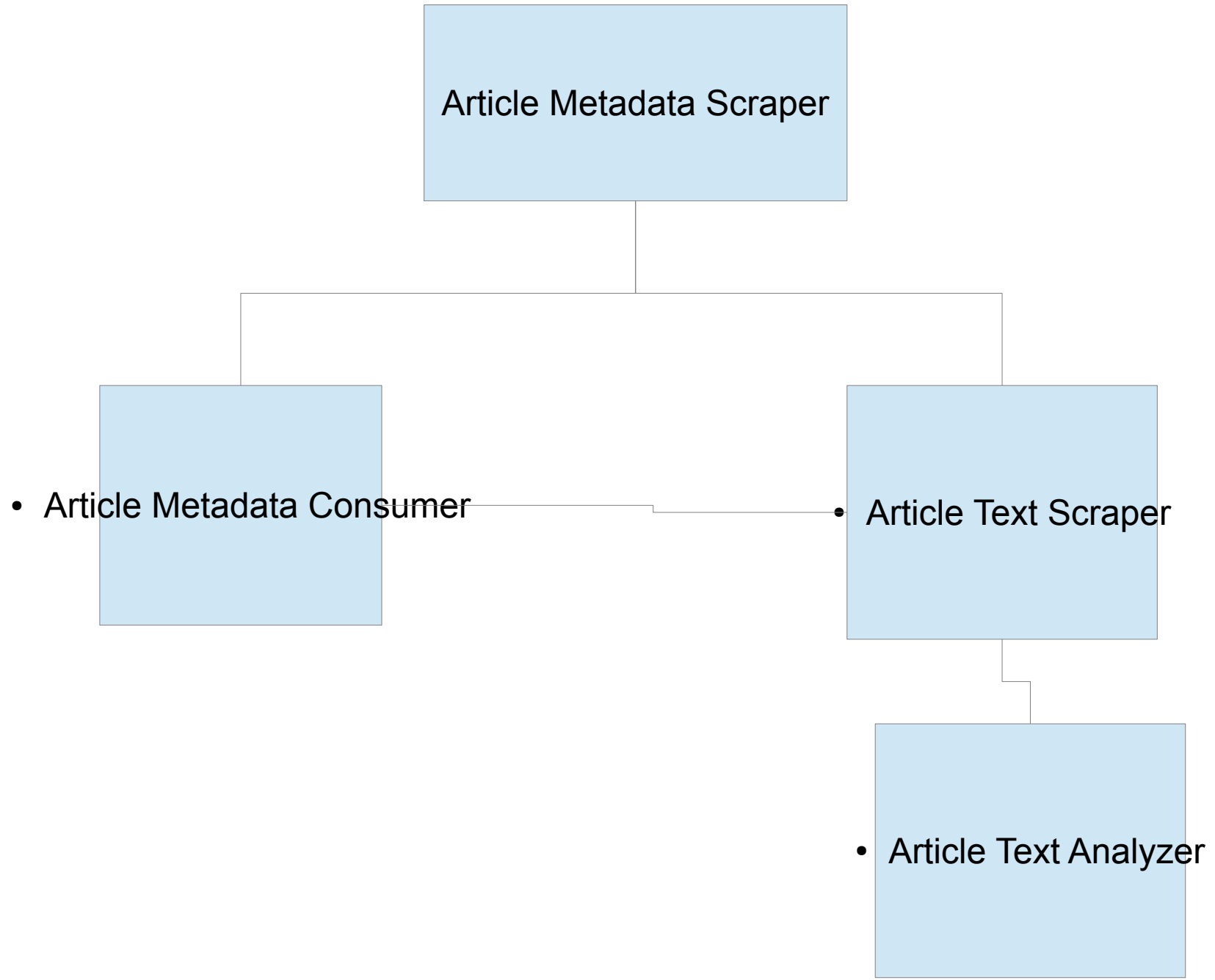
Recommendation System for Opinion Articles in Turkish Newspapers

Üstün Özgür

System Components

- Article Metadata Scraper
- Article Metadata Consumer
- Article Text Extractor
- Article Text Analyzer


Article Metadata Scraper
• Article Metadata Consumer
• Article Text Scraper
• Article Text Analyzer

Article Metadata Scraper

AMK Spor — Tüm Yazarlar

Akşam — Tüm Yazarlar

Aydınlık — Tüm Yazarlar

Birgün — Tüm Yazarlar

Bugün — Tüm Yazarlar

Cumhuriyet — Tüm Yazarlar

Dünya — Tüm Yazarlar

Evrensel — Tüm Yazarlar

Fanatik — Tüm Yazarlar

Fotomaç — Tüm Yazarlar

Güneş — Tüm Yazarlar

Habertürk — Tüm Yazarlar

Hürriyet — Tüm Yazarlar

Milliyet — Tüm Yazarlar

OdaTV — Tüm Yazarlar

Posta — Tüm Yazarlar

Radikal — Tüm Yazarlar

Sabah — Tüm Yazarlar

Star — Tüm Yazarlar

Sözcü — Tüm Yazarlar

T24 — Tüm Yazarlar

Takvim — Tüm Yazarlar

Taraf — Tüm Yazarlar

Türkiye — Tüm Yazarlar

Vatan — Tüm Yazarlar

Yeni Şafak — Tüm Yazarlar

Yeniçağ — Tüm Yazarlar

Yurt — Tüm Yazarlar

Zaman — Tüm Yazarlar

Article Metadata Scraper (contd)

- Rewritten in node.js
- Due to impedance mismatch between developer tools an Python
- Outputs a JSON document containing an array of documents
- Each document has several metadata, such as author name, newspaper name, article link

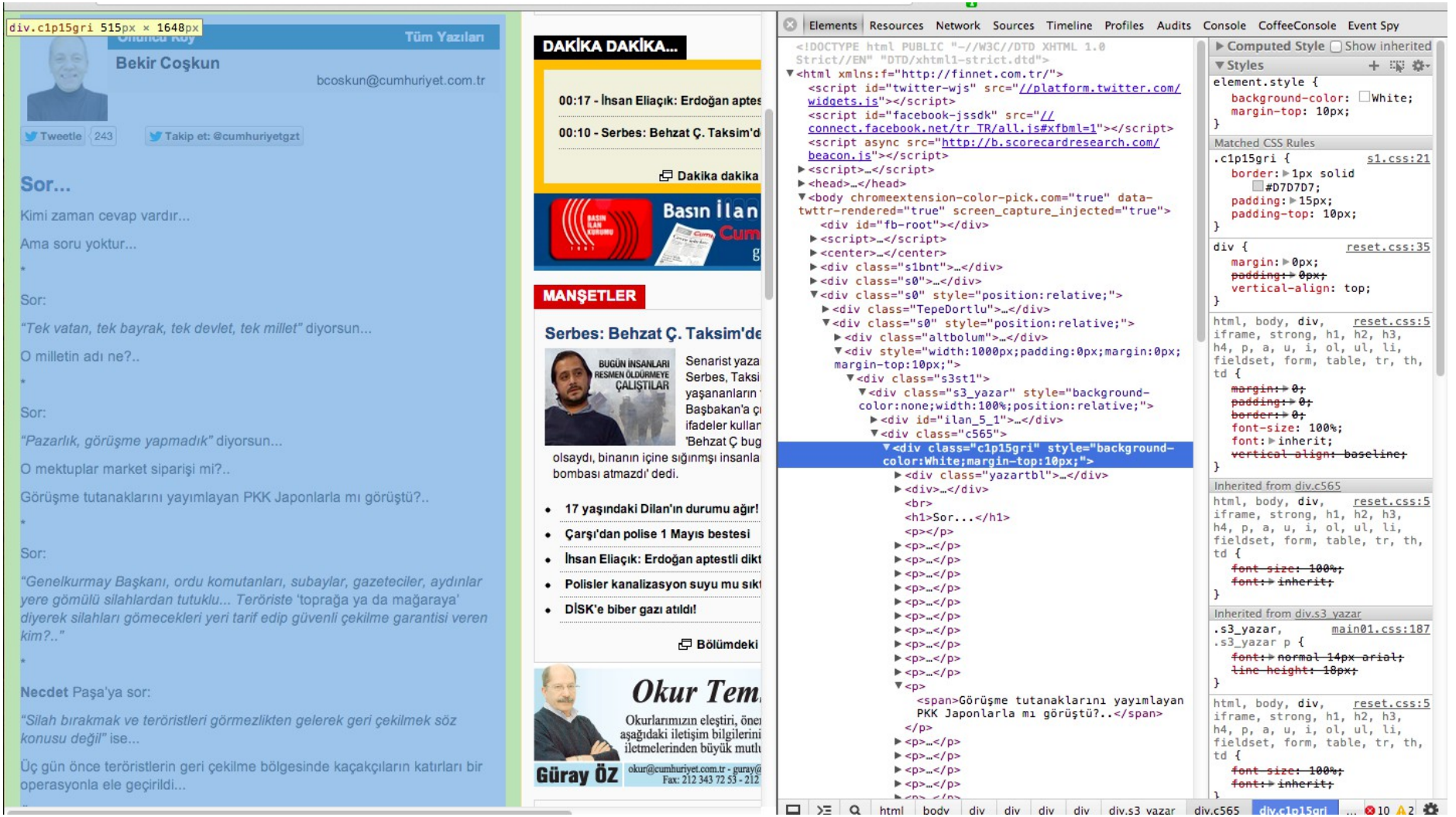
```
articles
▼ Object {articles: Array[29]}
  ▼ articles: Array[29]
    ▼ 0: Array[40]
      ▼ 0: Object
          gazeteAdi: "Akşam"
          gazeteAdiSlug: "aksam"
          yazarAdi: "İsmail Küçükkaya"
          yazarAdiSlug: "ismail-kucukkaya"
          yaziAdi: "Muhalefet ralliye hazır mı?"
          yaziAdiSlug: "muhalefet-ralliye-hazir-mi"
        ▶ yaziDate: Object
          yaziKategori: ""
          yaziKategoriSlug: ""
          yaziUrl: "http://www.aksam.com.tr/yazarlar/muhalefet-ne-yapacak/haber-200602"
        ▶ __proto__: Object
      ▼ 1: Object
          gazeteAdi: "Akşam"
          gazeteAdiSlug: "aksam"
          yazarAdi: "Deniz Gökçe"
          yazarAdiSlug: "deniz-gokce"
          yaziAdi: "Kemer sıkma yerine yapısal reform"
          yaziAdiSlug: "kemer-sikma-yerine-yapisal-reform"
        ▶ yaziDate: Object
          yaziKategori: ""
          yaziKategoriSlug: ""
          yaziUrl: "http://www.aksam.com.tr/yazarlar/kemer-sikma-yerine-yapisal-reform/haber-201314"
        ▶ __proto__: Object
      ▼ 2: Object
          gazeteAdi: "Akşam"
          gazeteAdiSlug: "aksam"
          yazarAdi: "Çiğdem Toker"
          yazarAdiSlug: "cigdem-toker"
          yaziAdi: "Görünmeden gidiyorlar"
          yaziAdiSlug: "gorunmeden-gidiyorlar"
        ▶ yaziDate: Object
          yaziKategori: ""
          yaziKategoriSlug: ""
          yaziUrl: "http://www.aksam.com.tr/yazarlar/gorunmeden-gidiyorlar/haber-201308"
        ▶ __proto__: Object
      ▶ 3: Object
      ▶ 4: Object
      ▶ 5: Object
```

•Article Metadata Consumer

- Existing Python codebase modified
- Data stored in RDMS
- Just consumes incoming data
- “Dumb” on purpose

•Article Text Extractor

- Consumes either the output of metadata scraper (currently implemented) or metadata consumer
- Separate scrapers for each article content




```
exports.selectorFn = function ($) {
    $('.yazartbl').remove();
    $('div.c1p15gri').find('div').eq(0).remove();
    $('a.smoothbox').remove();
    var txt = $('div.c1p15gri').text();
    return txt;
};
```

•Article Text Analyzer

```
(defn extract-lowercase-word-frequencies [full-text]
  "Given a text, return a map of lower-cased words excluding the stop words
and their occurrence frequency"
  (frequencies
   (map clojure.string/lower-case
        (remove short-word-p
                (remove stop-words
                        (whitespace-splitter full-text))))))
```

```
(defn n-common-words-at-indices [i j]
  (count (intersection (memoized-words-at-index i) (memoized-words-at-index j))))
```

Demo

- http://localhost:3000/yazi-short/286
- http://localhost:3000/yazi-short/100

  http://localhost:3000/yazi-short/3

Remaining Work

- More sophisticated comparison methods
- Other similarity measures
- Most common words and phrases for categorization
  - Documents containing those