Apache Mahout Making data analysis easy





Isabel Drost

Nighttime:

Co-Founder, committer Apache Mahout. Organiser of Berlin Hadoop Get Together.

Daytime:

Software developer. Guest lecturer at TU Berlin. Co-Organiser Berlin Buzzwords 2010.

- "Mastering Data-Intensive Collaboration and Decision Making"
- EU funded research project
 - Number of partners: 8
 - Coordinator: Research Academic Computer Technology Institute (CTI), Greece









Machine learning background?



Agenda

• Data Mining/ Machine Learning?

• Why is scaling hard?

• Going beyond simple statistics.

Data Mining Applications

- Marketing.
- Surveillance.
- Fraud Detection.
- Scientific Discovery.
- Discover items usually purchased together.

= Extracting patterns from data.

Machine Learning Applications

- E-Mail spam classification.
- News-topic discovery.
- Building recommender systems.

= Extracting prediction models from data.

Machine learning – what's that?



Image by John Leech, from: The Comic History of Rome by Gilbert Abbott A Beckett. Bradbury, Evans & Co, London, 1850s Archimedes taking a Warm Bath

Archimedes model of nature

 $\frac{Density of Object}{Density of Fluid} = .$

Weight

Weight – Apparent immersed weight











of the local division of the local divisiono

June 25, 2008 by chase-me http://www.flickr.com/photos/sasy/2609508999



An SVM's model of nature



The challenge

Mission

Provide scalable data mining algorithms.



when the survey and an interest law 3

- Colorist Mistory at the Internet and is increasingly Fraktamatte Fatura

And there we are a set of the set

The 1940s, 1950s, and 1940s. An about 5 to 1940s seemi 2000bb, and have sumproduces movies events much 2000bb takenesses Andree and 100 bb, and and much base base while a face seemide control 2000bb takenesses and and the termination of the set of the set of the second control 2000bb takeness while the set of the second termination takeness to the providence to any the second termination of the set of the set of the second termination of the second termination of the set of the set of the second termination of the second termination of the set of the second termination of the second termination of the second termination terms of present for the second termination of the second termination of the set of the second termination of the second termination of the second termination termination of the second termination of The 1940s, 1950s, and 1960s. As the first first want 1 when, where Development for Ander Startingen und Warter für stream in 1947 und 1988 of restriction for Ander Starting and the development of development restriction are streamed for the Angelop of Angelopment set and Statement Notice of Fourier Bully a second on Junior and Junior Printing adapt over 1988. mandreaf a mander or success rate for herboring to produce and and mands for flat they also they derived special of annually a single strengther wath growing parents, builds this any way out out deviating a strategied of making equivalent on provide the star, and shad from a form constal the entropy of of the same have and on the same parts of semiconductor. In achafternet to representations, either elevates a companies into mostly an representation. expansions and studies could be made using the same process. and materials (Handsont, Ming It is important to make that since the posts, the manifest of

transitions per unit area fun been doubling every one and a half process about the restance of the process of the state of the provide state of the new of encour Interaction is called Moore's Law runned after transfer Mouve a pleasary in the unregrated cocout right and Knowler of the Intel Corporation (deal.)

> Name and Address of the Owner, where the Name of Concession, or other states of the

http://www.flickr.com/photos/honou/2936937247/

a low New

- I'm Afre Face

and the lesson

and all a constants

HowTo: From data to information.

January 3, 2006 by Matt Callow http://www.flickr.com/photos/blackcustard/81680010

COMMUNITY NEWS

Finishing touches still to come

A glimpse of today, yesterday

http://www.flickr.com/photos/redux/409356158/









By Patrick Lanke. The second inderstanding of the rationale underpinning

heckpoints and current best







ect, and how to their project, and how to web authors/developers excellent starting point i Specification (PAS) 78 Gu

in Commissioning A





Rund 7500 Demonstranten nahmen an dem Prote "Freiheit statt Angst – Stoppt den Überwachungs





experteer^{de}

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Mo "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

seachungeetaat







http://www.flickr.com/photos/redux/409356158/





By Particle carrier By the access of the second se Content Accessibility Guidelines (WCAC WAYAW3.org/WAI/Intro/wcag provide foundation for development and evalu. However, authors still need an actual understanding of the rationale underplinning development and current best s checkpoints and current best





Development of the accession of the second s understanding of the rationale underpinning heckpoints and current best









ect, and how t their project, and how to web authors/developers excellent starting point i Specification (PAS) 78 Gu

in Commissioning A



Rund 7500 Demonstranten nahmen an dem Prote "Freiheit statt Angst – Stoppt den Überwachungs





experteer^{de}

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Mo "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

In web accessibility, you'll often hi being placed on the duty of web a create accessible content. Howeve one part of the web accessibility Right at the planning stage, cli owners should gain a basic under Right at the planning stage, clients owners should gain a basic understand what accessibility 6, whether accessibility 6, whether accessibility 6, specification (RAS) 7 in Commissioning Accessibility Rights Commission (RKC). Web authors and developer in aware of potential accessibility in corners to developing their content Content Accessibility Guidelines (WCAG) at works/W3-007/WAI/Intro/Acsa provide a solid foundation for development and evaluation-tiowerk, authors still need an actual understanding of the rationale undergranding understanding chorents and current text understanding of the rationale underpinning checkpoints and current best



Internet Dater



http://www.flickr.com/photos/redux/409356158/















DATUM 11.8.2010 - 17:20 Uw QUELLE ZEIT ONLINE, AFP, das QUENERZARE 4 EMPTEMENT E-Mail verschicken | 1 Tausende demonstrieren für Bürgerrechte im Netz Für einen besseren Arbeitnehmerdatenschutz die Gesundheitskarte: 130 Organisationen ha Demonstration aufgerufen. Sie fürchten den and gegen

ZEIT CONLINE | DATENSCHUTZ

reiheit

Finally, there ternet

agents are a

larger te

and how

nartic

themselve

24 Parcicle Joseph Construction Paradocessamily Web Parcicle Joseph Construction Paradocessamily Construction Paradocessamily Research Construction Paradocessamily Research Mark accessibility Is, why it's important for web accessibility Is, why it's important for web authors/developers to be to build Specific Accessibility (Section Paradocessibility) Specific Construction Paradocessibility Specific Construction (Construction) Paradocessibility (Section Paradocessibility) Section Paradocessibility (Section Paradocessibility) Specific Constructions and developers need to be consets to developing their constructions and Section (Construction) Accessibility (Sarded as actual Construction) (Construction) (Construction) Accessibility (Condense (VCAC)) and Construction Still need as actual Condition for development and evaluation. However, authors still need a current that authors/advention of current that their Accessibility of accessibility (Sarded and Construction) Accessibility (Condense (VCAC)) and Accessibility (Condense (VCAC)) and Accessibility (Condense (VCAC)) and Accessibility (Condense (VCAC)) and Accessibility (Construction) Accessibility (Condense (VCAC)) and Accessibility (Construction) Accessibility (Co

understanding of the rationale underpinning

nen hatten zur

Stat



ton straitet über k.

ieren für Dürperrecht

Rund 7500 Demonstranten nahmen an dem Prote "Freiheit statt Angst – Stoppt den Überwachungsv experteer^{de}

Patriersuche Immobilien Automarkt Jobs Reiseangebo STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN INGITAL STUDIUM KARRIERE LEBENBART REISEN AUTO Internet Datenachutz Mobil Games DATUM 11.9.2010 - 17.20 Unr 1- QUELLE 2017 ONLINE, APP, doa WOMMENTARE 4 * EMPERILEN E-Mail verschicken | Facebook, Twitter, Daren Tausende demonstrieren für Bürgerrechte im Netz Für einen besseren Arbeitnehmerdatenschutz und gegen DRUCKEN Dru die Gesundheitskarte: 120 O isationen hatten zur onstration aufgerufen. Sie fürchten der Überwachungsstaa NEU IM RESSORT ZDNE Kuner darf ORF-Portal Futurezone kaufen IRRECHTE "Wir haben als Kind gelennt, Teilen ist NT PAM AUF FACEBOOK "Ein iPad umsonst, ich halte es en" LIPAD APPR Menr Freiheit im Ann-Store EU AUF ZEIT ONLIN reiheit

ZEIT/MONLINE DATENSCHUTZ

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Mo "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.



experteer^{de}



Finally, there

es. They need t nts are available to th nd how to configure icular needs



understanding of the rationale underpinning checkpoints and current best









CRUITINGUT UTIMAMIN GASSING 2008 SEMILADARDE SANTAL STUDIER AND ANTICIAL STUDIER ANTICAL STUDIER ANTICIAL ST

Internet Dater

Taucondo domonetrioron für

reiheit

Bürgerrechte im Netz Für einen besseren Arbeitnehmerdatenschutz und geg die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.

AN ALE EACEDORY "Co. Dad unacoust, ich habe as i

Entremuche Immobilien Autometici John Beinenrugeboie

DATUM 11.9.2010 - 17:20 Ukr
1 - QUELLE ZEIT ONLINE, AFP, dos
q KOMMENTARE 4
 EMPFERLEN E-Mail verschicken | F

EL DRUCKEN D

E & IPAD APPS Meter Evoluti in Ann Store

experteer^{de}

0

stat

Rund 7500 Demonstranten nahmen an dem Prote 'Freiheit statt Angst – Stoppt den Überwachungsv

http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

http://www.flickr.com/photos/redux/409356158/





Rights Commission It Web authors and comes to developing Content Accessibility Guidelines (W www.w3.org/WAI/intro/vscag pro foundation for development and However, authors still need an ac derstanding of the rationale u heckpoints and current best



By Plattick Latter In web accessibility, we'll denn here emphasised being placed on the days of severe and one part of the web accessibility regarding. Right at the planning stage, diversation web atternor, diversion of the severe web atternor, diversion of the severe web atternor, diversion of the severe the commission of the severe the commission of the Web authors and 4 aware of potential accession methods. Web authors and a avare of potential accessionsy accessions comes to developing their content. Content Accessibility Guidelines (W www.w3.org/NAI/Intro/wcag prov foundation for development and However, authors still need an act-deventation of the rationale und understanding of the rationale underpinning understanding of the rationale underpinning





derstanding of the rationale underpinnin unious checkpoints and current best



ZEIT CONLINE DATENSCHUTZ mobilien Autometht Jobs Beiser STARTSEITE POUT KARRIERE LEBENSART REISEN AUTO Internet Datenschutz



Für einen besseren Arbeitnehmerdat die Gesundheitskarte: 130 Organisat Demonstration aufgerufen. Sie fürch





0

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Me "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil. experteer ZEIT CONLINE DATENSCHUTZ

Eathersuche Immobilien Automatist John Beiseangebote STARTSETTE POLITIK WIRTSCHAFT MEIMING GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUN KARRERE LEBENGART REISEN AUTO Daterschutz Mobil Game

Tausende demonstrieren für Bürgerrechte im Netz

on hoosonon Arhoit



Rund 7500 Demonstranten nahmen an d "Freiheit statt Angst – Stoppt den Überw



men an dem Protestzug unter dem M en Pherwachungswahn" in Berlin teil. experteer^{de}



Who's response By Patrick Lauke

Merivale mu By JUSTIN S. CAMPARIL classi

ZEIT DATENSCHUTZ

Partnersuche

DATUN

E. QUELL

KOMM

* EMPFE

SCHLA

NEU IM R I. DATEN

im Neta

2. FUTUR URHEB

gut" SPAM

den Hä 5. IPHON

NEU AUF

Datens

Buzz

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIL SPORT

Internet Datenschutz Mobil Games

DATENSCHUTZ

Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.



In Berlin demonstrierten tausende Demonstranten für mehr Datenschutz

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Motto "Freiheit statt Angst - Stoppt den Überwachungswahn" in Berlin teil.

I. NACHR http://www.flickr.com/photos/topsy/204929063/ 2. CDU Un 3. GEDEN Resser

1. BUNDESLIGA Der BVB 5. DAAD Das Ende der Åra Bode





& Harbor , among the most imaginative architec-The National Aquatics Center lies on th

described file projects as bullish nation's budding

buildings are

of China's emb the Modernist c swooping form suggests angs placed si has been con dragon. Yet cedent Airport monument conceived Tim Griffith/PTW Architects Speer in 1 gateway ent ceremonial axis. Europe, Both are part o bile society that exte

lew China

although at times ter

fying in their aggress

scale, they also ref

the country's effor

give shape to an en

ing national identit Foster's airport

nal, the world's

is the purest exp

GE.

Grand Central Term the great train halls Like Tempelhof, nal boasts a swee evokes the glame closing a surp



The HDFS filesystem is not restricted to MapReduce jobs. It can be used for other applications, many of which are under way at Apache. The list includes the HBase database, the Apache Mahout machine learning system, and matrix operations. http://www.flickr.com/photos/29143375@N05/3344809375/in/photostream/

http://www.flickr.com/photos/redux/409356158/



Second Second and an and an an APPS Meter Freiheit im Ann-Store

0













Extremuche Immobiler Autometri John Beinengebol



Rund 7500 Demonstranten nahmen an dem Prote "Freiheit statt Angst – Stoppt den Überwachungst experteer^{de}



experteer^{de}

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Mor "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin tell.



checkpoints and current best

larger text or particul

V Patrick Lauke

foundation for development and evalu However, authors still need an actual understanding of the rationale underpinning

s checkpoints and current best

Finally, there has to be an o themselves. They need to u agents are available to the and how to configure particular needs. For exa larger text or particular

From data to information.

Collect data and define your learning problem.

• Data preparation.

• Training a prediction model.

• Checking the performance of your model.

ZEIT CONLINE DATENSCHUTZ

Partnersuche Immobilien Automarkt Jobs Reiseangebote

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE LEBENSART REISEN AUTO

Internet Datenschutz Mobil Games

Anmelden | Registrieren

Suchen

DATENSCHUTZ

SPORT

Tausende demonstrieren für **Bürgerrechte im Netz**

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.



In Berlin demonstrierten tausende Demonstranten für mehr Datenschutz

Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Motto "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

esteanton wandton siah untoe andoean soran dia Valleasählung Dia Da

DATUM 11.9.2010 - 17:20 Uhr

- QUELLE ZEIT ONLINE, AFP, dpa
- KOMMENTARE 4
- * EMPFEHLEN E-Mail verschicken | Facebook, Twitter, Buzz .
- ARTIKEL DRUCKEN Druckversion | PDF SCHLAGWORTE Datenschutz | Demonstration |
- Datensicherheit | Medienpolitik

NEU IM RESSORT

- I. DATENSCHUTZ Tausende demonstrieren für Bürgerrechte im Netz
- 2. FUTUREZONE Kurier darf ORF-Portal Futurezone kaufen 3. URHEBERRECHTE "Wir haben als Kind gelernt, Teilen ist
- gut" 1. SPAM AUF FACEBOOK "Ein iPad umsonst, ich halte es in den Händen
- 5. IPHONE & IPAD APPS Mehr Freiheit im App-Store

NEU AUF ZEIT ONLINE I. NACHRUF Die Freie - Bärbel Bohley ist tot

- 2. CDU Union streitet über konservatives Profil
- 3. GEDENKTAG 9/11 Obama warnt vor religiösen Ressentiments
- 1. BUNDESLIGA Der BVB und die Freiburger siegen 5. DAAD Das Ende der Åra Bode





ZEIT CONLINE | DATENSCHUTZ

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE LEBENSART REISEN AUTO

Internet Datenschutz Mobil Games

DATENSCHUTZ

SPORT

Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.



Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Motto "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

Die Domonstranten wandten eich unter anderem zoren die Vellezrählung

Anmelden | Registrieren

Partnersuche Immobilien Automarkt Jobs Reiseangebote

- KOMMENTARE 4
 EMPFEHLEN E-Mail verschicken | Facebook, Twitter,
 Buzz
- ARTIKEL DRUCKEN Druckversion | PDF
 SCHLAGWORTE Datenschutz | Demonstration |
 Datensicherheit | Medienpolitik

NEU IM RESSORT I. DATENSCHUTZ Tausende demonstrieren für Bürgerrechte

im Netz 2. FUTUREZONE Kurier darf ORF-Portal Futurezone kaufen 3. URHEBERRECHTE "Wir haben als Kind gelernt, Teilen ist

gut" 5. SPAM AUF FACEBOOK "Ein iPad umsonst, ich halte es in den Händen" 5. IPHONE & IPAD APPS Mehr Freiheit im Apo-Store

NEU AUF ZEIT ONLINE 1. NACHRUF Die Freie – Bärbel Bohley ist tot 2. CDU Union streitet über konservatives Profil 3. GEDENIKTAG 9/11 Oberna warnt vor religiösen Ressentiments

BUNDESLIGA Der BVB und die Freiburger siegen
 DAAD Das Ende der Ära Bode

experteer^{de}

ANZEIGE

• Remove noise.

ZEIT CONLINE | DATENSCHUTZ

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE LEBENSART REISEN AUTO

Internet | Datenschutz | Mobil | Games

DATENSCHUTZ

Tausende demonstrieren für Bürgerrechte im Netz

Für einen besseren Arbeitnehmerdatenschutz und gegen die Gesundheitskarte: 130 Organisationen hatten zur Demonstration aufgerufen. Sie fürchten den Überwachungsstaat.



Rund 7500 Demonstranten nahmen an dem Protestzug unter dem Motto "Freiheit statt Angst – Stoppt den Überwachungswahn" in Berlin teil.

Die Domonstranton unndten eich unter anderem zogen die Vellezählung

Anmelden | Registrieren

Partnersuche Immobilien Automarkt Jobs Reiseangebote

- EMPFEHLEN E-Mail verschicken | Facebook, Twitter, Buzz ...
 ARTIKEL DRUCKEN Druckversion | PDF
- SCHLAGWORTE Dataschutz | Demonstration |
 Datensicherheit | Medienpolitik
 NEU IM RESSORT

I. DATENSCHUTZ Tausende demonstrieren für Bürgerrechte

im Netz

FUTUREZONE Kurier darf ORF-Portal Futurezone kaufen

URHEBERRECHTE "Wir haben als Kind gelemt, Teilen ist

- gut" **§ SPAM AUF FACEBOOK** "Ein iPad umsonst, ich halte es in den Händen"
- 5. IPHONE & IPAD APPS Mehr Freiheit im App-Store

NEU AUF ZEIT ONLINE

NACHRUF Die Freie - Bärbel Bohley ist tot
 COU Union streitlet über konservatives Profil
 GEDENKTAG 9/11 Obama warnt vor religiösen
 Ressentiments
 BUNDESLIGA Der BVB und die Freiburger siegen

BUNDESLIGA Der BVB und die Freiburger siegen
 S. DAAD Das Ende der Åra Bode

ANZEIGE

• Remove noise.

• Convert text to vectors.

From texts to vectors

If we looked at two words only:








Binary bag of words

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Entry in vector is one, if word occurs in text.

$$b_{i,j} = \begin{cases} 1 \forall x_i \in d_j \\ 0 \, else \end{cases}$$

- Problem:
 - Number of word occurrences not accounted for.

Term Frequency

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Entry in vector equal to the words frequency.

$$b_{i,j} = n_{i,j}$$

- Problem:
 - Common words dominate vectors.

TF with stop wording

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the words frequency.

$$b_{i,j} = n_{i,j}$$

- Problem:
 - Common and uncommon words with same weight.

TF- IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.

$$b_{i,j} = n_{i,j} \times \log\left(\frac{|D|}{|[d:t_i \in d]|}\right)$$

- Problem:
 - Long texts get larger values.

Normalized TF- IDF

- Imagine a n-dimensional space.
- Each dimension = one possible word in texts.
- Filter stopwords.
- Entry in vector equal to the weighted frequency.
- Normalize vectors.

$$b_{i,j} = \frac{n_{i,j}}{\sum_{k} n_{k,j}} \times \log\left(\frac{|D|}{|[d:t_i \in d]|}\right)$$

- Problem:
 - Additional domain knowledge ignored.

Reality

- There are a few more words in news.
- Use all relevant features/ signals available.
 - Words.
 - Header fields.
 - Characteristics of publishing url.
 - •
- Usually pipeline of feature extractors.

From data to information.

Collect data and define your learning problem.

• Training a prediction model.

• Checking the performance of your model.

Algorithm choice

- Naive Bayes.
 - Cannot reliably indicate how certain its classification is.

Algorithm choice

- Naive Bayes.
 - Cannot reliably indicate how certain its classification is.
- Logistic Regression.
- Complement. NB.
- Random Forests.

Algorithm choice

- Do you
 - want to interpret the resulting model?
 - want to update the model in an online fashion?
- The data you are working with
 - lives in high-dim feature space but is sparse?
 - has features that might depend on each other?
 - has outliers?
 - has missing values?

From data to information.

Collect data and define your learning problem.

Data preparation.
Training a prediction model.

• Checking the performance of your model.





Goals

• Did I use the best model parameters?

• How well will my model perform in the wild?



Prepare data

Compute expected performance



• Use same data for training and testing.

- Problem:
 - Highly optimistic.
 - Model generalization unknown.



Model generalization unknown.



- Use just a fraction for training.
- Set some data aside for testing.

- Problems:
 - Pessimistic predictor: Not all data used for training.
 - Result may depend on which data was set aside.



- Partition your data into n fractions.
- Each fraction set aside for testing in turn.

- Problem:
 - Still a pessimistic predictor.





- Use just a fraction for training.
- Set some data aside for tuning and testing.

- Problems:
 - Highly optimistic.
 - Parameters manually tuned to testing data.



• Parameters manually tuned to testing data.



- Use just a fraction for training.
- Set some data aside for tuning.
- Set another set of data aside for testing.

- Problems:
 - Pretty pessimistic as not all data is used.
 - May depend on which data was set aside.

Performance Measures

Correct prediction: Green Correct prediction: Orange Model prediction: Orange Model prediction: Green

Accuracy

$ACC = \frac{true \ positive + true \ negative}{true \ positive + false \ positive + false \ negative + true \ negative}$

- Problems:
 - What if class distribution is skewed?



Precision/ Recall

 $Precision = \frac{true \ positive}{true \ positive + false \ positive}$

 $Recall = \frac{true \ positive}{true \ positive + false \ negative}$

- Problem:
 - Depends on decision threshold.

















AUC – area under ROC



From data to information.

Collect data and define your learning problem. Data preparation. Training a prediction model. Checking the performance of your model.

http://www.flickr.com/photos/generated/943078008/
ttp://www.flickr.com/photos/eschipul/4160817135/



What else does Mahout have to offer.

Identify dominant topics

• Given a dataset of texts, identify main topics.

Algorithms: Parallel LDA

- Examples:
 - Dominant topics in set of mails.
 - Identify news message categories.

Discover groups of items

Group items by similarity.



- Examples:
 - Group news articles by topic.
 - Find developers with similar interests.

Ú.K.	World
Top Stories World U.K. Business Sci/Tech Entertainment Sports Health Spotlight	Wideo: Too early for US to withdraw from Iraq You RT Al-Qaida linked group claims Baghdad attacks The Associated Press Alizacera.net - BBC News - Sky News - Washington Post - Wikipedia: 25 October 2009 Bashdad Humbings all 3,834 news articles > 1 Email this story
Most Popular All news Headlines Images	Times Online Obama vows no rush on Afghanistan BBC News - 3 hours ago US President Barack Obama has said he will "never rush" a decision to send more troops to Afghanistan, as he comes under pressure to set out a new policy. Video: Obama resists pressure on Afghan war strategy - 27 Oct 09 Image: Al Jazeera Obama refuses to rush troops decision ABC Online New York Times - Bouters India - The Associated Press - AFP all 1,665 news articles » Email this story
	 Times Online Karadzic court case due to resume BBC News - 1 hour ago The genocide and war crimes trial of former Bosnian Serb leader Radovan Karadzic is due to resume in The Hague, a day after it was adjourned. ✓ Video: Karadzic is a surrogate Milosevic in The Hague W RT Karadzic snubs his war crimes trialbut it will go ahead without him Mirror.co.uk guardian ee uk. New York Times - The Associated Press - Independent all 1.214 news articles » C Email this story

Recommendation mining.

• Collaborative filtering.



Show most relevant ads

clearasil"	
erwandte Su	chbegriffe: <mark>clearasil pickelstift, world of warcraft</mark> .
este Ergebniss	e Zurück Seit
	Clearasil 44161 Tiefenreinigung Antibakterielle Reinigungspads, 60e Neu kaufen: EUR 5,99 2 Angebote ab EUR 5,15 Lieferung bis Samstag, 22. März: Bestellen Sie innerhalb der nächsten 23 Stunden Kostenlose Lieferung möglich. ARARA (1) Drogerie & Bad: Alle 13 Artikel ansehen
	Clearasil Ultra Anti-Pickel Reinigungspads, 65 Stück von Clearasil (Ba Neu kaufen: EUR 7,99 Gewöhnlich versandfertig in 1 bis 3 Wochen. Kostenlose Lieferung möglich. Drogerie & Bad: Alle 13 Artikel ansehen
3.	World of WarCraft: Wrath of the Lich King (Add-on) von Vivendi Unive Vista / XP) Neu kaufen: EUR 39,99 Vorbestellbar

Vorbestellbar Kostenlose Lieferung möglich. Games: Alle Artikel ansehen

Show most relevant ads



Publisher: learn how customers can search inside this book. ~ <u>Otis Gospodnetic</u> (Author), <u>Erik Hatcher</u> (Author)

List Price: \$44.95

Price: \$29.67 & this item ships for FREE with Super Saver Shipping. Details You Save: \$15.28 (34%)

In Stock.

Ships from and sold by Amazon.com. Gift-wrap available.

23 new from \$19.99 15 used from \$17.11



Frequently Bought Together

Customers buy this book with Building Search Applications: Lucene, LingPipe, and Gate by Manu Konchady



Customers Who Bought This Item Also Bought



Search Server by David







Hibernate Search in Action by Emmanuel







Collective Intelligence in Action by Satnam Alag

Recommending places

http://www.flickr.com/photos/sebastian_bergmann/1244514498

http://www.flickr.com/photos/jfclere/4061801735













Thanks to Falko Menge for the pictures of Brussels.







Recommending people













Frequent pattern mining

 Given groups of items, find commonly cooccurring items.

- Examples:
 - In shopping carts find items bought together.
 - In query logs find queries issued in one session.



rypto/3201254932/sizes/l/

By libraryman, http://www.flickr.com/photos/libraryman/78337046/sizes/l/

By quinnanya, http://www.flickr.com/photos/quinnanya/2806883231/



By crypto, http://www.flickr.com/photos/crypto/3201254932/sizes/l/

By libraryman, http://www.flickr.com/photos/libraryman/78337046/sizes/l/

Requirements to get started

March 14, 2009 by Artful Magpie http://www.flickr.com/photos/kmtucker/3355551036/





• AWS	· Products	· Developers	🗸 Community	🗸 Support	·· Account				
Products & Services	Amazon Elastic Compute Cloud (Amazon EC2)								
Amazon EC2 Details	Amazon Elastic resizable comp	Compute Cloud (Amazon ute capacity in the cloud. I	provides Sign Up F	or Amazon EC2 🕠					
EC2 Overview	computing easi	computing easier for developers.							
FAQs	Amazon EC2's	simple web service interfac	ce allows you to obtain and						
Amazon EC2 SLA	configure capac	configure capacity with minimal friction. It provides you with complete							
 EC2 Instance Types 	computing env	obtain							

Amazon Elastic MapReduce

Amazon Elastic MapReduce is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3).

Using Amazon Elastic MapReduce, you can instantly provision as much or as little capacity as you like to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, machine learning, financial



(Thanks to Thilo for helping set up the cluster, Thanks to packet and masq for two of the three machines.)





Why go for Apache Mahout?

Jumpstart your project with proven code.

January 8, 2008 by dreizehn28 http://www.flickr.com/photos/1328/2176949559

Discuss ideas and problems online.

November 16, 2005 [phil h] http://www.flickr.com/photos/hi-phi/64055296





http://www.flickr.com/photos/whassupbud55/4581192168



Sebastian Schelter Jake Mannix Benson Margulies Robin Anil David Hall AbdelHakim Deneche Karl Wettin Sean Owen Grant Ingersoll Otis Gospodnetic Drew Farris Jeff Eastman Ted Dunning Isabel Drost



Become a committer: Of Apache Mahout









Emeritus:

Niranjan Balasubramanian Erik Hatcher Ozgur Yilmazel Dawid Weiss *-user@mahout.apache.org *-dev@mahout.apache.org



Image by: Patrick McEvoy

Interest in solving hard problems. Being part of lively community. Engineering best practices.

Bug reports, patches, features. Documentation, code, examples.

Thanks to Tim Lossen et. al for taking amazing pictures of the conf.

ho

10

Herzlich Willkommen!

l love lelvetic ٦ I can't recommend this conference enough. Top industry speakers, top developers and fantastic organisation. Mark this event on your sponsoring calendar!" - Scott Robinson, Senior Marketing Manager, neofonie GmbH

> Great variety of talks, smart people (speakers & audience), nice location!

Really good conference and very exciting to have so much solr and nosql and knowledge concentrated right on our door step. *Nokia*

Great to have this kind of conferences here in Berlin. Enjoyed to get a good and overview about the various NoSQL options.

Berlin Buzzwords

The conference gave me a good overview on all kinds of scalable open-source projects.

The Buzzwords conference last year put Berlin on the map as Europe's perhaps most important hub for startups and cutting edge web technology today. Already looking forward to the next! - Eric Wahlforss, Soundcloud

Search/ Store/ Scale

June 2011

I enjoyed it very much: Very good location, decent-sized auditoriums, very good wifi, practically all talks were very good: deep expertise and mostly very good presenting skills. I will definitely try to attend again if the event is continued next year. *Nokia*

Berlin Buzzwords 2010 was a great opportunity to showcase our initial release of Lily - a NoSQL content repository based on HBase and SOLR. The event organisation was top-notch: from badges to bag inserts, bannering, food, videotaping - the organisers went out of their way to accommodate both audience, speaker and sponsors in a highly professional way, while still achieving the easy-going, content-above-form atmosphere of a grassroots conference." *Steven Noels - managing partner - Outerthought*

Thanks to Tim Lossen et. al for taking amazing pictures of the conf.

*-user@mahout.apache.org *-dev@mahout.apache.org



Image by: Patrick McEvoy

Interest in solving hard problems. Being part of lively community. Engineering best practices.

Bug reports, patches, features. Documentation, code, examples.