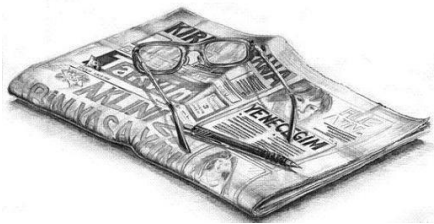




CS533-Information Retrieval Systems

TURKISH NEWS PAGE CONTENT EXTRACTION



Gülsüm Ece BIÇAKCI
Abdurrahman YAŞAR

PROBLEM DESCRIPTION



- ✓ Usage of internet increases
- ✓ Time is precious
- ✓ Criticising point of view

MOTIVATION TO PROBLEM



- ✓ **Web sites have**
 - ✓ *Advertisements*
 - ✓ *Banners*
 - ✓ *Hyperlinks*
 - ✓ *Reader Comments*

**Extract main content
from the web sites of Turkish news**

METHODOLOGY



- ✓ Analysis of HTML documents
 - ✓ Tags: `<p>..</p>`, `<div>..</div>`
 - ✓ Relevant picture of the article

- ✓ Analysis of keywords
 - ✓ Words start with capital letter
 - ✓ Verbs, Subjects etc.

- ✓ Analysis of articles
 - ✓ Finding introduction and conclusion
 - ✓ Paragraph relations

EXPECTED RESULTS



- ✓ **Successful content extraction**
 - ✓ **High accuracy**
- ✓ **Successful figure extraction**
 - ✓ **Picture of the article**



THANK YOU...