TRAITOR - Associating Concepts using the World Wide Web

Lesley Wevers Oliver Jundt Wanno Drijfhout
Djoerd Hiemstra

Hadoop Summit, March 2013

Norvig Web Data Science Award

show what you can do with 6 billion web pages

by SURFsara and Common Crawl

SURF SARA

Common Crawl

66 Nodes

460 TB
Filesystem

3.6 Billion Documents

> 100 TB
Uncompressed

"Apache Hadoop is awesome!"

(apache, hadoop, 1)
(apache, awesome, 1)
(awesome, hadoop, 1)
...

(apache, hadoop, 483)
(apache, awesome, 8392)
(awesome, hadoop, 5930)
...

Search associations

java Search

Found ≥ 100 results. Search query for `java` completed in 0.02 seconds.

Word	Score
code	41525.0
applications	36354.0
application	32494.0
games	30694.0
software	29266.0
programming	28815.0
language	28641.0
sun	27826.0
development	26629.0
server	25798.0
hosting	24813.0
platform	22116.0

Search associations

php Search

Found ≥ 100 results. Search query for php completed in 0.02 seconds.

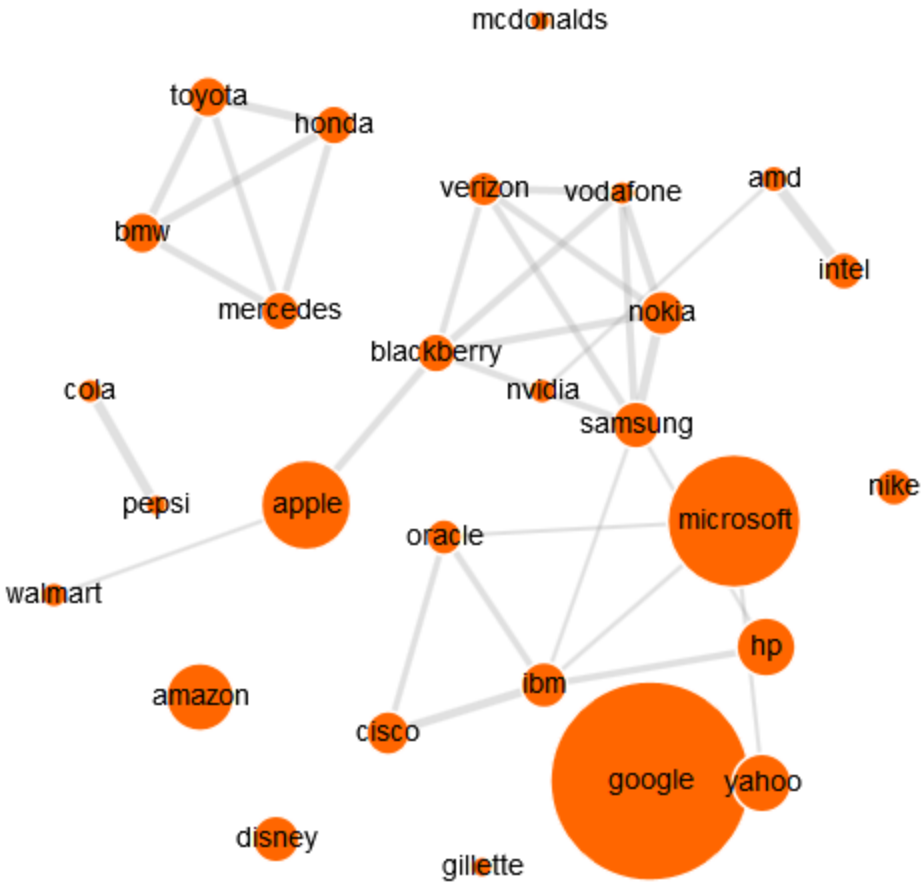
Word	Score
error	121102.0
mysql	116138.0
encountered	113211.0
code	53203.0
performance	52229.0
zend	43114.0
hosting	41823.0
server	40031.0
script	38327.0
optimizer	34884.0
html	34462.0
development	30635.0

Search associations

java -php Search

Found ≥ 100 results. Search query for `java` `-php` completed in 0.03 seconds.

Word	Score
applet	15996.0
applets	11697.0
alden	11193.0
jvm	8062.0
coffee	5871.0
marketplace	5625.0
east	5223.0
concentration	4788.0
flashcards	4779.0
servlet	4756.0
indonesia	4399.0
central	4357.0


mcdonalds
toyota
honda
verizon
vodafone
amd
bmw
intel
mercedes
nokia
blackberry
cola
nvidia
samsung
nike
pepsi
apple
microsoft
oracle
walmart
hp
amazon
ibm
cisco
google
yahoo
disney
gillette

Thank You!

traitor.imperamus.eu