

Twitter クライアント “Termtter” の紹介

と収集したソーシャルデータを Fluentd + Hadoop で分析する話

自己紹介

ソフトウェアエンジニア

渋谷家 / 平田ホスピタル 住人 (家がたくさんある)

<http://id774.net>

↑ここに色々書いてあるから読んで

ブログやツイッターのリンクは全部上から辿れるので省略

はじめに

Twitter クライアント作ったよ

Termtter

```
Terminal — ruby — 90x28
(16:10:43) [$by] daveohoots: Join the Owl Club as Community Manager for Europe & Asia #HootSuite #jobs #socmed #cmgr http://owl.li/9XYMf HootSuite
(16:11:26) [$bz] cesarmillan: Meet Beyonce,( Rescue!) the Coffee Cup-sized Puppy ... http://ow.ly/9XY0A web
(16:13:13) [$ca] debbiemillman: Emily Oberman shares examples from 17 years of Saturday Night Live opening titles http://t.co/4enZqyJ1 Twitter for iPhone
(16:14:55) [$cb] foodee: We just launched our Facebook Timeline page! Check it out and let us know what you'd like to see: https://www.facebook.com/Foodee HootSuite
(16:15:10) [$cc] brainpicker: Did you catch this? ▫ Physicist Brian Cox explains entropy and the arrow of time with sandcastles and glaciers http://j.mp/Hg7yrU Buffer
(16:17:37) [$cd] debbiemillman: Bill and Emily http://t.co/KPIl6EWd Twitter for iPhone
(16:19:25) [$ce] core77: Full house at the sappi fine paper talk w/ @kithinrichs @826_vale ncia @sampotts -- learning more abt special fx printing in 21st c. design Twitter for BlackBerry®
(16:20:07) [$cf] VIAwesome: Wait... @harvestunion is open?! Have you been? HootSuite
(16:21:14) [$cg] ujm: http://api.mongodb.org/scala/casbah/2.1.5.0/scaladoc/com/mongodb/casbah/Imports.html Termtter
(16:24:33) [$ch] IAMRUBY: RHCP = MY FAVE!! RT @katelesstodd: Tell me baby, what's your story? HootSuite
(16:24:47) [$ci] eyemagazine: RT @typofonderie: @typegirl So, in the 90s you had the designers and the super-strange? Will the latter do a comeback? #typewars Twitter for iPhone
(16:25:22) [$cj] abduzeedo: Video Inspiration 62 http://abduzeedo.com/video-inspiration-62 Twitter for Mac
(16:25:49) [$ck] eyemagazine: RT @CerealRecords: Drawings of Joost Swarte @ Martel Gallery, Paris http://www.domusweb.it/en/art/joost-swarte-total-swarte/ http://twitter.com/CerealRecords/status/185472716063772672/photo/1 via @Kerryrurcell Twitter for iPhone
(16:26:59) [$cl] core77: Superhero sketching for @SappiNA The Standard 5 with Kit Hinrichs @826national #graphicdesign http://yfrog.com/h0afjrjbtj Twitter for BlackBerry®
```

特徴

ターミナルから Twitter できる (べんり)

実装言語は Ruby

v1.0.0 リリースは 4 年前

2013 年 2 月からコミットに

最新版 v2.2.2 を 9/5 にリリース！

<http://blog.id774.net/post/2013/09/05/397/>

↑ここに色々書いてある

みんな使ってみてね！

termtter 2.2.2

Termtter is a terminal based Twitter client.

INSTALL > `gem install termtter`

 Download  Documentation  Badge  Subscribe  RSS

Authors

jugyo, ujihisa, koichiroo, id774

43,873 60
total downloads for this version

Owners



Licenses

MIT

おしまい

ご清聴ありがとうございました

補足

これだけじゃさみしいので...

ここから先はすべて「補足」です

Termtter + Fluentd

最新版 Termtter v2.2.2 は何がすごいのか

→ Fluentd と連携できるようになった

Fluentd ってなに

→ データをなんでも **JSON** 形式で扱うログ収集ツール

→ 入出力はプラグインで自由に拡張できる

→ 実装言語は Ruby

→ **とにかく最近流行ってる**

JSON ってなに

こういうやつ

```
{  
  "name": "山田太郎",  
  "age": 20  
}
```

JSON の特徴

JavaScript Object Notation

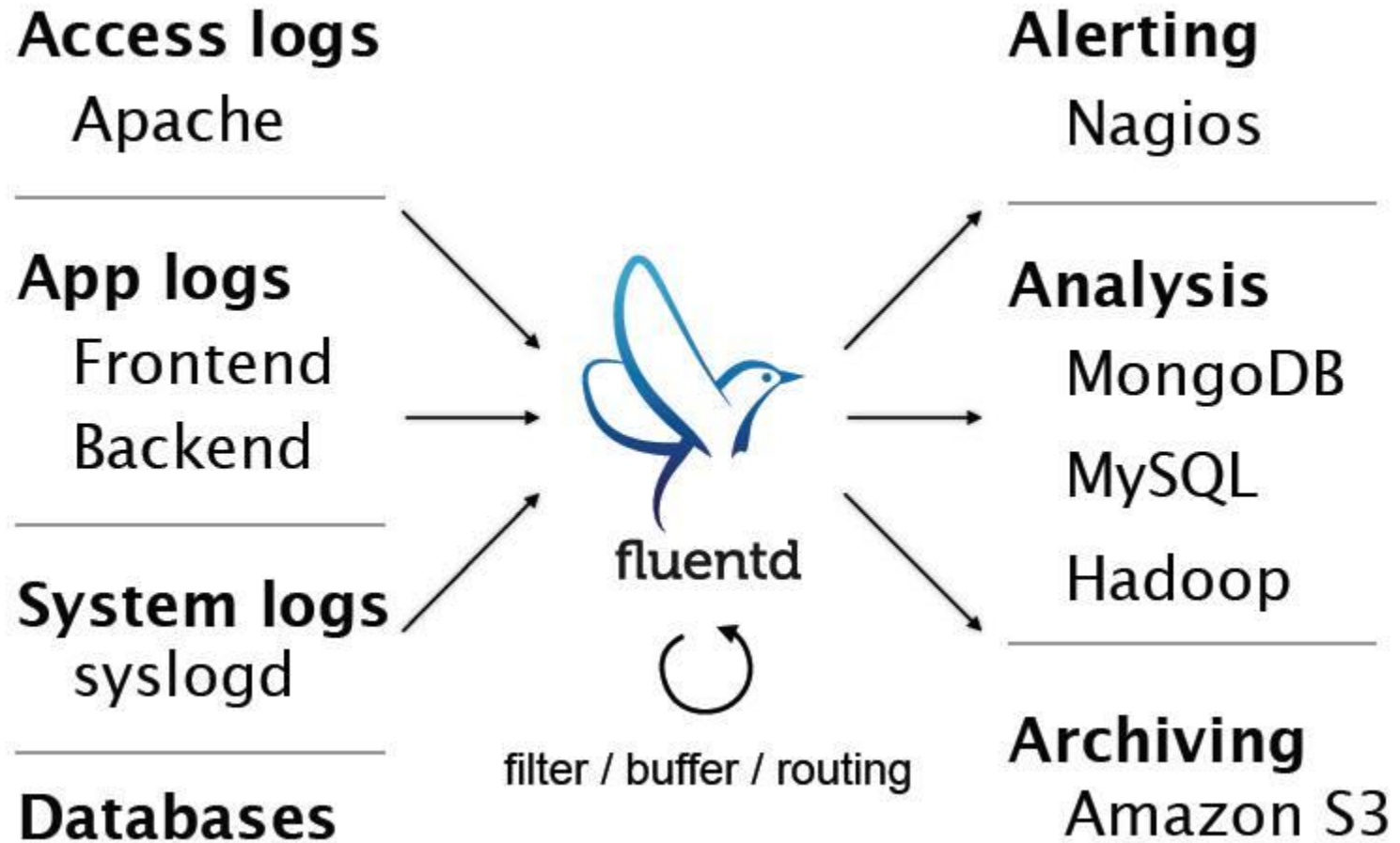
元々は JavaScript のサブセットだが言語を選ばない

キーと値のペアという形式

値に数値、文字列、配列、オブジェクト等の型を持てる

データの受け渡しとしてモダンな形式

Fluentd まじ便利



Fluentd の何がすごいのか

- ・プラグインで拡張できる

→ 新しい入出力方法に対応する場合もプラグインを書くだけで対応完了！

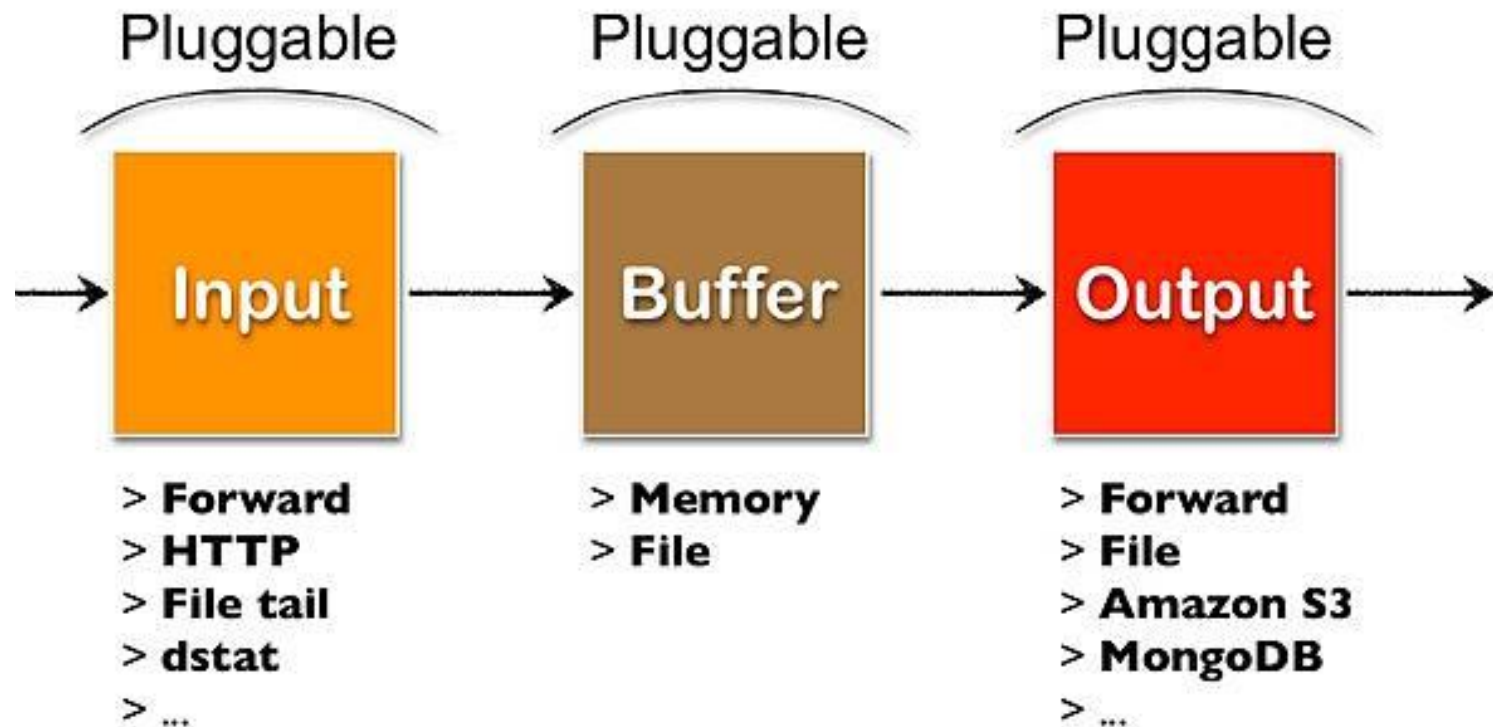
- ・なんでも JSON で扱う

→ どんな問題領域にも応用可能

とにかく Fluentd にデータを渡してしまえば良い！！！！

安心して思考停止できる

プラグインを書けば何でもできる



昔あったよね、そういうやつ...



引用元:

5分でネットがわかるシリーズ(6):一足早く「Plagger」の便利さを実感してみよう(2/5)

[江原顕雄, @IT]

http://www.atmarkit.co.jp/ait/articles/0611/21/news109_2.html

Plagger

- だいぶ前 (2006 年頃) 流行った
- Perl で実装されている
→ 多大な数の CPAN モジュールを利用する
- レシピと呼ばれる YAML ファイルで挙動を定義する
- Plagger というツールとしても有益であったが、何よりプラグブルなツールを作ると便利という見本になった

自分でも作った

Ruby General Automation Framework <http://blog.id774.net/post/2013/06/29/381/> — Edit

642 commits 2 branches 16 releases 8 contributors

branch: master automaticruby

This branch is 0 commits ahead and 0 commits behind master Pull Request Compare

Commit Message	Author	Latest Commit
358 Clean up Gemfile	id774 authored a month ago	latest commit 9bb5f0d28d
assets	236 Create new plugin Filter::FullFeed	8 months ago
bin	354 Fix footer	2 months ago
config	203 Refactoring	a year ago
db	5 Reconstruction the directory structure	2 years ago
doc	356 Version bump to 13.7.0	a month ago
lib	357 Version bump to the next release	a month ago
plugins	Merge pull request #49 from soramugi/chantoru-keywords	2 months ago
script	336 Bundle update in the CI script	2 months ago
spec	357 Version bump to the next release	a month ago
test	351 Clean Up	2 months ago
vendor	68 #40 Ruby Pattern like GitHub Codebase	2 years ago
gitignore	Add subscription-weather	4 months ago

SSH clone URL
git@github.com:auto

You can clone with HTTPS, SSH, Subversion, and other methods.

<https://github.com/automaticruby/automaticruby>

Automatic Ruby

- Plagger インスパイア
- 2012 年に **Ruby** で初めて実装
- 現時点の最新版 v13.7.0 (2013/07) リリース
- インターネットの情報を何でも集めてどこにでも出力

\$ gem install automatic
でインストールできます

Automatic Ruby の特徴

- ・フィードの配列をプラグイン間で受け渡す
- ・レシピと呼ばれる YAML ファイルで挙動を定義する
→ レシピの形式は Plagger とまったく同じ
- ・レシピの解釈は順列
→ Plagger と違い無限に YAML を順に解析し続ける
- ・たくさんのライブラリに依存するのは Plagger と一緒
→ Gemfile と bundler があるから管理がラクかな
- ・協力者募集してますので pull request 送ってね！

Fluentd と連携可能

automaticruby + fluentd でウェブのあらゆる情報を構造化データとして収集する

Automatic Ruby v13.6.0 を 6/28 にリリースした。

Fluentd との連携

今回のリリースの目玉としては `Publish::Fluentd` プラグインが標準添付となった。これにより Automatic Ruby で収集したあらゆるデータを `Fluentd` により構造化データとして扱うことができるようになる。ブログやウェブサイトの情報を集めて `Fluentd` 経由で `Hadoop` や `Amazon S3` などに収集、その後 `MapReduce` するなどしてデータを分析するといったことも自由にできるようになる。その際には現在まで利用してきたプラグインなどの資産をそのまま生かすことができる。

こんな感じのレシピで `Fluentd` にデータを出力できる。

```
plugins:
- module: SubscriptionFeed
  config:
    feeds:
      - http://blog.id774.net/post/feed/

- module: PublishFluentd
  config:
    host: localhost
    port: 9999
    tag: automatic.feed
```

<http://blog.id774.net/post/2013/06/29/381/>

Automatic Ruby まじ便利

Ruby で金髪美女のお宝画像をまとめてぶっこ抜き

さっそく実行

レシピを指定して Automatic Ruby を起動する。

```
automatic -c get_image.yml
```

するとどうだろう。君のデスクトップの上がたちまち金髪美女のあられもない画像で次々と埋め尽くされていくではないか！



よかったですね。

Tumblr の過去の画像までまとめてダウンロードする

自動で単純に Tumblr をスクレイピングするだけでは、まとめて大量にリブログされた場合に一定件数以上のものは流れてしまい拾えないことがある。更新件数の多い Tumblr の画像を抜け漏れなく入手するには、このプラグインで pages を 2 ~ 5 など適切な値に指定しておく和良好的。これで一度にたくさん画像が流れてきても逃さず入手できるようになる。



このプラグインを正式版に含めて Automatic Ruby v12.10 をリリースした。Ruby を使っている人は以下のコマンドで簡単にインストールできるので試してみると良いだろう。

```
gem install automatic
```

<http://blog.id774.net/post/2012/06/20/98/>

<http://blog.id774.net/post/2012/10/16/267/>

みんな使ってみてね！

automatic 13.7.0

Ruby General Automation Framework

INSTALL > `gem install automatic`

 Download  Documentation  Badge  Subscribe  RSS

Authors

id774

8,315 185
total downloads for this version

Owners



Licenses

GPL

おしまい

ご清聴ありがとうございました

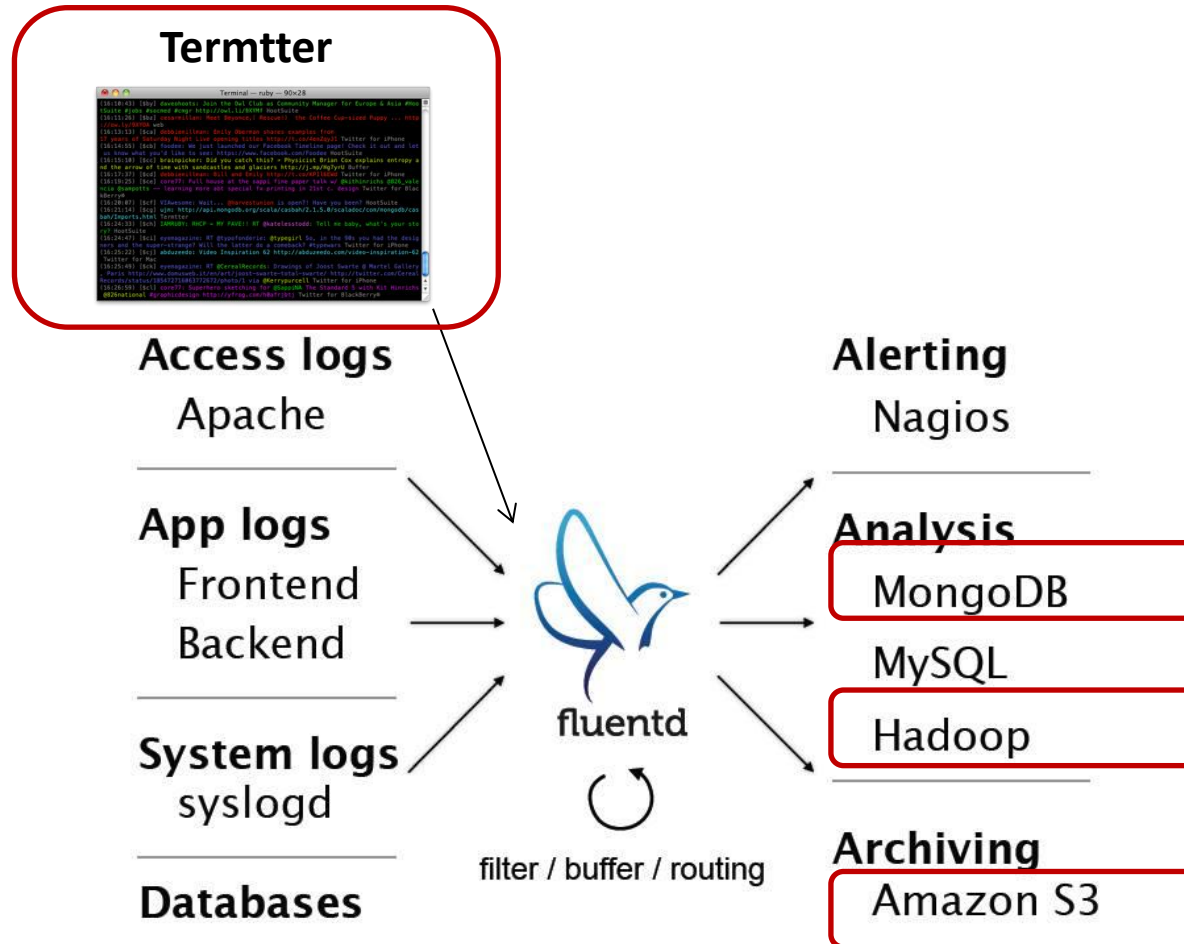
おしまいじゃないです

終わってないよ

話を元に戻すと...

収集したツイートを Fluentd で収集するという話

先ほどの絵で表すと



Fluentd 最強伝説

Fluentd にボタンさえ渡せばその先好きなようにできる

Amazon S3 にためる → 全ツイートを無限に永久保存

MongoDB にためる → いろんなアプリから利用する

Hadoop にためる → ソーシャルデータ分析

Hadoop に貯めてみた

対象 ... ここ約 3 年位のフォローしている人のツイート

期間 ... 2011 ~ 2013 年

フォロー数 ... 約 10,000 人 (1 分に 1 回程度クロール)

ツイート数 ... 約 3,666 万件

データサイズ ... 約 20GB

もっとクロールしまくればテラバイトやそれ以上の単位にもなるだろうけど取りあえず自分の観測範囲を分析

ここで MapReduce の基本方針

Key, Value のうち Value 部分に **JSON** 文字列を格納
→ 型付きのデータ構造をそのまま格納できる

Key 部分に Mapper/Reducer で利用する**キー**を格納
→ 一行ごとに JSON Parser を動かすのは非効率

Mapper を分散させて Reducer の数を **1 つ**に
→ 分散して処理 → 中間ソート後に集約するだけ

Hadoop Streaming を利用

- だって Ruby 使いたいよね

- 実行方法が面倒

```
hadoop jar ~/hadoop/contrib/hadoop-current-streaming.jar ¥  
-D hadoop.options ¥  
-input httpd_logs ¥  
-output logc_output ¥  
-mapper /home/who/work/hadoop/script/map.rb ¥  
-reducer /home/who/work/hadoop/script/reduce.rb ¥  
-inputformat TextInputFormat ¥  
-outputformat TextOutputFormat
```

Hadoop Streaming をラップするフレームワークを作成

→ 設定ファイルで挙動を一元管理、実行ログ出力、結果出力、ジョブ成否判定

Hadoop バッチの成否確認

- ・ジョブの先行後続関係

→ 出力ディレクトリの `_SUCCESS` の存在を確認する

```
hadoop fs -get $HDFS_OUT/_SUCCESS>>$JOBLOG 2>&1
```

```
test -f _SUCCESS && 成功時処理
```

Hadoop バッチの連結

(入力)

(出力)

/user/who/job1 → /user/who/job2 # job1

/user/who/job2 → /user/who/job3 # job2

/user/who/job3 → /user/who/out # job3

#!/bin/sh

/home/who/job1/bin/run > /home/who/job1/log/job.log

/home/who/job2/bin/run > /home/who/job2/log/job.log

/home/who/job3/bin/run > /home/who/job3/log/job.log

ワードカウント (単語を数える)

- ・最も基本的な MapReduce ジョブ

「艦隊」という単語の登場数

2012/04 ~ 2012/08 ... 338

2013/04 ~ 2013/08 ... 2039

去年のおよそ 6 倍に！！！！ 一体何が起こったのか...

(※ 艦これリリース = 2013/04)

ワードカウント (単語を数える)

「**駆逐**」という単語の登場数

2012/04 ~ 2012/08 ... **427**

2013/04 ~ 2013/08 ... **1666**

昨年のおよそ **4 倍**に！

調査結果を勝手に送り付ける事例 →



Fluentd + Hadoop まじ便利

- ・ひとつのツイートごとに特徴となる語彙を抽出
→ 特徴ベクトルとして JSON に配列を付与

後続ジョブで特徴ベクトルから

- 類似度を算出して似たユーザーを探す (クラスタリング)
- ネガ・ポジを判定してある映画の評判を探る (ベイズ分類)
- 特定の人発言からメンタル病み具合を診断 (線形判別)

いろんなことができる！！！！！！べんり！！！！！！1

みんな使ってみてね！

fluentd 0.10.38

Fluentd is an event collector system. It is a generalized version of syslogd, which handles JSON objects for its log messages

INSTALL > `gem install fluentd`

 Download  Documentation  Badge  Subscribe  RSS

Authors

Sadayuki Furuhashi

Owners



58,717 801
total downloads for this version

おしまい

ご清聴ありがとうございました

おしまいじゃないです

せつかなのでまだ続きます

巨大なデータの収集 → 機械学習

みたいなメチャクチャ流行ってる

機械学習 (Google Trends)

ウェブ検索の人気度: 機械学習。すべての国, 2004年 - 現在。

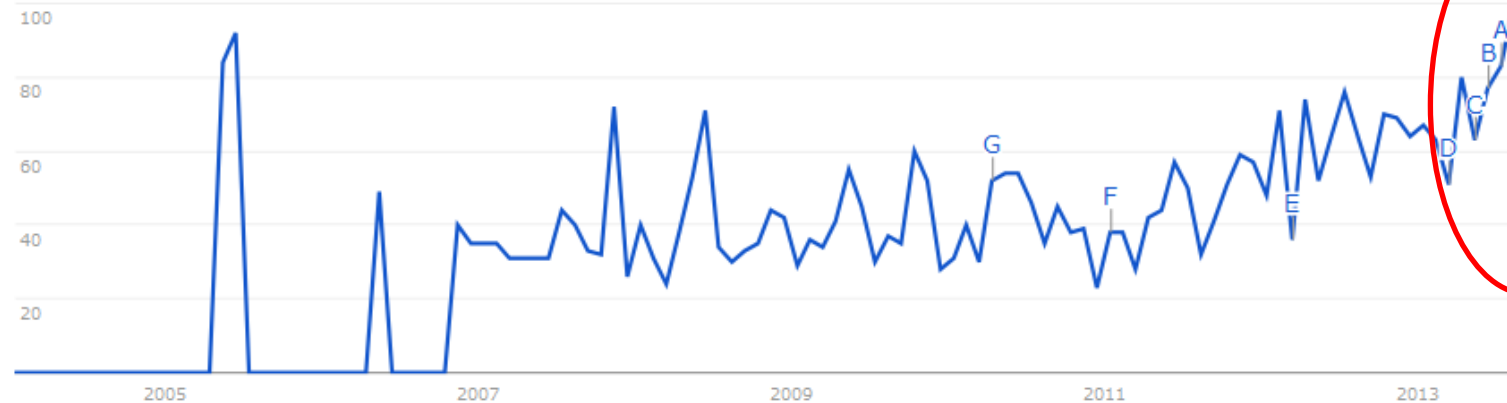


人気度の動向

100 が検索インタレストの最大値です

☒ ニュースのヘッドライン

☐ 予測 



埋め込む

エンジニア護送船団の予感！？！？

日経電子版 | 電子書籍 | Bizアカデミー | BizGate | 住宅 | レストラン | 転職 | 日経BP

日本経済新聞 8月12日 月曜日

English 中文

Web刊 | 速報 | ビジネスリーダー | マーケット | マネー | テクノロジー | ライフ | スポーツ | 朝刊

トップ | 特集 | コラム | 読者アンケート | 紙面連動 | 社説・書評 | 映画 | アジアBiz | ウーマン

ビッグデータ分析に人材の壁、25万人不足見通し

IT各社、育成へ本腰
2013/7/17 1:54

IT（情報技術）業界など産業界が注目する「ビッグデータ」ビジネスで、専門家不足が顕在化し始めた。データを分析する「データサイエンティスト」と呼ばれる人材で、現在日本には千人程度しかいないといわれる。政府は関連ビジネスの経済効果を7兆円超と試算するが、将来的には25万人が不足する見通しだ。危機感を募らせる業界は研究機関などと連携して人材育成へ本腰を入れ始めた。

ビッグデータは新聞約50億ページ分にも及ぶ大量のデータを指す。小売業界などが消費者の嗜好を精緻に分析してマーケティングや商品開発に生かそうとしたり、自動車各社が膨大な走行データの解析などを通してより安全でエネルギー効率の高いクルマの開発につなげようとしていたりするなど、ビッグデータ応用の動きが産業界全体に広がっている。

■産学連携で組織

だがここに来て、分析を担うデータサイエンティスト不足が顕著になっている。大学や大学院で統計学や数理学などを修めた人材で、豊富なIT知識も求められる。企業・組織の課題を的確に把握し、解決策を提案するコンサルティング能力なども不可欠だ。

データサイエンティスト IT各社の現状

アクセンチュア	データ分析専任部隊 300人のうち100人
日本ユニシス	50人。さらなる増員を検討
日立製作所	データ分析専任部隊 300人のうち80人
富士通	100人。2015年までに倍増させる
N E C	100人。2015年3月までに倍増させる
NTTデータ	データ分析専任部隊 120人のうち90人
SASインスティテュート ジャパン	100人。さらなる増員を検討

画像の拡大

©日本経済新聞

急に協会ができた



一般社団法人

データサイエンティスト協会

[📄 サイトマップ](#) [✉ お問い合わせ](#)

HOME

データサイエンティスト
協会とは

ニュース

活動内容

資料ダウンロード

データサイエンスに精通した
プロフェッショナルの育成を目指して

2013年7月16日スタート



データサイエンティスト協会について

当協会は、新しい職種であるデータサイエンティストに必要なスキル・知識を定義し、育成のカリキュ

メルマガ登録
MAIL MAGAZINE



機械学習の普及要因

タスクと手法の分離

→各タスク固有の問題を抽象化

学習方法とタスクを分離可能となった

文書
信号
画像
行動履歴
...等等

特徴を抽象化したデータ
(※分野に依存しない)

特徴ベクトル
(0,0,0,0,0,0,1,0,0,0,3,0 ...)
(1,0,1,0,0,0,0,0,-1,0,0,0 ...)
(0,1,0,0,0,2,0,0,0,0,0,0 ...)

グラフィカルモデル
(※ = 確率変数を頂点、変数間の
依存関係を枝としたグラフ構造)

+ - | - + - + - + - + - + - + | + -
| - - + - + - - - + - - - - - |

様々な手法や理論を適用

分類/回帰:
SVM (サポートベクトルマシン),
ナイーブベイズ, ロジスティック
回帰 ...

クラスタリング: K-means, MMC,
LSI, LDA, GM, ...

構造分析: HMM, MRF, CRF, ...

流行っているので作った

id774.net

Home

Apps

Blog

Tumblr

Twitter

GitHub

前の日

20130905

次の日

20130905 の注目キーワード (件数)

発表(116) 発売(55) 社会(55) 世界(48) 東京(47) 汚染(45) 韓国(42) 情報(41) 国際(37) 中国(36) 問題(35) 監督(34) 対応(30) 映画(29) 理由(29) 公開(28) 時間(27) 外子(27) 向け(26) 原発(26) 研究(26) 福島(25) 追害(25) 可能(24) 方法(24) 新聞(23) 招致(23) 宮崎(22) 最新(22) 端末(22) 科学(21) 女性(21) 開発(21) 無料(21) 市場(20) 決定(20) 人間(19) 機能(19) 簡単(18) 企業(18) 人生(18) 登場(18) 政治(18) 経済(18) 搭載(17) 利用(17) 教育(17) 環境(17) 海外(17) 投資(16) 逮捕(16) 提供(16) 竜巻(16) 相続(16) 終了(16) 国内(16) 差別(16) 対策(16) 時代(16) 導入(15) 関係(15) 報道(15) 男性(14) 開始(14) 連携(14) 大統領(14) 漏れ(14) 判断(14) 検索(14) 意味(14) 佐賀(14) 株式会社(14) 技術(13) 中村(13) 影響(13) 女子(13) 悲報(13) 進化(13) 日記(13) 漫画(13) 若者(13) 地下(13) 月額(12) 正式(12) 新型(12) 最高(12) 山形(12) 引退(12) 配信(12) 操作(12) 家族(12) 流出(12) 成功(12) 消滅(12) 以上(12) 腕時計(12) 調査(11) 販売(11) 山本(11) 現在(11) 効果(11) 判明(11) 注意(11) 代表(11) 強化(11) 雇用(11) 県立(10) 頭痛(10) 使用(10) 日テレ(10) 委員(10) 共同(10) 衝撃(10) 検討(10) 教科書(10) 会見(10) 作成(10) 支持(10) 未来(10) 介入(10) 議員(10) 作業(10) 移行(10) 政府(10) 発生(10) 最大(10) 起業(10) 衛星(10) 炎上(10) 誕生(10) 法則(10)

ID	タイトル	カテゴリ	SCORE
1	東京新聞: 違憲判断 婚外子相続差別 家族の多様化考慮 最高裁 : 社会(TOKYO Web)	電脳	500
2	Samsung、カメラ付き腕時計型端末「GALAXY Gear」を発表 日本でも10月に発売へ - ITmedia ニュース	電脳	414
3	サムスン、腕時計型端末を発表 日米では10月発売 : 日本経済新聞	電脳	410
4	ソニー、F2.0レンズ搭載、メタルフレームの「Xperia Z1」発表--国内でも発売予...	電脳	408
5	朝日新聞デジタル: 五輪招致、汚染水漏れで東京守勢 会見で「安心」空回り - 社会	エンタメ	408
6	PhotoshopとLightroomユーザー向けに月額1,000円の新たな提供方法を発表	電脳	400
7	日本でも発売予定: ソニーモバイル、Gレンズ+2070万画素カメラ搭載の「Xperia Z1」を発表 - ITmedia Mobile	電脳	396
8	No Sugar 〜私はどのようにしてRailsに貢献したか〜 - 2013/08/31関西ruby会議05発表資料	電脳	393
9	ネスレ日本、Android KitKatの国内発売を発表 juggly.cn	電脳	378
10	引退発表の宮崎駿監督『風立ちぬ』動員700万人突破でV7! - シネマトゥデイ	エンタメ	376

<http://newscloud.id774.net/newscloud>

ニュースクラウド

- 1) Automatic Ruby でニュースを収集
- 2) Fluentd で JSON 形式に変換して蓄積
- 3) Web アプリ (Rails) から参照可能に

ニュースランキング → 人気語彙の登場順にソート
ニュースツリー → 似たニュースを近い位置に配置

※ どのへんがクラウドなのかはよくわからない

どのへんが機械学習なの

id774.net

Home

Apps

Blog

Tumblr

Twitter

GitHub

ワードカウントの結果
(10件以上)

前の日

20130905

次の日

20130905 の注目キーワード (件数)

発表(116) 発売(55) 社会(55) 世界(48) 東京(47) 汚染(45) 韓国(42) 情報(41) 国際(37) 中国(36) 問題(35) 監督(34) 対応(30) 映画(29) 理由(29) 公開(28) 時間(27) 外子(27) 向け(26) 原発(26) 研究(26) 福島(25) 違憲(25) 可能(24) 方法(24) 新聞(23) 招致(23) 宮崎(22) 最新(22) 端末(22) 科学(21) 女性(21) 開発(21) 無料(21) 市場(20) 決定(20) 人間(19) 機能(19) 簡単(18) 企業(18) 人生(18) 登場(18) 政治(18) 経済(18) 搭載(17) 利用(17) 教育(17) 環境(17) 海外(17) 投資(16) 逮捕(16) 提供(16) 竜巻(16) 相続(16) 終了(16) 国内(16) 差別(16) 対策(16) 時代(16) 導入(15) 関係(15) 報道(15) 男性(14) 開始(14) 連携(14) 大統領(14) 漏れ(14) 判断(14) 検索(14) 意味(14) 佐賀(14) 株式会社(14) 技術(13) 中村(13) 影響(13) 女子(13) 悲報(13) 進化(13) 日記(13) 漫画(13) 若者(13) 地下(13) 月額(12) 正式(12) 新型(12) 最高(12) 山形(12) 引退(12) 配信(12) 操作(12) 家族(12) 流出(12) 成功(12) 消滅(12) 以上(12) 腕時計(12) 調査(11) 販売(11) 山本(11) 現在(11) 効果(11) 判明(11) 注意(11) 代表(11) 強化(11) 雇用(11) 県立(10) 頭痛(10) 使用(10) 日テレ(10) 委員(10) 共同(10) 衝撃(10) 検討(10) 教科書(10) 会見(10) 作成(10) 支持(10) 未来(10) 介入(10) 議員(10) 作業(10) 移行(10) 政府(10) 発生(10) 最大(10) 起業(10) 衛星(10) 炎上(10) 誕生(10) 法則(10)

10

タイトル

- 1 東京新聞: 違憲判断 婚外子相続差別 家族の多様化考慮 最高裁 :社会(TOKYO Web)
- 2 Samsung、カメラ付き腕時計型端末「GALAXY Gear」を発表 日本でも10月に発売へ - ITmedia ニュース
- 3 サムスン、腕時計型端末を発表 日米では10月発売 :日本経済新聞
- 4 ソニー、F2.0レンズ搭載、メタルフレームの「Xperia Z1」発表--国内でも発売予...
- 5 朝日新聞デジタル: 五輪招致、汚染水漏れで東京市勢 会見で「安心」空間の 社会
- 6 PhotoshopとLightroomユーザー向けに月額1,000円の新たな提供方法を発表
- 7 日本でも発売予定: ソニーモバイル、Gレンズ+2070万画素カメラ搭載の「Xperia Z1」を発表 - ITmedia Mobile
- 8 No Sugar 〜私はどのようにしてRailsに貢献したか〜 - 2013/08/31関西ruby会議発表資料
- 9 ネスレ日本、Android KitKatの国内発売を発表 | juggly.cn
- 10 引退発表の宮崎駿監督『風立ちぬ』動員700万人突破でV7! - シネマトゥデイ

ニュースとそのリンク

ここ

スコアとカテゴリを表示

カテゴリ SCORE

パソコン	500
パソコン	414
パソコン	410
パソコン	408
エンタメ	408
パソコン	400
パソコン	396
パソコン	393
パソコン	378
エンタメ	376

ナイーブベイズ

カテゴリについてはナイーブベイズ分類器で判定

(例)

野球、サッカー、オリンピックの話題 → スポーツ

参院選、婚外子相続の話題 → 政治

iPS 細胞の話題 → 科学

殺人事件、放火、家宅搜索など → 社会

ナイーブベイズ分類器

ベイズの定理に基づいた教師あり学習

$P(B)$ = 事象 B が発生する確率 (事前確率)

$P(B|A)$ = 事象 A 発生後の事象 B の確率 (事後確率)

$P(A) > 0$ のとき以下の公式が成立する

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

単独でも試せる

id774.net

Home

Apps

Blog

Tumblr

Twitter

GitHub

カテゴリ判定器

カテゴリ判定をするテキストを入力してください

送信する

項目	内容
カテゴリ分類	
社会	
政治	
国際	
経済	
電腦	
スポーツ	
エンタメ	
科学	

実装の説明

テキスト一覧を見る

<http://classify.id774.net/classify/>

ライブラリ作ったのでよろしく

naivebayes 0.0.2

Naive Bayes classifier

INSTALL > `gem install naivebayes`

[Download](#) [Documentation](#) [Badge](#) [Subscribe](#) [RSS](#)

Authors
id774

Owners


Licenses
GPL

361 **202**
total downloads for this version

ニュースツリー

id774.net

Home

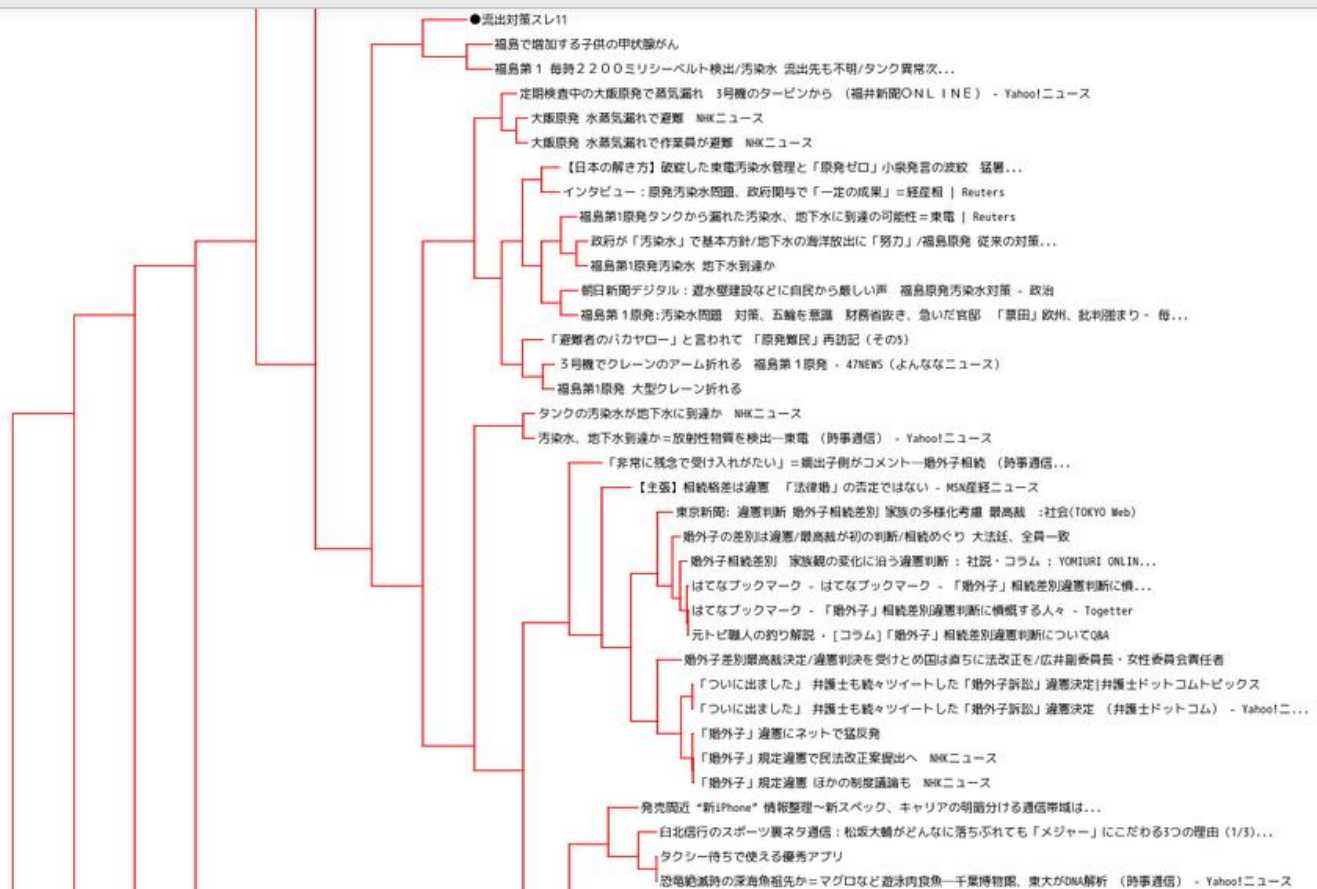
Apps

Blog

Tumblr

Twitter

GitHub



ニュースツリー(似たニュースを近くに)

id774.net

Home

Apps

Blog

Tumblr

Twitter

GitHub

原発関連

婚外子関連

●高圧対策スレ11

福島で増加する子供の甲状腺がん

福島第1 毎時2200ミリシーベルト検出/汚染水 流出先も不明/タンク異常次...

定期検査中の大飯原発で蒸気漏れ 3号機のタービンから (福井新聞ONLINE) - Yahoo!ニュース

大飯原発 水蒸気漏れで避難 NHKニュース

大飯原発 水蒸気漏れで作業員が避難 NHKニュース

【日本の解き方】破綻した東電汚染水管理と「原発ゼロ」小泉発言の波紋 猛暑...

インタビュー：原発汚染水問題、政府関係で「一定の成果」=経産相 | Reuters

福島第1原発タンクから漏れた汚染水、地下水に到達の可能性=東電 | Reuters

政府が「汚染水」で基本方針/地下水の海洋放出に「努力」/福島原発 従来対策...

福島第1原発汚染水 地下水到達か

朝日新聞デジタル：遊水壁建設などに自民から厳しい声 福島原発汚染水対策 - 政治

福島第1原発：汚染水問題 対策、五輪を意図 財務省抜き、急いだ官邸 「栗田」欧州、批判強まり - 毎...

「避難者のバカヤロー」と言われて 「原発難民」再訪記 (その3)

3号機でクレーンのアーム折れる 福島第1原発 - 47NEWS (よんななニュース)

福島第1原発 大型クレーン折れる

タンクの汚染水が地下水に到達か NHKニュース

汚染水、地下水到達か=放射性物質を検出=東電 (時事通信) - Yahoo!ニュース

「非常に残念で受け入れがたい」=輸出手前のコスト=婚外子相続 (時事通信)...

【主張】相続格差は違憲 「法律婚」の否定ではない - MSN経産ニュース

東京新聞：違憲判断 婚外子相続差別 家族の多様化考慮 最高裁：社会(TOKYO Web)

婚外子の差別は違憲/最高裁が初の判断/相続めぐり 大法院、全員一致

婚外子相続差別 家族観の変化に沿う違憲判断：社説・コラム：YOMIURI ONLINE...

はてなブックマーク - はてなブックマーク - 「婚外子」相続差別違憲判断に機...

はてなブックマーク - 「婚外子」相続差別違憲判断に憤慨する人々 - Together

元トビ職人の約9割が「コラム」「婚外子」相続差別違憲判断についてQ&A

婚外子差別最高裁決定/違憲判決を受けとめ国は直ちに法改正を/広井副委員長、女性委員会責任者

「ついに出来ました」 弁護士も続々ツイートした「婚外子訴訟」違憲決定|弁護士ドットコムトピックス

「ついに出来ました」 弁護士も続々ツイートした「婚外子訴訟」違憲決定 (弁護士ドットコム) - Yahoo!ニ...

「婚外子」違憲にネットで猛反発

「婚外子」規定違憲で民法改正案提出へ NHKニュース

「婚外子」規定違憲 ほかの制度議論も NHKニュース

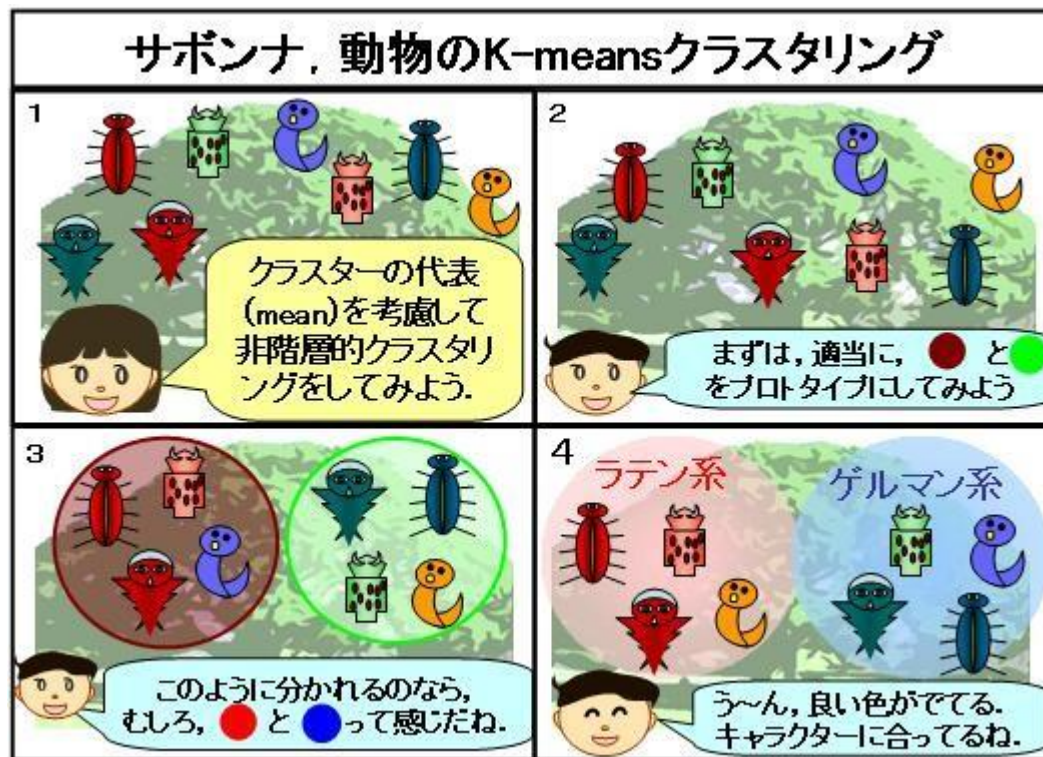
発売間近「新iPhone」情報整理~新スペック、キャリアの朝顔分ける通信帯域は...

白北信行のスポーツ裏ネタ通信：松坂大輔がどんなに落ちぶれても「メジャー」にこだわる3つの理由 (1/3)...

タクシー待ちで使える優秀アプリ

恐竜絶滅時の深海魚祖先が=マグロなど遊泳肉食魚=千葉博物館、東大がDNA解析 (時事通信) - Yahoo!ニュース

K 平均法クラスタリング

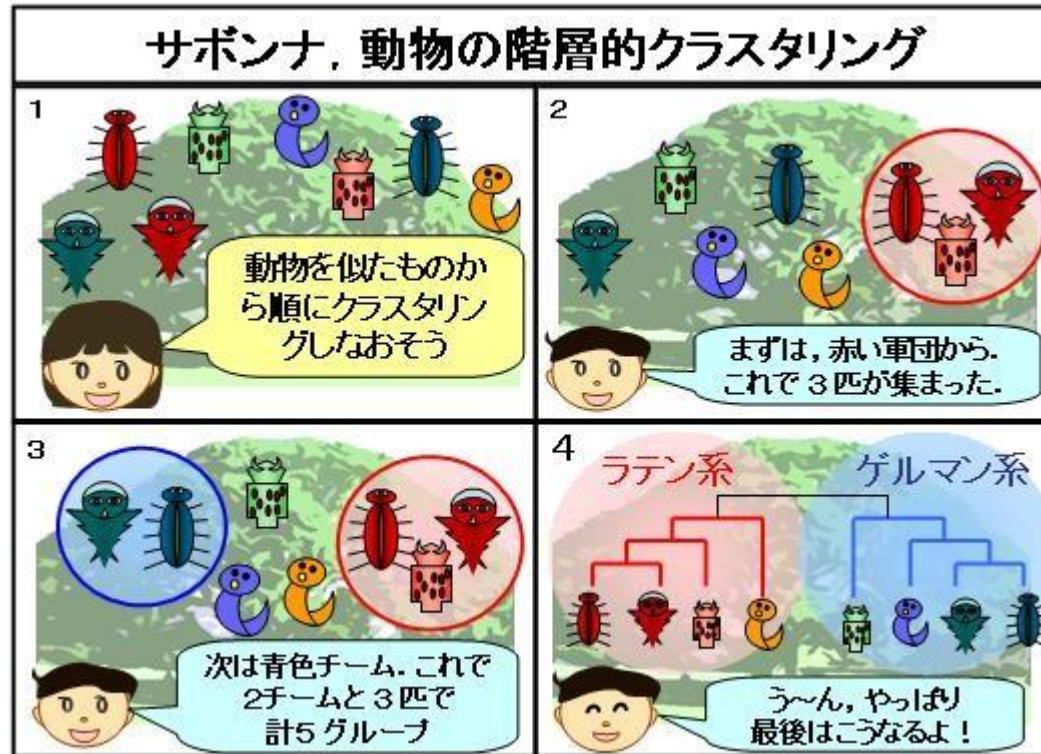


引用元:

大阪府立大学大学院工学研究科 電気・情報系専攻知能情報工学分野 教授
本多克宏氏のホームページ

<http://www.cs.osakafu-u.ac.jp/hi/honda/k-means.htm>

階層型クラスタリング



引用元：

大阪府立大学大学院工学研究科 電気・情報系専攻知能情報工学分野 教授
本多克宏氏のホームページ

<http://www.cs.osakafu-u.ac.jp/hi/honda/k-means.htm>

デンドログラム (樹状図) の描画

- 1) ニュース内の特徴語彙を形態素解析にて抽出
- 2) 語彙全体内における語彙の分布マップを生成
- 3) 分布をベクトルとしたピアソン積率相関係数を算出
- 4) 類似度の近いニュースを階層型クラスタリング
- 5) 生成されたクラスタをもとにデンドログラムを生成

とってもかんたん！！！！1

ライブラリ作ったのでよろしく

kmeans 0.1.0

K-means clustering

INSTALL > `gem install kmeans`

[Download](#) [Documentation](#) [Badge](#) [Subscribe](#) [RSS](#)

Authors
id774

Owners


Licenses
GPL

3,594 207
total downloads for this version

まとめ

だいぶ話がそれたので話を戻すと

- 1) Termtter や Automatic Ruby から Fluentd を経由してデータを収集するとあれこれ分析できてべんり
- 2) 機械学習はべんり
- 3) せっかく Hadoop にデータをためたのだから色々やりたいよね → 夢ひろがりまくりんぐ

さいごに

今日紹介したライブラリ、ウェブサービスの類はすべてホームページから辿れるので参照してみてください

<http://id774.net/>

またブログの個別の記事にも詳しい説明があります

おしまい

今度こそ本当におしまいです
ご清聴ありがとうございました