CS533 – Information Retrieval Systems

SINCAP: SIMPLE NEWS' CONTENT eXTRACTION APPLICATION

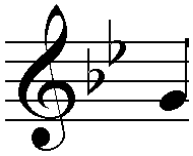
Gülsüm Ece BIÇAKCI
Abdurrahman YAŞAR

Outline

- Introduction
- Methodology
- Algorithm
- Infrastructure
- Experiments
- Results
- Conclusion
- Future Work

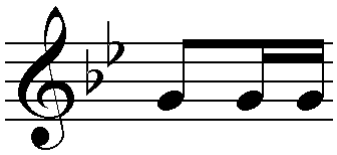
Introduction

- Usage of internet increases
- Newspapers have lots of noise
- Aim: erase the noise and get the content with an application



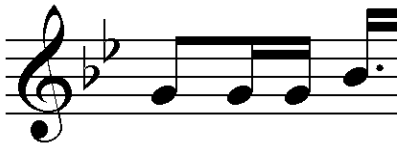
Methodology

- Parsing HTML documents
  - Removing unnecessary parts
  - Get the header and related part
- Get the content
- Send it to database
- Showing the data with SINCAP app



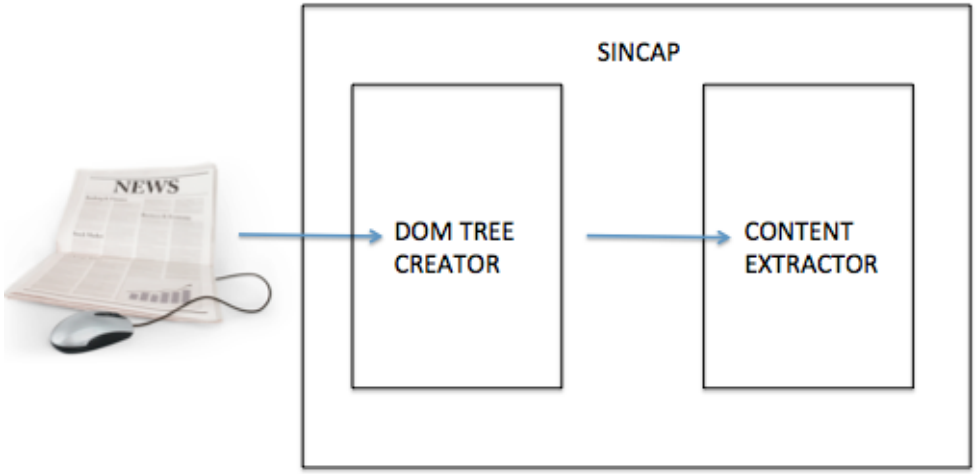
Algorithm

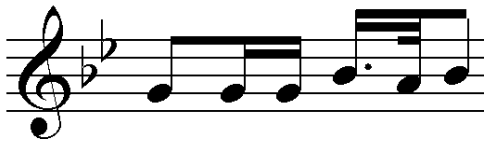
- Get URLs of news
- Parsing their HTML docs
  - Creating DOM tree
- Find published date
- Look for title
- Get content and header



Infrastructure

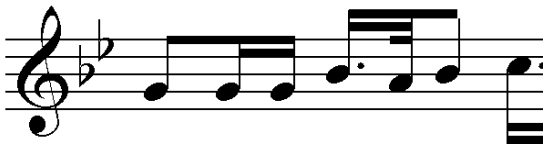
- Java, php, SqlLite





Experiments

- With several newspapers
- News from at most 3 days



Newspapers…

Hürriyet

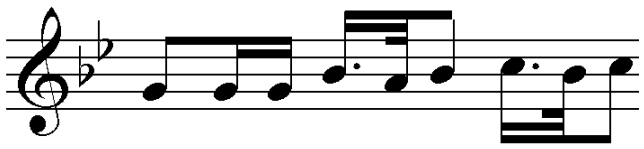
Milliyet
BASINDA GÜVEN

Radikal

SABAH

Sözcü

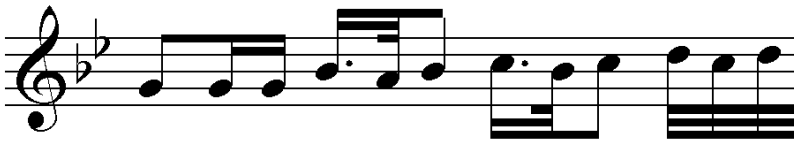
Yeni Şafak
TÜRKİYE'NİN BİRİKİMİ



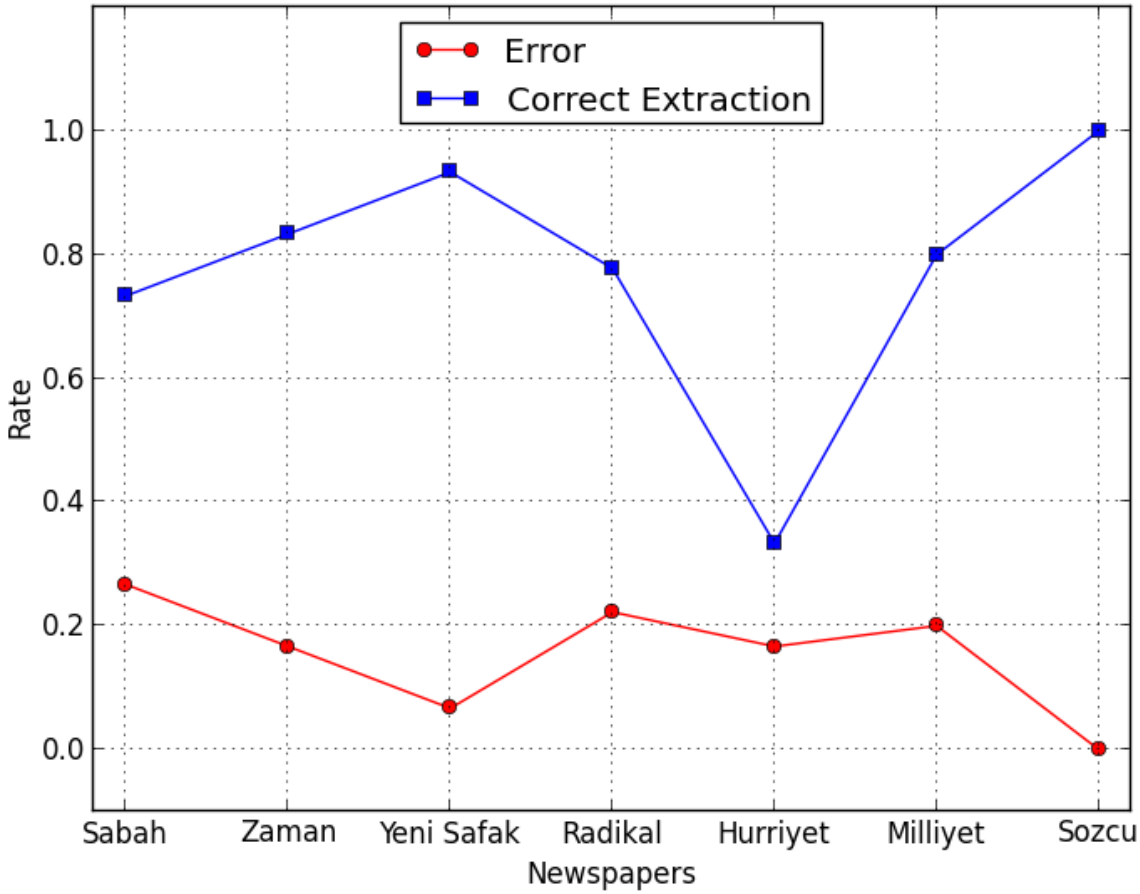
ZAMAN

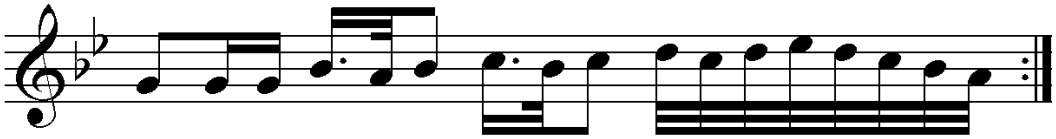
Results

- Get good results in well structured news
  - Sabah, Yeni Şafak, Milliyet
- Avg results in news have many banners



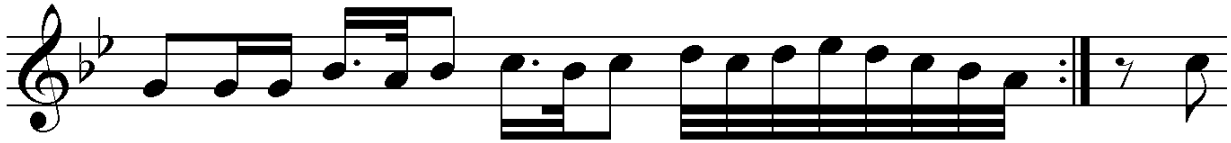
Results


Error
Correct Extraction
1.0
0.8
0.6
0.4
0.2
0.0
Rate
Sabah
Zaman
Yeni Safak
Radikal
Hurriyet
Milliyet
Sozcu
Newspapers



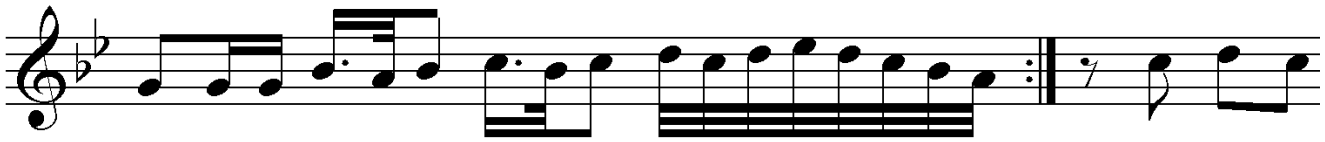
Conclusion

- Solving a problem with Information Retrieval perspective
- Doing a real life application

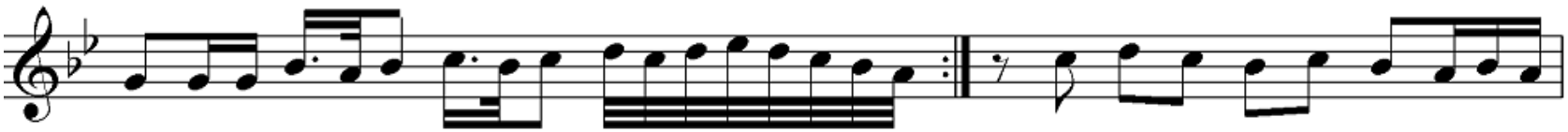


Future Work

- Extend the project
- Mobile application form



Demo…



Thank You…


he will be remembered.
Funny people, folk.
with the exhilarating Grand
Prix, for example, and his
spat with his dangerous