

CS 6675: Programming Assignment 1

Abhishek Sen, GTID: 902898709

January 30, 2016

WRITE A WEB CRAWLER OF YOUR OWN

DESIGN OF THE CRAWLER

MOTIVATION

I love technology and I love to read. 'The Google Story' by David A Vise is a great book that recounts Google's journey through the years, its inception and rapid rise. I had enjoyed reading the book a lot and was keen on finding similar books to read. I wrote a simple web crawler of my own using Python to do just that.

DESIGN

Goodreads is a social cataloging website for books with the tagline 'Meet your next favorite book'. The website has a very comprehensive catalog of books where readers can rate and review books. An interesting feature of the website is the 'Readers also enjoyed' section which I think is an attempt at recommending similar books using collaborative filtering. However, the section shows only five to ten similar books at the most on a page. I would rather have a huge list of books similar to my favorite book. My spider accepts a seed url, gets the source HTML and parses that to grab the 'Readers also enjoyed' section. I achieve this using BeautifulSoup which is a very handy Python HTML parsing library. Then the spider goes on to perform an aggressive Breadth First Search(BFS) by selecting three similar books from the 'Readers also enjoyed' section and expands the Crawl-Graph in realtime. I call it 'aggressive' because it selects only a limited number of similar book pages to crawl, (parameterized as 'branching-factor', I choose 3 in my code), as opposed to a 'crawl all similar book approach' as a regular BFS might have done. It stores the link structures in a text file in a 'BookA ->

BookB' format which implies 'BookA' led to 'BookB' in the web crawl. This allows me to retrieve all the book names from the file once the spider has finished crawling. Never again will I have to be confused about what book to read next, or so I wish.

VISUALIZATION

I create a visualization of the web-graph crawled by the spider using D3, a visualization Javascript library. This allows me to visualize and understand similarities in the book graph. Every Breadth first layer (nodes connected to the same node) are similar in themes. The themes vary as we go from one layer to another.

PROS

The aggressive BFS discovers books varying in subject matter yet it stays true to themes like Technology, Silicon Valley, Software, Entrepreneurship, Internet Computing and similar subjects. The idea is very simple but solves the problem I intended it to solve.

On going through the crawl graph, I found that uncovered books not only like 'The Second Coming of Steve Jobs' and 'The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia' that are very related to our origin book subject matter, but also books like 'The Human Brain: A Guided Tour' and 'An Imaginary Tale: The Story of the Square Root of Minus One', that are not very related yet are relevant to the central themes. This brings variety to my reading list. The branching-factor can be controlled to control the variety in the subject matter. The higher the branching-factor, the more similar the books are to each other, while, the lower the branching-factor the higher the variety in the themes of the books.

I also investigated using a Depth First style Crawler, the pages crawled in that case was, as one would expect had a much greater variety in subjects since DFS is a more aggressive search than BFS.

CONS AND POSSIBLE FIXES

The spider is slow. It takes about thirty minutes to crawl 1100 pages. The statistics are presented in the crawl statistics section. A caching technique could be used to speed it up.

The spider discovers some unrelated books at times. The farther we go from the origin (seed) book, the more the difference. For instance, 'Manliness and Civilization: A Cultural History of Gender and Race in the United States, 1880-1917' is not really a book I would believe is similar to 'The Google Story', but was uncovered by the spider. However this can be fixed by fine tuning the branching factor.

SOME SCREENSHOTS

```

abhishekp@abhishekp-E1-510: ~/FunProjects/Crawler
abhishekp@abhishekp-E1-510:~/FunProjects/Crawler$ cd Crawler/
abhishekp@abhishekp-E1-510:~/FunProjects/Crawler$ python crawler.py
Start
Beginning the crawl with The Google Story: Inside the Hottest Business, Media, and Technology Success of Our Time
The Google Story: Inside the Hottest Business, Media, and Technology Success of Our Time -> The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture
The Google Story: Inside the Hottest Business, Media, and Technology Success of Our Time -> Planet Google: One Company's Audacious Plan To Organize Everything We Know
The Google Story: Inside the Hottest Business, Media, and Technology Success of Our Time -> The Google Way: How One Company Is Revolutionizing Management as We Know It
The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture -> The Googlization of Everything: (And Why We Should Worry)
The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture -> The Perfect Store: Inside eBay
The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture -> Ambient Findability: What We Find Changes Who We Become
Planet Google: One Company's Audacious Plan To Organize Everything We Know -> The Perfect Thing: How the iPod Shuffles Commerce, Culture, and Coolness
Planet Google: One Company's Audacious Plan To Organize Everything We Know -> Google Speaks: Secrets of the Worlds Greatest Billionaire Entrepreneurs, Sergey Brin and Larry Page
The Google Way: How One Company Is Revolutionizing Management as We Know It -> The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia
The Google Way: How One Company Is Revolutionizing Management as We Know It -> The Apple Way
The Googlization of Everything: (And Why We Should Worry) -> The Filter Bubble: What the Internet is Hiding From You
The Googlization of Everything: (And Why We Should Worry) -> Consent of the Networked: The Worldwide Struggle For Internet Freedom
The Googlization of Everything: (And Why We Should Worry) -> Who Controls the Internet?: Illusions of a Borderless World
The Perfect Store: Inside eBay -> The PayPal Wars: Battles with eBay, the Media, the Mafia, and the Rest of the Planet Earth
The Perfect Store: Inside eBay -> eBoys: The First Inside Account of Venture Capitalists at Work
The Perfect Store: Inside eBay -> Code Name Ginger: The Story Behind Segway and Dean Kamen's Quest to Invent a New World
Ambient Findability: What We Find Changes Who We Become -> Designing Interfaces: Patterns for Effective Interaction Design
Ambient Findability: What We Find Changes Who We Become -> Designing for Interaction: Creating Smart Applications and Clever Devices
Ambient Findability: What We Find Changes Who We Become -> Designing Social Interfaces
The Perfect Thing: How the iPod Shuffles Commerce, Culture, and Coolness -> Infinite Loop: How Apple, the World's Most Insanely Great Computer Company, Went Insane
The Perfect Thing: How the iPod Shuffles Commerce, Culture, and Coolness -> Return to the Little Kingdom: Steve Jobs, the Creation of Apple, and How It Changed the World
Google Speaks: Secrets of the Worlds Greatest Billionaire Entrepreneurs, Sergey Brin and Larry Page -> The Scavengers' Manifesto
The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia -> The Public Domain: Enclosing the Commons of the Mind
The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia -> Click: What Millions of People Do Online and Why It Matters

```

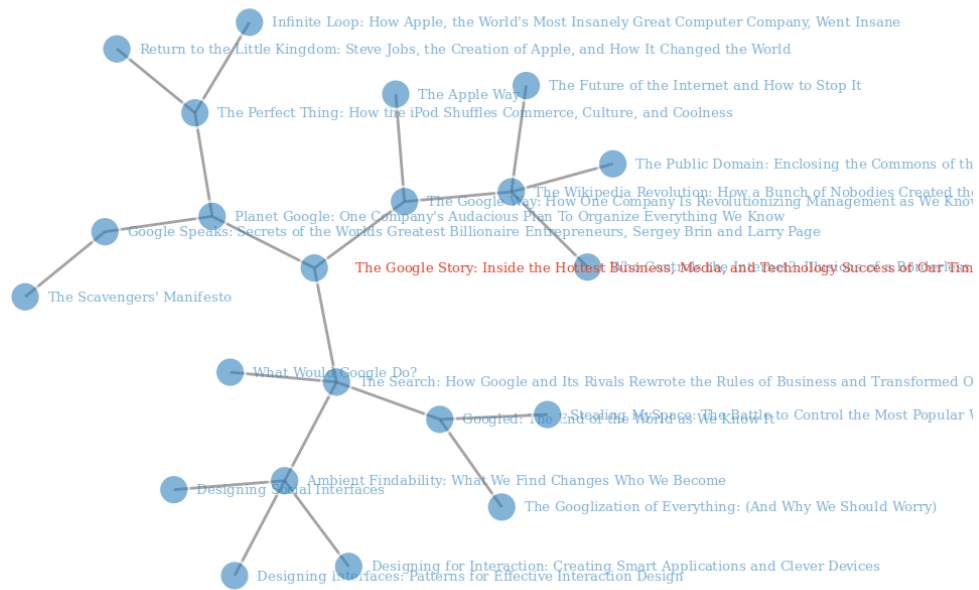
The crawl begins with our seed book url and then proceeds to find similar books

```

abhishekp@abhishekp-E1-510: ~/Test
abhishekp@abhishekp-E1-510:~/Test$ python Test
Border Crossing -> One of the Survivors
The Disciplined Trader: Developing Winning Attitudes -> Trading for a Living: Psychology, Trading Tactics, Money Management
Technical Analysis of Stock Trends -> Stock Market Wizards: Interviews with America's Top Stock Traders
Technical Analysis of Stock Trends -> Technical Analysis: The Complete Resource for Financial Market Technicians
The Aggressive Conservative Investor -> Value Investing: Tools and Techniques for Intelligent Investment
The Little Book of Valuation: How to Value a Company, Pick a Stock and Profit -> The Little Book of Value Investing
Contrarian Investment Strategies: The Classic Edition -> The Five Rules for Successful Stock Investing: Morningstar's Guide to Building Wealth and Winning in the Market
The Intelligent Asset Allocator: How to Build Your Portfolio -> All about Asset Allocation
The Intelligent Asset Allocator: How to Build Your Portfolio -> Common Sense on Mutual Funds: New Imperatives for the Intelligent Investor
The Intelligent Asset Allocator: How to Build Your Portfolio -> Unconventional Success: A Fundamental Approach to Personal Investment
Mouse Under Glass: Secrets of Disney Animation and Theme Parks -> Since the World Began: Walt Disney World--The First 25 Years
Mouse Under Glass: Secrets of Disney Animation and Theme Parks -> The Imagineering Field Guide to Magic Kingdom at Walt Disney World
Mouse Under Glass: Secrets of Disney Animation and Theme Parks -> The Haunted Mansion: From the Magic Kingdom to the Movies
Fireball: Carole Lombard and the Mystery of Flight 3 -> Lana: The Lady the Legend the Truth
Fireball: Carole Lombard and the Mystery of Flight 3 -> True Hollywood Noir: Filmland Mysteries and Murders
Jazz Cleopatra: Josephine Baker in Her Time -> Paris Noir: African-Americans in the City of Light
Jazz Cleopatra: Josephine Baker in Her Time -> Charmed Circle: Gertrude Stein and Company
Jean Arthur: The Actress Nobody Knew -> The Name Above the Title
Jean Arthur: The Actress Nobody Knew -> If This Was Happiness: A Biography of Rita Hayworth
Life is a Banquet -> This 'n That
Myrna Loy: Being and Becoming -> Self-Portrait
Myrna Loy: Being and Becoming -> Swanson on Swanson
Book name is not ascii encoded -> The Partnership: The Making of Goldman Sachs
Book name is not ascii encoded -> A Colossal Failure of Common Sense: The Inside Story of the Collapse of Lehman Brothers
The Lost Bank: The Story of Washington Mutual-The Biggest Bank Failure in American History -> Bull by the Horns: Fighting to Save Main Street from Wall Street and Wall Street from Itself
The Lost Bank: The Story of Washington Mutual-The Biggest Bank Failure in American History -> Crash of the Titans: Greed, Hubris, the Fall of Merrill Lynch, and the Near-Collapse of Bank of America
Last Man Standing: The Ascent of Jamie Dimon and JPMorgan Chase -> In FED We Trust: Ben Bernanke's War on the Great Panic
Last Man Standing: The Ascent of Jamie Dimon and JPMorgan Chase -> In an Uncertain World: Tough Choices from Wall Street to Washington
Social Darwinism in American Thought -> A History of American Law
Social Darwinism in American Thought -> Where Have All the Soldiers Gone?: The Transformation of Modern Europe
Programming Pearls -> The Little Schemer
Advanced Programming in the UNIX Environment -> The UNIX Programming Environment
Advanced Programming in the UNIX Environment -> The Implementation (TCP/IP Illustrated, Volume 2)
real 43m9.381s
user 20m7.396s
sys 0m21.508s
abhishekp@abhishekp-E1-510:~/Test$

```

It took about 43 minutes to crawl 1500 pages

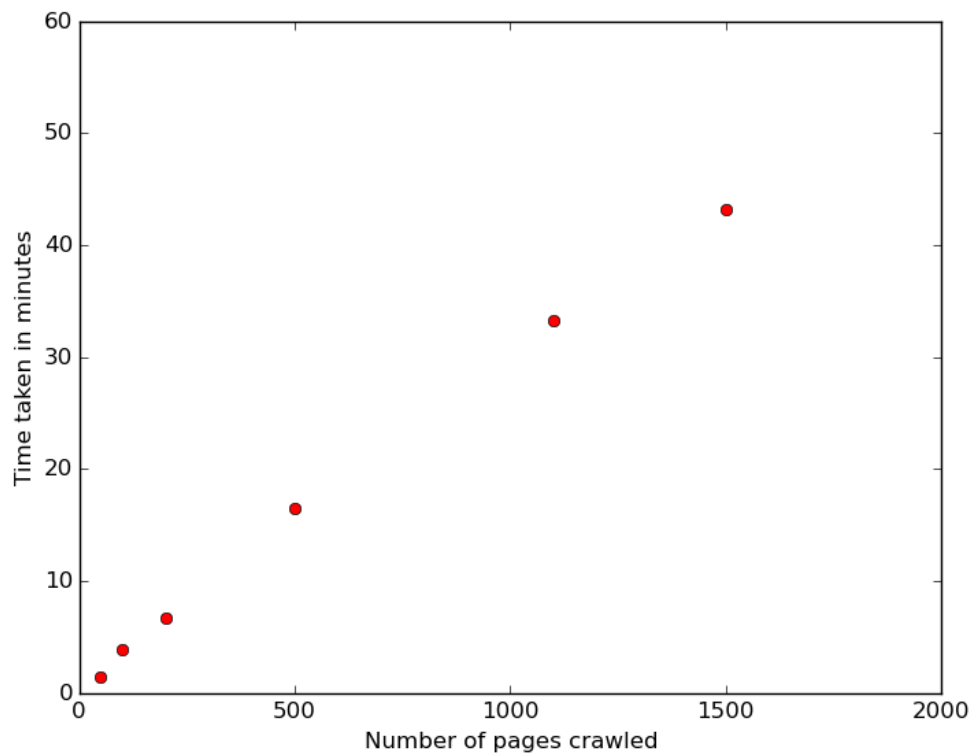


Visualizing the crawl graph using D3

CRAWL STATISTICS

The spider crawls at the rate of 35 pages a minute on an average.

Number of pages crawled	Time Taken(minutes)
50	1.5
100	3.9
200	6.67
500	16.5
1100	33.3
1500	43.16



Time taken to crawl as a scatter plot

Precision for a focused crawler is the percentage of the relevant web-pages it has crawled out of the total pages crawled. Recall is what percentage of relevant web-pages were actually crawled. The precision and recall for this particular focused crawler is very subjective because currently I don't have a quantitative measure for measuring the relevance of a web-page. Two books that seem seemingly unrelated from the title might actually be related. One would have to read them to understand.

EXPERIENCES AND LESSONS LEARNT

I found it particularly interesting to write a focused crawler because I used it to solve a problem of my interest. The crawler would crash often, particularly on calling `requests.get` method. I could make it more robust using exception handling. I applied a modified version of BFS to add variety to my crawled pages set. I also learnt to use BeautifulSoup which is a very useful HTML scraping tool.