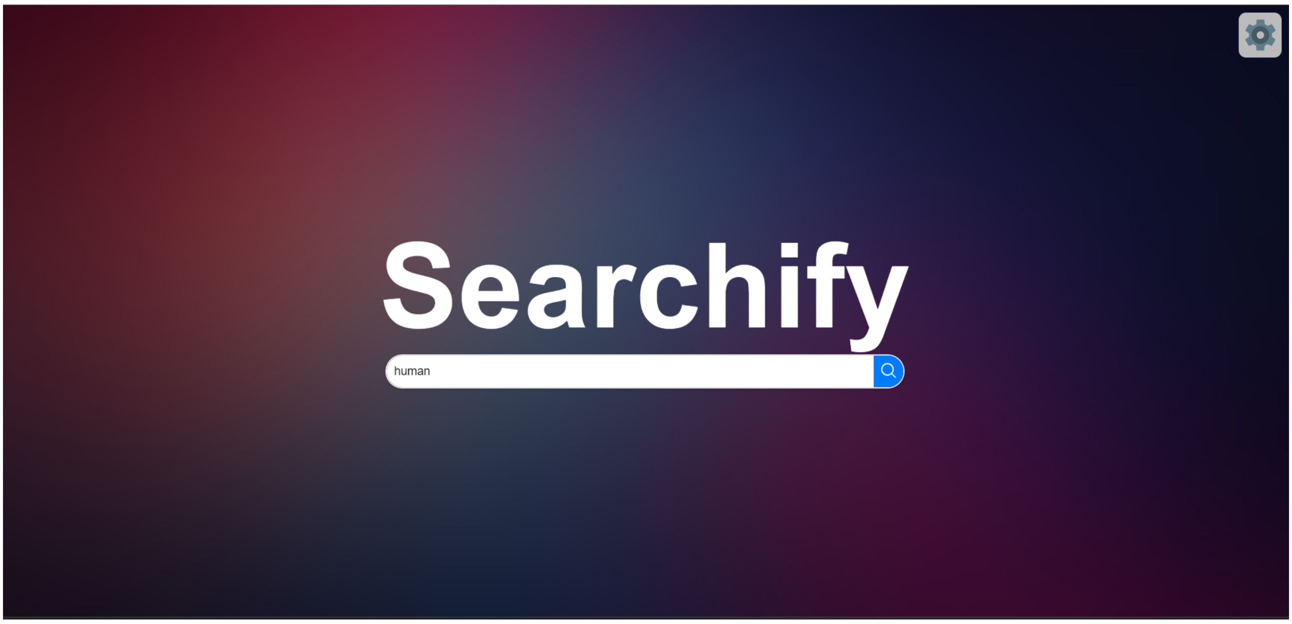
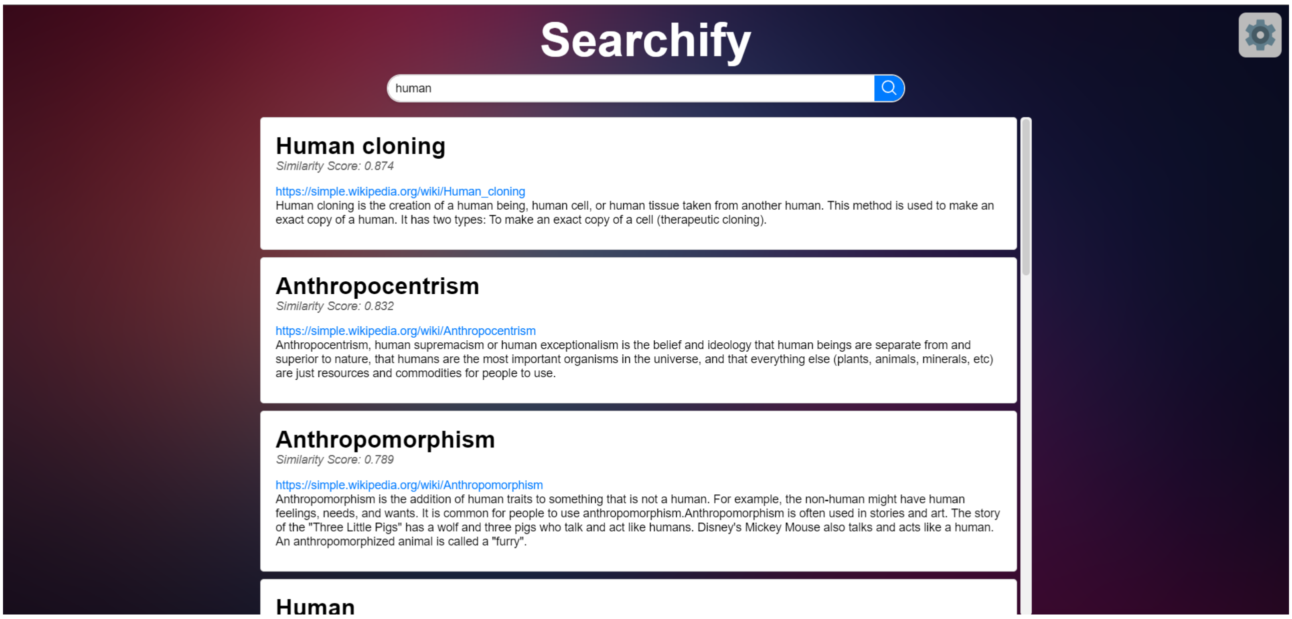


Figure 2.1: Search page



Human cloning

Similarity Score: 0.874

https://simple.wikipedia.org/wiki/Human_cloning

Human cloning is the creation of a human being, human cell, or human tissue taken from another human. This method is used to make an exact copy of a human. It has two types: To make an exact copy of a cell (therapeutic cloning).

Anthropocentrism

Similarity Score: 0.832

https://simple.wikipedia.org/wiki/Anthropocentrism

Anthropocentrism, human supremacism or human exceptionalism is the belief and ideology that human beings are separate from and superior to nature, that humans are the most important organisms in the universe, and that everything else (plants, animals, minerals, etc) are just resources and commodities for people to use.

Anthropomorphism

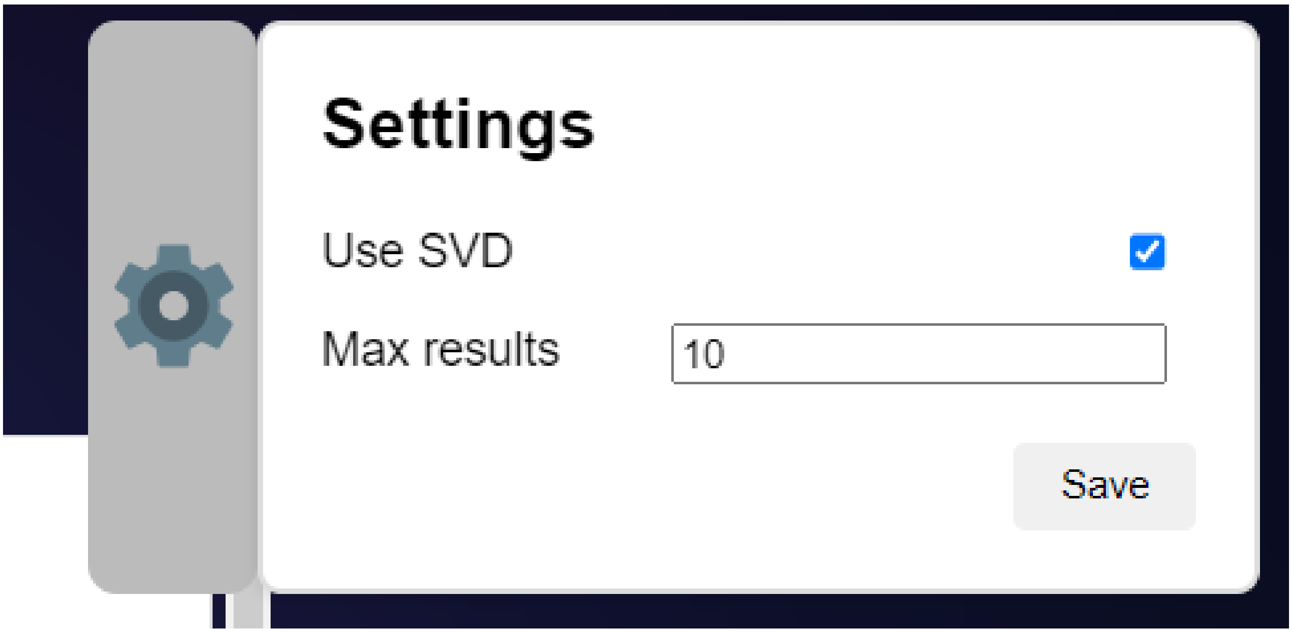
Similarity Score: 0.789

https://simple.wikipedia.org/wiki/Anthropomorphism

Anthropomorphism is the addition of human traits to something that is not a human. For example, the non-human might have human feelings, needs, and wants. It is common for people to use anthropomorphism.Anthropomorphism is often used in stories and art. The story of the "Three Little Pigs" has a wolf and three pigs who talk and act like humans. Disney's Mickey Mouse also talks and acts like a human. An anthropomorphized animal is called a "furry".

Human

Figure 2.2: Search results



Settings

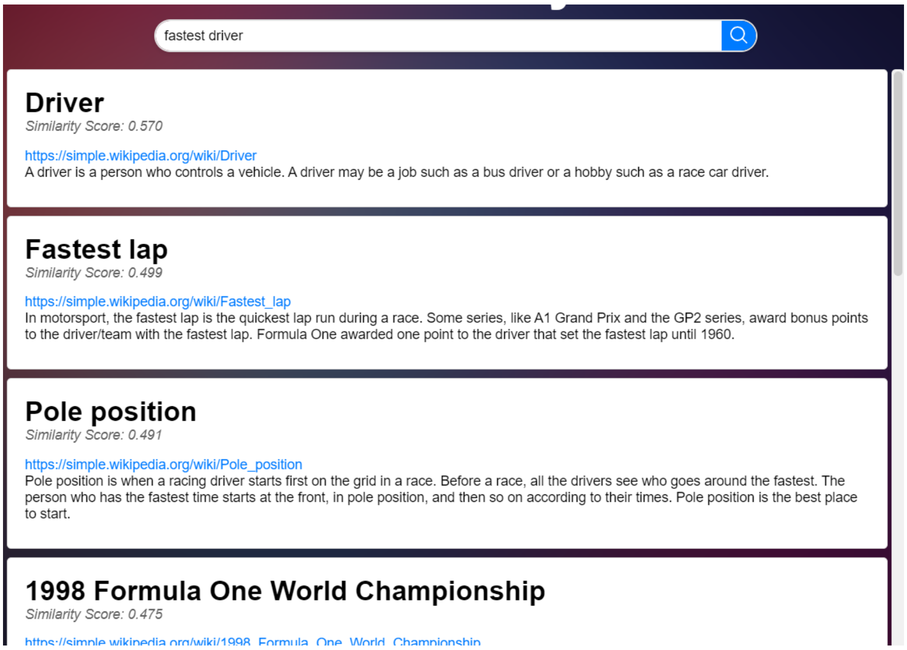
Use SVD ☑

Max results 10

Figure 2.3: Settings

3 Search results

3.1 fastest driver



fastest driver

Driver

Similarity Score: 0.570

https://simple.wikipedia.org/wiki/Driver

A driver is a person who controls a vehicle. A driver may be a job such as a bus driver or a hobby such as a race car driver.

Fastest lap

Similarity Score: 0.499

https://simple.wikipedia.org/wiki/Fastest_lap

In motorsport, the fastest lap is the quickest lap run during a race. Some series, like A1 Grand Prix and the GP2 series, award bonus points to the driver/team with the fastest lap. Formula One awarded one point to the driver that set the fastest lap until 1960.

Pole position

Similarity Score: 0.491

https://simple.wikipedia.org/wiki/Pole_position

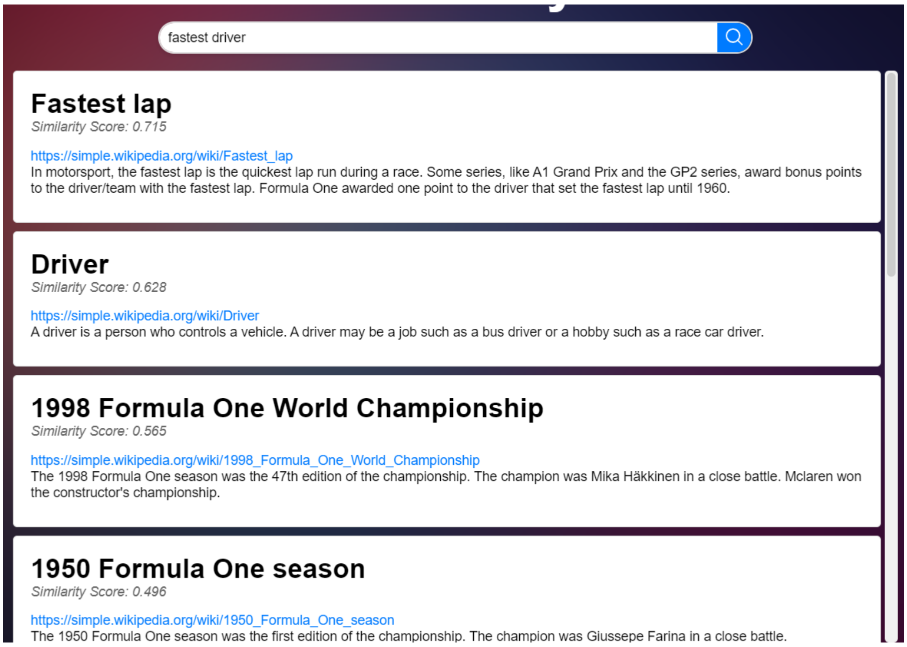
Pole position is when a racing driver starts first on the grid in a race. Before a race, all the drivers see who goes around the fastest. The person who has the fastest time starts at the front, in pole position, and then so on according to their times. Pole position is the best place to start.

1998 Formula One World Championship

Similarity Score: 0.475

https://simple.wikipedia.org/wiki/1998_Formula_One_World_Championship

Figure 3.1: no SVD



fastest driver

Fastest lap

Similarity Score: 0.715

https://simple.wikipedia.org/wiki/Fastest_lap

In motorsport, the fastest lap is the quickest lap run during a race. Some series, like A1 Grand Prix and the GP2 series, award bonus points to the driver/team with the fastest lap. Formula One awarded one point to the driver that set the fastest lap until 1960.

Driver

Similarity Score: 0.628

https://simple.wikipedia.org/wiki/Driver

A driver is a person who controls a vehicle. A driver may be a job such as a bus driver or a hobby such as a race car driver.

1998 Formula One World Championship

Similarity Score: 0.565

https://simple.wikipedia.org/wiki/1998_Formula_One_World_Championship

The 1998 Formula One season was the 47th edition of the championship. The champion was Mika Häkkinen in a close battle. Mclaren won the constructor's championship.

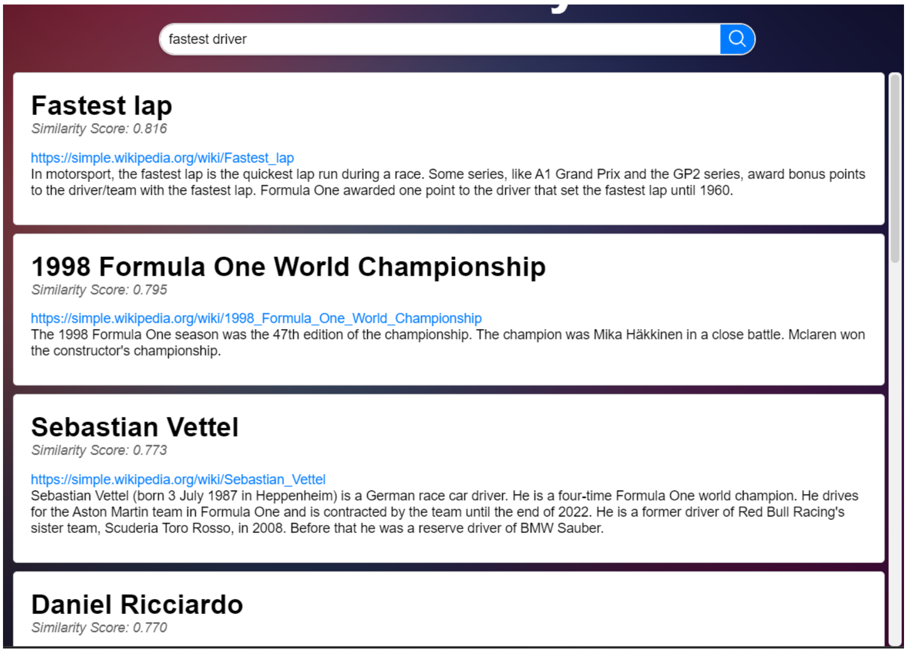
1950 Formula One season

Similarity Score: 0.496

https://simple.wikipedia.org/wiki/1950_Formula_One_season

The 1950 Formula One season was the first edition of the championship. The champion was Giuseppe Farina in a close battle.

Figure 3.2: SVD, k=5000



fastest driver

Fastest lap
Similarity Score: 0.816

https://simple.wikipedia.org/wiki/Fastest_lap

In motorsport, the fastest lap is the quickest lap run during a race. Some series, like A1 Grand Prix and the GP2 series, award bonus points to the driver/team with the fastest lap. Formula One awarded one point to the driver that set the fastest lap until 1960.

1998 Formula One World Championship
Similarity Score: 0.795

https://simple.wikipedia.org/wiki/1998_Formula_One_World_Championship

The 1998 Formula One season was the 47th edition of the championship. The champion was Mika Häkkinen in a close battle. Mclaren won the constructor's championship.

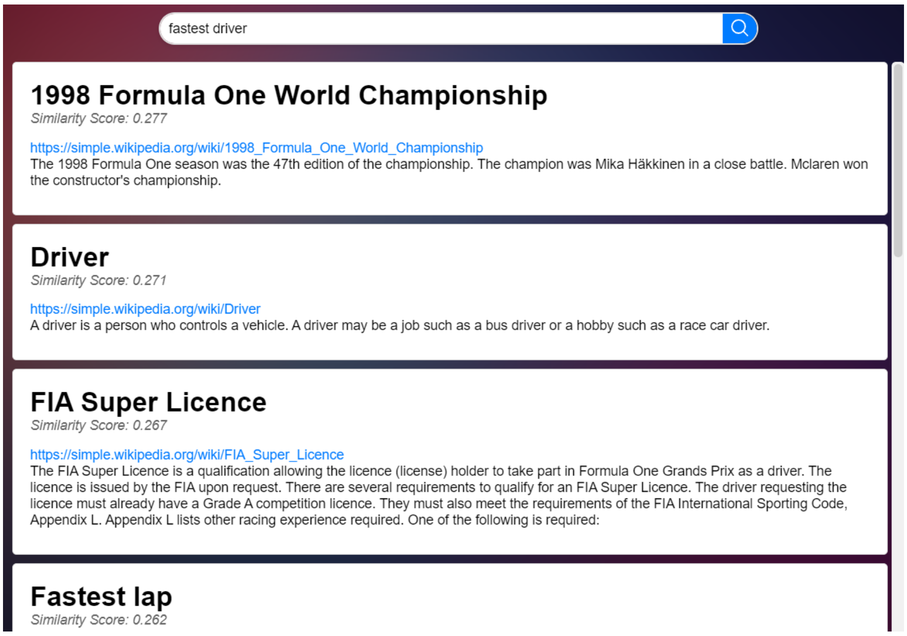
Sebastian Vettel
Similarity Score: 0.773

https://simple.wikipedia.org/wiki/Sebastian_Vettel

Sebastian Vettel (born 3 July 1987 in Heppenheim) is a German race car driver. He is a four-time Formula One world champion. He drives for the Aston Martin team in Formula One and is contracted by the team until the end of 2022. He is a former driver of Red Bull Racing's sister team, Scuderia Toro Rosso, in 2008. Before that he was a reserve driver of BMW Sauber.

Daniel Ricciardo
Similarity Score: 0.770

Figure 3.3: SVD, k=1000



fastest driver

1998 Formula One World Championship
Similarity Score: 0.277

https://simple.wikipedia.org/wiki/1998_Formula_One_World_Championship

The 1998 Formula One season was the 47th edition of the championship. The champion was Mika Häkkinen in a close battle. Mclaren won the constructor's championship.

Driver
Similarity Score: 0.271

https://simple.wikipedia.org/wiki/Driver

A driver is a person who controls a vehicle. A driver may be a job such as a bus driver or a hobby such as a race car driver.

FIA Super Licence
Similarity Score: 0.267

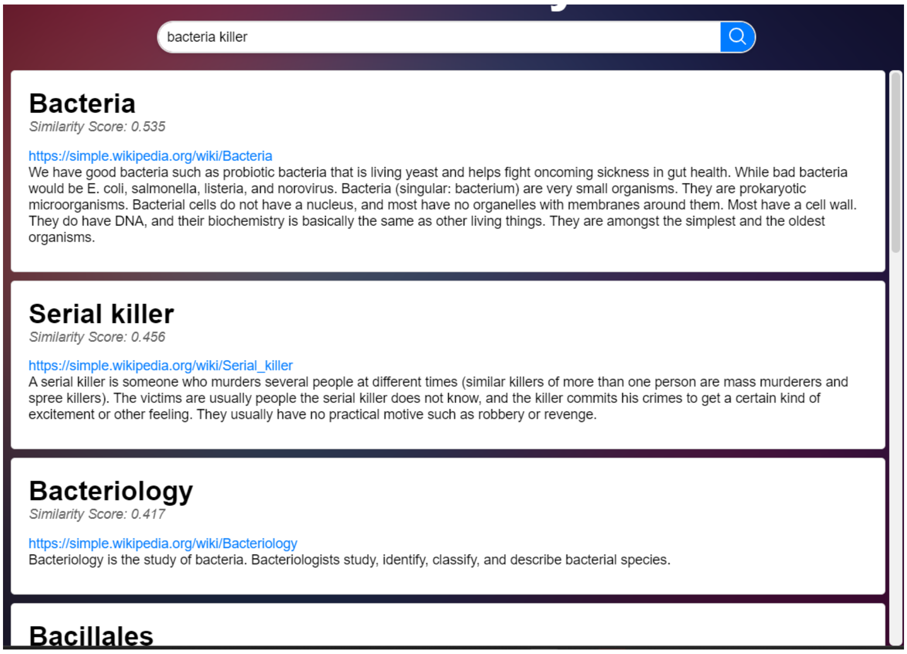
https://simple.wikipedia.org/wiki/FIA_Super_Licence

The FIA Super Licence is a qualification allowing the licence (license) holder to take part in Formula One Grands Prix as a driver. The licence is issued by the FIA upon request. There are several requirements to qualify for an FIA Super Licence. The driver requesting the licence must already have a Grade A competition licence. They must also meet the requirements of the FIA International Sporting Code, Appendix L. Appendix L lists other racing experience required. One of the following is required:

Fastest lap
Similarity Score: 0.262

Figure 3.4: SVD, k=500

3.2 bacteria killer



bacteria killer

Bacteria

Similarity Score: 0.535

https://simple.wikipedia.org/wiki/Bacteria

We have good bacteria such as probiotic bacteria that is living yeast and helps fight oncoming sickness in gut health. While bad bacteria would be E. coli, salmonella, listeria, and norovirus. Bacteria (singular: bacterium) are very small organisms. They are prokaryotic microorganisms. Bacterial cells do not have a nucleus, and most have no organelles with membranes around them. Most have a cell wall. They do have DNA, and their biochemistry is basically the same as other living things. They are amongst the simplest and the oldest organisms.

Serial killer

Similarity Score: 0.456

https://simple.wikipedia.org/wiki/Serial_killer

A serial killer is someone who murders several people at different times (similar killers of more than one person are mass murderers and spree killers). The victims are usually people the serial killer does not know, and the killer commits his crimes to get a certain kind of excitement or other feeling. They usually have no practical motive such as robbery or revenge.

Bacteriology

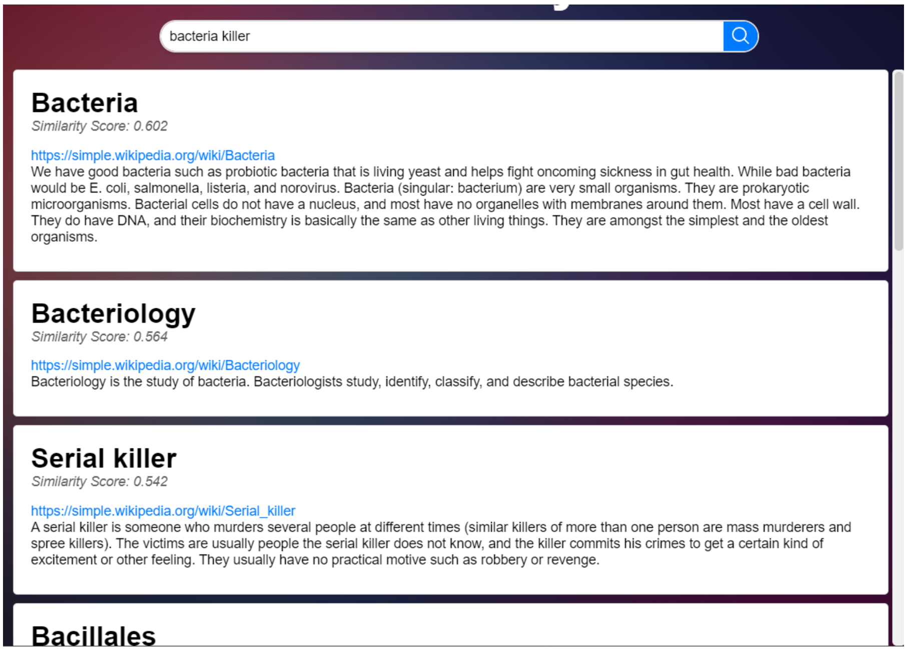
Similarity Score: 0.417

https://simple.wikipedia.org/wiki/Bacteriology

Bacteriology is the study of bacteria. Bacteriologists study, identify, classify, and describe bacterial species.

Bacillales

Figure 3.5: no SVD



bacteria killer

Bacteria

Similarity Score: 0.602

https://simple.wikipedia.org/wiki/Bacteria

We have good bacteria such as probiotic bacteria that is living yeast and helps fight oncoming sickness in gut health. While bad bacteria would be E. coli, salmonella, listeria, and norovirus. Bacteria (singular: bacterium) are very small organisms. They are prokaryotic microorganisms. Bacterial cells do not have a nucleus, and most have no organelles with membranes around them. Most have a cell wall. They do have DNA, and their biochemistry is basically the same as other living things. They are amongst the simplest and the oldest organisms.

Bacteriology

Similarity Score: 0.564

https://simple.wikipedia.org/wiki/Bacteriology

Bacteriology is the study of bacteria. Bacteriologists study, identify, classify, and describe bacterial species.

Serial killer

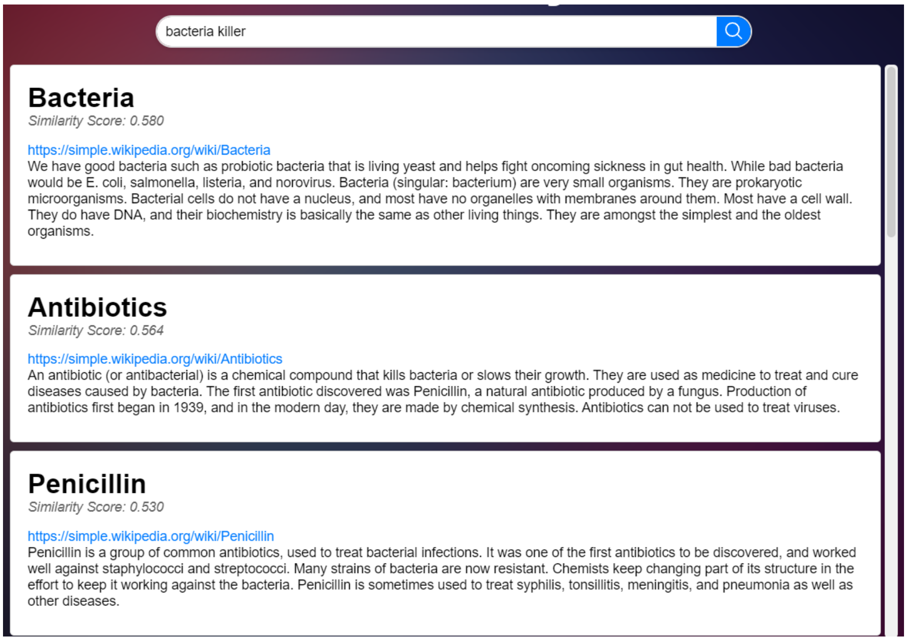
Similarity Score: 0.542

https://simple.wikipedia.org/wiki/Serial_killer

A serial killer is someone who murders several people at different times (similar killers of more than one person are mass murderers and spree killers). The victims are usually people the serial killer does not know, and the killer commits his crimes to get a certain kind of excitement or other feeling. They usually have no practical motive such as robbery or revenge.

Bacillales

Figure 3.6: SVD, k=5000



bacteria killer

Bacteria

Similarity Score: 0.580

https://simple.wikipedia.org/wiki/Bacteria

We have good bacteria such as probiotic bacteria that is living yeast and helps fight oncoming sickness in gut health. While bad bacteria would be E. coli, salmonella, listeria, and norovirus. Bacteria (singular: bacterium) are very small organisms. They are prokaryotic microorganisms. Bacterial cells do not have a nucleus, and most have no organelles with membranes around them. Most have a cell wall. They do have DNA, and their biochemistry is basically the same as other living things. They are amongst the simplest and the oldest organisms.

Antibiotics

Similarity Score: 0.564

https://simple.wikipedia.org/wiki/Antibiotics

An antibiotic (or antibacterial) is a chemical compound that kills bacteria or slows their growth. They are used as medicine to treat and cure diseases caused by bacteria. The first antibiotic discovered was Penicillin, a natural antibiotic produced by a fungus. Production of antibiotics first began in 1939, and in the modern day, they are made by chemical synthesis. Antibiotics can not be used to treat viruses.

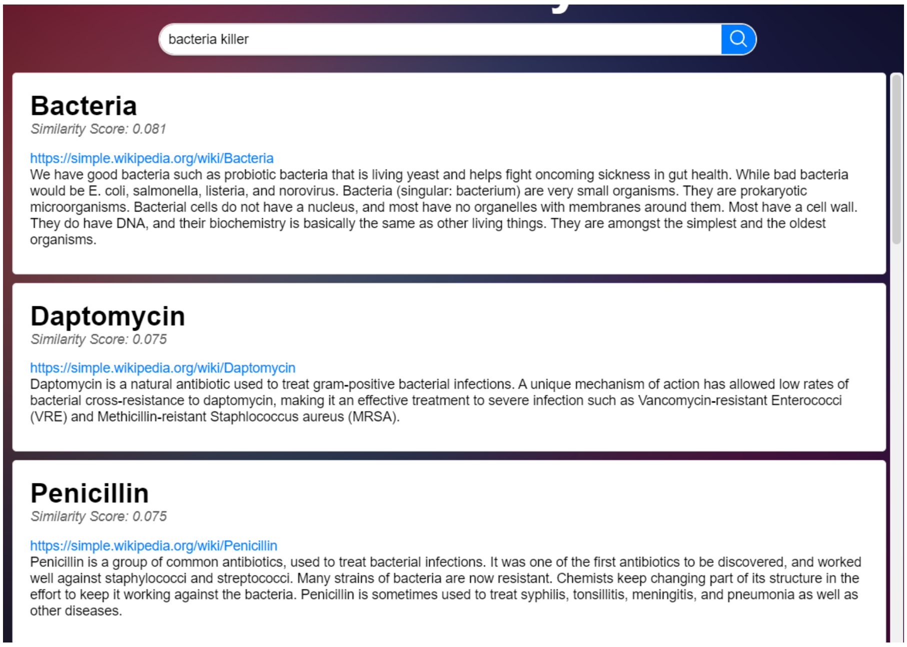
Penicillin

Similarity Score: 0.530

https://simple.wikipedia.org/wiki/Penicillin

Penicillin is a group of common antibiotics, used to treat bacterial infections. It was one of the first antibiotics to be discovered, and worked well against staphylococci and streptococci. Many strains of bacteria are now resistant. Chemists keep changing part of its structure in the effort to keep it working against the bacteria. Penicillin is sometimes used to treat syphilis, tonsillitis, meningitis, and pneumonia as well as other diseases.

Figure 3.7: SVD, k=1000



bacteria killer

Bacteria

Similarity Score: 0.081

https://simple.wikipedia.org/wiki/Bacteria

We have good bacteria such as probiotic bacteria that is living yeast and helps fight oncoming sickness in gut health. While bad bacteria would be E. coli, salmonella, listeria, and norovirus. Bacteria (singular: bacterium) are very small organisms. They are prokaryotic microorganisms. Bacterial cells do not have a nucleus, and most have no organelles with membranes around them. Most have a cell wall. They do have DNA, and their biochemistry is basically the same as other living things. They are amongst the simplest and the oldest organisms.

Daptomycin

Similarity Score: 0.075

https://simple.wikipedia.org/wiki/Daptomycin

Daptomycin is a natural antibiotic used to treat gram-positive bacterial infections. A unique mechanism of action has allowed low rates of bacterial cross-resistance to daptomycin, making it an effective treatment to severe infection such as Vancomycin-resistant Enterococci (VRE) and Methicillin-reistant Staphlococcus aureus (MRSA).

Penicillin

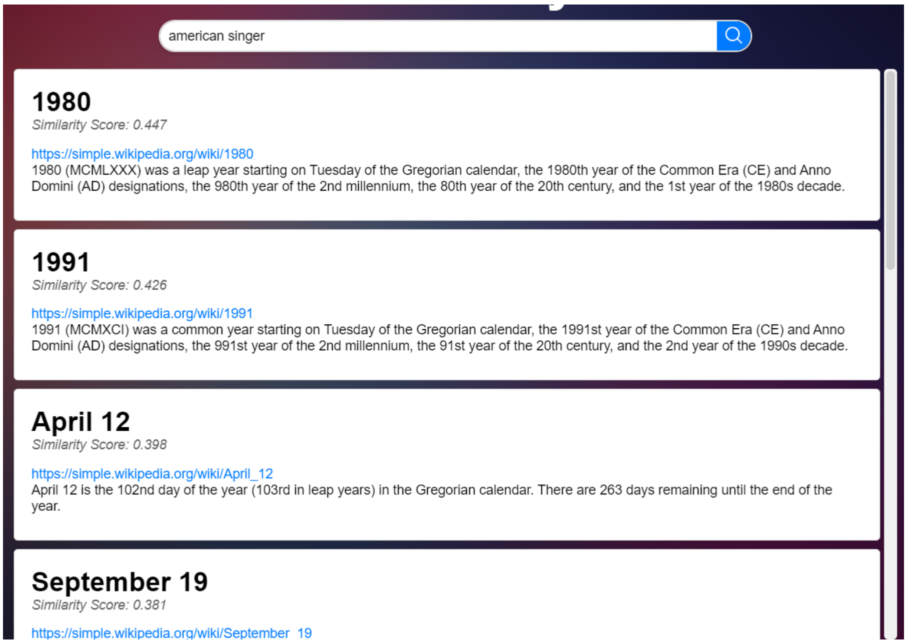
Similarity Score: 0.075

https://simple.wikipedia.org/wiki/Penicillin

Penicillin is a group of common antibiotics, used to treat bacterial infections. It was one of the first antibiotics to be discovered, and worked well against staphylococci and streptococci. Many strains of bacteria are now resistant. Chemists keep changing part of its structure in the effort to keep it working against the bacteria. Penicillin is sometimes used to treat syphilis, tonsillitis, meningitis, and pneumonia as well as other diseases.

Figure 3.8: SVD, k=500

3.3 american singer



american singer

1980

Similarity Score: 0.447

https://simple.wikipedia.org/wiki/1980

1980 (MCMLXXX) was a leap year starting on Tuesday of the Gregorian calendar, the 1980th year of the Common Era (CE) and Anno Domini (AD) designations, the 980th year of the 2nd millennium, the 80th year of the 20th century, and the 1st year of the 1980s decade.

1991

Similarity Score: 0.426

https://simple.wikipedia.org/wiki/1991

1991 (MCMXCI) was a common year starting on Tuesday of the Gregorian calendar, the 1991st year of the Common Era (CE) and Anno Domini (AD) designations, the 991st year of the 2nd millennium, the 91st year of the 20th century, and the 2nd year of the 1990s decade.

April 12

Similarity Score: 0.398

https://simple.wikipedia.org/wiki/April_12

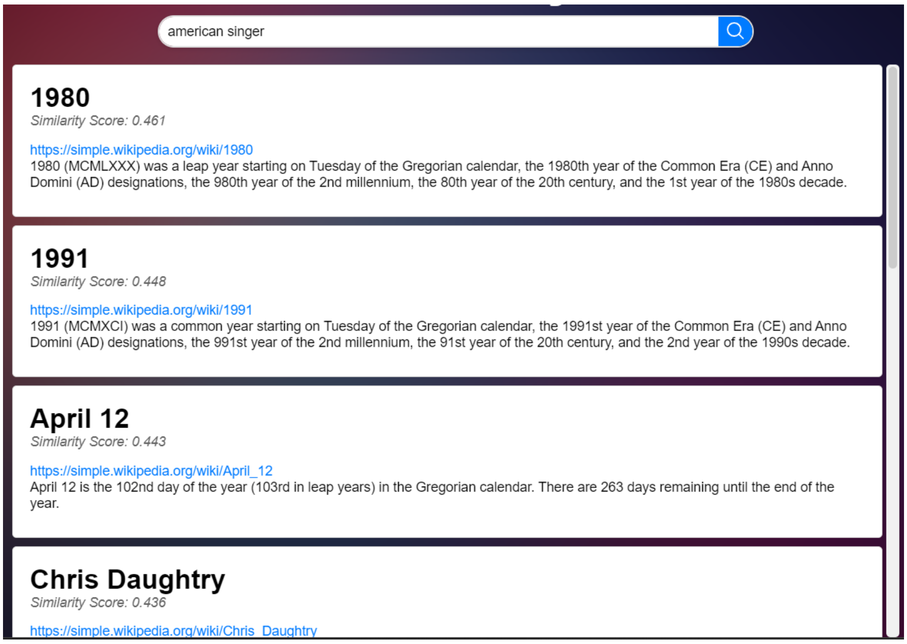
April 12 is the 102nd day of the year (103rd in leap years) in the Gregorian calendar. There are 263 days remaining until the end of the year.

September 19

Similarity Score: 0.381

https://simple.wikipedia.org/wiki/September_19

Figure 3.9: no SVD



american singer

1980

Similarity Score: 0.461

https://simple.wikipedia.org/wiki/1980

1980 (MCMLXXX) was a leap year starting on Tuesday of the Gregorian calendar, the 1980th year of the Common Era (CE) and Anno Domini (AD) designations, the 980th year of the 2nd millennium, the 80th year of the 20th century, and the 1st year of the 1980s decade.

1991

Similarity Score: 0.448

https://simple.wikipedia.org/wiki/1991

1991 (MCMXCI) was a common year starting on Tuesday of the Gregorian calendar, the 1991st year of the Common Era (CE) and Anno Domini (AD) designations, the 991st year of the 2nd millennium, the 91st year of the 20th century, and the 2nd year of the 1990s decade.

April 12

Similarity Score: 0.443

https://simple.wikipedia.org/wiki/April_12

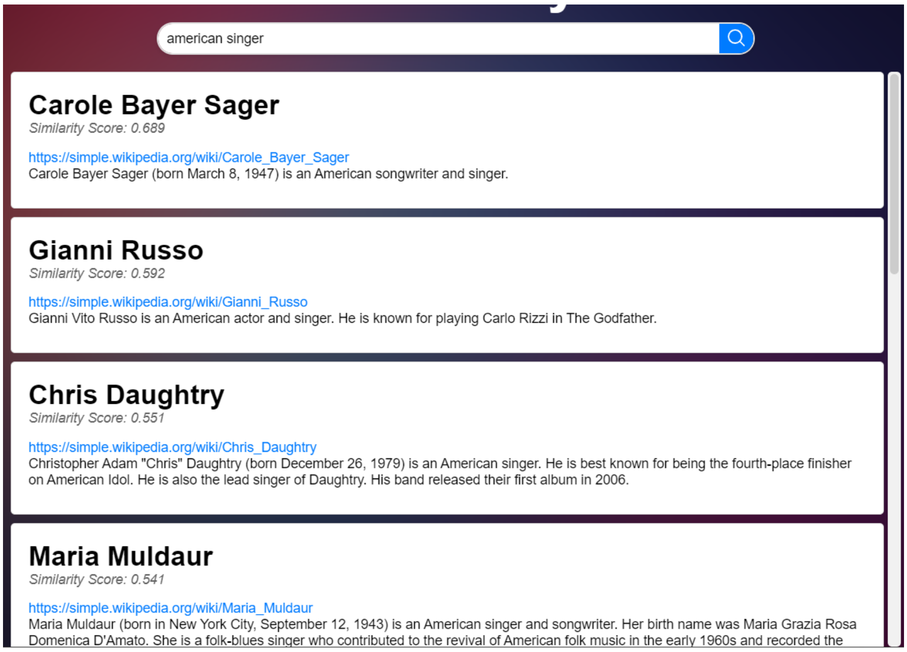
April 12 is the 102nd day of the year (103rd in leap years) in the Gregorian calendar. There are 263 days remaining until the end of the year.

Chris Daughtry

Similarity Score: 0.436

https://simple.wikipedia.org/wiki/Chris_Daughtry

Figure 3.10: SVD, k=5000



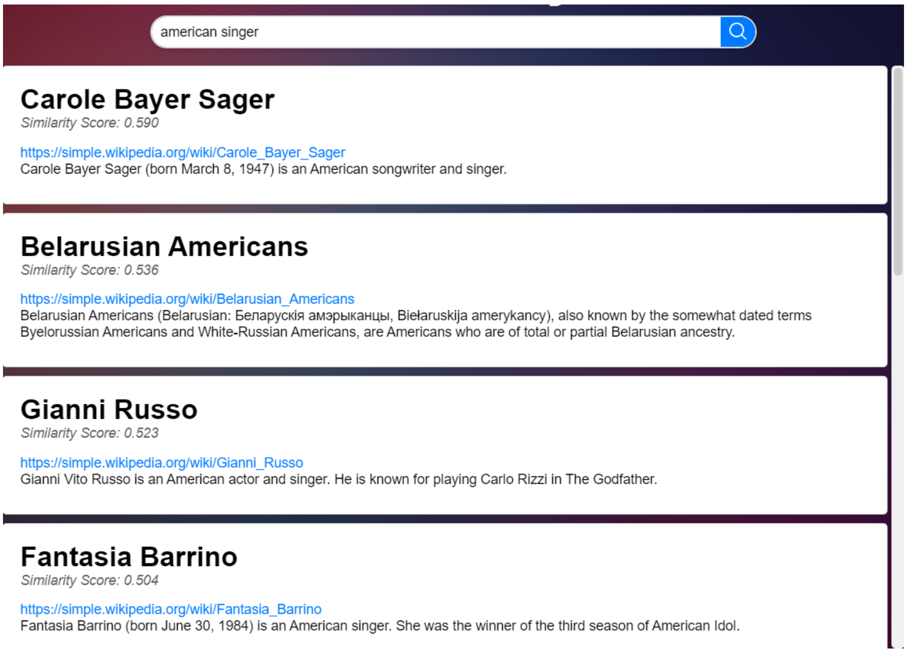
Carole Bayer Sager
Similarity Score: 0.689
https://simple.wikipedia.org/wiki/Carole_Bayer_Sager
Carole Bayer Sager (born March 8, 1947) is an American songwriter and singer.

Gianni Russo
Similarity Score: 0.592
https://simple.wikipedia.org/wiki/Gianni_Russo
Gianni Vito Russo is an American actor and singer. He is known for playing Carlo Rizzi in The Godfather.

Chris Daughtry
Similarity Score: 0.551
https://simple.wikipedia.org/wiki/Chris_Daughtry
Christopher Adam "Chris" Daughtry (born December 26, 1979) is an American singer. He is best known for being the fourth-place finisher on American Idol. He is also the lead singer of Daughtry. His band released their first album in 2006.

Maria Muldaur
Similarity Score: 0.541
https://simple.wikipedia.org/wiki/Maria_Muldaur
Maria Muldaur (born in New York City, September 12, 1943) is an American singer and songwriter. Her birth name was Maria Grazia Rosa Domenica D'Amato. She is a folk-blues singer who contributed to the revival of American folk music in the early 1960s and recorded the

Figure 3.11: SVD, k=1000



Carole Bayer Sager
Similarity Score: 0.590
https://simple.wikipedia.org/wiki/Carole_Bayer_Sager
Carole Bayer Sager (born March 8, 1947) is an American songwriter and singer.

Belarusian Americans
Similarity Score: 0.536
https://simple.wikipedia.org/wiki/Belarusian_Americans
Belarusian Americans (Belarusian: Беларускія амэрыканцы, Biełaruskija amerykancy), also known by the somewhat dated terms Byelorussian Americans and White-Russian Americans, are Americans who are of total or partial Belarusian ancestry.

Gianni Russo
Similarity Score: 0.523
https://simple.wikipedia.org/wiki/Gianni_Russo
Gianni Vito Russo is an American actor and singer. He is known for playing Carlo Rizzi in The Godfather.

Fantasia Barrino
Similarity Score: 0.504
https://simple.wikipedia.org/wiki/Fantasia_Barrino
Fantasia Barrino (born June 30, 1984) is an American singer. She was the winner of the third season of American Idol.

Figure 3.12: SVD, k=500

3.4 summary

Testing showed that when we use SVD for searching, especially with high values of k, the results are pretty similar to searches without SVD. Sometimes, though, SVD can understand the search better. But if we use low values of k, the results are often not very related to what we're searching for. In my opinion, the k value of 1000 performed the best.