good_kensaku();

tool_for_busy_engineer

Presented by C20++

Member of “C20++”

Maeda Koki

Presenter

Project Manager

Fujita Hayato

Developer

Nakamura Shukai

Data Analyst

Masui Takashi

Researcher

Uchida Yusuke

Researcher

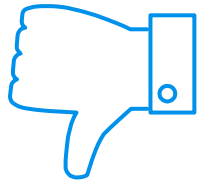
Nagata Reiji

Data Analyst

日々良い 検索がしたい…

日頃からそう思っている人は…
少ない。

こんな経験は?



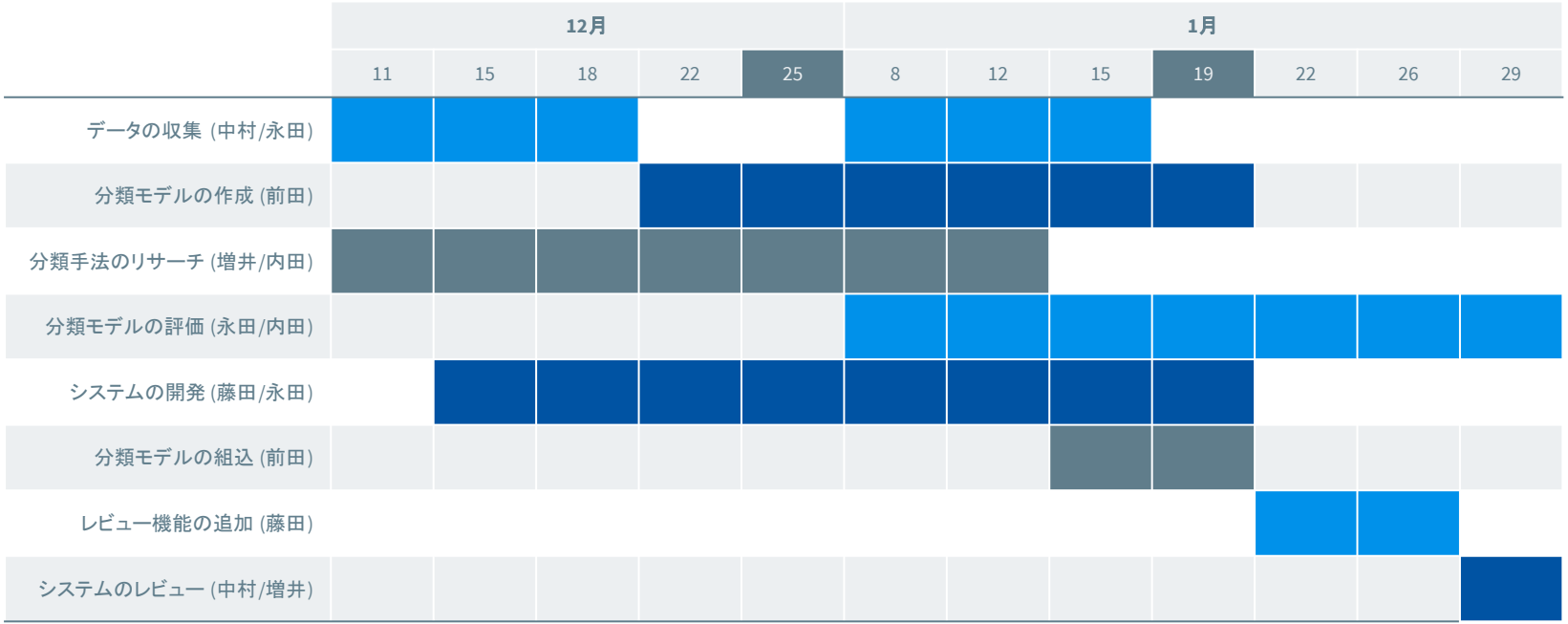
わからないことを検索

あっという間に1時間...

“良く”検索しよう

検索の質を上げると,
自由な時間が生まれる

Gantt chart


12月
1月
11
15
18
22
25
8
12
15
19
22
26
29
データの収集 (中村/永田)
分類モデルの作成 (前田)
分類手法のリサーチ (増井/内田)
分類モデルの評価 (永田/内田)
システムの開発 (藤田/永田)
分類モデルの組込 (前田)
レビュー機能の追加 (藤田)
システムのレビュー (中村/増井)

good_kensaku(); 開発に利用した技術

Slack

チームの連絡手段

GitHub

ソースコードの共有

Google Colaboratory

分類モデルの作成

Docker

どんな環境でも動作

Redis

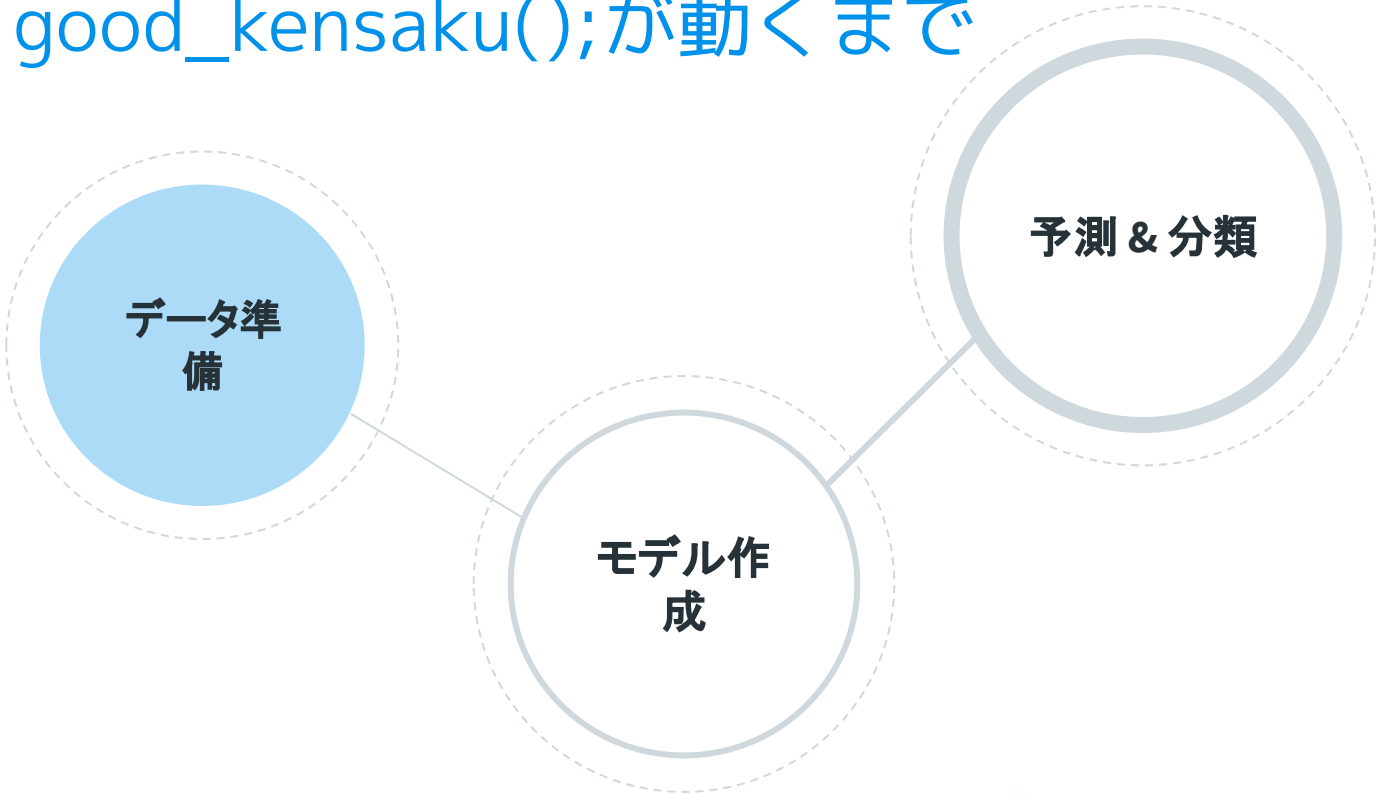
読み込み速度の改善

OpenAPI Generator

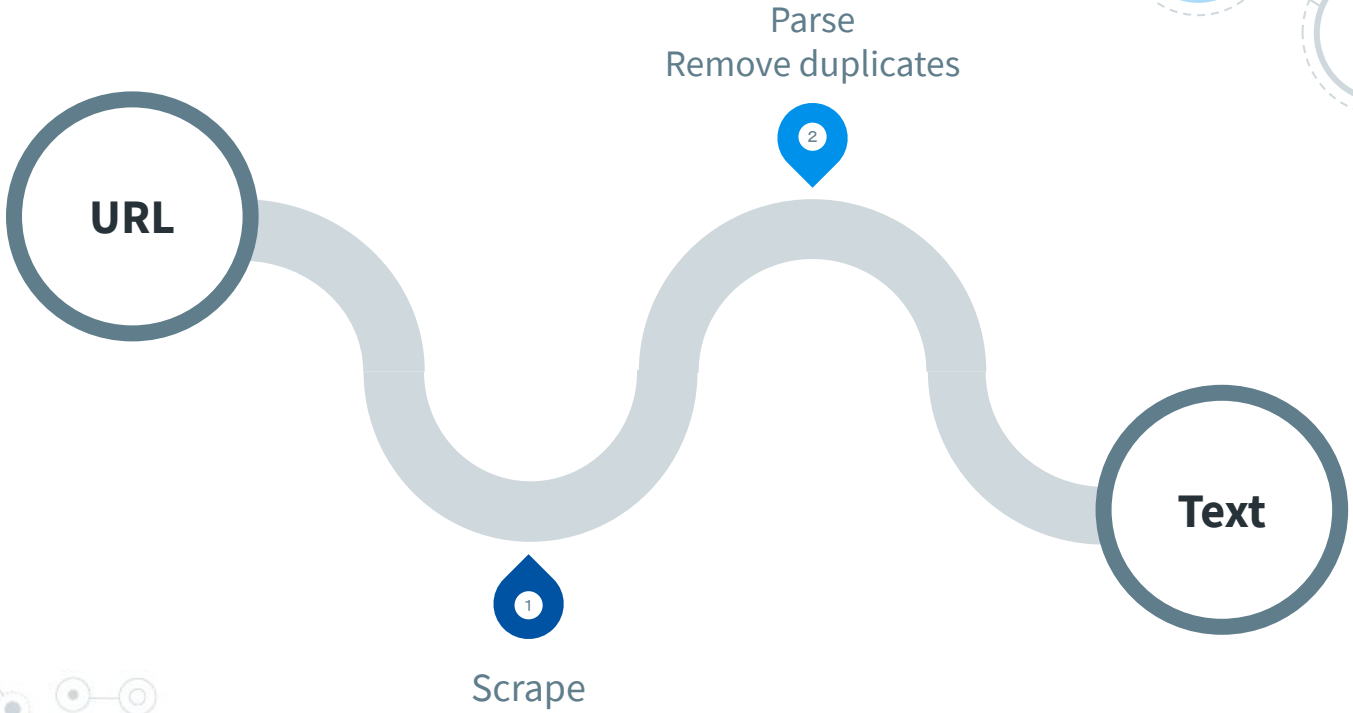
APIサーバ開発をより楽に


どう動く？
デモンストレーションを見てみよう

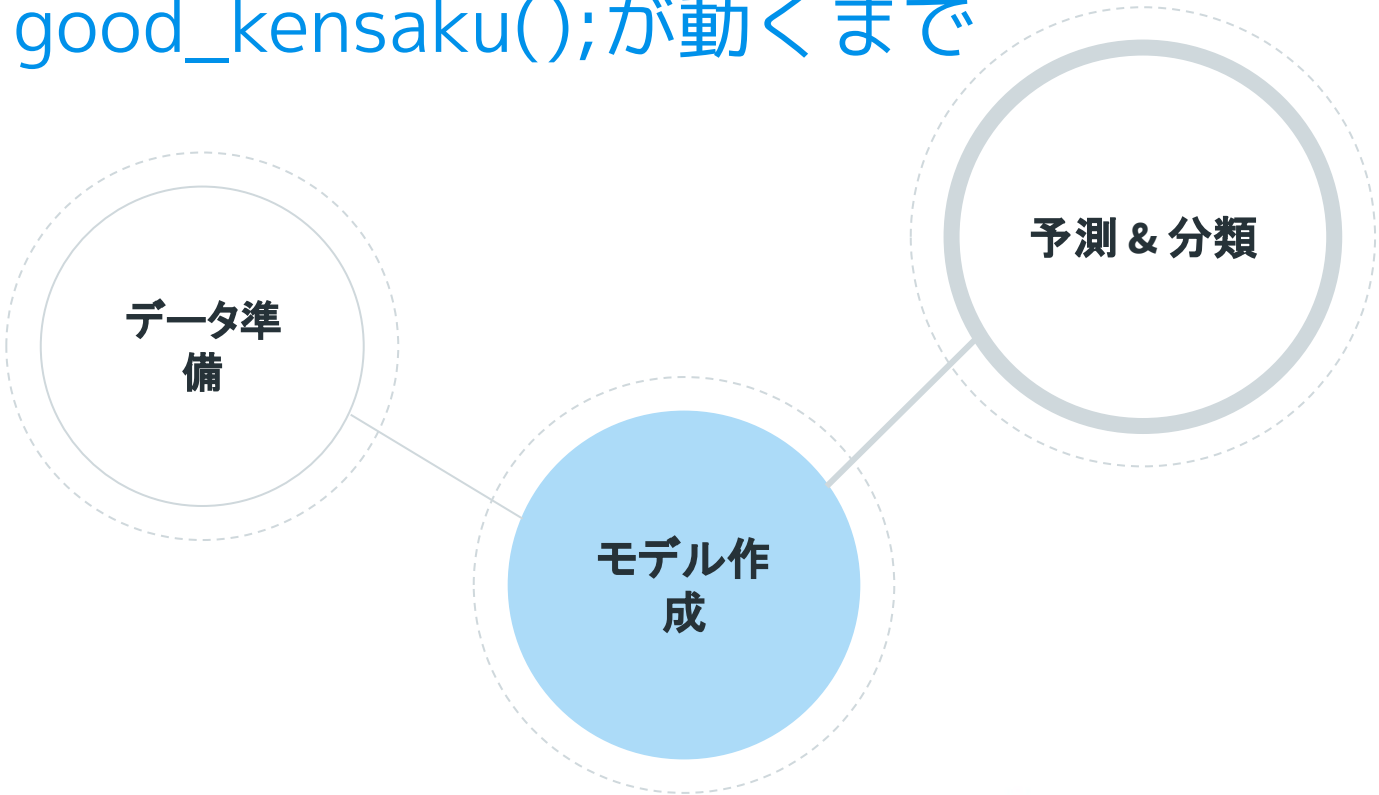
good_kensaku();が動くまで



#1. Digging, Polishing.



good_kensaku();が動くまで



#2. Embedding, Training.

Bag of Words

単語列に分解
出現頻度を計算

前後関係の欠落

LSI

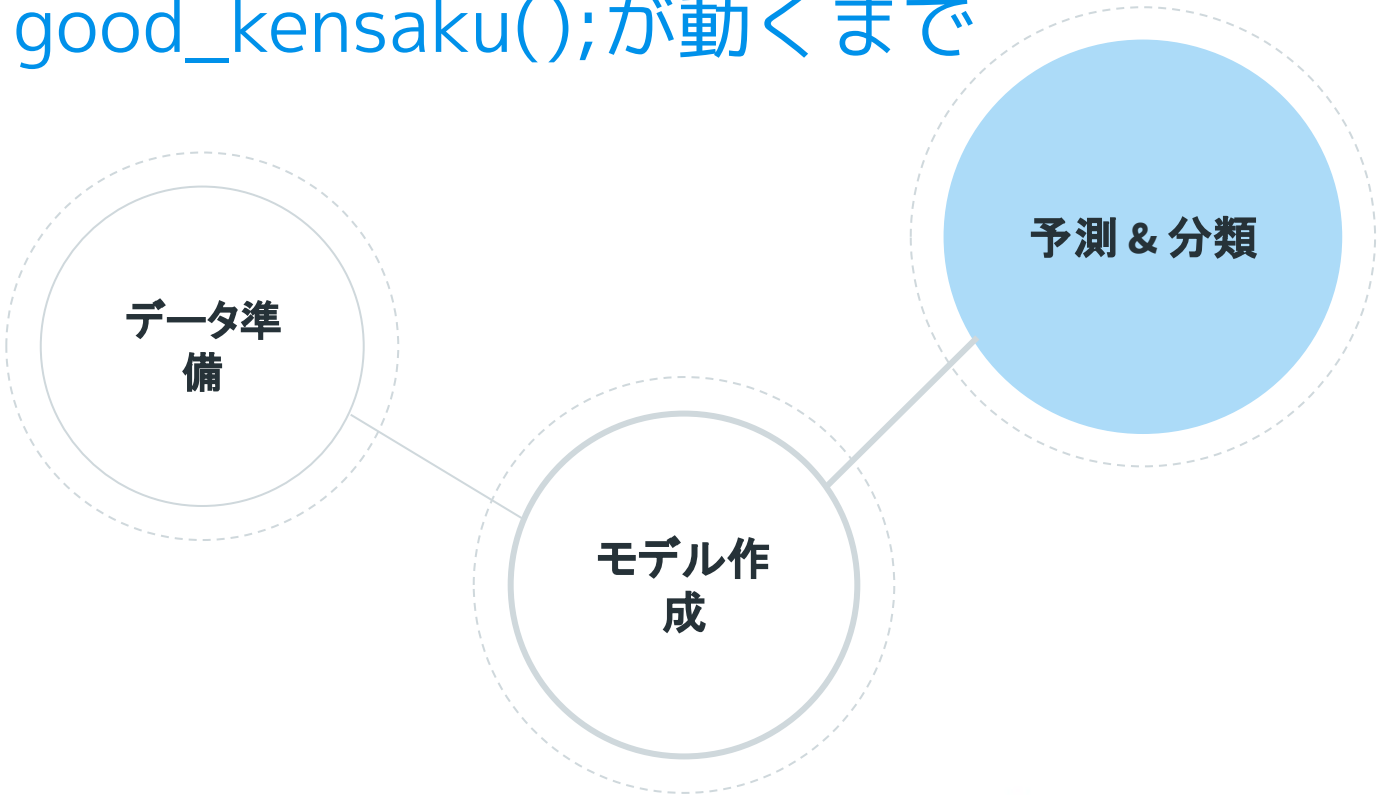
疎行列を次元圧縮
20,000超
→**300次元**に

LightGBM

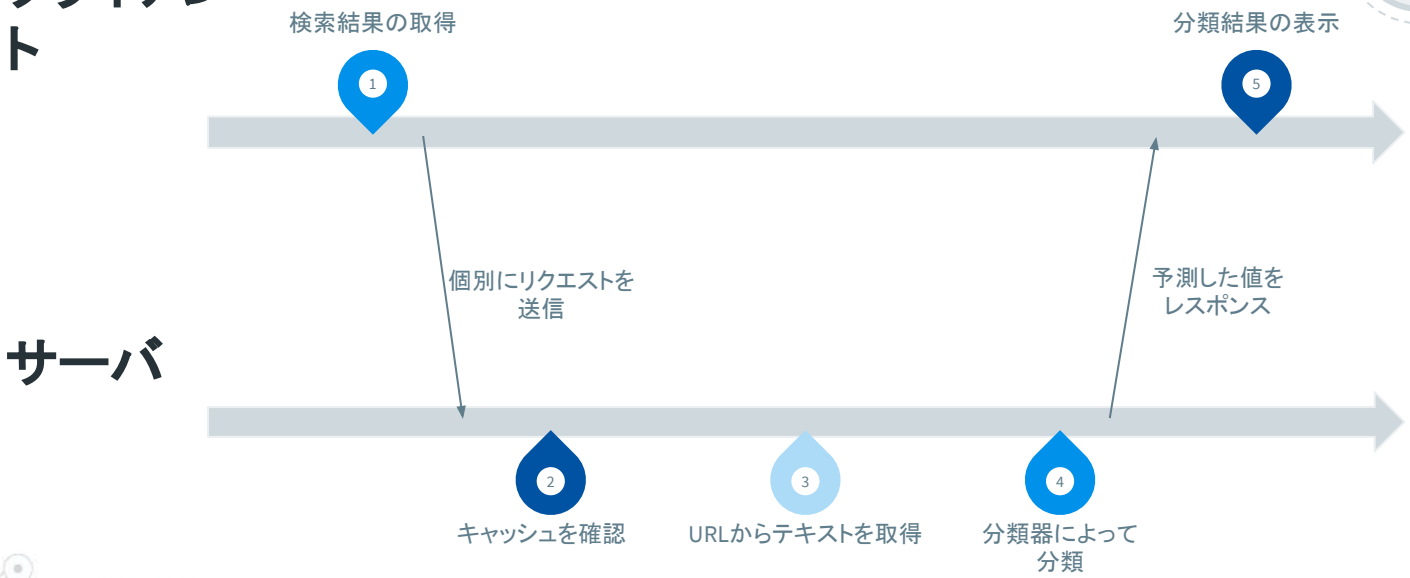
決定木を用いた学習

高速な学習

good_kensaku();が動くまで



#3. Posting, Responding

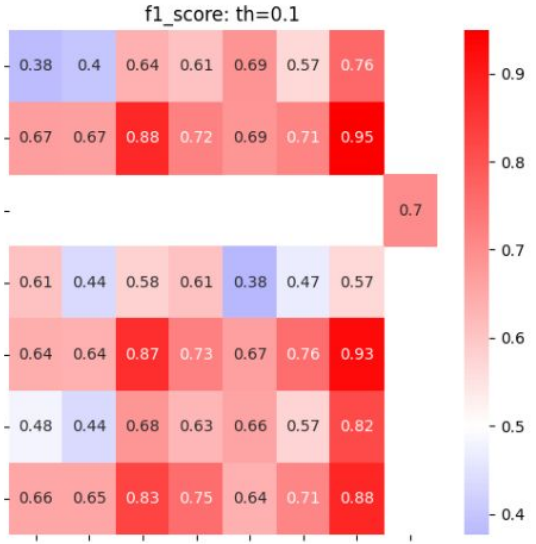


評価

対象ドメインで性能が大きく変化

類似した未知サイトが得意

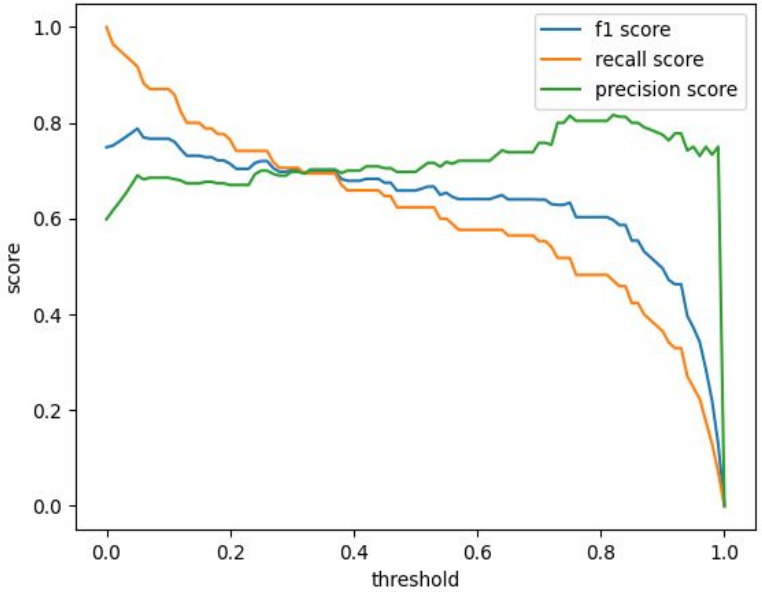
翻訳サイトの分類 / **敬体**は苦手



ユーザーからの評価

閾値で
分類の特性を変える

threshold	F1 score	recall	precision
0.1	**0.77**	**0.87**	0.67
0.5	0.66	0.62	0.7
0.9	0.5	0.36	**0.78**



評価

既知のドメイン

ラベリングされたドメインや類似した記事は正しく分類された

未知のドメイン

特色の異なるサイトを判別することが難しい
フィードバック機能で学習し直し, 改善することを期待する

作業評価

作業の分担

進捗管理

ユーザビリティ

性能向上の難しさ

サイトの処理