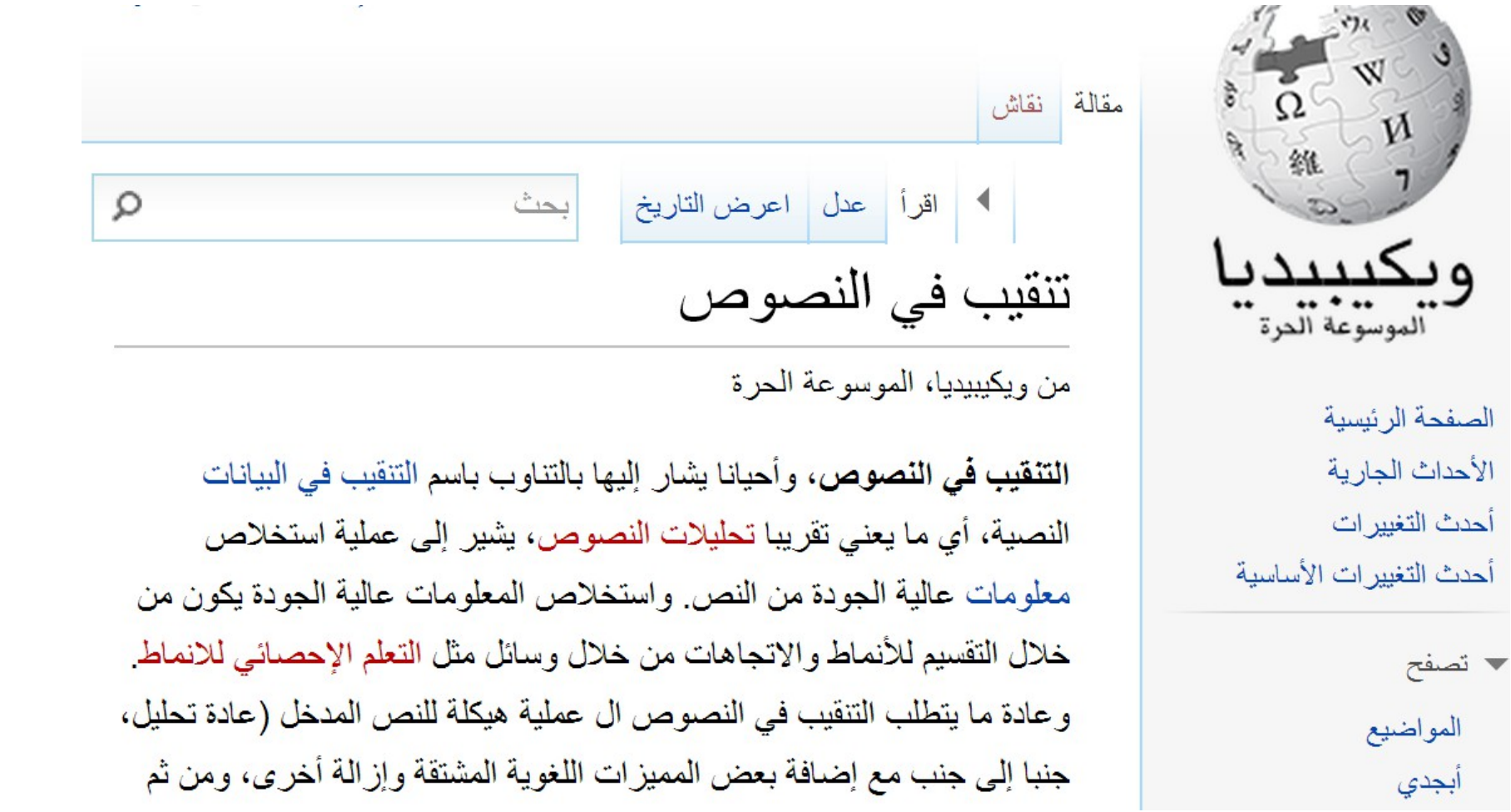
The Case of Text and Data Mining

Mathias Schindler
Wikimedia Deutschland e.V.

Wikipedia


WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Interaction
Article
Talk
Read
Edit
Search
Text mining
From Wikipedia, the free encyclopedia
Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving

encyclopedia


ويكيبيديا
الموسوعة الحرة
الصفحة الرئيسية
الأحداث الجارية
أحدث التغييرات
أحدث التغييرات الأساسية
تصفح
المواضيع
أبجدي
مقالة
نقاش
اقرأ
عدل
اعرض التاريخ
بحث
تنقيب في النصوص
من ويكيبيديا، الموسوعة الحرة
التنقيب في النصوص، وأحيانا يشار إليها بالتناوب باسم التنقيب في البيانات النصية، أي ما يعني تقريبا تحليلات النصوص، يشير إلى عملية استخلاص معلومات عالية الجودة من النص. واستخلاص المعلومات عالية الجودة يكون من خلال التقسيم للأنماط والاتجاهات من خلال وسائل مثل التعلم الإحصائي للانماط.
وعادة ما يتطلب التنقيب في النصوص ال عملية هيكلة للنص المدخل (عادة تحليل، جنبا إلى جنب مع إضافة بعض المميزات اللغوية المشتقة وإزالة أخرى، ومن ثم

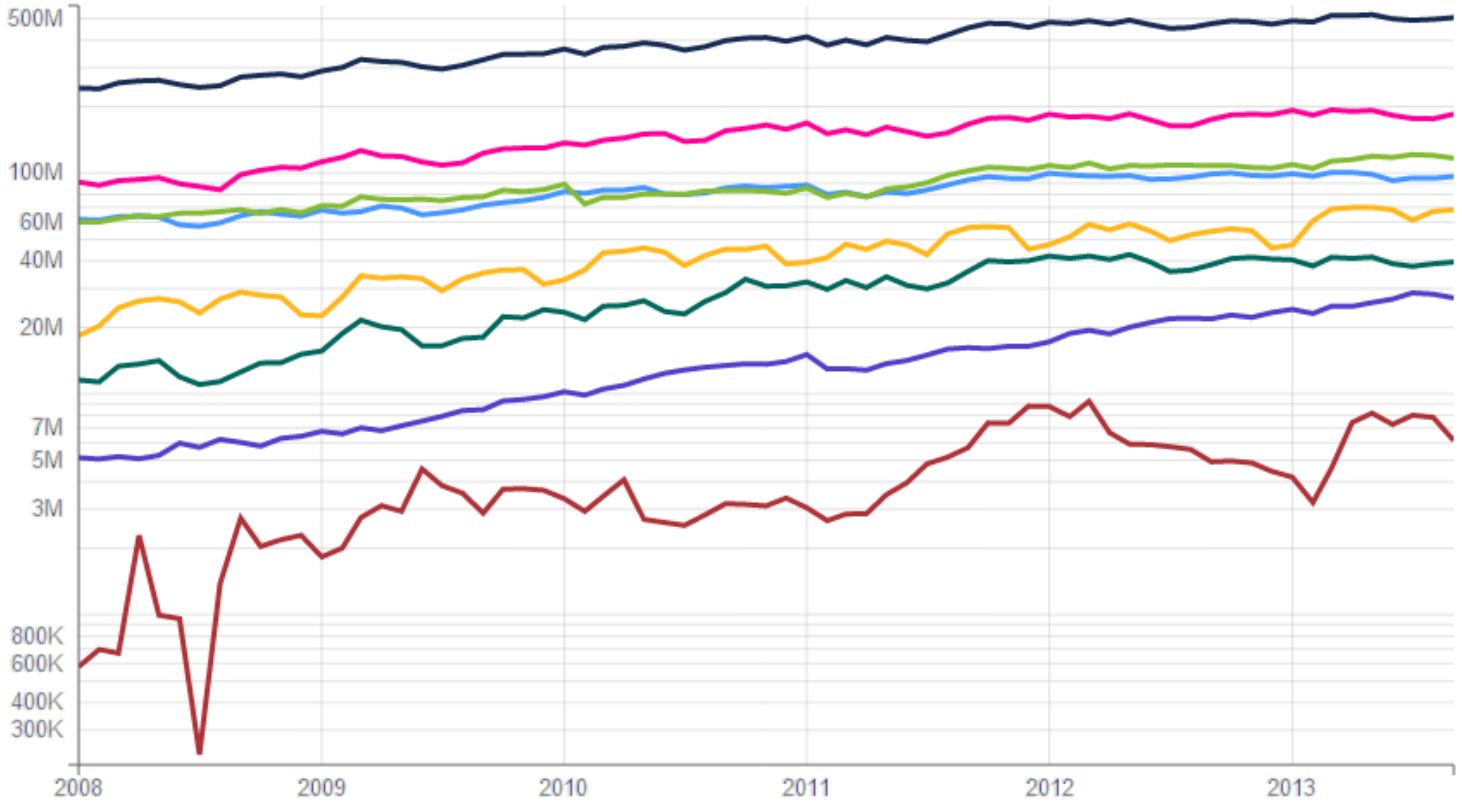
276 languages

505.90 Million

Sep 12 — Sep 13	6.54%
Aug 13 — Sep 13	1.82%

Unique Visitors per Region (comScore)

Sep 2013	
World	505.90M
China	6.15M
Europe	184.95M
India	27.18M
Latin America	68.18M
Middle East/Africa	39.52M
North America	96.52M
Asia Pacific	116.73M



http://stats.wikimedia.org

free

gratis ≠ libre

copy

modify

publish

Creative Commons

Dump complete

Verify downloaded files against the MD5 checksums to check for corrupted files.

2013-10-11 17:28:31 **done** Articles, templates, media/file descriptions, and primary meta-pages, in multiple bz2 streams, 100 pages per stream

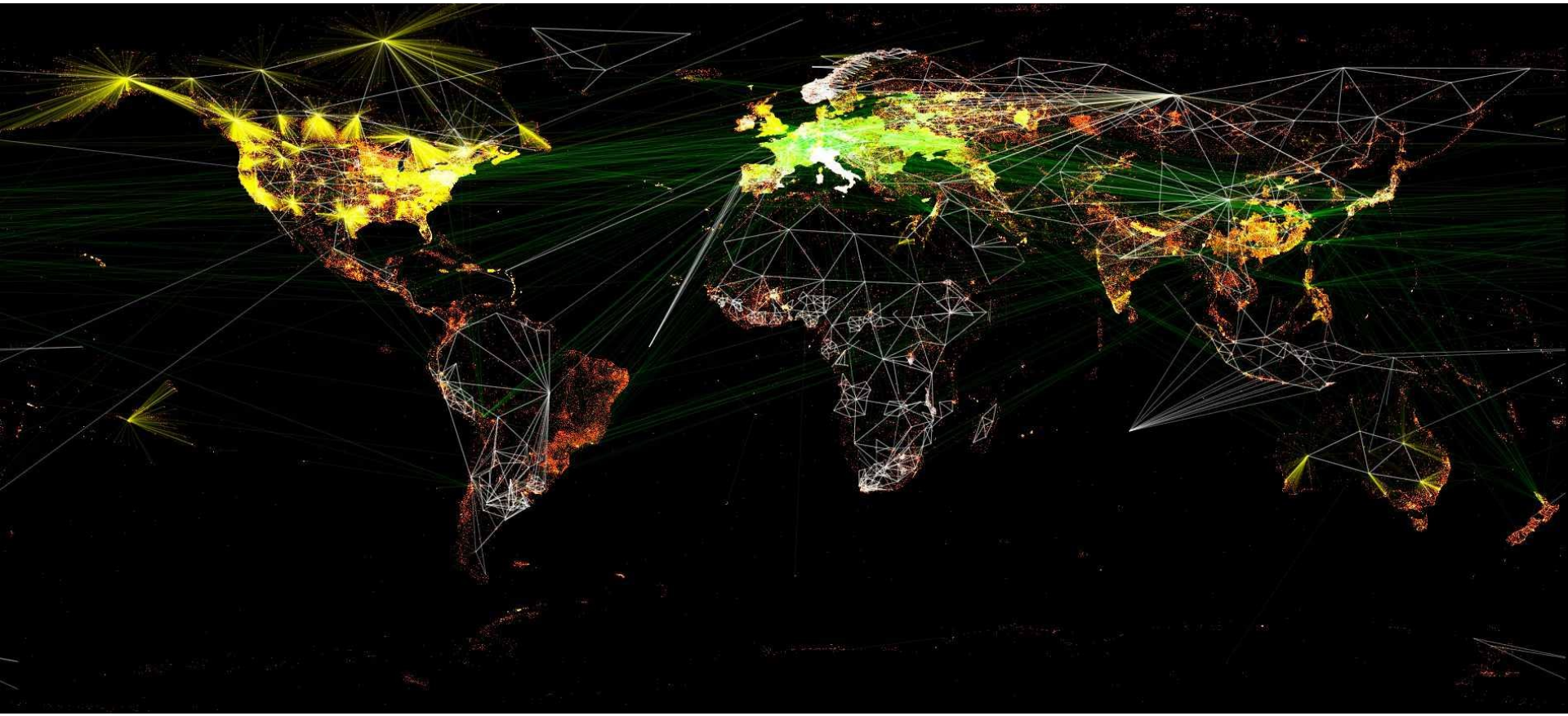
enwiki-20131001-pages-articles-multistream.xml.bz2 10.2 GB
enwiki-20131001-pages-articles-multistream-index.txt.bz2 143.9 MB

2013-10-11 14:43:32 **done** All pages with complete edit history (.7z)

enwiki-20131001-pages-meta-history1.xml-p000000010p000003390.7z 221.8 MB
enwiki-20131001-pages-meta-history1.xml-p000003392p000005816.7z 209.1 MB
enwiki-20131001-pages-meta-history1.xml-p000005817p000008695.7z 221.1 MB
enwiki-20131001-pages-meta-history1.xml-p000008697p000010000.7z 104.2 MB
enwiki-20131001-pages-meta-history2.xml-p000010001p000012958.7z 218.6 MB
enwiki-20131001-pages-meta-history2.xml-p000012959p000015299.7z 208.7 MB
enwiki-20131001-pages-meta-history2.xml-p000015301p000016780.7z 134.2 MB
enwiki-20131001-pages-meta-history2.xml-p000016781p000018724.7z 135.5 MB
enwiki-20131001-pages-meta-history2.xml-p000018726p000020162.7z 128.3 MB
enwiki-20131001-pages-meta-history2.xml-p000020163p000021849.7z 141.7 MB
enwiki-20131001-pages-meta-history2.xml-p000021850p000023768.7z 151.5 MB
enwiki-20131001-pages-meta-history2.xml-p000023769p000025000.7z 84.9 MB
enwiki-20131001-pages-meta-history3.xml-p000025001p000027276.7z 196.4 MB
enwiki-20131001-pages-meta-history3.xml-p000027277p000030292.7z 219.0 MB
enwiki-20131001-pages-meta-history3.xml-p000030294p000032780.7z 205.4 MB

http://download.wikimedia.org

download.wikimedia.org



http://commons.wikimedia.org/wiki/File:Wikidata_POTW_Candidate_01_(Geocoord_map).JPG (CC0)

outside of licenses

freedom to read

freedom to read fast

freedom to Ctrl+F

freedom to grep

grep

From Wikipedia, the free encyclopedia

Grep is a command-line utility for searching plain-text data sets for lines matching a regular expression. Grep was originally developed for the Unix operating system, but is available today for all Unix-like systems. Its name comes from the ed command *g/re/p* (***g****lobally search a* ***r****egular* ***e****xpression* and ***p****rint*), which has the same effect: doing a global search with the regular expression and printing all matching lines.[1][2]

WIKIMEDIA DEUTSCHLAND

freedom to sed

sed

From Wikipedia, the free encyclopedia

For other uses, see Sed (disambiguation).

sed (*stream editor*) is a Unix utility that parses and transforms text, using a simple, compact programming language. sed was developed from 1973 to 1974 by Lee E. McMahon of Bell Labs,[1] and is available today for most operating systems.[2] sed was based on the scripting features of the interactive editor ed ("editor", 1971) and the earlier qed ("quick editor", 1965–66). sed was one of the earliest tools to support regular expressions, and remains in use for text processing, most notably with the substitution command. Other options for doing "stream editing" include AWK and Perl.

freedom to gawk

AWK

From Wikipedia, the free encyclopedia

This article is about the programming language. For other uses, see AWK (disambiguation).

AWK is an interpreted programming language designed for text processing and typically used as a data extraction and reporting tool. It is a standard feature of most Unix-like operating systems. AWK was very popular in the late 1970s and 1980s, but from the 1990s has largely been replaced by Perl, on which AWK had a strong influence.

TDM

fix copyright

access

legal deposit

Thank you

mathias.schindler@wikimedia.de