Framgia Vietnam Co.,ltd

Professional Company

Friendly working environment

Recruitment

- Ruby Developer
- Senior PHP Developer
- Senior Java Developer
- Senior .NET Developer
- BrSE
- New Developer

- hr_team@framgia.com

Extract Content from Articles / News Pages

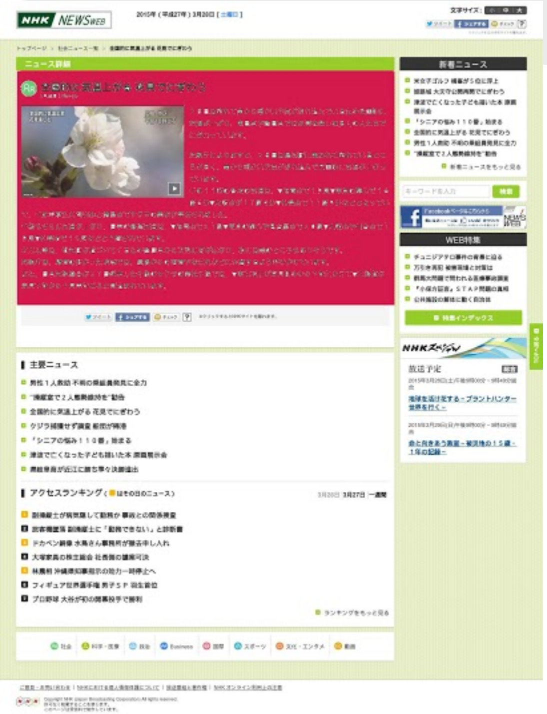
Giang Nguyen @nguyenducgiang

background

- API for a news reader app
  - web pages content extraction is vital

unify content structure

- title
- description
- readability content
- images



NHK NEWSweb 2015年（平成27年）3月28日［土曜日］

文字サイズ： 小 中 大

トップページ ＞ 社会ニュース一覧 ＞ 全国的に気温上がる 花見でにぎわう

ニュース詳細

全国的に 気温上がる 花見でにぎわう

[illegible]

新着ニュース

- 米女子ゴルフ 横峯が５位に浮上
- 姫路城 大天守公開再開でにぎわう
- 津波で亡くなった子ども描いた本 原画展示会
- 「シニアの悩み１１０番」始まる
- 全国的に気温上がる 花見でにぎわう
- 男性１人救助 不明の乗組員発見に全力
- "操縦室で２人態勢維持を"勧告

新着ニュースをもっと見る

キーワードを入力 検索

WEB特集

- チュニジアテロ事件の背景に迫る
- 万引き再犯 被害実態と対策は
- 群馬大問題で問われる医療事故調査
- 『小保方証言』ＳＴＡＰ問題の真相
- 公共施設の解体に動く自治体

特集インデックス

NHKスペシャル

放送予定

地球を遊び尽くす～プラントハンター 世界を行く～

命と向きあう教室 ～被災地の１５歳・１年の記録～

主要ニュース

- 男性１人救助 不明の乗組員発見に全力
- "操縦室で２人態勢維持を"勧告
- 全国的に気温上がる 花見でにぎわう
- クジラ捕獲せず調査 船団が帰港
- 「シニアの悩み１１０番」始まる
- 津波で亡くなった子ども描いた本 原画展示会
- [illegible]

アクセスランキング（はその日のニュース）

3月28日 3月27日 一週間

1. 副操縦士が病気隠して勤務か 事故との関係捜査
2. 医師機関等 副操縦士に「勤務できない」と診断書
3. ドカベン銅像 水島さん事務所が撤去申し入れ
4. 大塚家具の株主総会 社長側の議案可決
5. 林農相 沖縄県知事指示の効力一時停止へ
6. フィギュア世界選手権 男子ＳＰ 羽生首位
7. プロ野球 大谷が初の開幕投手で勝利

ランキングをもっと見る

社会 | 科学・医療 | 政治 | Business | 国際 | スポーツ | 文化・エンタメ | 動画

every site is different

- frontend styles
- backend configurations
- not every site is of news / article type

approach

- a general parser to extract desire information
  - any kind of pages, focus on articles / news
- custom parsers for specific cases
- post-proceed extracted content for extra goodness

notice when fetching urls

- mime types

- redirections

- user agents

- http / https

- ~~authentication~~

~~- proxies~~

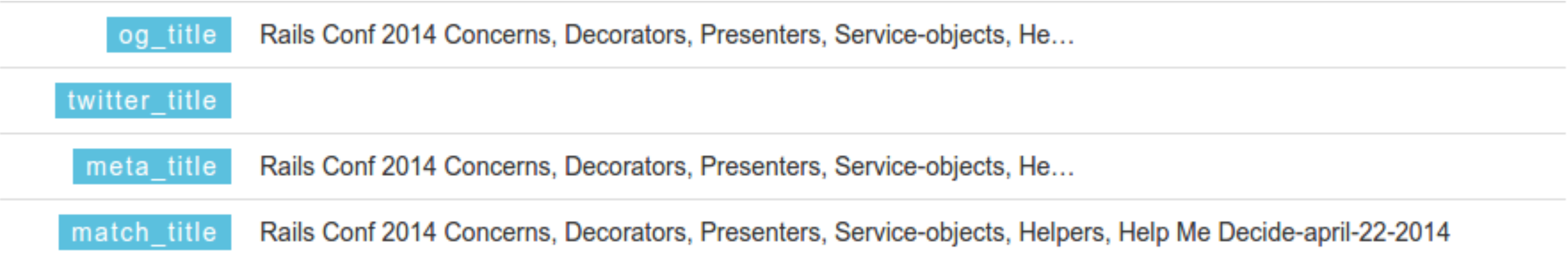
NET::HTTP is good enough!

tools & techniques

- Nokogiri FTW
- ruby-readability
- HTML metas
    - w3c specified / Open Graph / Twitter Card

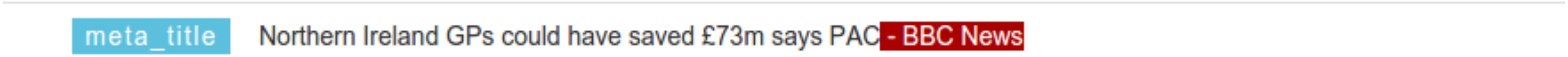
title - what could be wrong?

- length (truncated title)



og_title	Rails Conf 2014 Concerns, Decorators, Presenters, Service-objects, He…
twitter_title	
meta_title	Rails Conf 2014 Concerns, Decorators, Presenters, Service-objects, He…
match_title	Rails Conf 2014 Concerns, Decorators, Presenters, Service-objects, Helpers, Help Me Decide-april-22-2014

- site name in title



meta_title	Northern Ireland GPs could have saved £73m says PAC - BBC News

title - approach

- pattern to recognize visible title
- longest common substring
- regex to remove site name from title

description

- og:description || meta description
- yes, it just this simple!

content - headaches

- encoding
  - work for a JP company? I know that feel, mate.
- ReactJS & heavy AJAX
- complex / alien structured articles
  - multiple pages content, slides, etc.

content - approach

- ruby-readability (monkey-patched) for general use
- custom methods for specific sites
- reject overcomplicated sites

images - issues

- no images!!!
- relative path
- size
- fixed og:image

select feature image

- og:image || first image in content || snapshot of page
- some sites need specific tune up

a possible (better) approach

- machine learning?
- feed sample correct content (per site)
- let the machine
    - guess the element it should get content from the base samples
    - improve itself from guess results

what next!

- further manipulation of extracted content
    - category & type recognition
    - language recognition
- involve machine learning
- increase users' experience
    - related & recommend news / articles

conclusion

- there is no absolute / perfect solutions
  - requirements / priorities / desire
- understand the general issues make it easier to work with specific one
- there are gems for almost everything, but writing you own lib helps to understand the issues in a higher level