日記と自然言語処理

宇都宮大学 工学部 2年
うすゆき

SELF-INTRODUCTION

プロフィール

- 工学部 基盤工学科 情報電子オプティクス 情報科学分野
- 鳥取県生まれ、島根県育ち
- 好きなフォントはKiwi Maru

最近のマイブーム

- VRChat楽しい
- キャラメルポップコーンおいしい

ポートフォリオ

pf.usuyuki.net

TOC

目次

TOC

目次

SETTINGS

LT会なので
ぱっぱと進みます。

SETTINGS

学術的な話はしません！！

正直、できません

SETTINGS

今回は実用上の話

自然言語処理を趣味でも使おう。

TOC

目次

NLPってなんだよ！！！

NLP

Natural Language Processing

自然言語処理

TOC

目次

GOAL

「自然言語処理、なんかええな」って思ってほしい

TOC

目次

PROLOGUE

日記webアプリ作ったんです

この話をすることになったきっかけでもある。

KADODE

かどで日記

日記を作成・管理・分析できる
webアプリ

KADODE

できること

KADODE

日記の作成、管理

KADODE

統計情報の生成

KADODE

This is 個人開発

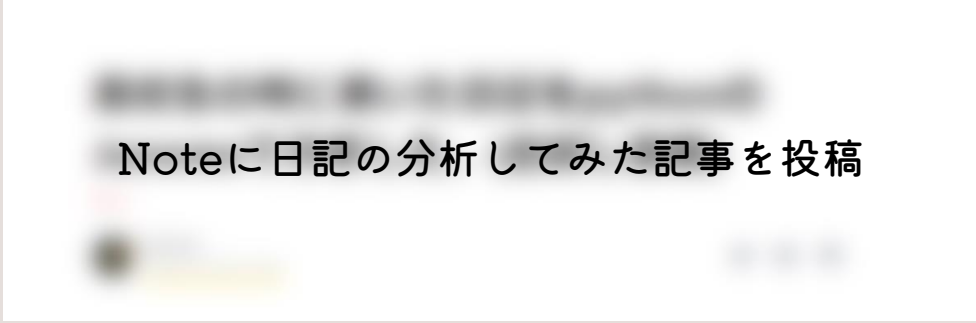
U-labとは一切関係ありません！！！

KADODE

きっかけ

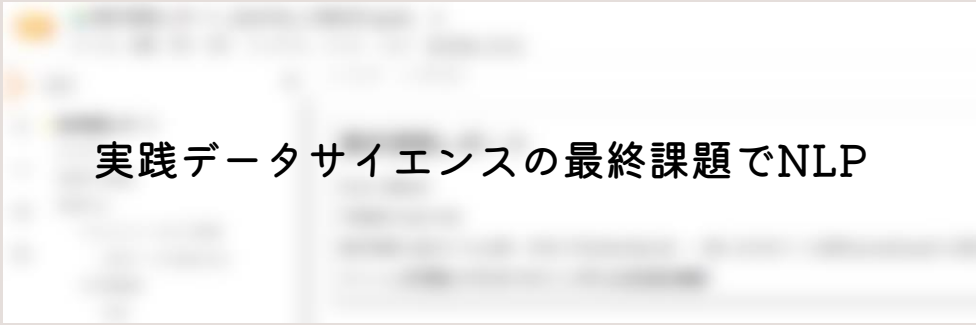
[後期教養科目]実践データサイエンス
最終課題で出したものを応用して作った

1年物のアイデアを実現！



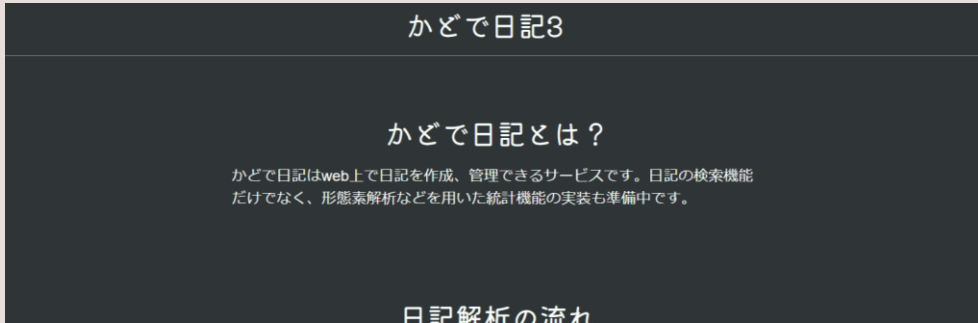
2020年7月

⇩



2021年2月

⇩



ここ発表しない

TOC

目次

ここ発表しない

KADODE

採用技術

詳細省きます

ここ発表しない

WEB_TECH

インフラ

サーバー：さくらVPS

ローカル：Docker

コード管理：GitHub

文章周り：GitHub Wiki

自動デプロイ：GitHub Actions

バックアップ：GCP Google Cloud Storage

WEB_TECH

バックエンド

言語：PHP,Python

フレームワーク：Laravel 8系

認証：Laravel Jetstream Livewire

DB：MySQL

PHPライブラリ：Goodby CSV

ここ発表しない

WEB_TECH

NLP

言語：Python

ライブラリ：GiNZA v5

固有表現ラベル：関根の拡張固有表現階層 ver7.1.2

感情極性辞書：日本語評価極性辞書北大学 乾・鈴木研究室

形態素解析辞書：Sudachi辞書

WEB_TECH

フロントエンド

言語：PHP,JS,HTM,CSS

CSSライブラリ：Tailwind CSS

JSライブラリ：Chart.js,D3-cloud

ここ発表しない

WEB_TECH

その他

CPUのジョブ割当制御：cpulimit

などなど

詳しいことは、かどで日記wikiをご覧ください。

ここ発表しない

TOC

目次

ここ発表しない

KODAWARI

かどで日記の技術的こだわり

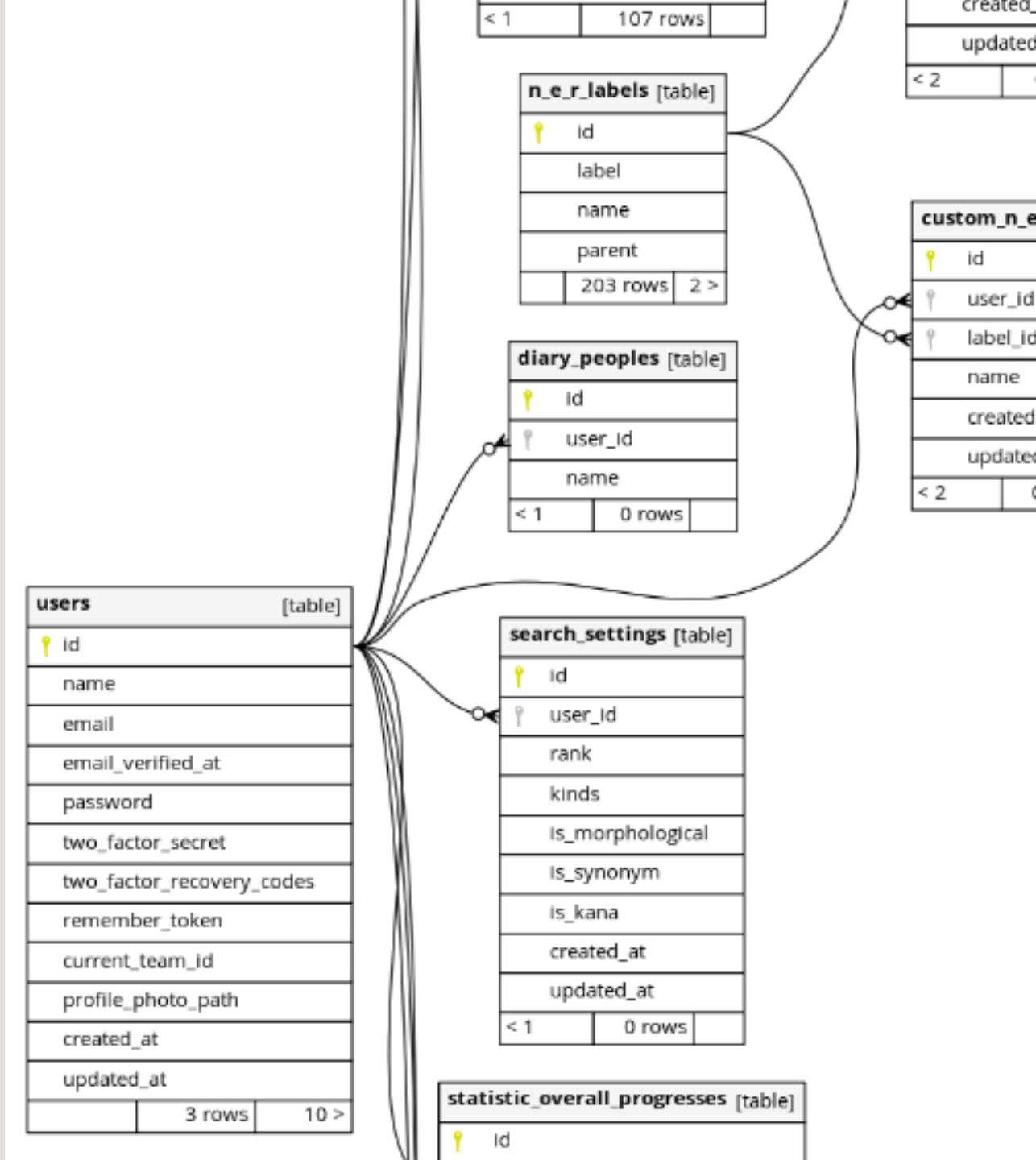
そんなの当たり前じゃんって言わないで……

ここ発表しない

KODAWARI

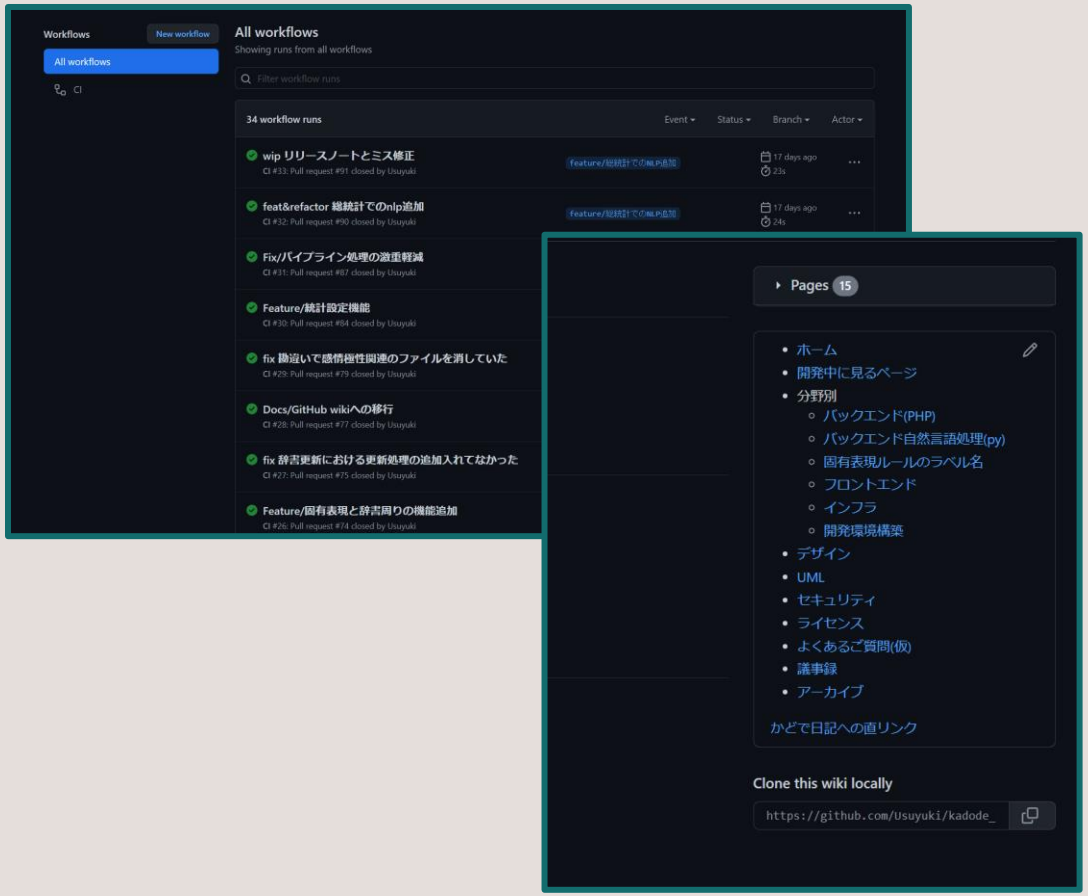
こだわり①

DB設計で正規化を
それなりにした



ここ発表しない

KODAWARI



こだわり②

環境周りの整備

- 機能ごとにブランチ切る
- 自動デプロイ
- 自動バックアップ
- Wiki化
- ローカル開発をDockerに

ここ発表しない

KODAWARI

こだわり③

ライセンス周りをちゃんと調べた

ここ発表しない

WEB_TECH

オープンソース≠自由に使っていい

ここ発表しない

WEB_TECH

ある意味危険なライセンスもあるよ

ここ発表しない

WEB_TECH

このライセンスで公開されているライブラリを使ったら……

LGPLライセンス

そのソフトウェアもこのライセンスにしないといけない

GPLライセンス

作成したソフトウェアはソースコードを公開しないといけない

ここ発表しない

WEB_TECH

かどで日記はMITライセンスにしたい……

Cabocha使えない！！

ここ発表しない

WEB_TECH

ライブラリやフレームワークは便利だけど、

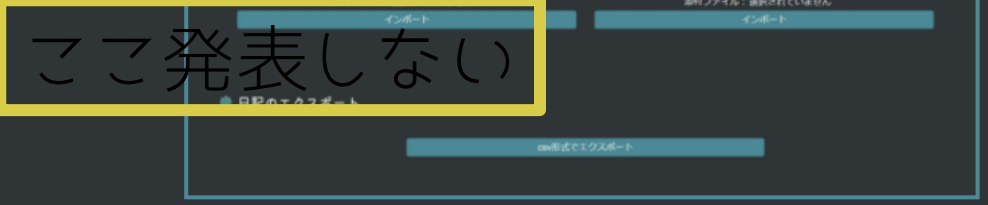
ライセンスはちゃんと見ましょう。

ここ発表しない

WEB_TECH

ネットからコピペする場合は

クレジット表記もしましょう



などなど……

新規登録

人が作った趣味の遊びサービスです。
理解の上、ご利用ください。

```
<h2 class="mt-24 mb-12 text-center text-3xl kiwi-maru">日記解析の流れ</h2>
▶<div>…</div>
<h2 class="mt-24 mb-12 text-center text-3xl kiwi-maru">できること</h2>
▶<div class="flex justify-center imtes-center flex-wrap my-8">…</div> flex
▶<div class="flex justify-center imtes-center flex-wrap my-8">…</div> flex
▶<div class="flex justify-center imtes-center flex-wrap my-8">…</div> flex
▶<div class="flex justify-center imtes-center flex-wrap my-8">…</div> flex
▶<div class="flex justify-center imtes-center flex-wrap my-8">…</div> flex
▶<div class="flex justify-center imtes-center flex-wrap my-8">…</div> flex
<p class="text-center mb-4 kiwi-maru">などなど……</p>
<!--
    Copyright (c) June 1, 2015 Tuomas Pöyry
    Released under the MIT license
    http://opensource.org/licenses/mit-license.php
    -->
<canvas id="top-animation" width="633" height="400"> == $0
▶<div class="mb-24 mt-4">…</div>
▶<div class="mt-4 mb-12">…</div>
▶<div class="mt-4 mb-24">…</div>
</div>
<script type="text/javascript" src="https://kadodenikki3.usuyuki.net/js/topPerticle.js"></script>
</div>
▶<footer class="py-4">…</footer>
<script type="text/javascript" src="https://kadodenikki3.usuyuki.net/js/kadodeMain.js?ver=19.2"></script>
</body>
</html>
```

ここ発表しない

WEB_TECH

当たり前だけど、忘れがちなので

ここ発表しない

KADODE

かどで日記

でした

現在Public Betaです

https://kadodenikki3.usuyuki.net/

NLPの魅力

TOC

目次

ここ発表しない

MODERN_NLP

モダンなNLPはかどで日記で導入してない

ここ発表しない

MODERN_NLP

そもそもモダンなNLPとは？

ここ発表しない

MODERN_NLP

具体的には……

ここ発表しない

MODERN_NLP

2017年から始まったNLP戦国時代

Transformer
Attention

ELMo
embedding

BERT
bidirectional

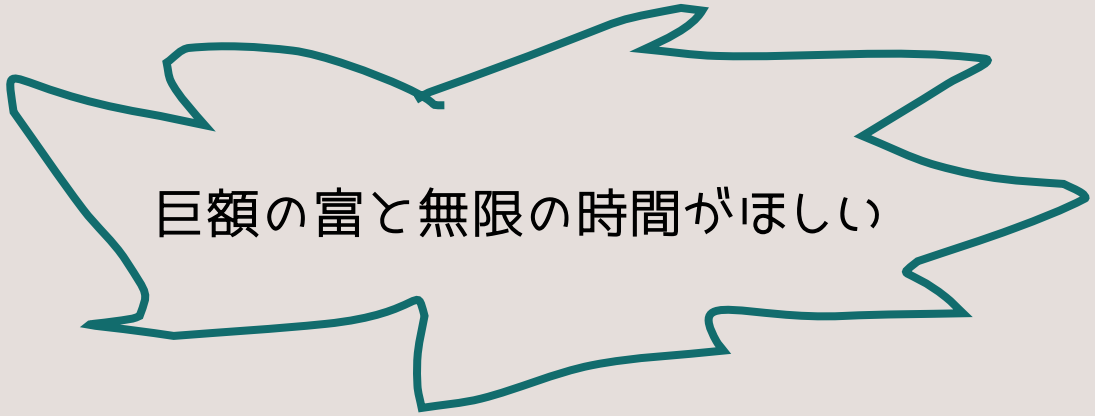
GPTシリーズ
Pre-training
Fine-tuning

ここ発表しない

MODERN_NLP

理由

1. 使用用途にそぐわない
2. 単純に理解が追いつけていない……
3. サーバーの性能不足でtransformersモデルとか動かせない
   →メモリ1GBのサーバーに対して、推奨メモリは16GB



ここ発表しない

MODERN_NLP

これも自然言語処理なので、紹介だけします

ここ発表しない

EXAMPLE

2019/12/10~
Google検索結果の精度向上

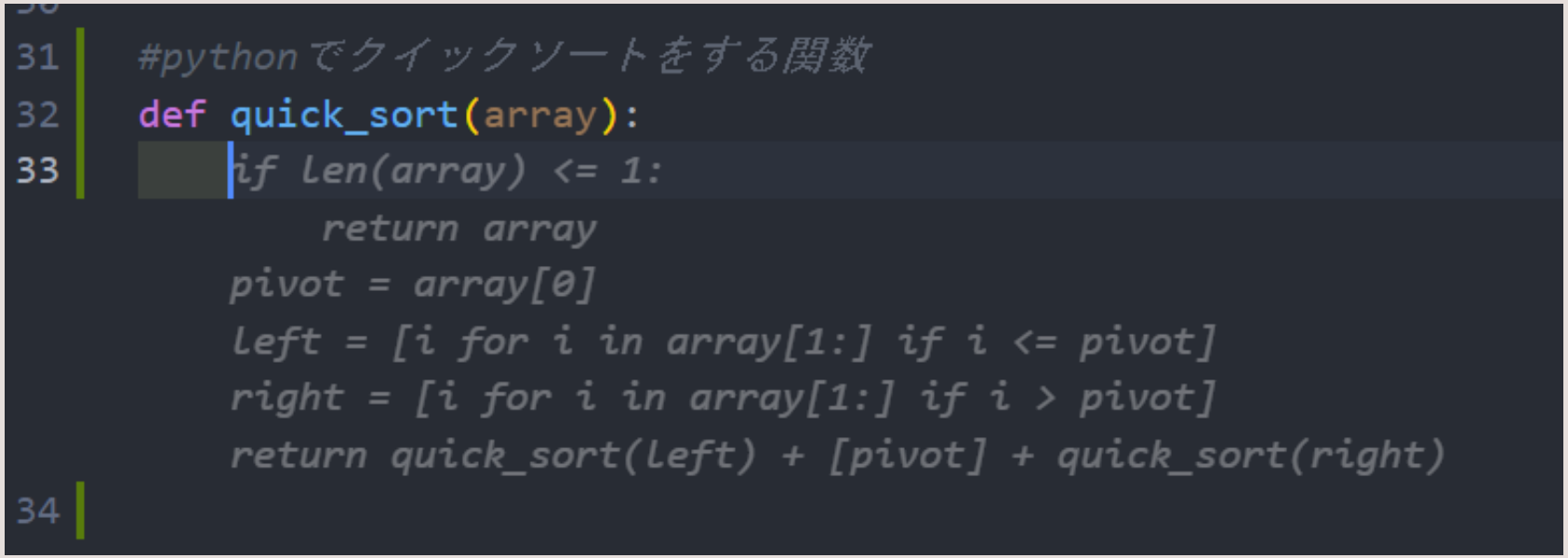
BERTの使用

微妙なニュアンスや、**文脈の理解をした結果に！**

ここ発表しない

EXAMPLE

GitHub Copilot

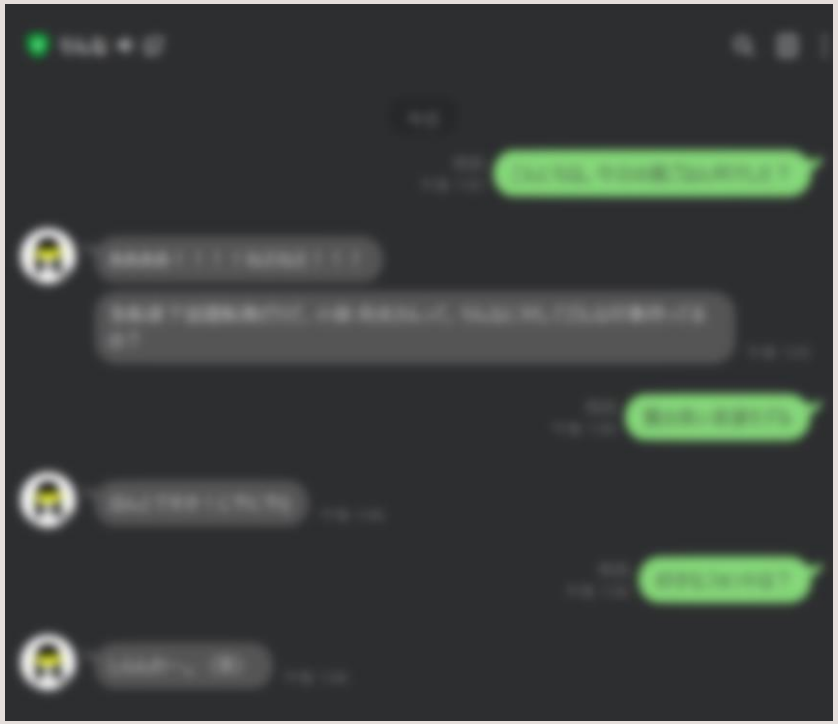


ソースコードの自動生成
(樋口先輩をも虜にするレベル)

ここ発表しない

EXAMPLE

りんな



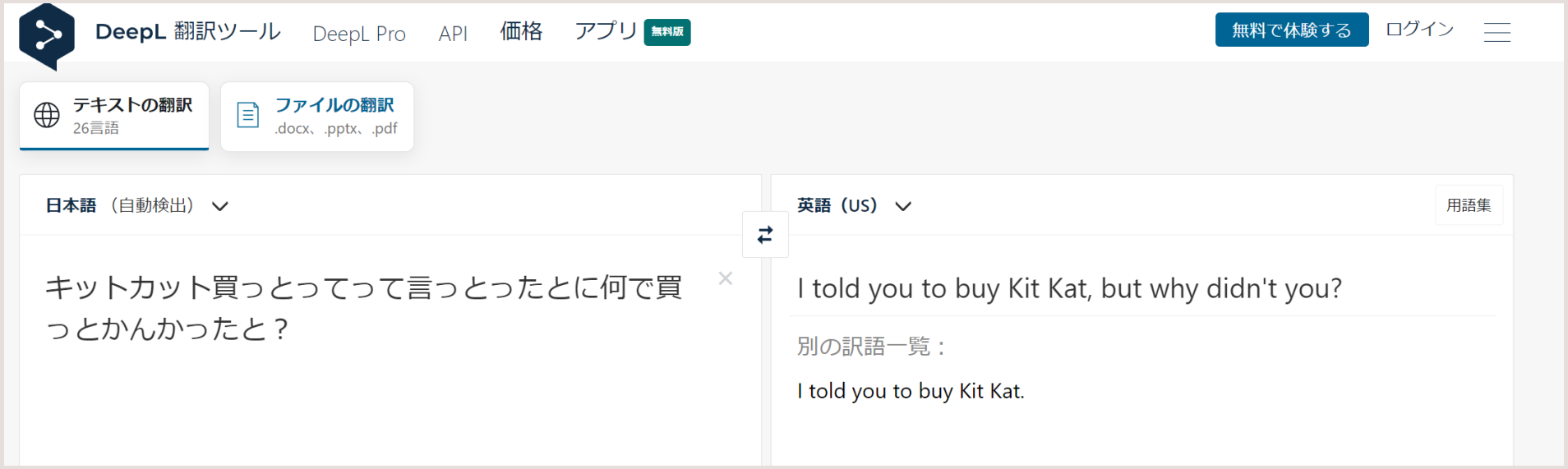
自分たちが中学生の頃バズったやつですね。

GPT-2のりんなモデルが公開されている

ここ発表しない

EXAMPLE

DeepL翻訳



方言にも対応する翻訳！

ここ発表しない

BUT

かどで日記でやりたいのは”分析”

TOC

目次

GINZA

GINZA

GiNZA is 日本語自然言語処理オープンソースライブラリ

↑Pythonで使える

GINZA

国立国語研究所
と
Universal Dependencies for Japanese
の
共同研究成果

論文:短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習

言語処理学会 第25回年次大会 発表論文集に掲載

GINZA

できること

- 形態素解析
- 係り受け解析
- 文章要約
- 文の類似度計算
- 固有表現の抽出

などなど

GINZA

できること

- 形態素解析
- 係り受け解析
- 文章要約
- 文の類似度計算
- 固有表現の抽出

などなど

MORPHOLOGICAL ANALYSIS

日本語は分割しないといけない。

MORPHOLOGICAL ANALYSIS

ビーカーに淹れた珈琲は格別だ。

1.Separing Words(wakatigaki)

ビーカー/に/淹れた/珈琲/は/格別/だ/。

2. Recognize part of speech

```
ビーカー　　名詞,一般,*,*,*,*,ビーカー,ビーカー,ビーカー
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
淹　名詞,一般,*,*,*,*,*
れ　動詞,接尾,*,*,一段,連用形,れる,レ,レ
た　助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
珈琲　名詞,一般,*,*,*,*,珈琲,コーヒー,コーヒー
は　助詞,係助詞,*,*,*,*,は,ハ,ワ
格別　名詞,一般,*,*,*,*,格別,カクベツ,カクベツ
だ　助動詞,*,*,*,特殊・ダ,基本形,だ,ダ,ダ
。　記号,句点,*,*,*,*,。,。,。
```

MORPHOLOGICAL ANALYSIS

It is like a “Hinsibunkai” in Koten

「何をかたてまつらむ。まめまめしき物は、まさなかりなむ。」

係助詞
格助詞
ラ行四段活用
意志の助動詞
形容詞
係助詞
形容詞
強意の助動詞
推量の助動詞

ここ発表しない

GINZA

できること

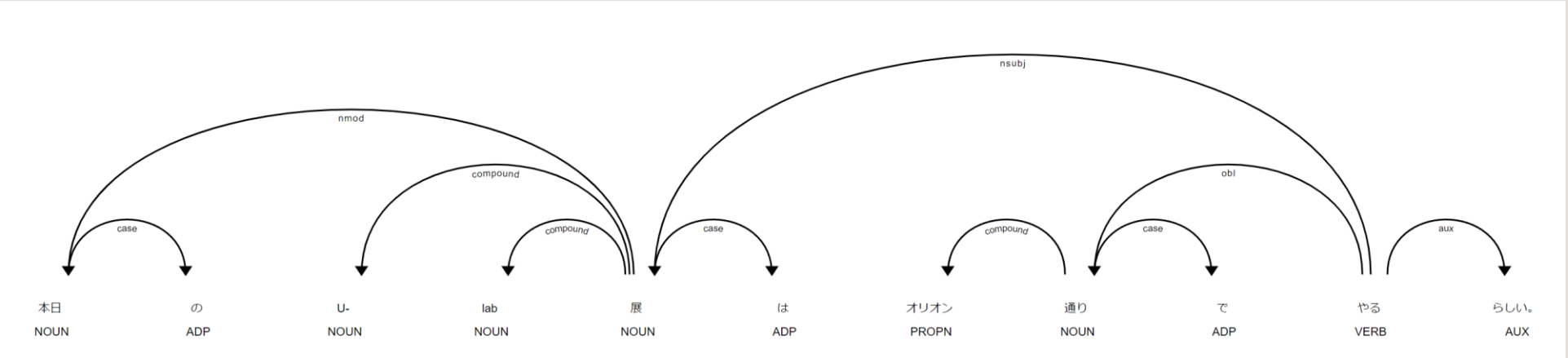
- 形態素解析
- 係り受け解析
- 文章要約
- 文の類似度計算
- 固有表現の抽出

などなど

ここ発表しない

DEPENDENCY ANALYSIS

係り受け解析



ここ発表しない

GINZA

できること

- 形態素解析
- 係り受け解析
- 文章要約
- 文の類似度計算
- 固有表現の抽出

などなど

ここ発表しない

GINZA

できること

- 形態素解析
- 係り受け解析
- 文章要約
- 文の類似度計算
- 固有表現の抽出

などなど

GINZA

できること

- 形態素解析
- 係り受け解析
- 文章要約
- 文の類似度計算
- 固有表現の抽出

などなど

NAMED ENTITY

固有表現の抽出

本日の U-lab展 **Occasion_Other** は オリオン通り **Road** でやるらしい。

TOC

目次

KADODE

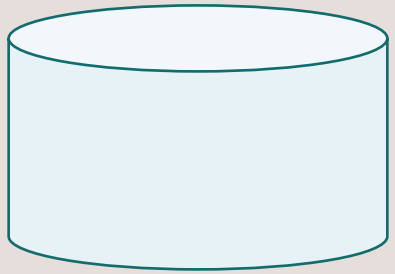
このスライドでの日記解析データ元

GINZA

※かどで日記の仕組み簡易版

ユーザーが日記を書く

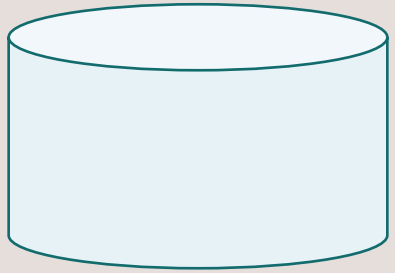
データベースに格納



GiNZAで解析

解析結果を加工

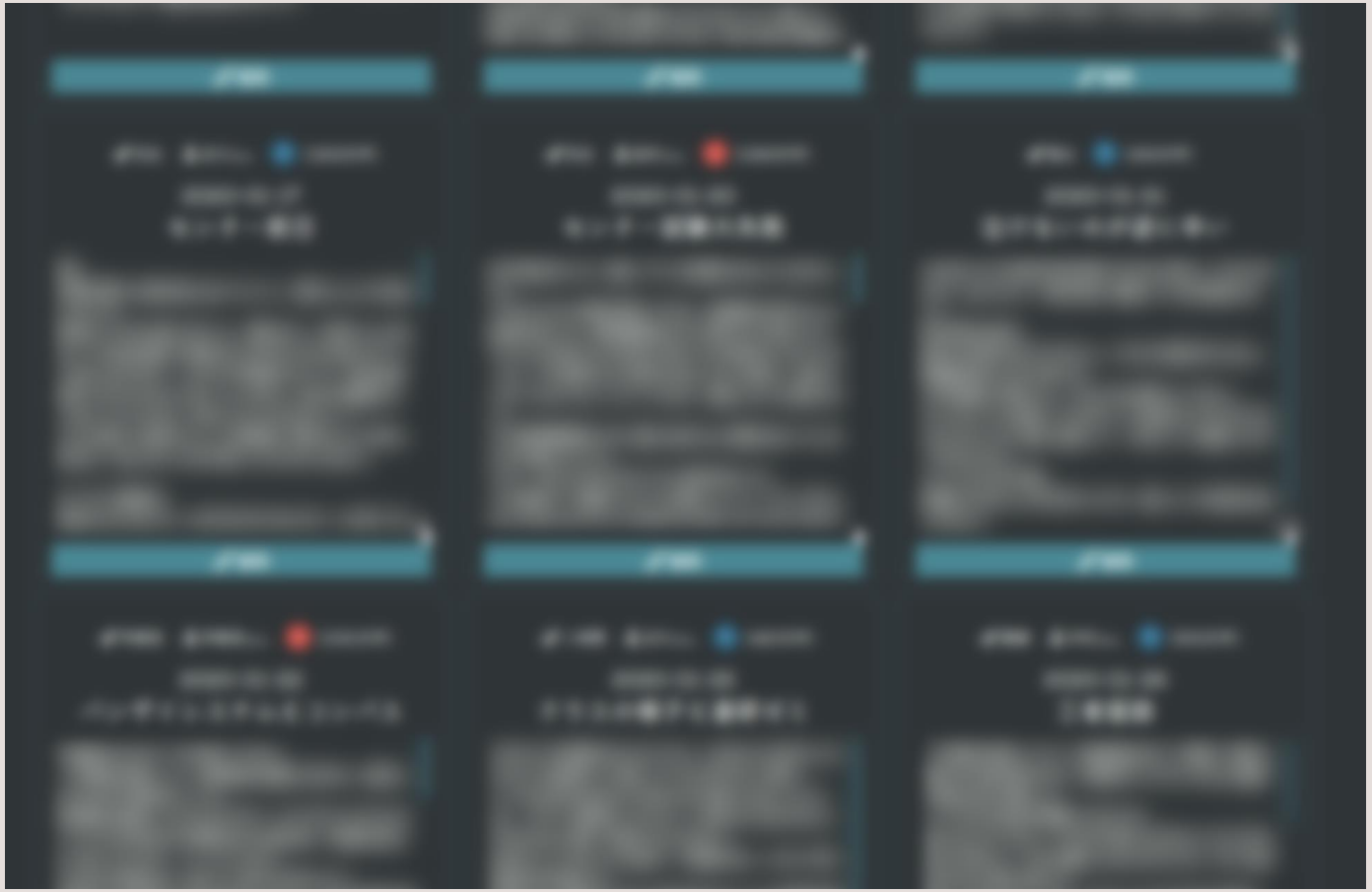
データベースに格納



画面に表示

KADODE

折角なので、センター試験目線で見る



KADODE

ほとんどの1年生にとっては今年の出来事なので、昔を思い出しながら的な……

KADODE

個別　　月別　　総計

INDIVIDUAL

個別　　月別　　総計

2020/1/17

センター試験前日

INDIVIDUAL

※一般公開用はぼかしています。

2020/1/18,19

センター試験日

2020/1/20

センター試験自己採点

INDIVIDUAL

※一般公開用はぼかしています。

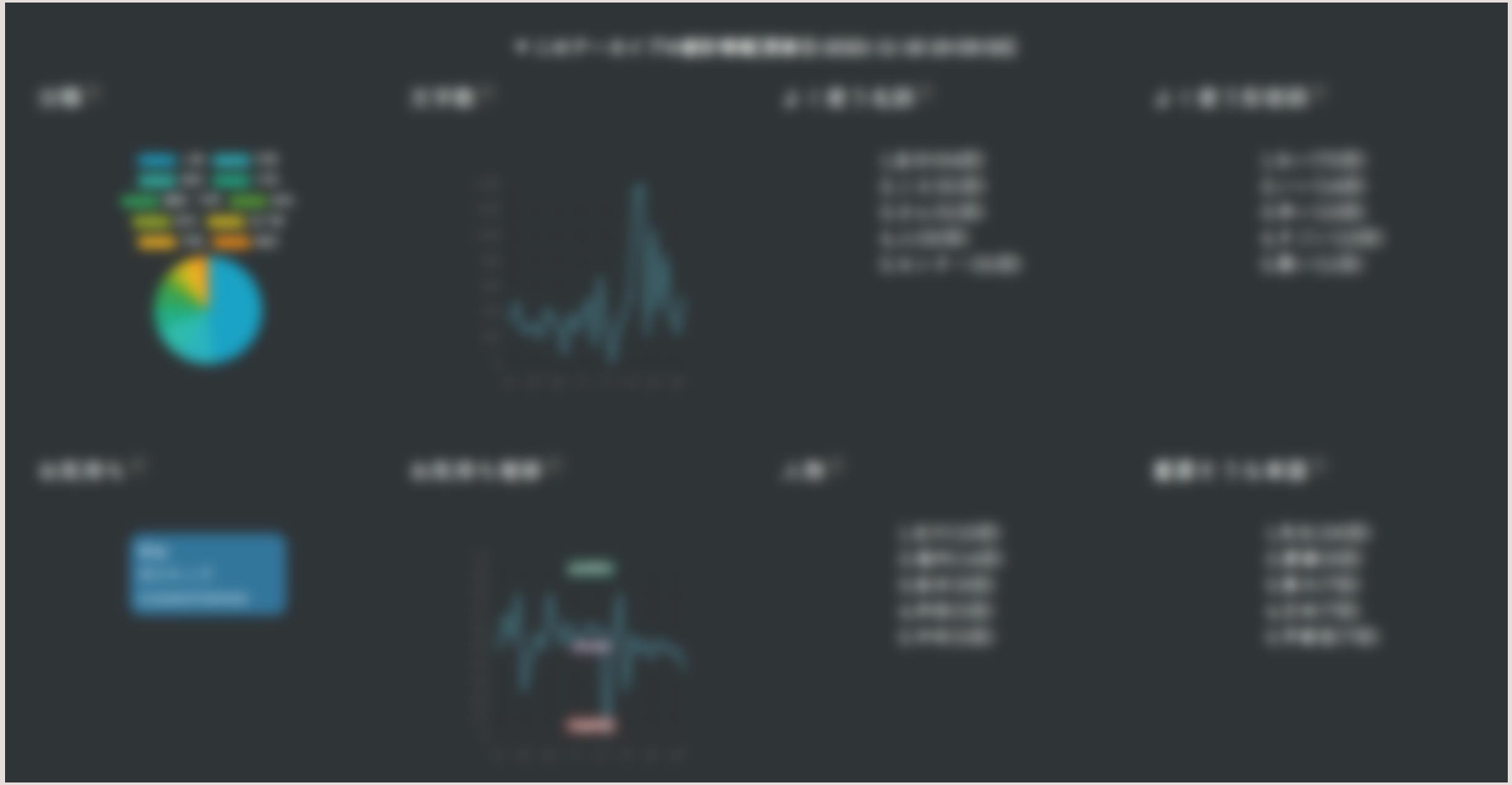
よく分からない

MONTHLY

個別　　月別　　総計

MONTHLY

この月を見てみると……



※一般公開用はぼかしています。

MONTHLY

この月を見てみると……

※一般公開用はぼかしています。

MONTHLY

この月を見てみると……

※一般公開用はぼかしています。

MONTHLY

この月を見てみると……

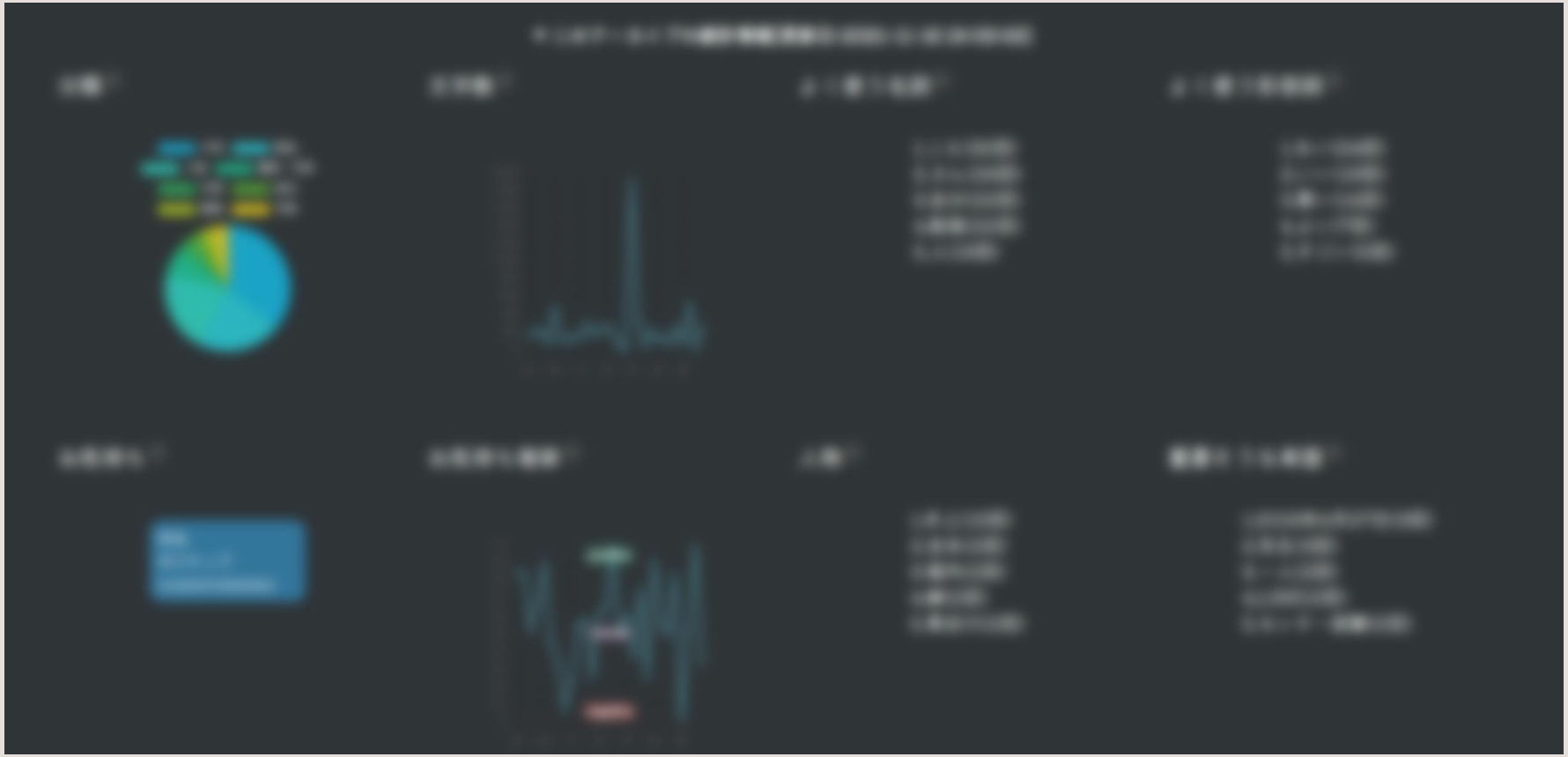
※一般公開用はぼかしています。

MONTHLY

この月を見てみると……

※一般公開用はぼかしています。

比較：2019年1月



※一般公開用はぼかしています。

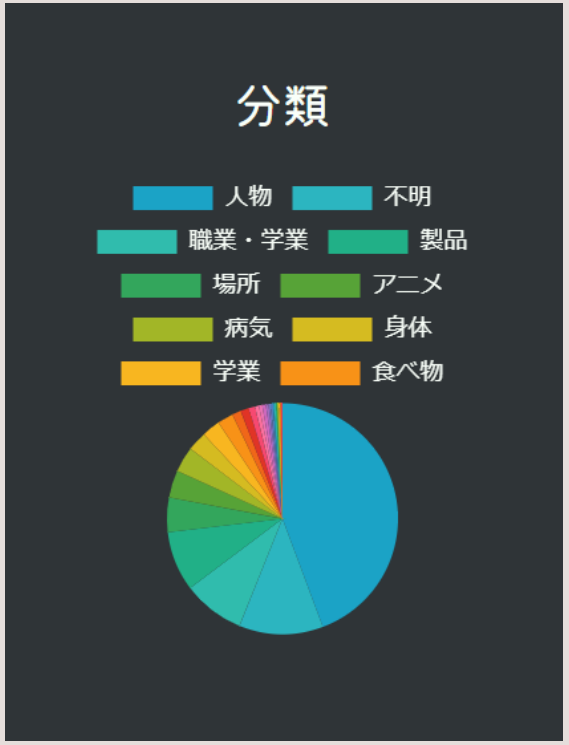
比較：2021年1月

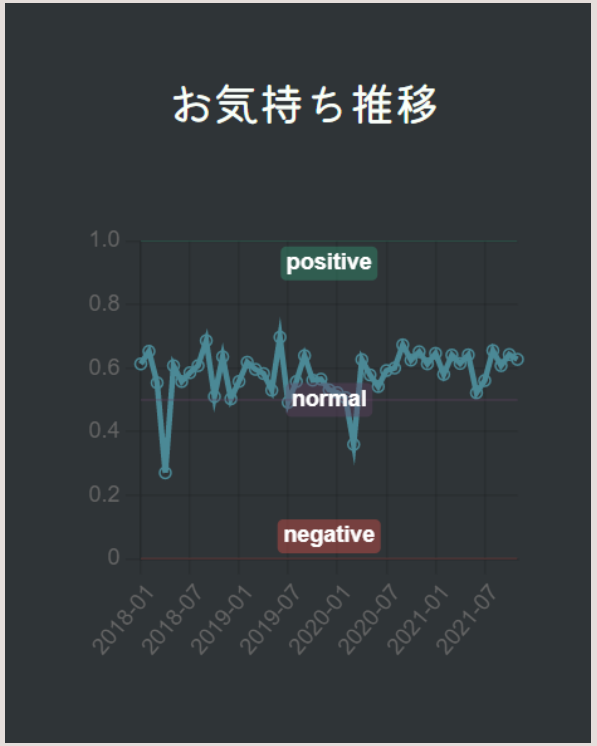
※一般公開用はぼかしています。

TOTAL

個別　　月別　　総計

TOTAL





重要そうな単語

1.先生(415回)
2.先輩(195回)
3.夜(160回)
4.日本(109回)
5.朝(101回)
6.大学生(89回)
7.LINE(84回)
8.ぶいちゃ(64回)
9.人間(62回)

TOTAL

文字数

月別

総計

CONCLUSION

結局データベースは結論を出せない。。

かどで日記 is 思い出の補助ツール

CONCLUSION

個別の解析では劣るが
過去の傾向は少しだけ役に立つかも……？

ここ発表しない

OMAKE

※一般公開用はぼかしています。

ここ発表しない

FROM NOW ON

今後やりたいこと

ここ発表しない

FROM NOW ON

現状、データのちょっとした加工しかできていない

ここ発表しない

FROM NOW ON

データを活用して別の情報を表示したい

ここ発表しない

FROM NOW ON

やりたいこと

- 共起ネットワークの表示
- 目次の自動生成
- 最近使わなくなった言葉表示

などなど

ここ発表しない

FROM NOW ON

日記からの予測は難しい

かどで日記 is 思い出の補助ツール

ここ発表しない

FROM NOW ON

フレームワークとライブラリでぶん回してるだけじゃん

分かってます……許してください……
“趣味”でしか無いので……

ここ発表しない

FROM NOW ON

今回頑張ったのは

NLP<<webとNLPをつなぐ部分やDB、UI周り

TOC

目次

END

文章から統計情報を取り出せる
自然言語処理、すごくない？

END

ご清聴ありがとうございました

うすゆき

END

参考にした資料や記事1

- Pythonで動かして学ぶ 自然言語処理入門
  ［著者:柳井 孝介, 庄司 美沙,発行年:2019年1月］
- 【深層学習】word2vec - 単語の意味を機械が理解する仕組み【ディープラーニングの世界 vol. 21】
  [URL:https://www.youtube.com/watch?v=0CXCqxQAKKQ, AIcia Solid Project]
- 【深層学習】Attention - 全領域に応用され最高精度を叩き出す注意機構の仕組み【ディープラーニングの世界 vol. 24】
  [URL:https://www.youtube.com/watch?v=bPdyuIebXWM, AIcia Solid Project]
- 【深層学習】Transformer - Multi-Head Attentionを理解してやろうじゃないの【ディープラーニングの世界vol.28】
  [URL:https://www.youtube.com/watch?v=50XvMaWhiTY, AIcia Solid Project]
- 【深層学習】ELMo - 複数粒度の文脈情報を持つ単語ベクトルで広範囲のタスク性能改善【ディープラーニングの世界vol.30】
  [URL:https://www.youtube.com/watch?v=hMrOcH5dcGM, AIcia Solid Project]
- 【深層学習】GPT - 伝説の始まり。事前学習とファインチューニングによるパラダイムシフト【ディープラーニングの世界vol.31】
  [URL:https://www.youtube.com/watch?v=wDXPXgn5hX4, AIcia Solid Project]
- 【深層学習】BERT - 実務家必修。実務で超応用されまくっている自然言語処理モデル【ディープラーニングの世界vol.32】
  [URL:https://www.youtube.com/watch?v=IaTCGRL41_k, AIcia Solid Project]
- 【深層学習】GPT-2 - 大規模言語モデルの可能性を見せ、社会もざわつかせたモデルの仕組み【ディープラーニングの世界vol.33】
  [URL:https://www.youtube.com/watch?v=3BUk7mtf10M, AIcia Solid Project]
- 形態素解析器Sudachiの「辞書」はどのように作られているか: 複数の分割単位を例として
  [URL:https://zenn.dev/sorami/articles/c9a506000fd1fbd1cf98, sorami]

END

参考にした資料や記事2

- はじめての自然言語処理 第4回 spaCy/GiNZA を用いた自然言語処理
  [URL:https://www.ogis-ri.co.jp/otc/hiroba/technical/similar-document-search/part4.html,オージス総研 技術部 アドバンストテクノロジセンター]
- GiNZA - Japanese NLP Library
  [URL:https://megagonlabs.github.io/ginza, megagonlabs]
- 日本語NLPライブラリGiNZAのすゝめ
  [URL:https://qiita.com/poyo46/items/7a4965455a8a2b2d2971, poyo46]

END

質疑応答タイムがあるらしい

かどで日記

kadodenikki3.usuyuki.net

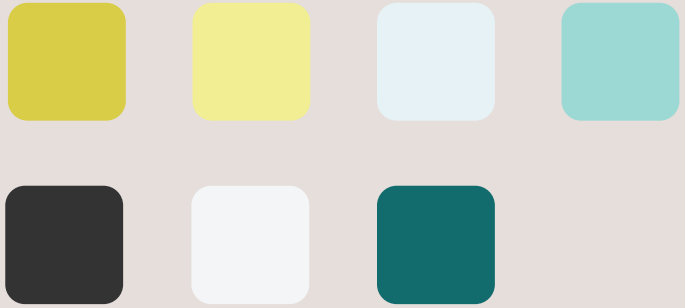
ポートフォリオ

pf.usuyuki.net

SETTINGS

カラー

和風カラーパレット 10 卯の花 base



フォント

場所	フォント
タイトル	ベストテン-CRT
見出し1	STARWAY
見出し2	HGPゴシックE
本文	Kiwi Maru Medium
注釈	JKゴシックL