“SINCAP: Simple News’ Content eXtraction Application”

CS533-Final Project Presentation

[Handout]

Bıçakcı E., Yaşar A.

Overview

In this project our aim is to develop a simple and efficient fully automatic application for extracting contents from online Turkis news. To represent web pages SINCAP uses a DOM tree. SINCAP firstly finds the node where some part of the new is wrapped firstly then track this node to find a summary node. In our application we will consider several different online newspapers like Milliyet, Radikal, Sabah, Zaman, Sozcu etc.

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

Experiments

- With several newspapers
- News from at most 3 days
- Hurriyet, Milliyet, Radikal, Sabah, Sozcu, Yeni Safak, Zaman

Evaluation & Conclusion

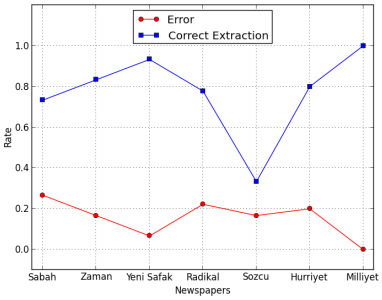
We have measured two different results of our application:

1. Error Rate
2. Correct Content Extraction

Environment Set-up

- We have coded SINCAP using Java
- Mac Book Pro i5 10Gb of Ram

Experimental Results:



High level description of SINCAP