

SearchPeople

实现功能:

- 1. 爬取wikipedia中的10041个人物信息，包括了计算机科学家，数学家，物理学家和化学家分类
- 2. 抽取页面上infobox中的内容，存储在文件中
- 3. 在每次启动时建立关键字到人的倒排列表
- 4. 使用Django渲染模板，对浏览器端提交的搜索请求进行响应，返回包含关键字链接的结果并高亮关键字
- 5. 实现了搜索结果的分页显示
- 6. 可以对人物按照姓名，国籍，研究领域等字段针对性查询

运行效果:

- 1. 搜索框

关键词

Name

Born

Nationality

Fields

- 2. 关键词搜索与高亮(第一页结果)

关键词 knuth	<input type="button" value="搜索"/>
Name	<input type="text"/>
Born	<input type="text"/>
Nationality	<input type="text"/>
Fields	<input type="text"/>
<input type="button" value="高级搜索"/>	
Noam Nisan	Gödel Prize (2012) <i>Knuth</i> Prize (2016)
Andrew Chi-Chih Yao 姚期智	Pólya Prize (SIAM) (1987) <i>Knuth</i> Prize (1996) Turing Award (2000)
Richard Lipton	<i>Knuth</i> Prize (2014)
Christos Papadimitriou	Von Neumann Medal (2016) EATCS Award (2015) Gödel Prize (2012) IEEE Computer Society Charles Babbage Award (2004) <i>Knuth</i> Prize (2002)
Vaughan Pratt	<i>Knuth</i> -Morris-Pratt algorithm Pratt certificate Pratt parser Donald <i>Knuth</i>
Leonid Anatolevich Levin	<i>Knuth</i> Prize (2012)
Mihalis Yannakakis	<i>Knuth</i> Prize (2005)
Leslie Valiant	ACM Turing Award (2010) EATCS Award (2008) <i>Knuth</i> Prize (1997) AAAI Fellow (1992) FRS (1991) Nevanlinna Prize (1986)
Miklos Ajtai	<i>Knuth</i> Prize (2003)[1]
László Babai	Gödel Prize (1993) <i>Knuth</i> Prize (2015) Dijkstra Prize (2016)
László Lovász	Kyoto Prize (2010) Hungary's Széchenyi Grand Prize (2008) Bolyai Prize (2007) John von Neumann Theory Prize (2006) Gödel Prize (2001) <i>Knuth</i> Prize (1999) Wolf Prize (1999) Fulkerson Prize (1982, 2012) Best Information Theory Paper Award (IEEE) (1981) Pólya Prize (SIAM) (1979)
Ravindran Kannan ரவீந்திரன் கண்ணன்	<i>Knuth</i> Prize (2011) Fulkerson Prize (1991)
Donald Knuth	Donald <i>Knuth</i> Donald Ervin <i>Knuth</i> (1938-01-10) January 10, 1938 (age 79) Milwaukee, Wisconsin, U.S. The Art of Computer Programming TeX, METAFONT <i>Knuth</i> -Morris-Pratt algorithm <i>Knuth</i> -Bendix completion algorithm MMIX Robinson-Schensted- <i>Knuth</i> correspondence
David S. Johnson	ACM Fellow (1995) <i>Knuth</i> Prize (2010)
Arthur Lee Samuel	Samuel Checkers-playing Program Alpha-beta pruning (an early implementation) Pioneer in Machine Learning [1] TeX project (with Donald <i>Knuth</i>)

第1页 [下一页](#)

- 3. 关键词搜索与高亮(第二页结果)

关键词

搜索

Name	
Born	
Nationality	
Fields	
高级搜索	

Jeffrey D. Ullman	Fellow of the Association for Computing Machinery (1994) ACM SIGMOD Contributions Award (1996) ACM SIGMOD Best Paper Award (1996) Karl V. Karlstrom outstanding educator award (1998) <i>Knuth Prize</i> (2000) ACM SIGMOD Edgar F. Codd Innovations Award (2006) ACM SIGMOD Test of Time Award (2006) IEEE John von Neumann Medal (2010)
Robert Sedgewick	Donald <i>Knuth</i>
Marshall Hall, Jr.	Robert Calderbank Donald <i>Knuth</i> Robert McEliece E. T. Parker

上一页 第2页

4. 根据不同字段针对性查找

关键词

搜索


Name	dijkstra
Born	
Nationality	
Fields	computer
高级搜索	

Edsger Wybe Dijkstra	<i>Dijkstra's</i> algorithm (single-source shortest path problem) DJP algorithm (minimum spanning tree problem) First implementation of ALGOL 60 (<i>Dijkstra-Zonneveld</i> ALGOL 60 compiler for the Electrologica X1) Structured analysis Structured programming Semaphore Layered approach to operating system design THE multiprogramming system Concept of levels of abstraction[1][2] Concept of layered structure in software architecture (layered architecture) Concept of cooperating sequential processes[3] Concept of program families[4] Multithreaded programming Concurrent programming Concurrent algorithms Principles of distributed computing Distributed algorithms Synchronization primitive Mutual exclusion Critical section Generalization of Dekker's algorithm Tri-color marking algorithm Call stack Fault-tolerant systems Self-stabilizing distributed systems Resource starvation Deadly embrace Deadlock prevention algorithms Shunting-yard algorithm Banker's algorithm Dining philosophers problem Sleeping barber problem Producer–consumer problem (bounded buffer problem) Dutch national flag problem Predicate transformer semantics Guarded Command Language Weakest precondition calculus Unbounded nondeterminism <i>Dijkstra</i> -Scholten algorithm Smoothsort Separation of concerns Program verification Program derivation Software crisis[5] Software architecture[6] Turing Award (1972) ACM Fellow (1994) <i>Dijkstra Prize</i> (2002) Computing science Theoretical <i>computer science</i> Communication with an Automatic <i>Computer</i> (1959)
----------------------	---

第1页

5. 点击搜索结果链接后的详细信息展示

- Knuth 大师



name	Donald Knuth
Born	Donald Ervin Knuth (1938-01-10) January 10, 1938 (age 79) Milwaukee, Wisconsin, U.S.
Nationality	American
Other names	simplified Chinese: 高德纳; traditional Chinese: 高德納; pinyin: Gāo dé nà
Alma mater	Case Institute of Technology (B.S., M.S.) California Institute of Technology (Ph.D.)
Known for	The Art of Computer Programming TeX, METAFONT Knuth–Morris–Pratt algorithm Knuth–Bendix completion algorithm MMIX Robinson–Schensted–Knuth correspondence
Awards	Grace Murray Hopper Award (1971) Turing Award (1974) National Medal of Science (1979) John von Neumann Medal (1995) Harvey Prize (1995) Kyoto Prize (1996) Computer History Museum Fellow (1998)[1] ForMemRS (2003)[2] Faraday Medal (2011) BBVA Foundation Frontiers of Knowledge Award (2010) Turing Lecture (2011)
Website	cs.stanford.edu/~uno
Fields	Mathematics Computer science
Institutions	Stanford University
Thesis	Finite Semifields and Projective Planes (1963)
Doctoral advisor	Marshall Hall, Jr.[3]
Doctoral students	Leonidas J. Guibas Michael Fredman Scott Kim Vaughan Pratt Robert Sedgewick Jeffrey Vitter Andrei Broder[3]

- Albert Einstein



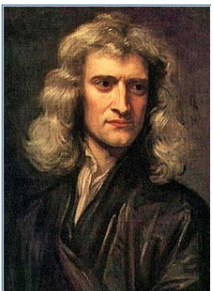
name	Albert Einstein
Pronunciation	/ˈaɪnʃtaɪn/^{[*]} German: [ˈalbɛʁt ˈaɪnʃtaɪn] (listen)
Born	(1879-03-14) 14 March 1879 Ulm, Kingdom of Württemberg, German Empire
Died	18 April 1955 (1955-04-18) (aged 76) Princeton, New Jersey, U.S.
Residence	Germany, Italy, Switzerland, Austria (present-day Czech Republic), Belgium, United States
Citizenship	Subject of the Kingdom of Württemberg during the German Empire (1879–1896)[note 1] Stateless (1896–1901) Citizen of Switzerland (1901–1955) Austrian subject of the Austro-Hungarian Empire (1911–1912) Subject of the Kingdom of Prussia during the German Empire (1914–1918)[note 1] German citizen of the Free State of Prussia (Weimar Republic, 1918–1933) Citizen of the United States (1940–1955)
Education	Swiss Federal Polytechnic (1896–1900; B.A., 1900) University of Zurich (Ph.D., 1905)
Known for	General relativity Special relativity Photoelectric effect E=mc² (Mass–energy equivalence) E=hf (Planck–Einstein relation) Theory of Brownian motion Einstein field equations Bose–Einstein statistics Bose–Einstein condensate Gravitational wave Cosmological constant Unified field theory EPR paradox List of other concepts
Spouse(s)	Mileva Marić (m. 1903; div. 1919) Elsa Löwenthal (m. 1919; d. 1936)[2][3]
Children	"Lieserl" Einstein Hans Albert Einstein Eduard "Tete" Einstein
Awards	Barnard Medal (1920) Nobel Prize in Physics (1921) Matteucci Medal (1921) ForMemRS (1921)[4] Copley Medal (1925)[4] Gold Medal of the Royal Astronomical Society (1926) Max Planck Medal (1929) Time Person of the Century (1999)
Fields	Physics, philosophy
Institutions	Swiss Patent Office (Bern) (1902–1909) University of Bern (1908–1909) University of Zurich (1909–1911) Charles University in Prague (1911–1912) ETH Zurich (1912–1914) Prussian Academy of Sciences (1914–1933) Humboldt University of Berlin (1914–1933) Kaiser Wilhelm Institute (director, 1917–1933) German Physical Society (president, 1916–1918) Leiden University (visits, 1920) Institute for Advanced Study (1933–1955) Caltech (visits, 1931–1933)
Thesis	Eine neue Bestimmung der Moleküldimensionen (A New Determination of Molecular Dimensions) (1905)
Doctoral advisor	Alfred Kleiner
Other academic advisors	Heinrich Friedrich Weber
Influenced	Ernst G. Straus Nathan Rosen Leó Szilárd

- 成果多到页面显示不下的 Euler



name	Leonhard Euler
Born	(1707-04-15)15 April 1707 Basel, Switzerland
Died	18 September 1783 (1783-09-18) (aged 76) [OS: 7 September 1783] Saint Petersburg, Russian Empire
Residence	Kingdom of Prussia Russian Empire Switzerland
Alma mater	University of Basel
Known for	See full list
Fields	Mathematics and physics
Institutions	Imperial Russian Academy of Sciences Berlin Academy
Thesis	Dissertatio physica de sono ("Physical dissertation on sound") (1726)
Doctoral advisor	Johann Bernoulli
Doctoral students	Nicolas Fuss Johann Hennert Stepan Rumovsky
Other notable students	Joseph-Louis Lagrange

- Isaac Newton



name	Sir Isaac Newton
Born	25 December 1642 [NS: (1643-01-04)4 January 1643][1] Woolsthorpe, Lincolnshire, England
Died	20 March 1726/7 (aged 84) [OS: (1726-03-20)20 March 1726 NS: (1727-03-31)31 March 1727][1] Kensington, Middlesex, England
Resting place	Westminster Abbey
Nationality	English
Alma mater	Trinity College, Cambridge
Known for	Newtonian mechanics Universal gravitation Calculus Newton's laws of motion Optics Binomial series Principia Newton's method
Awards	FRS (1672)[2] Knight Bachelor (1705)
Fields	Physics Natural philosophy Alchemy Theology Mathematics Astronomy Economics
Institutions	University of Cambridge Royal Society Royal Mint
Academic advisors	Isaac Barrow[3] Benjamin Pulleyn[4][5]
Notable students	Roger Cotes William Whiston

模块划分：

爬虫(./crawler文件夹下)：

- 这个爬虫是此次大作业里让我花费了最多时间的一个部分。爬取一万人，把每个分类下面的人物悉数爬取到，做到尽可能少地遗漏确实感觉不容易。
- 反反爬手段：使用chrome浏览器请求头的user-agent信息，两个网页间隔1~5s爬取。
- 防断网，防反爬，防崩溃手段：每爬取一定数量的网站自动把已经爬取的网址，将要爬取的网址，已经爬取的人物数量写进log中，下次启动的时候从记录恢复

倒排列表(./InvertedList文件夹下)：

- 通过正则表达式将句子替换标点，按空格分词找到关键词，建立关键词到文档的对应列表

网站后台(`./SearchPeople`文件夹下):

- 基本就是写一堆处理请求，渲染页面并返回的函数，照着网上的教程一步步写就好

网页模板(`./templates`文件夹下):

- 打开百度，搜狗，google的主页各种借鉴

大作业感想与心路历程

1. python写起来好舒服啊，很多情况下完全不用写一堆烦人的for(.....){}
2. 用户界面好难啊，随便凑合一个吧
3. 小学期终于结束了，完结撒花！