

第7回産業日本語研究会・シンポジウム

ニュースアプリケーションに おける自然言語処理

スマートニュース株式会社

徳永拓之

SmartNews

ニュース、サクサク。



App Store
からダウンロード

App Store でダウンロード

for iPhone/iPad



ANDROID アプリ
Google play

for Android



多彩なジャンル



圏外でも読める



話題の記事を厳選



スマートニュースのミッション

MISSION

世界中の良質な情報を
必要な人に送り届ける

Delivering the world's quality information to
the people who need it



ニュースアプリとの比較*

1,335万時間
SPENDS



A

B

C



総利用時間
(千時間)

1位
13,350

2位
4,343

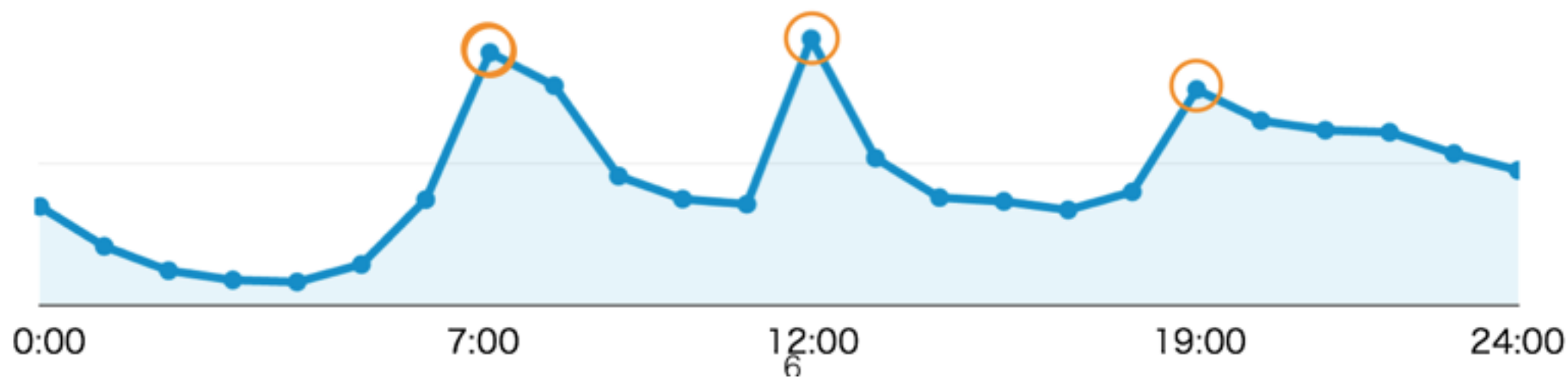
3位
4,116

4位
3,067

*Nielsen Mobile NetView 2015年3月



アクセスのピークは1日3回



アプリを開くと 記事一覧が表示される

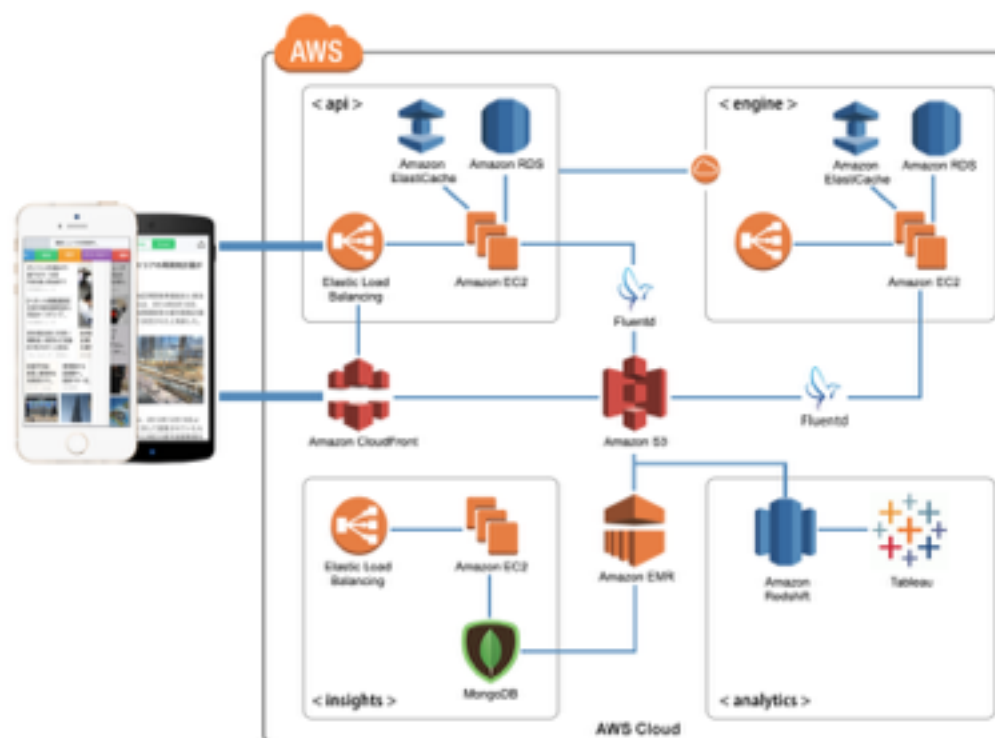
- ・ 見た目は非常にシンプル
- ・ しかしその裏側では...



その裏側では...

- 多くのサーバーが記事の解析・配信のために稼働している

「スマートニュース様システム構成図」



スマートニュースの難しさ

- ・ **どこにどの記事**を出すのか？
- ・ 記事を複数の媒体社様から頂いている
 - ・ 人手で全部チェックするのは不可能
 - ・ ソフトウェアによる自動的な解析が重要

記事解析において重要な技術

- 自然言語処理
- 画像処理
- ログ解析

NLP以外の技術も活用しています

- 画像処理
 - サムネイル画像のクロッピング
 - セクシー画像の自動判定
- ログ解析
 - 記事クオリティの判定
 - アプリの改善

自然言語処理をどこで使うか？

- ・ 記事カテゴリの分類
- ・ NG記事の分類
- ・ 関連記事の推薦
- ・ 記事主題の抽出

記事カテゴリの分類

- 典型的な文書分類
 - 本文、タイトルから記事カテゴリを当てる
 - 機械学習 + ルールベース
- Paragraph Vector (Le et al., 2014) を使用

NG記事分類

- ・ 掲載基準を満たさない記事を分類する
- ・ こちらも典型的な文書分類

関連記事の推薦

- ・ ある記事を読んだユーザーに対して別の記事をおすすめる
- ・ 現在はタイトルと本文の情報が類似する記事を推薦している

記事主題の抽出

- 記事の中心となるトピック（を表現するキーワード）を抽出する
- 曖昧性の問題が難しい
 - 「羽生」は「羽生善治」か「羽生結弦」か

記事解析と記事配信の関係

- ・ 記事解析の結果は重要なシグナルとなり得る
 - ・ 多くの記事が言及しているトピックは重要である可能性が高い
 - ・ 他にも用途は考えられる
- ・ どのカテゴリに配信するかも記事解析の結果で決まる

機械学習を使うことについて

- ・ 記事解析システムは下記の性質が重要である
 - ・ 高速に動作する
 - ・ 挙動の予想が付きやすい
 - ・ トラブル時に原因がわかりやすい
- ・ 機械学習を使うと下の2つが満たされない事が多いが、NLPでは機械学習は重要である

機械学習とのつきあい方

- ・ シンプルなケースは予めルールで処理する
 - ・ ややこしいものだけ機械学習
- ・ 機械学習の結果を上書きできる仕組みを用意する
- ・ 機械学習の性質について周知が有効な場合は、周知を怠らない

まとめ

- ・ スマートニュースは、ニュース記事を自動的に解析し配信するプラットフォームである
- ・ 記事解析を行う上で最も重要なパーツの1つが自然言語処理である
- ・ 記事カテゴリ分類をはじめとして多くの部分に自然言語処理、機械学習が使われている