Computational Methods - Search Engine

Report

Filip Zieliński

2024

1 Introduction

The aim of this task was to gather a large dataset of text files (over 1000) from webpages and implement a custom search engine. The search engine allows the user to input a query and returns the most similar pages from the collected data.

2 Theory Background

Here is a detailed explanation of each step of the algorithm and the rationale behind it.

2.1 Definition of Keywords

First, we need to define keywords (or key terms) - these are the important words that our search engine will focus on. The simplest approach is to define keywords as all the words used in the data set.

2.2 Term Frequency

Let j be an article in our data set and k a keyword in our dictionary. We define $D_{tf}(j,k) = \frac{counter(k,j)}{|j|}$, where $counter(k,j)$ is the number of occurrences of k in j, and $|j|$ is the number of words in article j.

2.3 Weights Vector

For every article j, we define its *weights vector* $d_j = (D_{tf}(j,k_1), D_{tf}(j,k_2), \ldots, D_{tf}(j,k_n))$, where $k_1, \ldots, k_n$ are all the keywords.

2.4 Term-by-Document Matrix

Taking each d_j as a column, we create the *term-by-document matrix* $A_{m \times n}$, where m is the number of keywords in the vocabulary and n is the number of articles gathered.

2.5 Document Frequency

Let k be a keyword. We define *document frequency* $DF(k)$ as the number of articles where k occurs at least once.

2.6 Inverse Document Frequency

Let N be the number of all documents (articles) and k a keyword.
We define *inverse document frequency* $IDF(k) = \log(\frac{N}{DF(k)})$. Since $DF(k)$ measures how common a keyword is throughout all texts, a high $IDF(k)$ means that the keyword is rather unique and rare. We use the logarithm function to smooth out the function, as we do not want the amplitude of IDF to be too high.

2.7 Term-by-Document Normalized Matrix

To improve the search engine's performance, instead of taking $A_{m \times n}$ as defined earlier, we multiply the row corresponding to keyword k by $IDF(k)$. This way, we "normalize"the whole matrix, reducing bias towards very common words.

2.8 Cosine Similarity

Let q_str be a string representing the query given as an input to the search engine. Let q be the weight vector of this query $q = d_{q\_str}$. This way, we have a simple representation of the query in the context of our keywords.
We define*cosine similarity* as:

$$\cos(\theta_j) = \frac{q^T d_j}{||q||||d_j||}$$

It is a measure of how similar document j is to the given query.
We can construct $|q^T A| = (|\cos(\theta_1)|, |\cos(\theta_2)|, \ldots, |\cos(\theta_n)|)$. It is a vector representing how well every article (document) corresponds to the query.
If we are interested in the best search results, we may simply sort this vector.

2.9 SVD Usage - Noise Reduction & Low Rank Approximation

To reduce noise and improve results, we may use Singular Value Decomposition (SVD) on matrix A with a set parameter of rank k. This way, we can also reduce used memory.
If we decide to use SVD, we simply calculate cosine similarity for our new matrix.

3 Implementation

In this section, I will present details about my implementation of the Search Engine. The entire implementation is in Python.

3.1 Data Set

To scrape data from websites, I've used the *Beautiful Soup* library. I've written (with some help from ChatGPT...) over 100 Google queries. For each query, I downloaded approximately 10 pages. I focused mainly on topics such as Mathematics, Mathematicians, History of Mathematics, Physics, Physicists, History of Physics, the Relationship between Mathematics and Physics, and the Philosophy of Mathematics.
The full list of queries can be found at the end of the report in Section 6.

3.2 Document Preprocessing

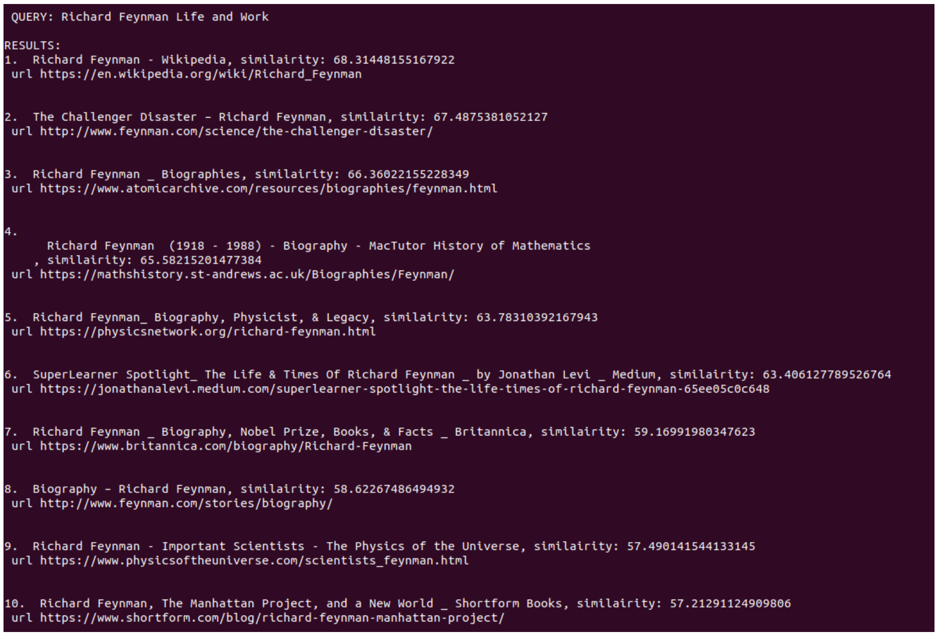
First, each document was parsed into separate words. Then, every non-alphabetical word was deleted. I also used the *nltk* Python library to stem words and remove common words such as conjunctions, because we do not want to count them as they occur very frequently regardless of context.

3.3 Creating Vocabulary, Calculating IDF, Calculating Search Matrix

As vocabulary I chose 5000 most common words in all texts. Later without any substantial changes to the theory shown, I used the given data set to calculate all crucial components for this algorithm. All data necessary for searching is saved in special files.

3.4 GUI

I do not do GUI. Ever. I have a reputation to maintain.
That's why my search engine works on a good old terminal!



```
 QUERY: Richard Feynman Life and Work

RESULTS:
1.  Richard Feynman - Wikipedia, similairity: 68.31448155167922
 url https://en.wikipedia.org/wiki/Richard_Feynman


2.  The Challenger Disaster – Richard Feynman, similairity: 67.4875381052127
 url http://www.feynman.com/science/the-challenger-disaster/


3.  Richard Feynman _ Biographies, similairity: 66.36022155228349
 url https://www.atomicarchive.com/resources/biographies/feynman.html


4.
      Richard Feynman  (1918 - 1988) - Biography - MacTutor History of Mathematics
    , similairity: 65.58215201477384
 url https://mathshistory.st-andrews.ac.uk/Biographies/Feynman/


5.  Richard Feynman_ Biography, Physicist, & Legacy, similairity: 63.78310392167943
 url https://physicsnetwork.org/richard-feynman.html


6.  SuperLearner Spotlight_ The Life & Times Of Richard Feynman _ by Jonathan Levi _ Medium, similairity: 63.406127789526764
 url https://jonathanalevi.medium.com/superlearner-spotlight-the-life-times-of-richard-feynman-65ee05c0c648


7.  Richard Feynman _ Biography, Nobel Prize, Books, & Facts _ Britannica, similairity: 59.16991980347623
 url https://www.britannica.com/biography/Richard-Feynman


8.  Biography – Richard Feynman, similairity: 58.62267486494932
 url http://www.feynman.com/stories/biography/


9.  Richard Feynman - Important Scientists - The Physics of the Universe, similairity: 57.490141544133145
 url https://www.physicsoftheuniverse.com/scientists_feynman.html


10.  Richard Feynman, The Manhattan Project, and a New World _ Shortform Books, similairity: 57.21291124909806
 url https://www.shortform.com/blog/richard-feynman-manhattan-project/
```

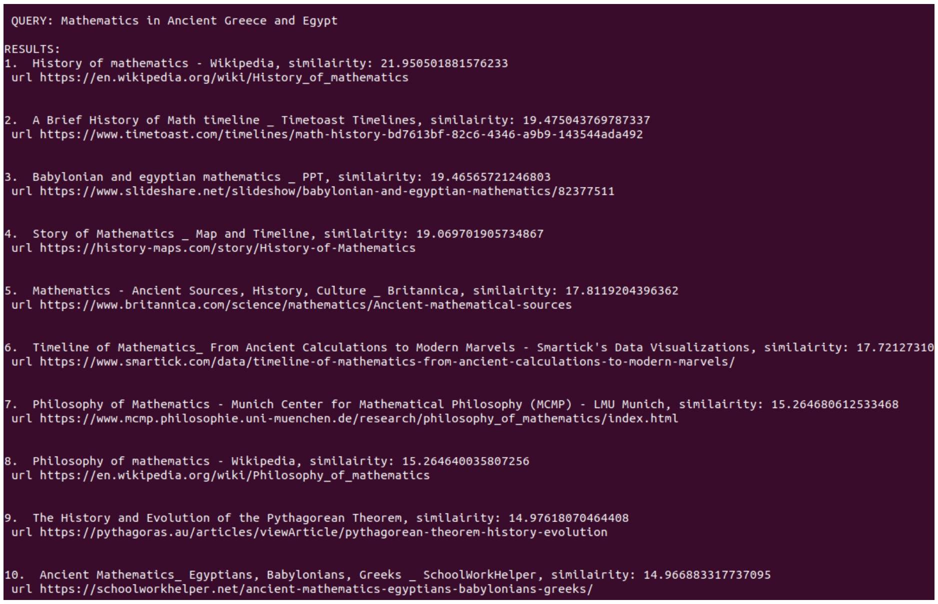
Rysunek 1: Example Query

4 Results

The results are not as satisfying as I had hoped. I believe 1050 files are not enough for a search engine. I also believe I made a mistake by downloading only 10 pages per query. Although the topics of the queries were related and similar, I think this led to a situation where we cannot ask something very specific because there is not enough data on any particular topic.

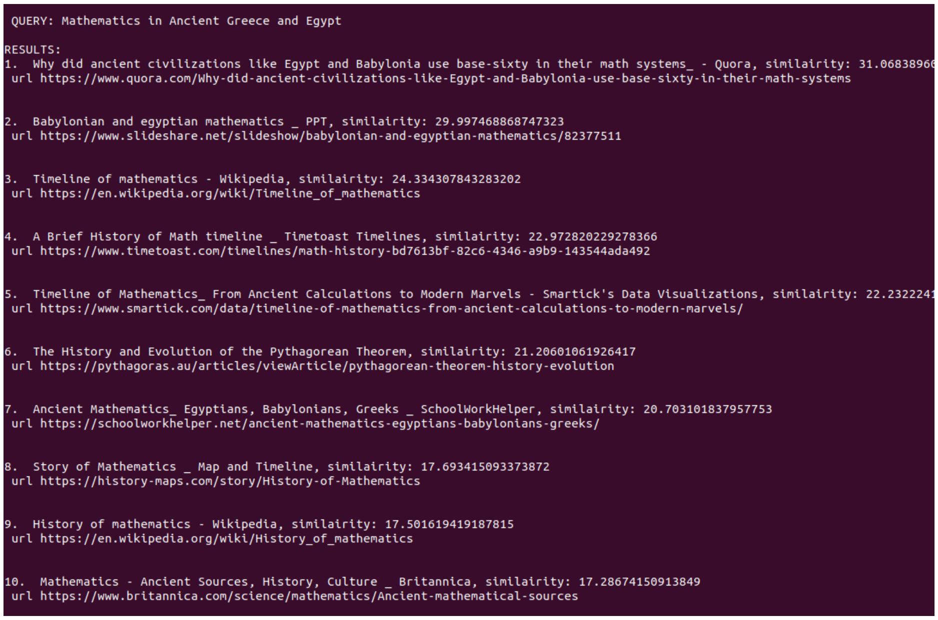
4.1 SVD vs No-SVD Approach

I used $k = 100$ as the set parameter.
From my observations, the results are mostly not very different whether I used SVD or not. Most changes are only cosmetic, with small similarity differences, and sometimes there is a slightly different order of results, but they remain the same. If I have to choose, I would say that non-SVD searches give better results, which is especially shown by Example 2.



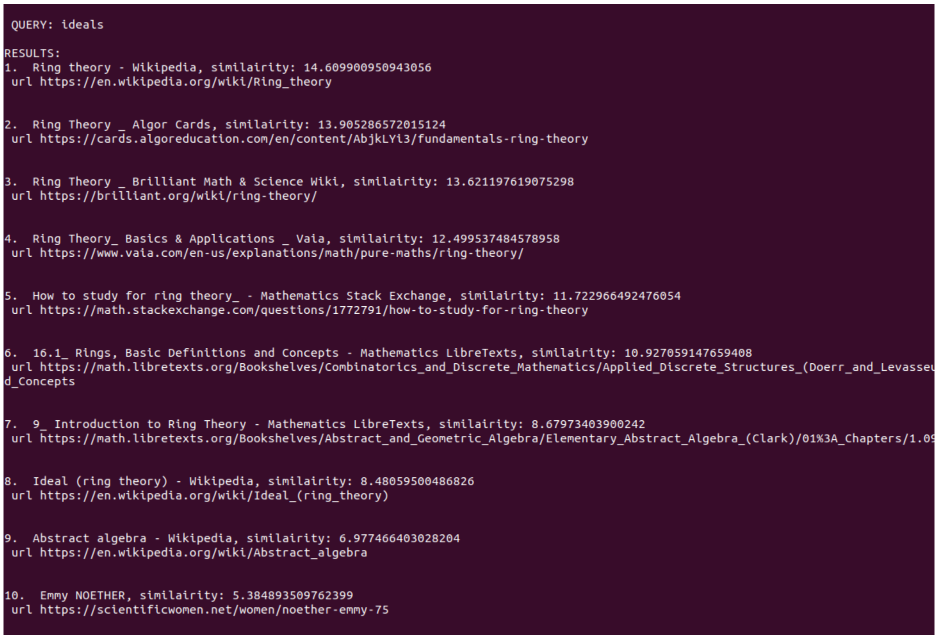
```
 QUERY: Mathematics in Ancient Greece and Egypt

RESULTS:
1.  History of mathematics - Wikipedia, similairity: 21.950501881576233
 url https://en.wikipedia.org/wiki/History_of_mathematics


2.  A Brief History of Math timeline _ Timetoast Timelines, similairity: 19.475043769787337
 url https://www.timetoast.com/timelines/math-history-bd7613bf-82c6-4346-a9b9-143544ada492


3.  Babylonian and egyptian mathematics _ PPT, similairity: 19.46565721246803
 url https://www.slideshare.net/slideshow/babylonian-and-egyptian-mathematics/82377511


4.  Story of Mathematics _ Map and Timeline, similairity: 19.069701905734867
 url https://history-maps.com/story/History-of-Mathematics


5.  Mathematics - Ancient Sources, History, Culture _ Britannica, similairity: 17.8119204396362
 url https://www.britannica.com/science/mathematics/Ancient-mathematical-sources


6.  Timeline of Mathematics_ From Ancient Calculations to Modern Marvels - Smartick's Data Visualizations, similairity: 17.72127310
 url https://www.smartick.com/data/timeline-of-mathematics-from-ancient-calculations-to-modern-marvels/


7.  Philosophy of Mathematics - Munich Center for Mathematical Philosophy (MCMP) - LMU Munich, similairity: 15.264680612533468
 url https://www.mcmp.philosophie.uni-muenchen.de/research/philosophy_of_mathematics/index.html


8.  Philosophy of mathematics - Wikipedia, similairity: 15.264640035807256
 url https://en.wikipedia.org/wiki/Philosophy_of_mathematics


9.  The History and Evolution of the Pythagorean Theorem, similairity: 14.97618070464408
 url https://pythagoras.au/articles/viewArticle/pythagorean-theorem-history-evolution


10.  Ancient Mathematics_ Egyptians, Babylonians, Greeks _ SchoolWorkHelper, similairity: 14.966883317737095
 url https://schoolworkhelper.net/ancient-mathematics-egyptians-babylonians-greeks/
```

Rysunek 2: Example 1: SVD usage search



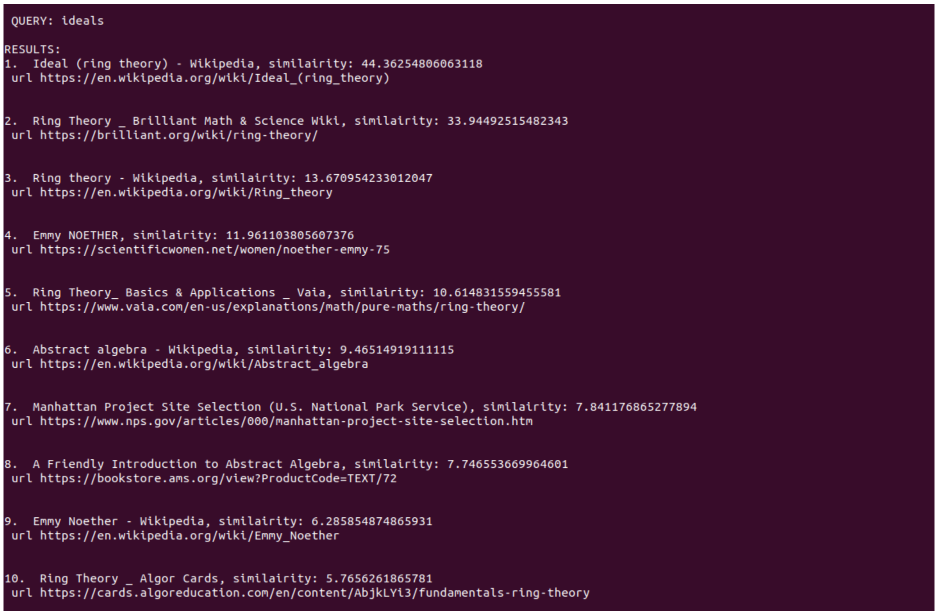
```
 QUERY: Mathematics in Ancient Greece and Egypt

RESULTS:
1.  Why did ancient civilizations like Egypt and Babylonia use base-sixty in their math systems_ - Quora, similairity: 31.06838966
 url https://www.quora.com/Why-did-ancient-civilizations-like-Egypt-and-Babylonia-use-base-sixty-in-their-math-systems


2.  Babylonian and egyptian mathematics _ PPT, similairity: 29.997468868747323
 url https://www.slideshare.net/slideshow/babylonian-and-egyptian-mathematics/82377511


3.  Timeline of mathematics - Wikipedia, similairity: 24.334307843283202
 url https://en.wikipedia.org/wiki/Timeline_of_mathematics


4.  A Brief History of Math timeline _ Timetoast Timelines, similairity: 22.972820229278366
 url https://www.timetoast.com/timelines/math-history-bd7613bf-82c6-4346-a9b9-143544ada492


5.  Timeline of Mathematics_ From Ancient Calculations to Modern Marvels - Smartick's Data Visualizations, similairity: 22.2322241
 url https://www.smartick.com/data/timeline-of-mathematics-from-ancient-calculations-to-modern-marvels/


6.  The History and Evolution of the Pythagorean Theorem, similairity: 21.20601061926417
 url https://pythagoras.au/articles/viewArticle/pythagorean-theorem-history-evolution


7.  Ancient Mathematics_ Egyptians, Babylonians, Greeks _ SchoolWorkHelper, similairity: 20.703101837957753
 url https://schoolworkhelper.net/ancient-mathematics-egyptians-babylonians-greeks/


8.  Story of Mathematics _ Map and Timeline, similairity: 17.693415093373872
 url https://history-maps.com/story/History-of-Mathematics


9.  History of mathematics - Wikipedia, similairity: 17.501619419187815
 url https://en.wikipedia.org/wiki/History_of_mathematics


10.  Mathematics - Ancient Sources, History, Culture _ Britannica, similairity: 17.28674150913849
 url https://www.britannica.com/science/mathematics/Ancient-mathematical-sources
```

Rysunek 3: Example 1: Search without SVD



```
 QUERY: ideals

RESULTS:
1.  Ring theory - Wikipedia, similairity: 14.609900950943056
 url https://en.wikipedia.org/wiki/Ring_theory


2.  Ring Theory _ Algor Cards, similairity: 13.905286572015124
 url https://cards.algoreducation.com/en/content/AbjkLYi3/fundamentals-ring-theory


3.  Ring Theory _ Brilliant Math & Science Wiki, similairity: 13.621197619075298
 url https://brilliant.org/wiki/ring-theory/


4.  Ring Theory_ Basics & Applications _ Vaia, similairity: 12.499537484578958
 url https://www.vaia.com/en-us/explanations/math/pure-maths/ring-theory/


5.  How to study for ring theory_ - Mathematics Stack Exchange, similairity: 11.722966492476054
 url https://math.stackexchange.com/questions/1772791/how-to-study-for-ring-theory


6.  16.1_ Rings, Basic Definitions and Concepts - Mathematics LibreTexts, similairity: 10.927059147659408
 url https://math.libretexts.org/Bookshelves/Combinatorics_and_Discrete_Mathematics/Applied_Discrete_Structures_(Doerr_and_Levasseu
d_Concepts


7.  9_ Introduction to Ring Theory - Mathematics LibreTexts, similairity: 8.67973403900242
 url https://math.libretexts.org/Bookshelves/Abstract_and_Geometric_Algebra/Elementary_Abstract_Algebra_(Clark)/01%3A_Chapters/1.09


8.  Ideal (ring theory) - Wikipedia, similairity: 8.48059500486826
 url https://en.wikipedia.org/wiki/Ideal_(ring_theory)


9.  Abstract algebra - Wikipedia, similairity: 6.977466403028204
 url https://en.wikipedia.org/wiki/Abstract_algebra


10.  Emmy NOETHER, similairity: 5.384893509762399
 url https://scientificwomen.net/women/noether-emmy-75
```

Rysunek 4: Example 2: SVD usage search



```
 QUERY: ideals

RESULTS:
1.  Ideal (ring theory) - Wikipedia, similairity: 44.36254806063118
 url https://en.wikipedia.org/wiki/Ideal_(ring_theory)


2.  Ring Theory _ Brilliant Math & Science Wiki, similairity: 33.94492515482343
 url https://brilliant.org/wiki/ring-theory/


3.  Ring theory - Wikipedia, similairity: 13.670954233012047
 url https://en.wikipedia.org/wiki/Ring_theory


4.  Emmy NOETHER, similairity: 11.961103805607376
 url https://scientificwomen.net/women/noether-emmy-75


5.  Ring Theory_ Basics & Applications _ Vaia, similairity: 10.614831559455581
 url https://www.vaia.com/en-us/explanations/math/pure-maths/ring-theory/


6.  Abstract algebra - Wikipedia, similairity: 9.46514919111115
 url https://en.wikipedia.org/wiki/Abstract_algebra


7.  Manhattan Project Site Selection (U.S. National Park Service), similairity: 7.841176865277894
 url https://www.nps.gov/articles/000/manhattan-project-site-selection.htm


8.  A Friendly Introduction to Abstract Algebra, similairity: 7.746553669964601
 url https://bookstore.ams.org/view?ProductCode=TEXT/72


9.  Emmy Noether - Wikipedia, similairity: 6.285854874865931
 url https://en.wikipedia.org/wiki/Emmy_Noether


10.  Ring Theory _ Algor Cards, similairity: 5.7656261865781
 url https://cards.algoreducation.com/en/content/AbjkLYi3/fundamentals-ring-theory
```

Rysunek 5: Example 2: Search without SVD

As you can see, for the query "ideals"in the SVD search, the Wikipedia page dedicated to ideals comes 8th, while in the non-SVD search, it is the first one, which is expected. That indicates that SVD is not very good for these results.

5 Conclusions

Cosine similarity is not the best measure in the context of search engines, but it may work with a large enough data set and specifically chosen topics.

6 List of Queries

- Mathematical concepts in quantum physics
- Applications of statistics in everyday life
- Mathematics behind computer algorithms
- John von Neumann and game theory
- The role of mathematics in economic models
- Euclid's elements and their influence on geometry
- Mathematics in ancient civilizations: Egypt and Babylon
- Famous female mathematicians throughout history
- Biography of Pythagoras and his theorem
- Applications of differential equations in physics
- The significance of prime numbers in cryptography
- Mathematical modeling in epidemiology
- Contributions of Hypatia of Alexandria to mathematics
- The role of topology in modern mathematics
- The history and importance of the number pi.

- Galois theory and its impact on abstract algebra
- Applications of linear algebra in computer graphics
- The life and work of Sophie Germain
- Famous unsolved mathematical problems
- The mathematics of fractals and chaos theory
- Contributions of Rene Descartes to mathematics
- The origins of the concept of infinity in mathematics
- Applications of graph theory in network analysis
- The mathematics of music: harmony and frequencies
- Mathematical challenges in artificial intelligence
- The life and work of Srinivasa Ramanujan
- The role of mathematics in climate modeling
- Fundamental concepts in set theory
- Properties of prime numbers
- Theorems of number theory
- Axiomatic systems in mathematics
- Introduction to abstract algebra
- Exploring Euclidean geometry
- Mathematical logic and proof techniques
- Principles of mathematical analysis
- Theory of equations and polynomials
- Topics in discrete mathematics
- Foundations of topology
- Exploring differential geometry
- Principles of combinatorics
- Introduction to group theory
- Understanding ring theory
- Concepts in field theory
- Exploring linear algebra
- Theory of vector spaces
- Properties of matrices and determinants
- Topics in functional analysis
- Theory of differential equations
- Principles of calculus of variations
- Concepts in probability theory

- Principles of stochastic processes
- Exploring graph theory
- Principles of algebraic geometry
- Topics in category theory
- Understanding homological algebra
- Principles of algebraic topology
- Concepts in differential topology
- Introduction to differential forms
- Topics in mathematical philosophy
- Topics in mathematical reasoning and logic
- Einstein's theory of relativity explained
- Famous contributions of Richard Feynman
- The life and work of Marie Curie
- Significant discoveries by Stephen Hawking
- Contributions of Niels Bohr to quantum mechanics
- The legacy of Isaac Newton in physics
- Major theories by James Clerk Maxwell
- The scientific achievements of Galileo Galilei
- Michael Faraday's discoveries in electromagnetism
- The influence of J. Robert Oppenheimer on physics
- Pioneering work of Lise Meitner in radioactivity
- Theoretical advancements by Paul Dirac
- Werner Heisenberg and the uncertainty principle
- The discoveries of Hans Bethe in nuclear physics
- Richard Feynman contributions to quantum mechanics
- Richard Feynman personal life and biography
- Richard Feynman and the Manhattan Project
- Manhattan Project history
- Oppenheimer's role in the Manhattan Project
- Scientists involved in the Manhattan Project
- Los Alamos during the Manhattan Project
- Robert Oppenheimer biography
- Impact of the Manhattan Project on World War II
- Key figures in the Manhattan Project
- Development of atomic bomb Manhattan Project

- Manhattan Project timeline
- Manhattan Project research facilities
- Contributions of Enrico Fermi to the Manhattan Project
- Leo Szilard and the Manhattan Project
- Manhattan Project engineering challenges
- Hiroshima and Nagasaki bombings Manhattan Project
- Manhattan Project legacy
- Security measures in the Manhattan Project
- Manhattan Project and the Cold War
- Manhattan Project physicists
- Ethical implications of the Manhattan Project
- Mathematical methods in theoretical physics
- Role of calculus in classical mechanics
- Linear algebra applications in quantum mechanics
- Differential equations in fluid dynamics
- Topology in general relativity
- Fourier analysis in signal processing
- Complex numbers in electromagnetism
- Mathematics behind string theory
- Statistical mechanics and probability theory
- Mathematical foundations of quantum field theory
- Symmetry and conservation laws in physics