Ruby World Conference 2021

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

TWtrends：ツイッターでの話題を可視化するシステム

中央大学 iTL 飯尾淳

自己紹介

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- 中央大学 国際情報学部（iTL） 教授
- 特定非営利活動法人 人間中心設計推進機構 理事
- 一般社団法人 ことばのまなび工房 理事
  - 博士（工学）技術士（情報工学部門）
  - 人間中心設計推進機構認定 人間中心設計専門家
- 経歴
  - 1994年4月
    ～2013年3月　株式会社三菱総合研究所勤務
  - 2013年4月～　中央大学文学部社会情報学専攻
  - 2019年4月～　現職
- 専門分野
  - システムと人間のインタラクションに関する研究に従事
    - ユーザインタフェース，感性情報学，画像情報処理，ソフトウェア工学，行動情報分析，など

TWtrends

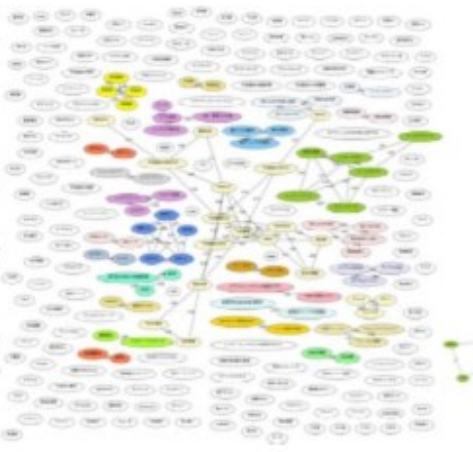
CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- TWtrends（日本語名「ツイトレ」）とは？
  - Twitterから世相を斬るためのシステム
- 2019年1月1日より稼働
  - https://twt.iiojun.com/
- （事情により，2021年1月～8月まで休止）

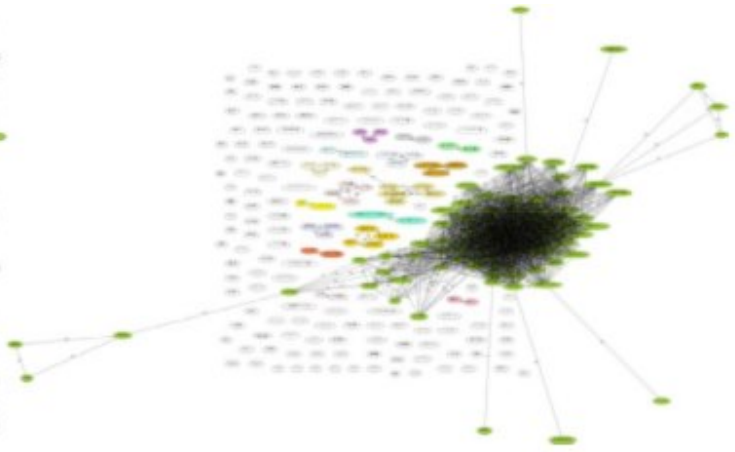
世相を反映？たとえばコレ

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

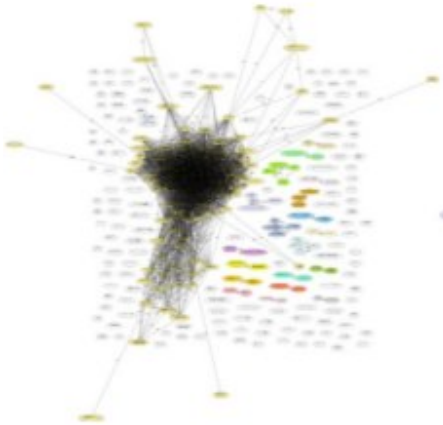
4月28日



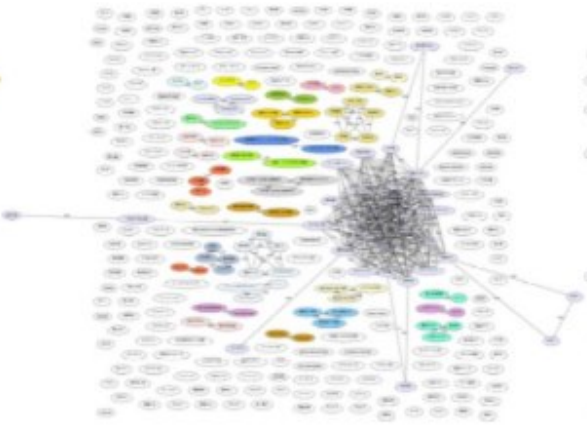
4月29日



4月30日

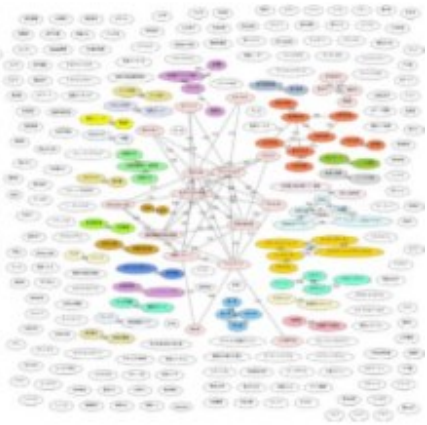


5月1日



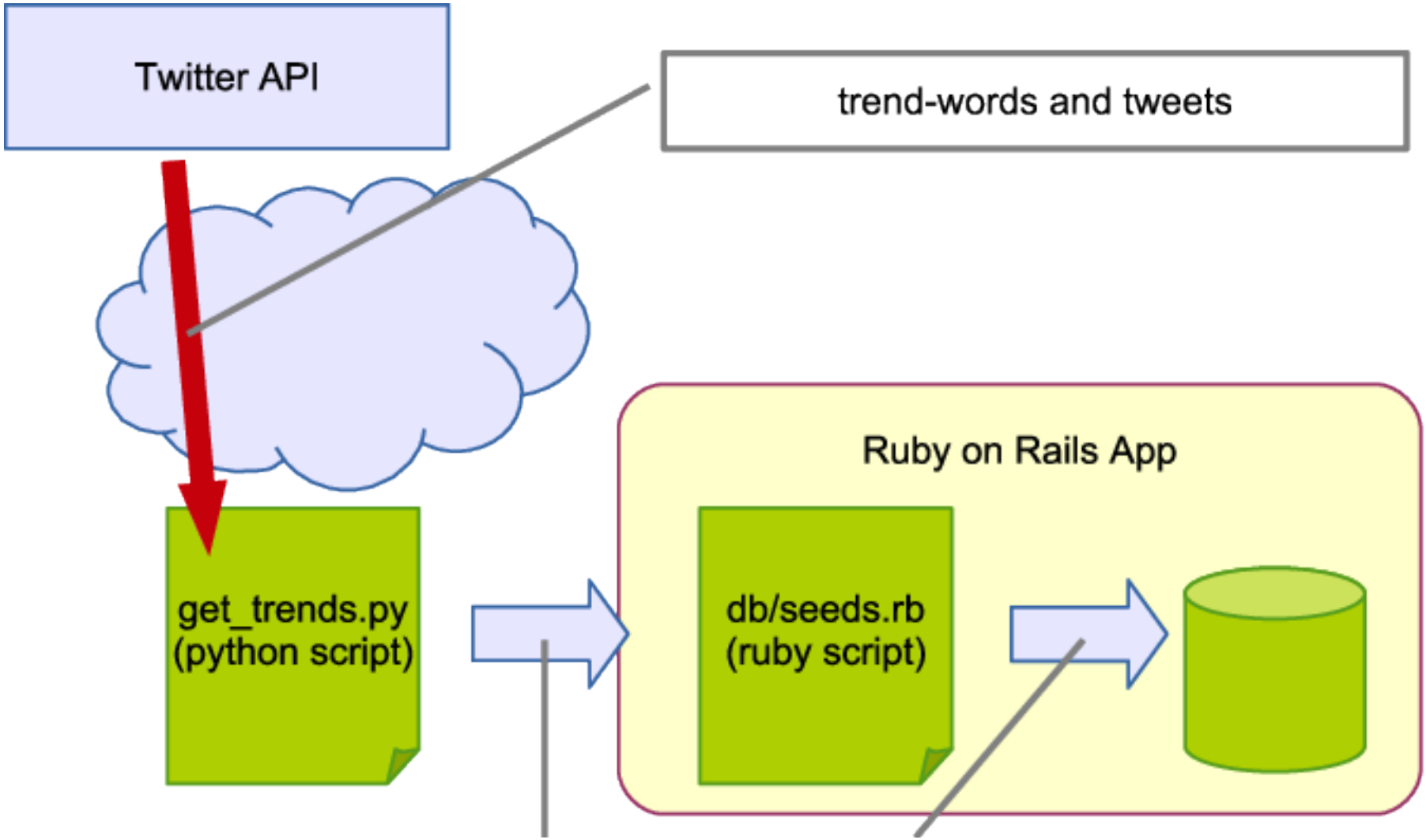
5月2日

5月3日



5月4日

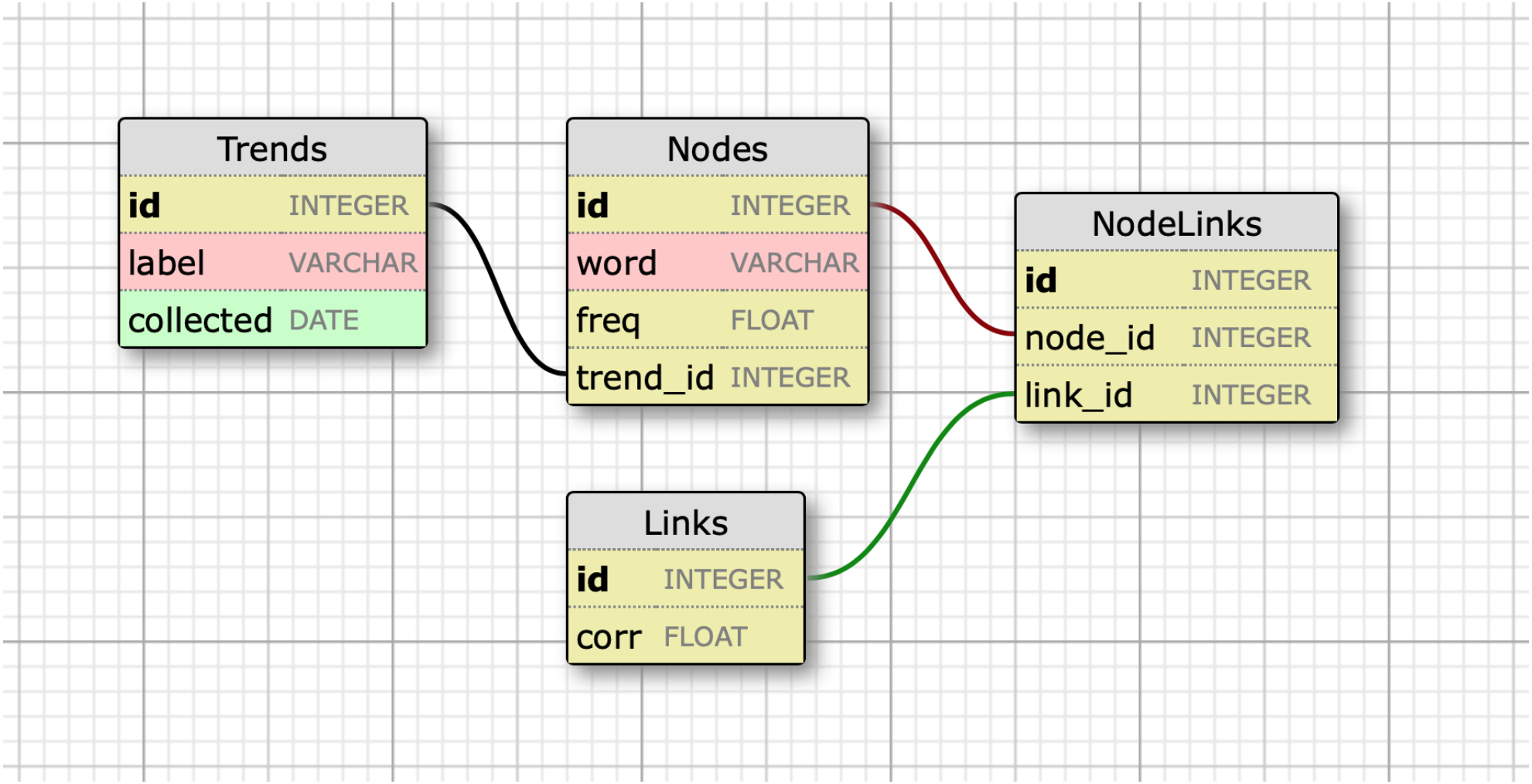
システムの概要



20分ごとにデータを収集し，データベースをアップデートする

データベースのテーブル構成

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS



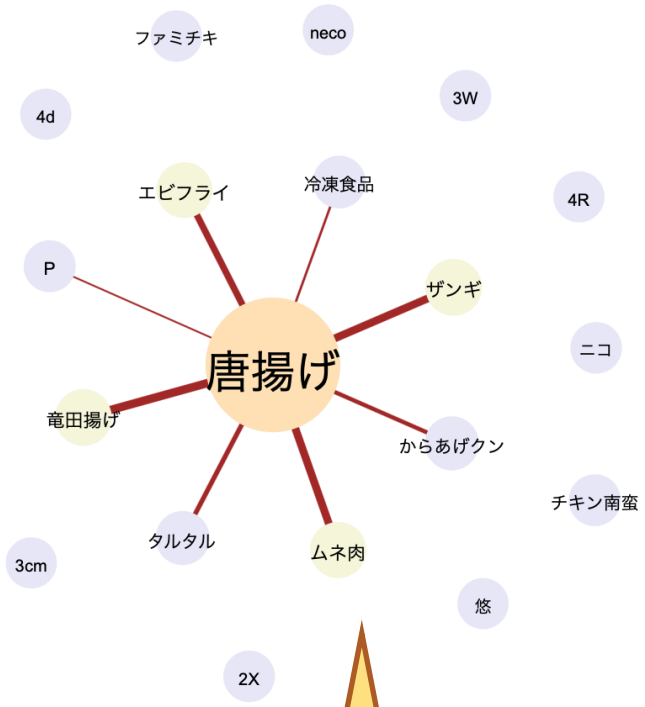
これどうやって作ってるの？

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

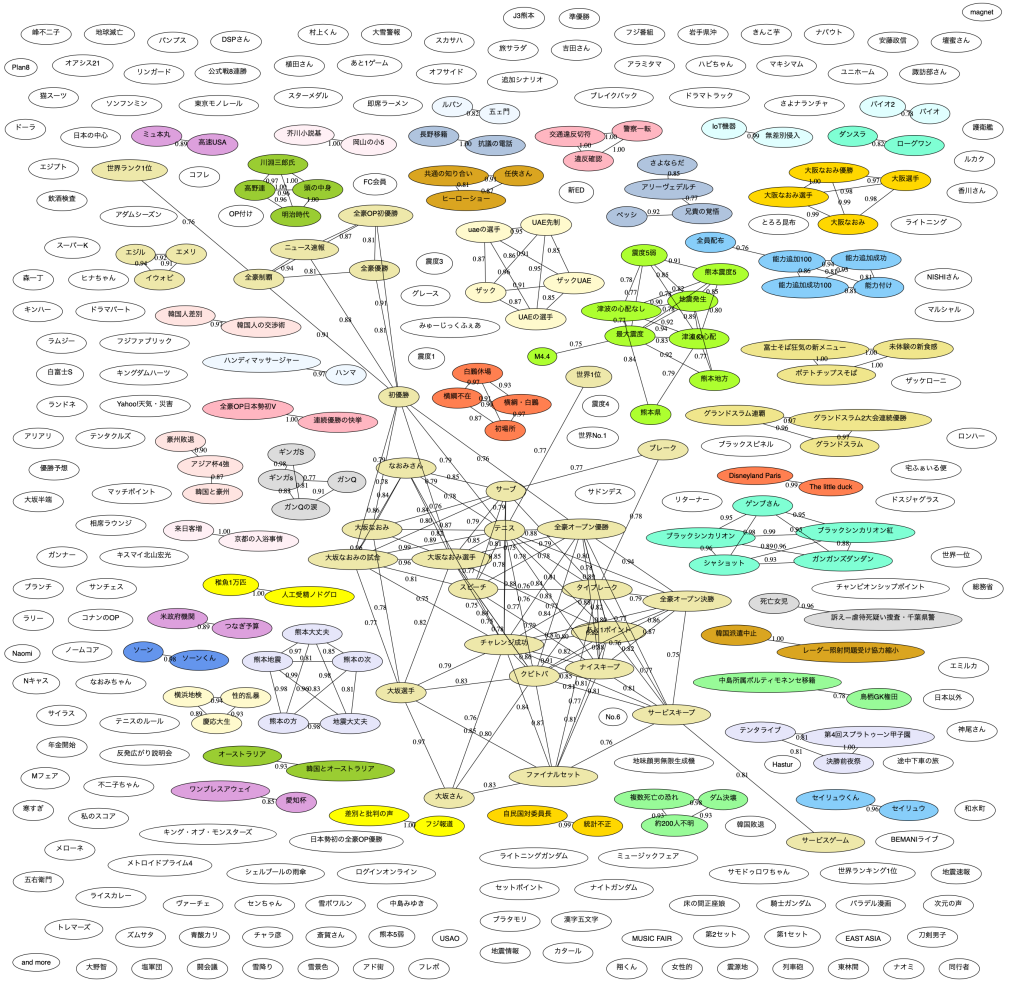
共起ネットワークグラフ

タルタルソース

取得日: 2019年09月26日 ← →



トピックマップ



このような，丸（ノード）を線（エッジ）で結んだ図形のことを「グラフ」という

TWtrendsで使われている技術

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- 形態素解析
  - → 日本語のテキストを分析するためには必須
- 共起ネットワーク分析
  - → 数学（確率統計）の話．さほど難しくない
- コサイン類似度
  - → 数学（線形代数）の話．
    でも（本来は）高校生が理解できるレベル

システムの動作の流れ

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- ツイッターのトレンドを取ってくる
  - Trend API を使う．東京のトレンドを取得
- 得られたトレンドをキーにして，関連ツイートを取得する
  - Standard Search APIを使う．リツイートは排除
- 集めたツイートに基づき共起ネットワーク図を作る
  - 共起ネットワーク図はそれぞれのトレンドごとに作成される
- 1日の関連データが集まったら，トピックマップを作る
  - トピックマップは1日ごとに作成（各トレンドの関連性を示すグラフだから）

トレンドごとの可視化

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・おおまかな流れ

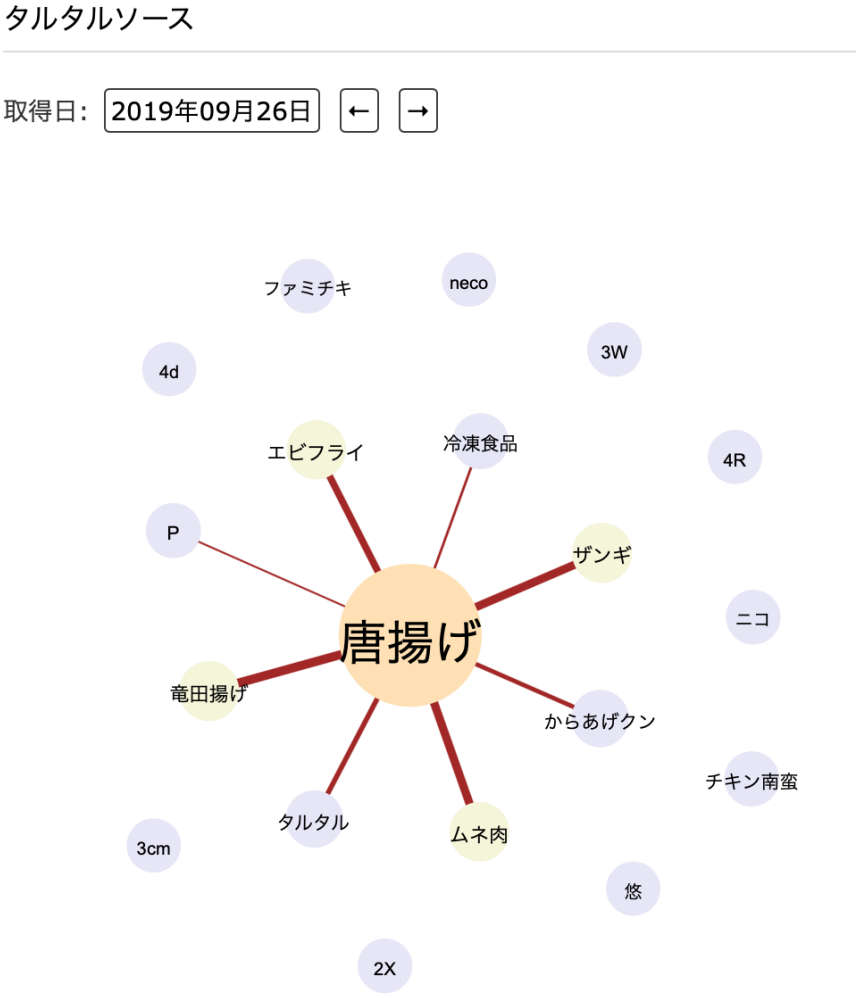
1.ツイート100個取得

2.ツイートを形態素解析でバラバラにする

3.固有名詞を抽出・数え上げる → 出現頻度

4.共起分析する

→ 共起確率



共起ネットワークによる
トレンドの可視化

形態素解析

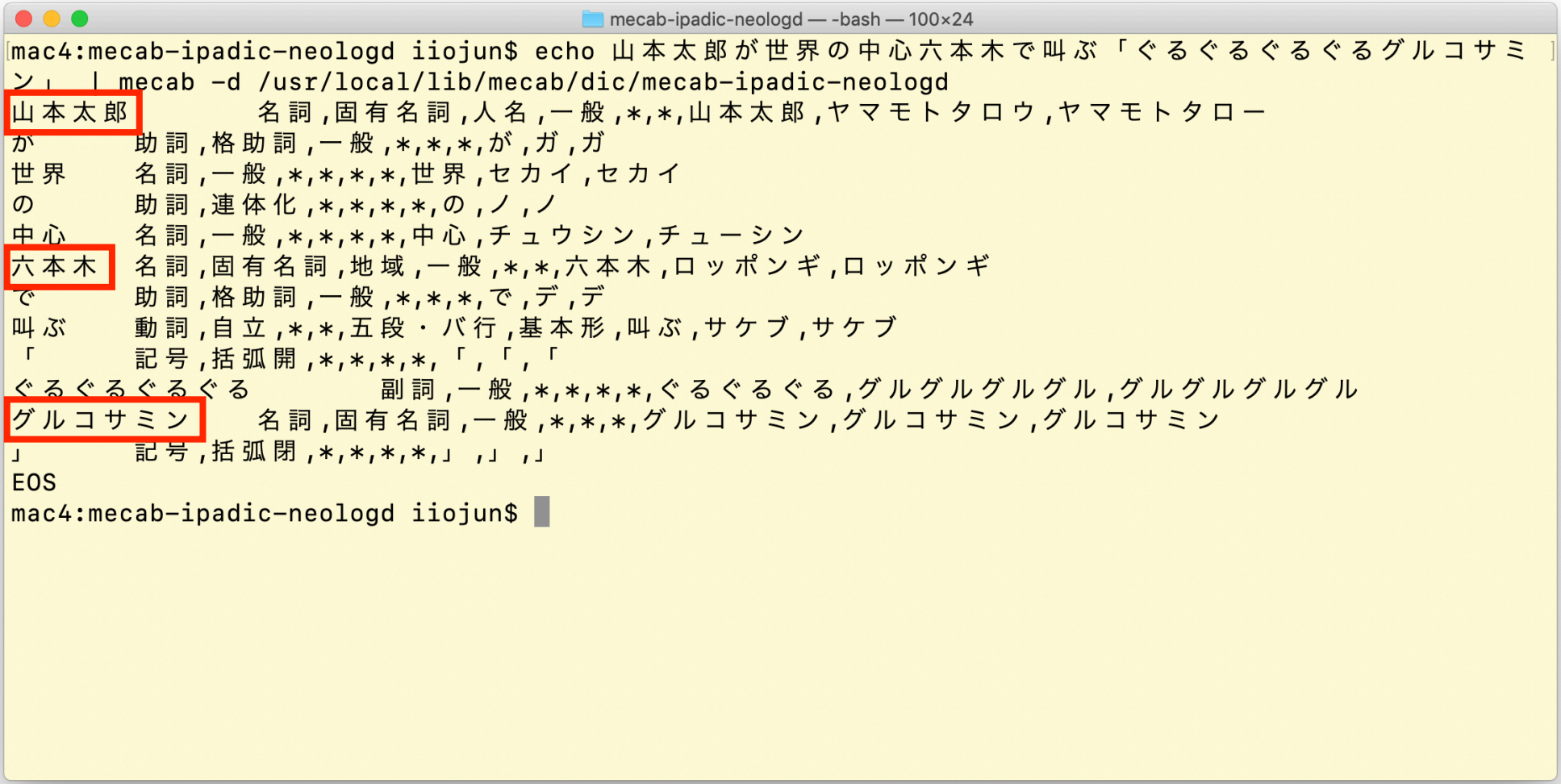
CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- 英文は簡単に分割可能（スペースでぶった切ればいいから）
  “A quick brown fox jumps over the lazy dog.”
- 日本語はそうはいかない
  - 「庭には二羽鶏がいます」

  - 「庭　に　は　二羽　鶏　が　い　ます」
- MeCab + NEologd:
  - 日本語の形態素解析器としてはスタンダード
  - 最近はよい辞書（NEologd）が出てきていい感じの分析ができるようになった

固有名詞抽出の例

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS



単語の数え上げ

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・抽出した固有名詞が出現する回数を数え上げる

・ある単語 t_i が取得したツイート全てに出現した回数を N_i，全ての単語に関する総出現回数を N_{all} とすると，その単語 t_i に関する出現頻度 Tf_i は以下で計算される（Tf …… Term frequency）

$$Tf_i = \frac{N_i}{N_{all}},$$ （なお $N_{all} = \Sigma_i N_i$ であり，$Tf_{all} = 1.0$ となる）

※ 実際には，$\bar{Tf}_i = \frac{Tf_i}{\max_j Tf_j} * 100.0$ として，正規化処理している

（各トレンドごとの出現頻度の多寡を一定にしたかったため）

数え上げの具体例

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・例文

- 長野の林檎は青森の林檎よりおいしいと信州人は信じている
- 青森の林檎は生食用よりジュースにされることが多いらしい
- ところで長野は林檎だけじゃなくて蕎麦もおいしい
- 江戸前の蕎麦もよいが，東京の蕎麦はちょびっとで高価い
- 東京から長野まではいまや新幹線ですぐに行ける

単語	出現回数	出現頻度
長野	3	0.158
青森	2	0.105
林檎	4	0.211
信州人	1	0.053
生食用	1	0.053
ジュース	1	0.053
蕎麦	3	0.158
江戸前	1	0.053
東京	2	0.105
新幹線	1	0.053

$$N_{all} = 19$$

（参考）Tf・Idf 値

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・一般に，自然言語処理を行う際には，Tf・Idf と呼ばれる値が使われることが多い

  ・Tf…… Term frequency

  ・Idf…… Inverse document frequency

$$Idf_i = -\log \frac{|\{d : t_i \in d\}|}{|D|}$$

t_iを含む文書の数

全文書数

・例えば，「の」という言葉

  ・頻出する（価値は高い）が，どの文書にも頻出する（価値は低い）

・「（その単語の文書中の頻度）*（その単語の全文書中の普遍性を鑑みた重み）」で，単語の価値を定める方法

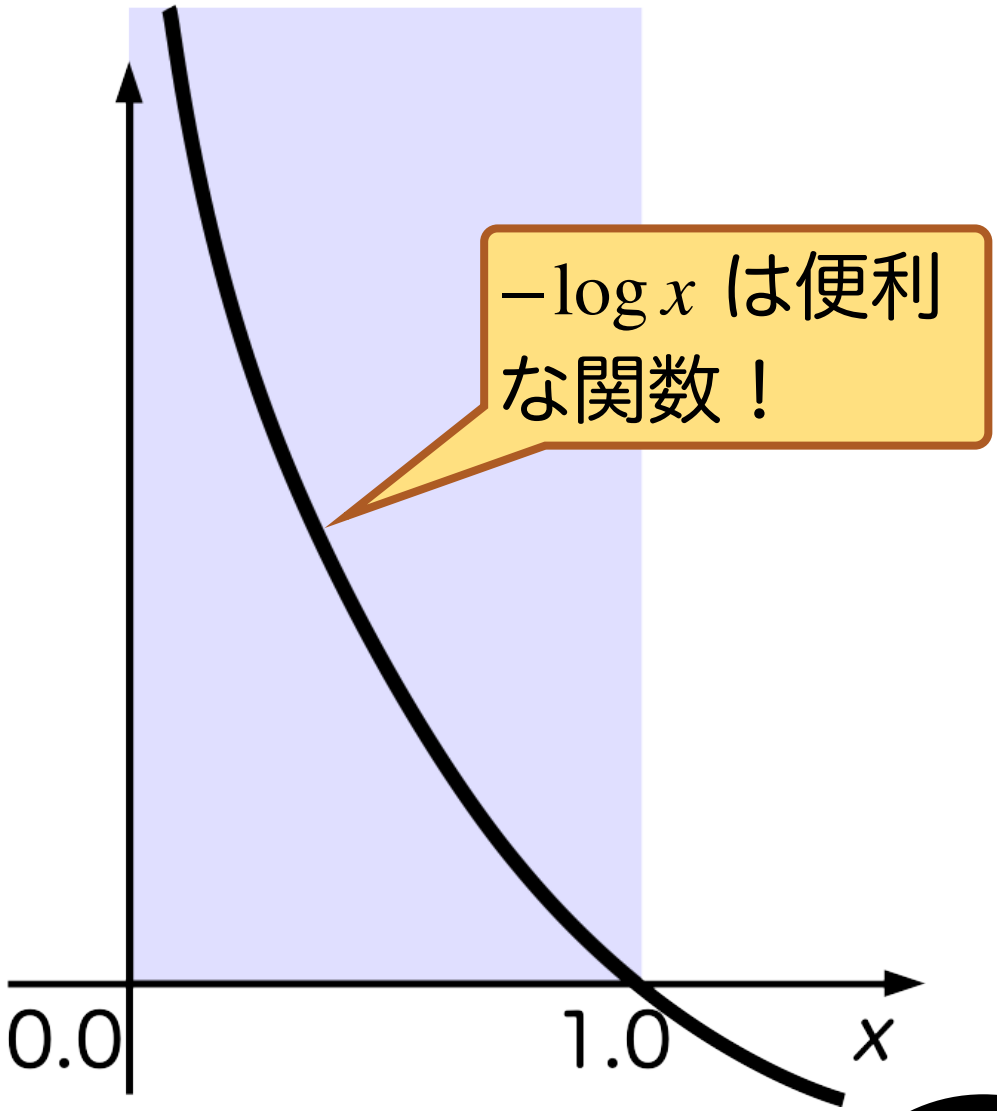
（参考）*Idf* 値の意味

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・*Idf* の取り得る値：

$x = \dfrac{|\{d : t_i \in d\}|}{|D|}$ のとき $-\log x$ のグラフ →

これを見れば，t_i が貴重なとき *Idf* 値は無限大，普遍的なとき *Idf* 値は0ということがわかる



共起関係

- 共起関係とは何か？まずは例文を……
  - これは例文です．日本語は難しい．「象は鼻が長い」この主語は何でしょう？日本語にはこのような曖昧さがあります．
  - “The quick brown fox jumps over the lazy dog.” この英語の例文には，全てのアルファベットが含まれています．

共起関係

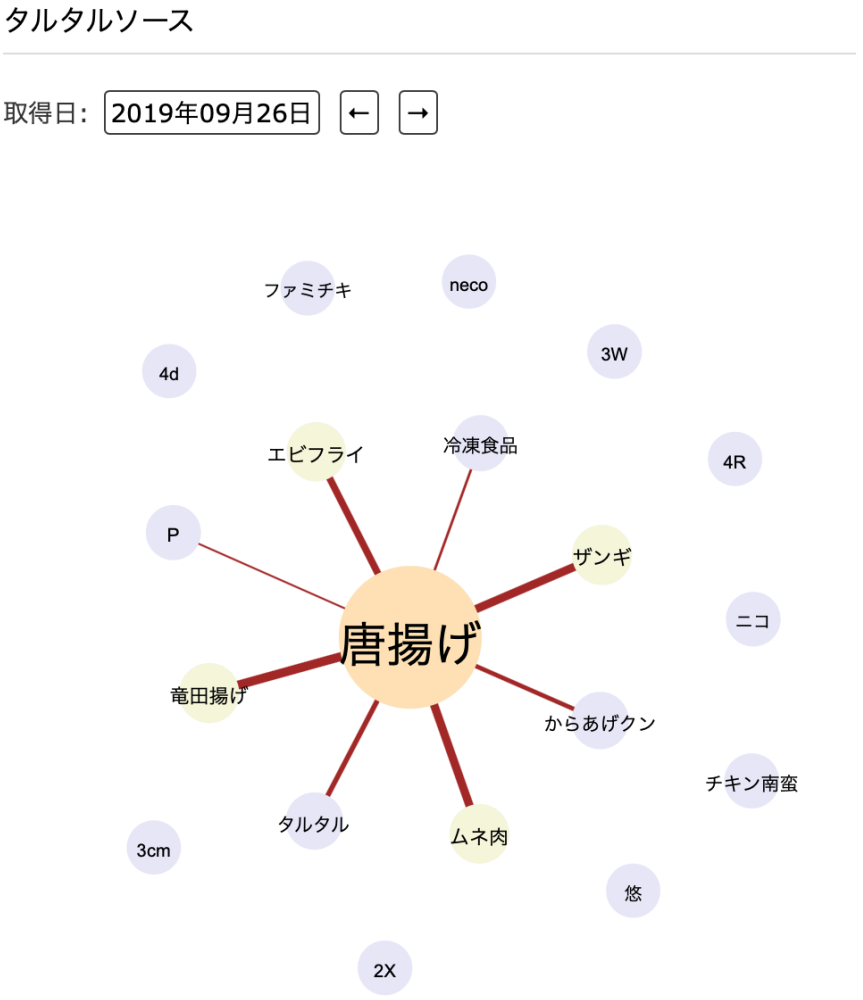
- 先の2つの段落（あるいは，ツイート，文書）に着目
  - 「例文」と「日本語」，あるいは「例文」と「英語」という単語は，同じ範囲（段落，ないしは，ツイート，もしくは，文書）に同時に出てくる（「共起する」という）…… 共起関係がある
  - 「日本語」と「英語」という単語は同時に出てこない …… 共起関係はない
- 共起確率 …… 共起関係の現れる頻度を計算したもの
  （出現頻度と同様なので数式による説明は省略）

トレンドごとの可視化

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・共起ネットワーク分析

- 出現頻度の大きさを，ノードの大きさと色で表現
  - 出現頻度の大きいほうから20個を選抜（20個に満たない場合もある）
- 共起関係の強さ（共起確率の大小）をエッジの幅で表現
  - 図が煩雑になるのを防ぐため，共起確率が計算できた関係のうち，上位1/4のみ描画

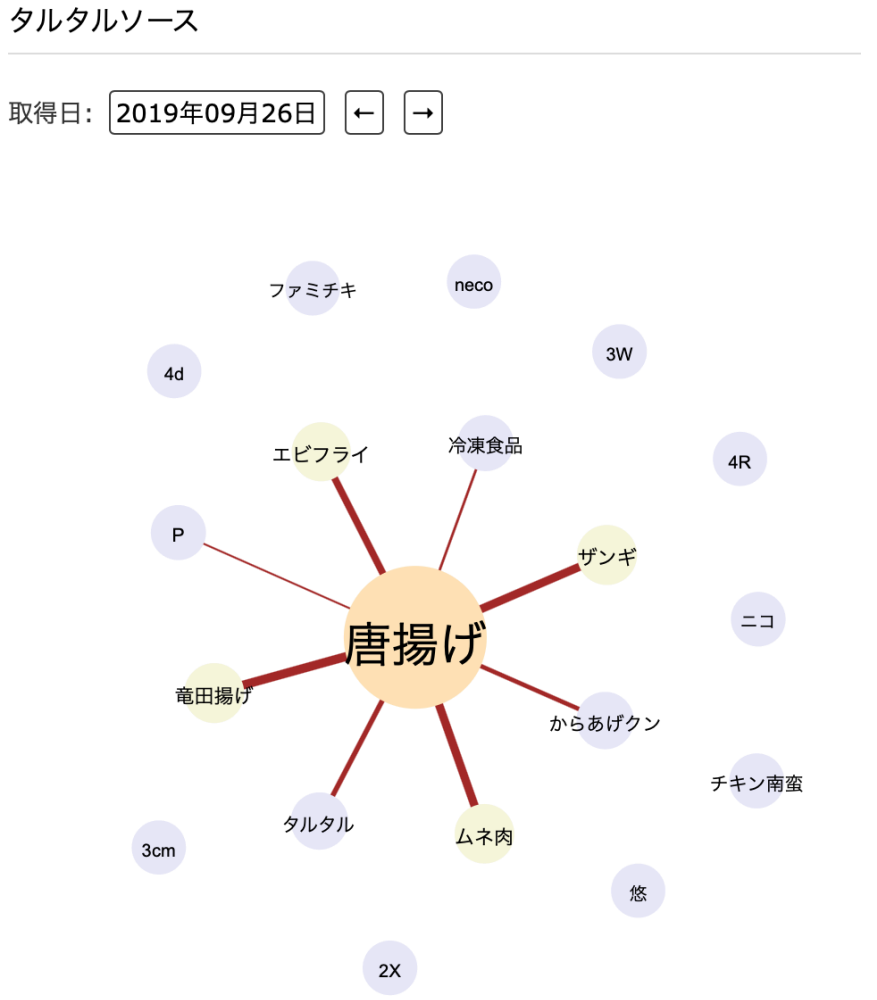


トレンドごとの可視化

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・右図でいえば……

- 「タルタルソース」というキーワードで得られたツイートに多く含まれていた単語は，「唐揚げ」「エビフライ」「ザンギ」「竜田揚げ」など
- それらは強く関連づけられており，ひとつのツイートのなかで同時に語られることが多かった



実装技術的な補足

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- リアルタイムに共起ネットワークを計算するのは，計算リソース的に厳しい
  - 20分おきに情報を取得するタイミングで，ノードとエッジの情報まで全て計算し，データベースに投入
- アクセス時はそれを読み出しているだけ
  - 描画には D3.js を利用，AJAXで描画
  - D3.js 描画時のデータソースにRailsのDBを指定
    - そのためRails側にもAJAX向けデータ提供用のAPIを用意
  - 描画方法は Spring Embedder というアルゴリズムを使用
    - これも興味深いが，今回は割愛します m(_ _)m

データ投入までの流れ

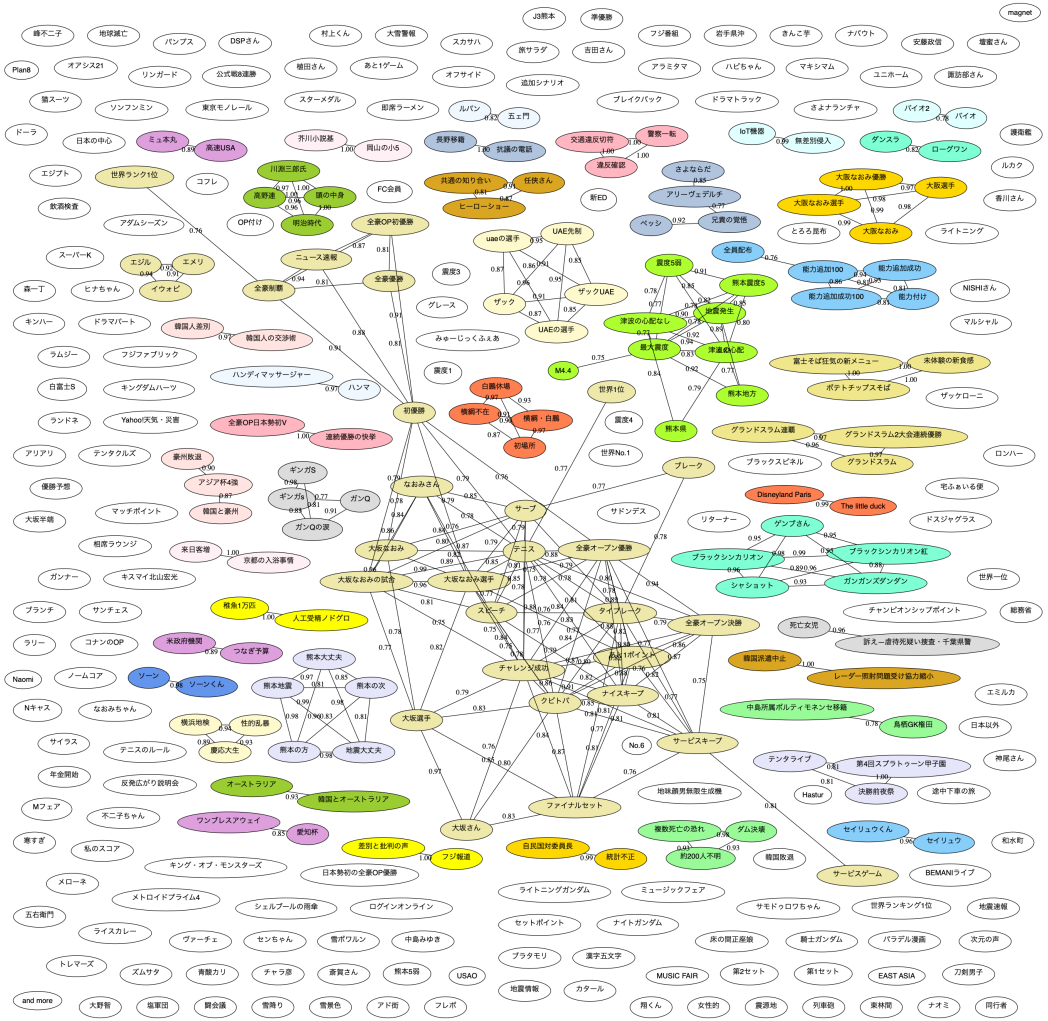
CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- Twitter APIを叩いてデータを取得，共起ネットワークを作成
  - Pythonスクリプトで実現→db/seed.rb を出力
  - bin/rails db:seed で更新
- なぜ前半部がPythonスクリプトなのか？
  - 試行錯誤で「使えるもの」を選んだから
- なぜ db/seed.rb 経由などとまどろっこしいことをやっているのか？
  - 開発を着手した時点ではまだRoRにあまり詳しくなかったから

トレンド傾向の可視化

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- おおまかな流れ
  - 各トピックに関する共起ネットワークのノードに描かれる全ての単語から「単語空間」を作る
  - その空間のなかで各トピックは「点」として表されるので，その近さを計算し，近いトピック同士をひとつのグループとしてまとめる



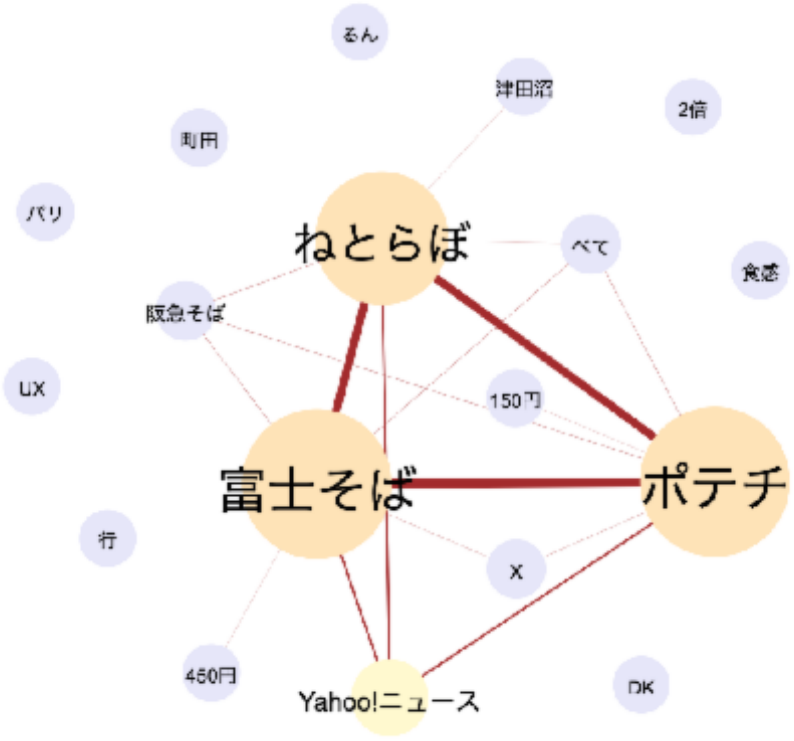
コサイン類似度を利用したトピックマップの作成

類似のトピック

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

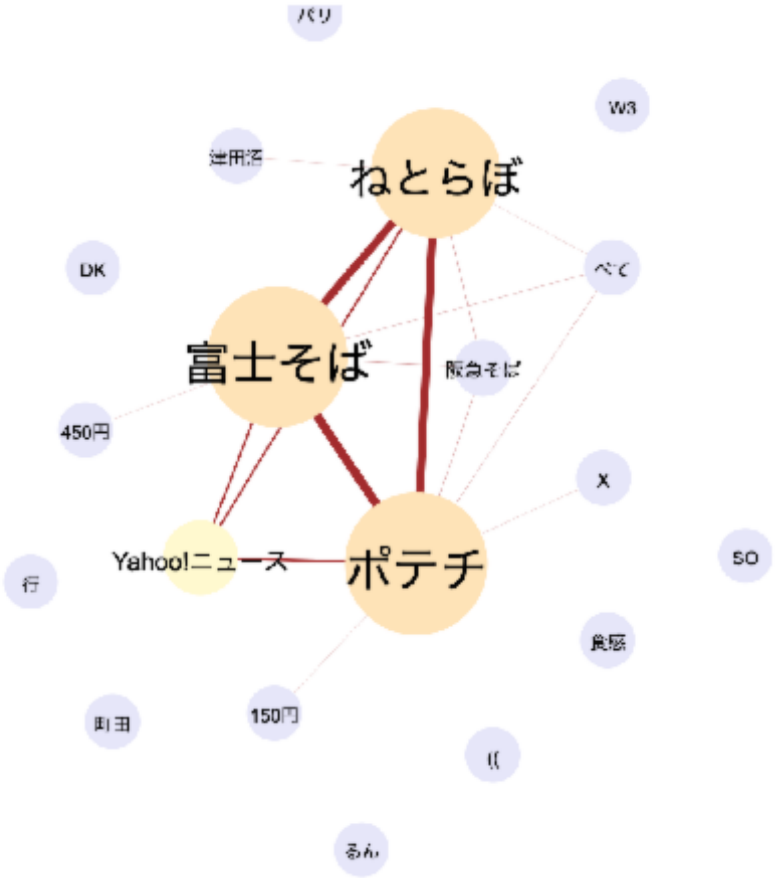
ポテトチップスそば

取得日: 2019年01月26日 ← →



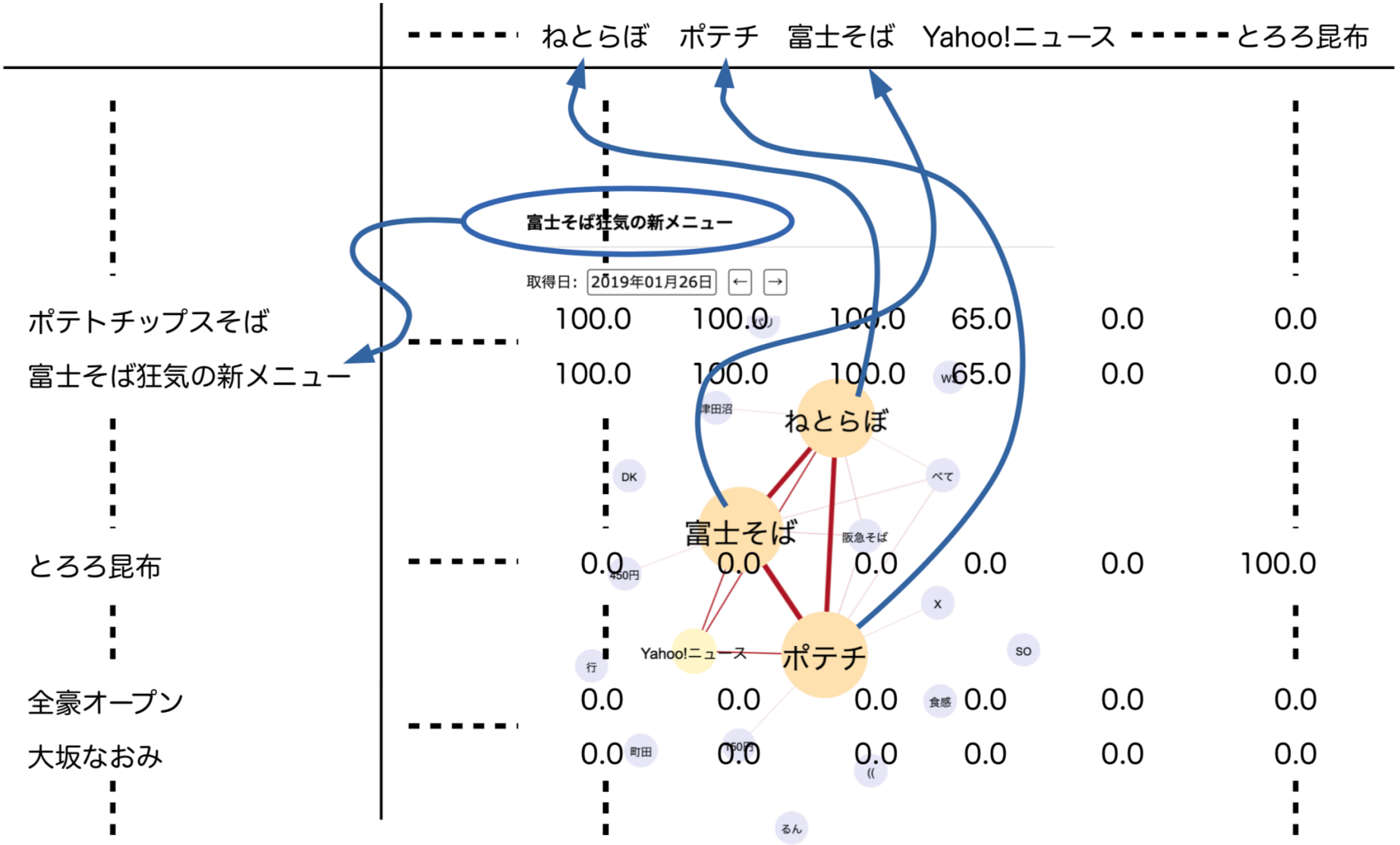
富士そば狂気の新メニュー

取得日: 2019年01月26日 ← →



単語空間

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

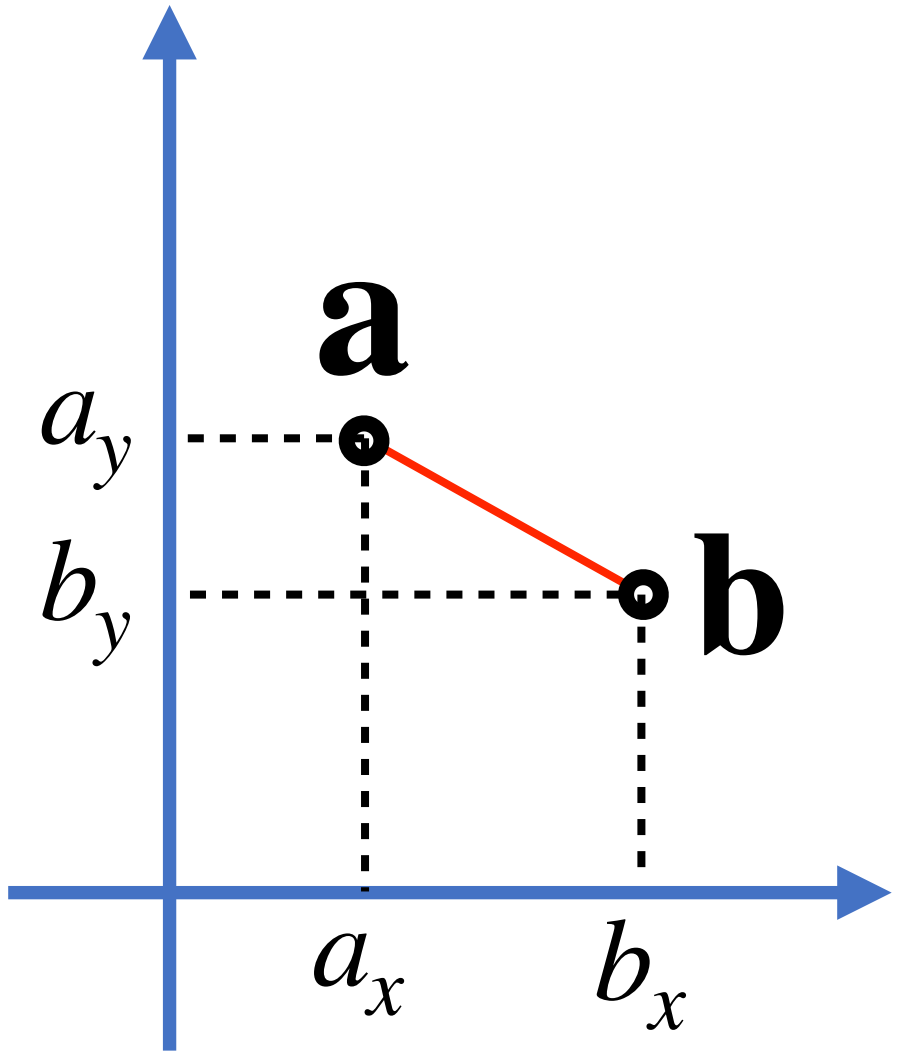


類似度をどう考えるか？

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・ユークリッド距離

$$\mathrm{dist}(\mathbf{a}, \mathbf{b}) = \sqrt{\Sigma(a_i - b_i)^2}$$



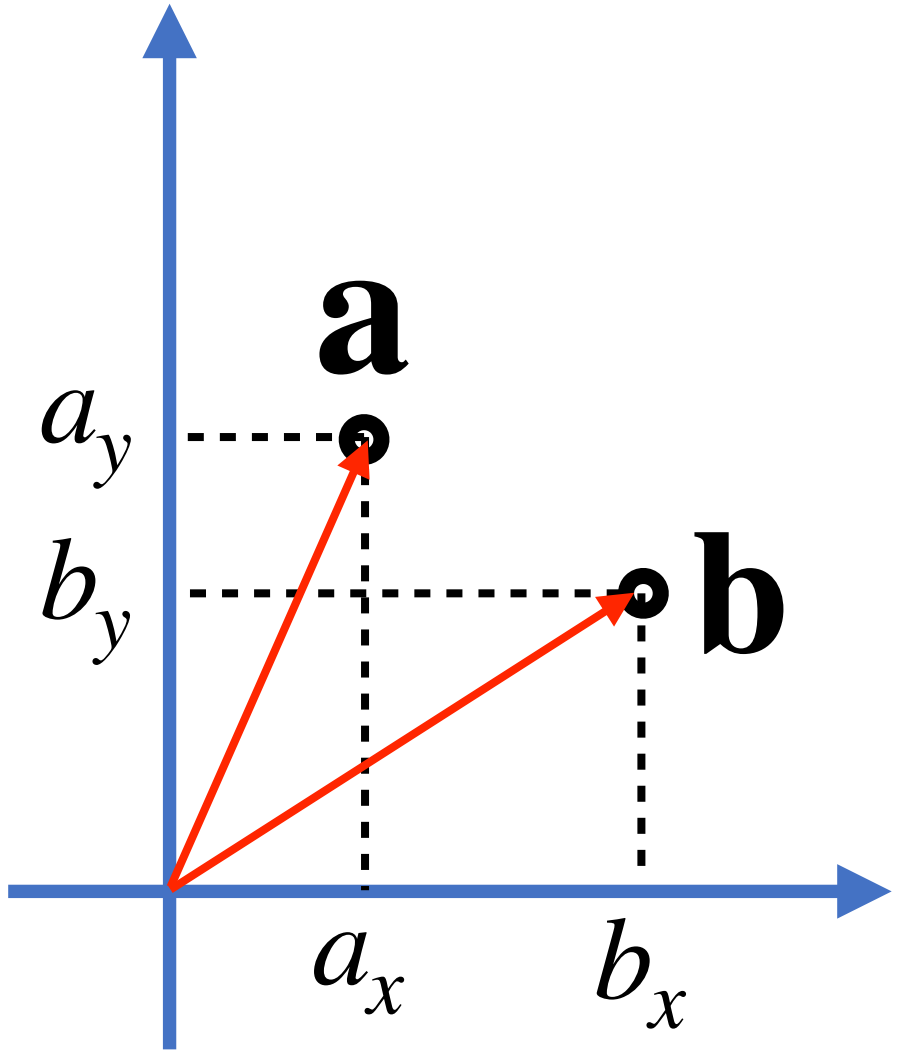
類似度をどう考えるか？

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・コサイン類似度

$$\mathrm{sim}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

$$= \frac{\Sigma a_i b_i}{\sqrt{\Sigma a_i^2}\sqrt{\Sigma b_i^2}}$$



2つのトピック間の類似度

・単語空間にマップされた2つのトピック

・ベクトル a, b がトピック a, b を表すとき

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}|\cos\theta$$

$$\therefore \ \mathrm{sim}(\boldsymbol{a}, \boldsymbol{b}) = \cos\theta = \boldsymbol{a} \cdot \boldsymbol{b} \,/\, |\boldsymbol{a}||\boldsymbol{b}|$$

トピックマップの作成

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

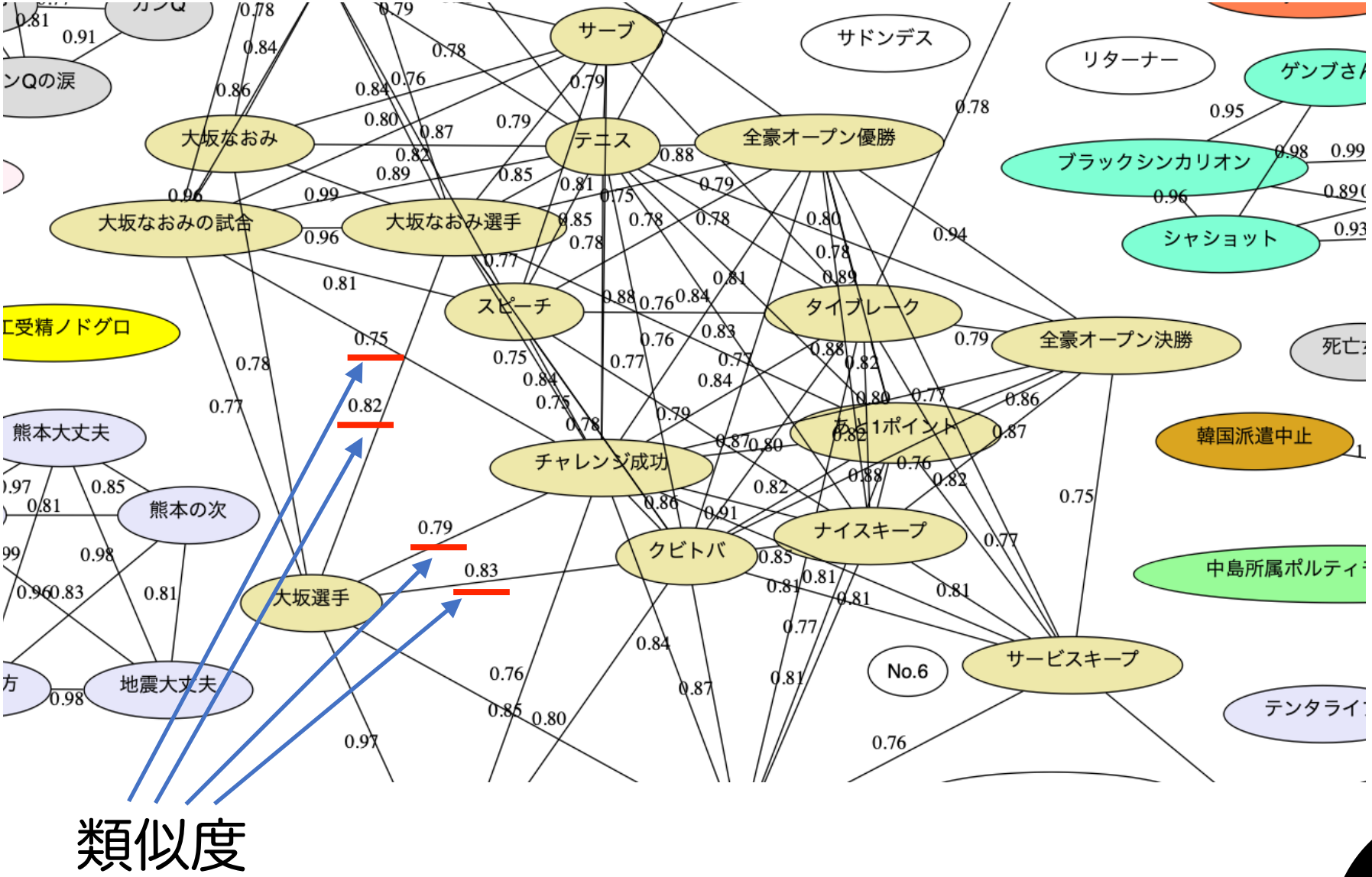
- 各トピック間の類似度を「総当たりで」計算
  - コサイン類似度 …… 0.0 ～ 1.0 の値をとる
    （本来は-1.0～1.0だが，データの性質上，負の値にはなりえない）
- しきい値（0.50）を超えるものを線で結び，同じクラスタとする

コサイン類似度を選んだ理由はコレを決めやすかったから

  - なお，クラスタの色に意味はない．違うクラスタに属することを分かりやすく表現するために色付けしているだけ

トレンド傾向の可視化

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS



トピックマップ作成の流れ

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- TWtrendsがAPIを提供
  - そこからトレンドのデータを取得
    （なので，トピックマップは別ホストで作成可能）
- Python，Rubyスクリプトの組み合わせでDotスクリプトを作成
  - Pythonでコサイン類似度を計算，各ノード間の距離を出力
  - その出力に基づいて，dotスクリプトを出力するRubyプログラム
- できあがったDotスクリプトに基づきGraphVizで画像を作成→サーバにアップロード

実装技術的な補足

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

・コサイン類似度の計算は Numpy で一発計算

・トピックマップの描画は GraphViz （dotコマンド）を利用

  ・get_data.py ${day}
    …… 日付を指定してその日のデータをサーバから取得

  ・calc_cos_sim.py …… 総当りでトピックの類似度を計算

  ・mk_net.rb …… 上記の情報に基づいてdotスクリプトを作成

・バッチスクリプトの例

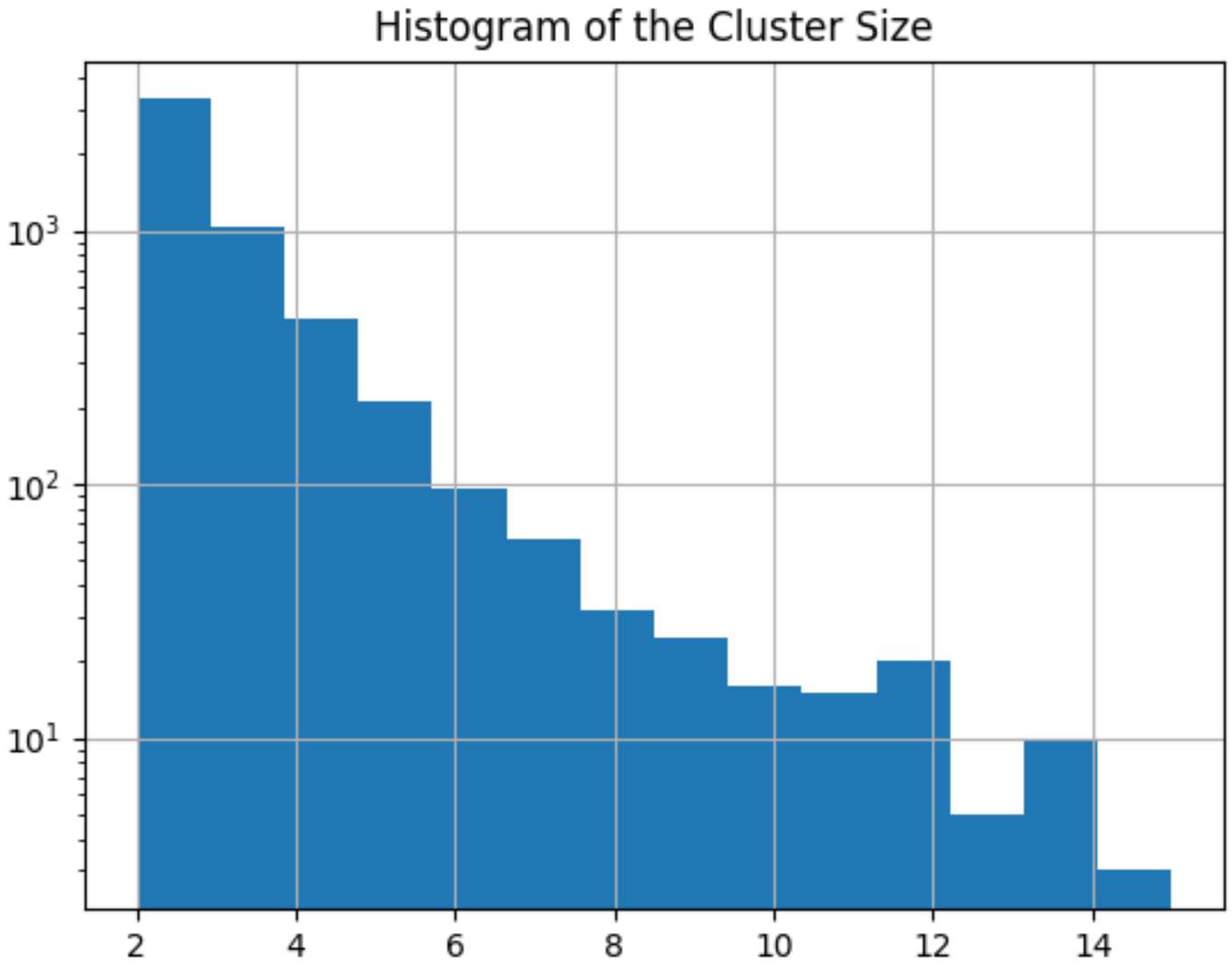
```
day=$(date -v-1d +%Y-%m-%d)
./get_data.py ${day} | ./calc_cos_sim.py | ./mk_net.rb > /tmp/graph.dot
dot -Tpng /tmp/graph.dot -o/tmp/graph_${day}.png
scp /tmp/graph_${day}.png twt.iiojun.com:twt/app/assets/images/graphs/
```

Rubyを用いた研究・教育

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

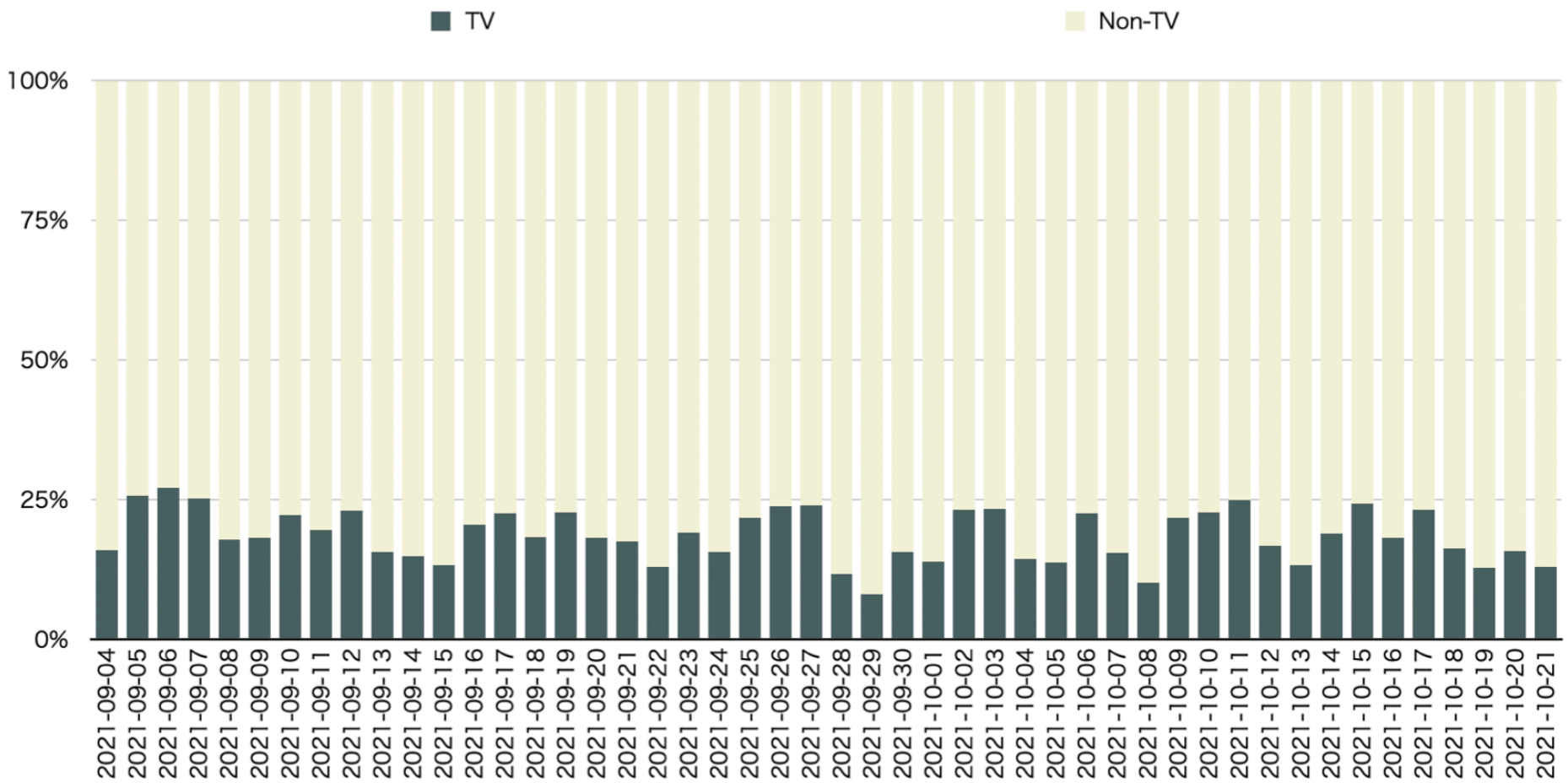
- TWtrends作成の経緯
  - 2018年度後期：大学院ゼミ
    - 「なんかRailsで面白いWebアプリ作らへん？」
  - 皆で試行錯誤しながら考えた
  - そのときのメンバー
    - Iio, J., Poppe, S., Aoki, Y., Nakamura, E., Kim, S., and Lee, T. (2019) Visualization of Twitter Trends using a Co-occurrence Network, The 12th IEEE Pacific Visualization Symposium (PacificVis2019) Poster Proceedings, pp. 321-322, Bangkok, Thailand.

話題クラスタの数理モデル研究



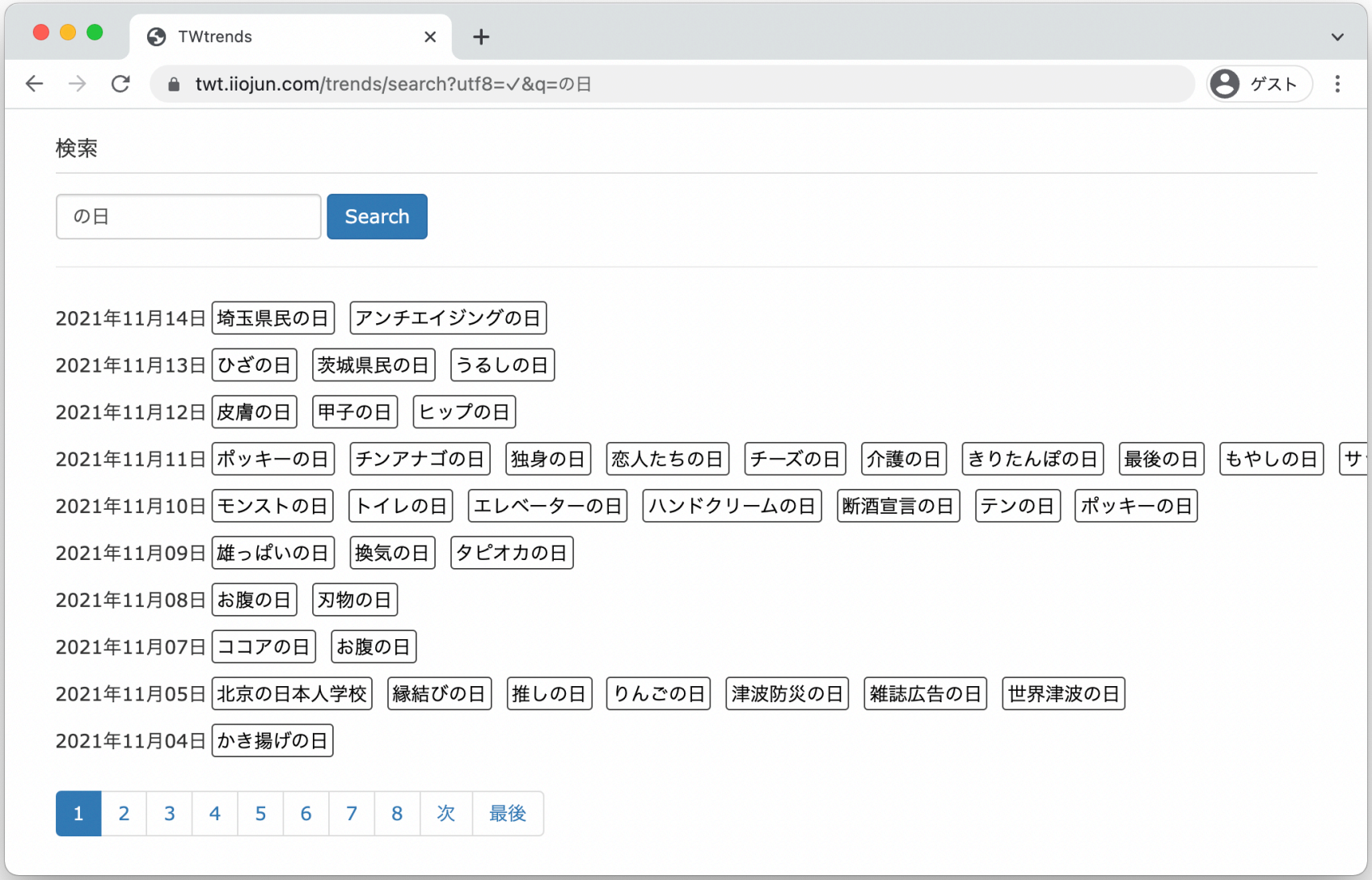
現在進行中のプロジェクト

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS



「の日」プロジェクト

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS



まとめ

CHUO UNIVERSITY
FACULTY OF
GLOBAL INFORMATICS

- TWtrendsで使われている技術を紹介しました
- システムとしては単純だが，面白い効果は示せている（と自負）
- 教育・研究の一環としてRailsのシステムを活用している事例を紹介しました
- まだ派生プロジェクトを実施しています
  - 応援してください＼(^o^)／