



Vague but exciting...

CERN DD/OC

Tim Berners-Lee, CERN/DD

Information Management: A Proposal

March 1989



and...etc objects. For Sept 5. Usage: 230,000 per month.

- Introduction and Statistics
- Search and Search Engines
- Known Bugs - CERN Proxy Server
- Help and New Resource
- Feedback page at W3C

Select:

- Search only in Title of citing documents
- Search only in Names of citing documents
- Search all Citation hypertext
- Search all Names of cited URLs

Keywords: Start Search

WEBCRAWLER™
LIGHTNING FAST WEB SEARCH

Enter some words and start your search

Find pages with of these words and return results

Help · Facts · Top 25 Sites · Submit URLs · Random Links · No-forms Search

Copyright © 1995 America Online, Inc. **POWERED BY NEXTSTEP**

last updated: September 26, 1995

ONLINE Chess Lessons
Digital DVD Series

The 1997 Rematch Game 3

Preview

KASPAROV VS DEEP BLUE

GAME ANALYSIS BY:
GM YASSER SEIRAWAN

HOSTED BY:
INTERNATIONAL GRANDMASTER
RON W. HENLEY

Kasparov vs Deep Blue - Real time Commentary by GM Yasser Seirawan

Google!

Search the web using Google!

10 results Google Search I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

About Google!

Stanford Search Linux Search

Get Google! updates monthly!

your e-mail Subscribe Archive

Copyright ©1997-8 Stanford University

Sentiment Mining in WebFountain

Jeonghee Yi Wayne Niblack

IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120

Abstract A key component of our research is the sentiment mining that ex

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

{sergey, page}@cs.stanford.edu

Computer Science Department, Stanford University, Stanford, CA 94305

Abstract

LINKS AS FOLLOWS.

We assume page A has pages $T_1 \dots T_n$ which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also $C(A)$ is defined as the number of links going out of page A . The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d (PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$$

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

PORN

DEEP WEB

HACKERS

WIKILEAKS

GOVERNMENT

UNDETECTED

ILLEGAL PORN

Verification of a human in the loop
OR
Identification via the Turing Test*

Moni Naor †

September 13th, 1996

Abstract

We propose using a “Turing Test” in order to verify that a human is the one making a query to a service over the web. Thus, before a request is processed the user should answer as a challenge an instance of a problem chosen so that it is easy for humans to solve but the best known programs fail on a non-negligible fraction of the instances. We discuss several scenarios where such tests are desired and several potential sources for problems instances. We also discuss the application of this idea for combatting junk mail.

1

4

7

9

7

6

2

5

4

4

3

9

0

8

7

6

0

9

7

2

```
URL="http://www2.planalto.gov.br/acompanhe-planalto/"\  
"discursos/discursos-do-presidente-da-republica/"\  
"discurso-do-presidente-da-republica-michel-temer-"\  
"durante-cerimonia-de-encerramento-do-ii-encontro-"\  
"da-carta-caiman"
```

```
import requests  
from selenium import webdriver
```

```
r=requests.get(URL, timeout=20)
```

```
timeout
```

Traceba

```
In [6]: wd=webdriver.PhantomJS()
```

```
In [7]: wd.get(URL)
```

```
In [8]: wd.save_screenshot("govTest.png")
```

```
Out[8]: True
```

The screenshot shows the Planalto website interface. At the top, there are navigation links for 'BRASIL', 'Serviços', 'Participar', 'Acesso à informação', 'Legislação', and 'Canais'. Below this is a search bar with the text 'Planalto PRESIDÊNCIA DA REPÚBLICA' and a search button. The main content area features a news article titled 'Discurso do Presidente da República, Michel Temer, durante cerimônia de Encerramento do II Encontro da Carta Caiman'. The article is dated 'Miranda/MS, 21 de outubro de 2017'. The text of the article begins with 'Olha eu quero começar cumprimentando o Roberto Klabin, pela gentileza de nos receber e promover este encontro em um momento que todos os senhores e as senhoras percebem que é um momento importantíssimo para o meio ambiente no nosso País. Que proteger o Pantanal é proteger uma parte do nosso País. Além, quando ouvi a exposição preambular, preliminar em que o Roberto mostrava que o Pantanal, com a sua área, poderia ocupar cerca de 4 países da Europa, nos podemos verificar a dimensão em um ato que na verdade preserva o meio ambiente no Pantanal.' The article is attributed to 'Eu quero cumprimentar o Reinaldo Azambuja, governador de Mato Grosso do Sul, e Marcela Cruz, ministro'.

Apache Nutch

Heritrix

Scrapy

PhantomJS

Selenium



deeplearn.js
a hardware-accelerated
machine intelligence
library for the web

scikit-learn
Machine Learning in Python

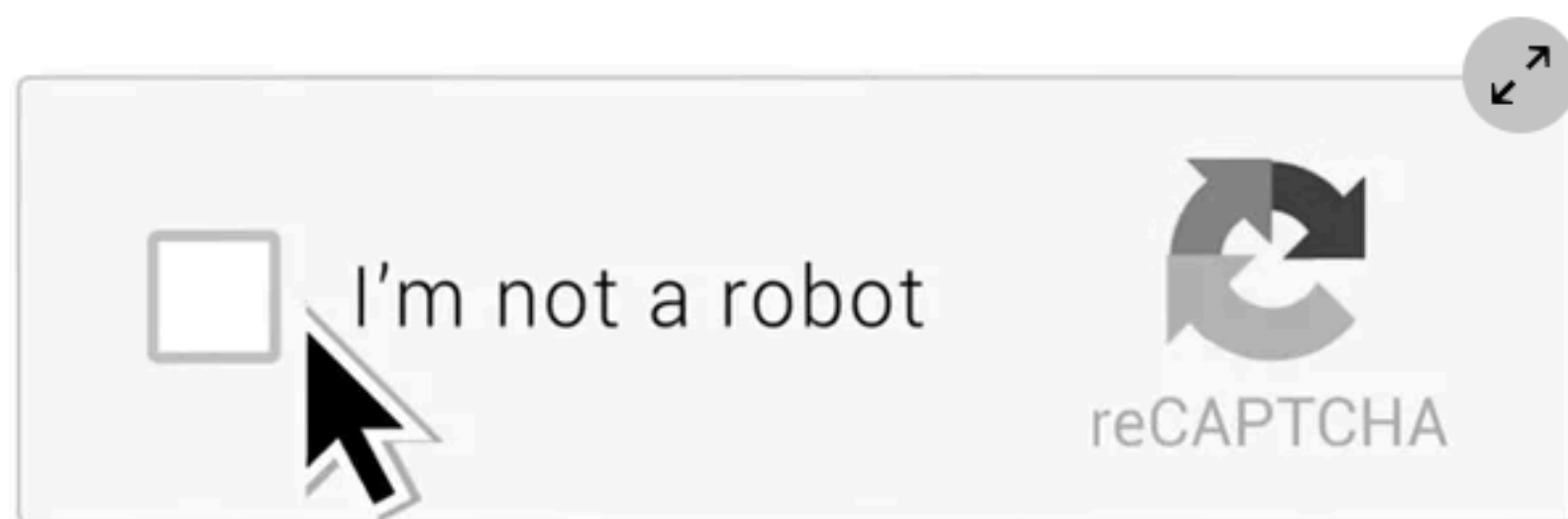
- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license





How the internet found a better way than illegible squiggles to prove you're not a robot

Captcha has evolved from identifying mangled letters to web users unwittingly training Google's AI. Now, finally, you won't have to do anything



NEW! The Free & Easy Way to Prevent Bots from Ruining Google Analytics

Are You a Human is now part of Distil Networks

[Read the Announcement](#) [Get Bot Filtered Analytics](#)

Bots	45,289	62%
Humans	117,940	38%



TECHNOLOGY

HOW IT WORKS

INTEGRATION OPTIONS

WEB CRAWLER BLOCKING SERVICE

Detect and Stop Suspicious Crawling Activities on Your Website With ShieldSquare

Protecting billions of Web pages for businesses spread across 70 countries



Information extraction — or automatically classifying data items stored as plain text — is a major topic of artificial-intelligence research.

Image: MIT News

Artificial-intelligence system surfs web to improve its performance

“Information extraction” system helps turn plain text into data for statistical analysis.

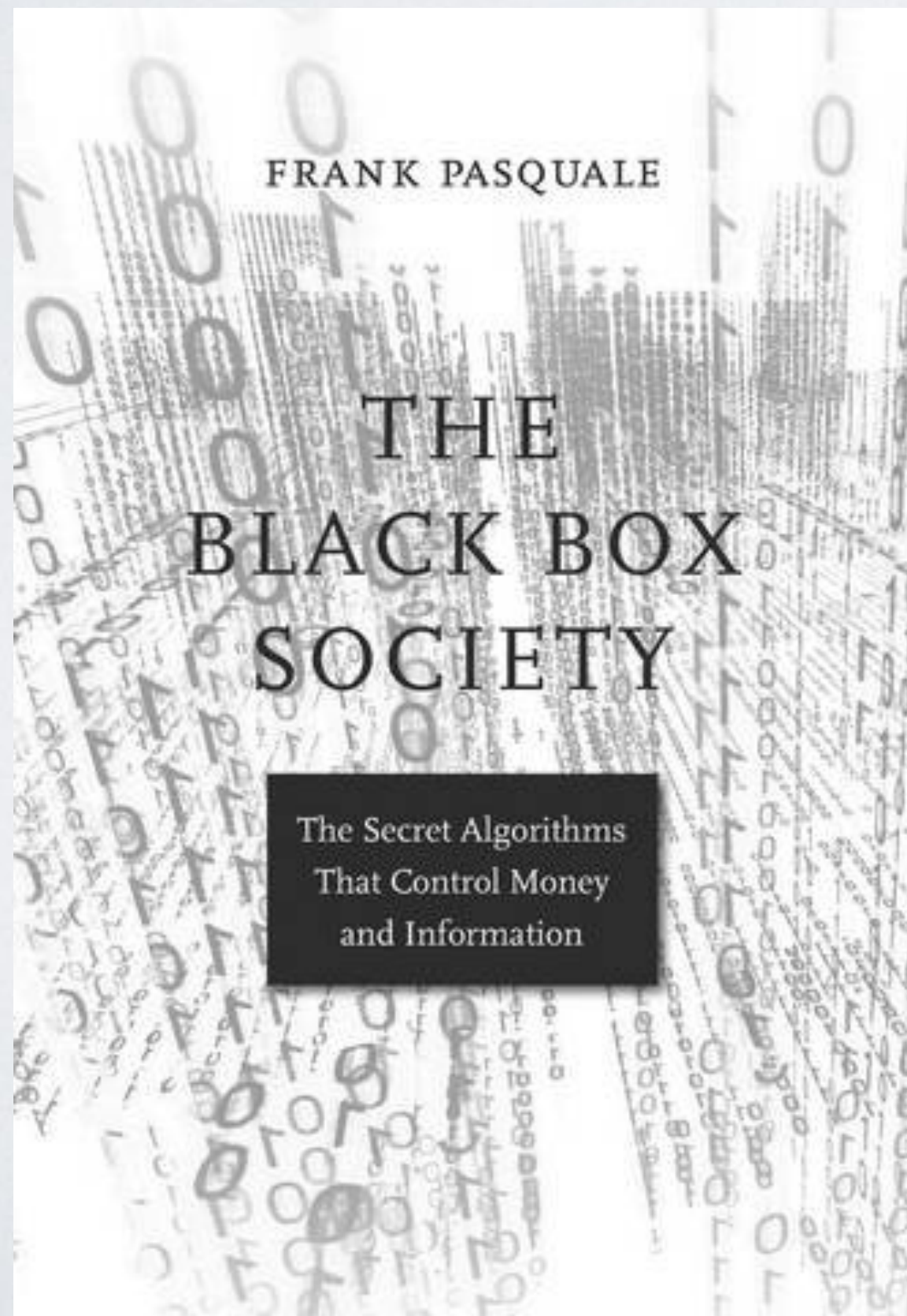
Larry Hardesty | MIT News Office
November 10, 2016

▼ Press Inquiries

RELATED

Of the vast wealth of information unlocked by the Internet, most is plain text. The data

Paper: “Improving information extraction by



TOM SIMONITE BUSINESS 10.18.17 03:00 PM

AI EXPERTS WANT TO END 'BLACK BOX' ALGORITHMS IN GOVERNMENT



Text Generation With LSTM Recurrent Networks in Python with Keras

by Jason Brownlee on August 4, 2016 in Long Short-Term Memory Networks



Recurrent neural networks can also be used as generative models.

This means that in addition to being used for predictive models (making predictions on the sequences of a problem and then generate entirely new plausible sequences from the domain.

Generative models like this are useful not only to study how well a model has learned to learn more about the problem domain itself.

Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by Will Knight April 11, 2017

Last year, a strange self-driving car was released onto the quiet roads of Monmouth County, New Jersey. The experimental vehicle, developed by researchers at the chip maker Nvidia, didn't look different from other autonomous cars, but it was unlike anything demonstrated by Google, Tesla, or General Motors, and it showed the rising power of artificial intelligence. The car didn't follow a single instruction provided by an engineer or programmer. Instead, it

Obrigado

Rodrigo Arrigoni

Twitter e Github: @VulcanoAhab