

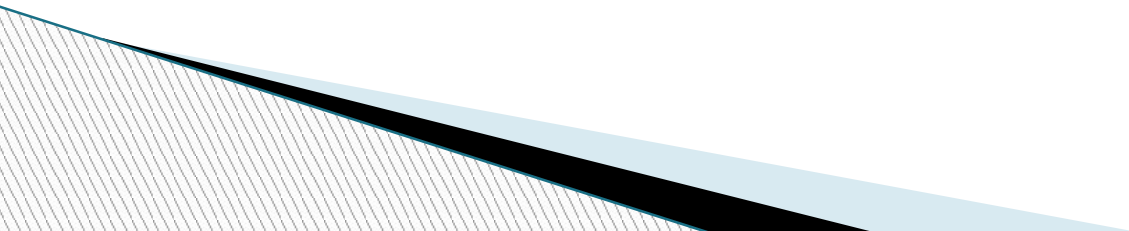
RSoC 2013 - DexOnline

Romanian literature crawler

Student: Alin Ungureanu
Mentor: Cătălin Frâncu

De unde am plecat?

Pornind de la discuțiile de pe mail purtate cu mentorul, am creat un document de specificații software și de design.



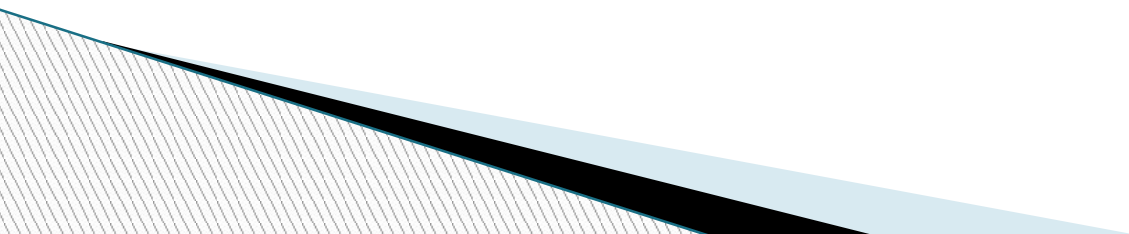
Ce am realizat?

- ▶ O aplicație pentru crawlat site-uri de literatură română (care știm noi că sunt de încredere).

Aplicația descarcă pagina respectivă, extrage textul fără elemente HTML, apoi urmează toate linkurile relevante care nu părăsesc site-ul respectiv. Crawlerul dispune și de o pagină în browser care îi monitorizează activitatea.

- ▶ O aplicație care inserează diacritice într-un text dat.

Aplicația se compune din 2 module:



► mecanismul de învățare:

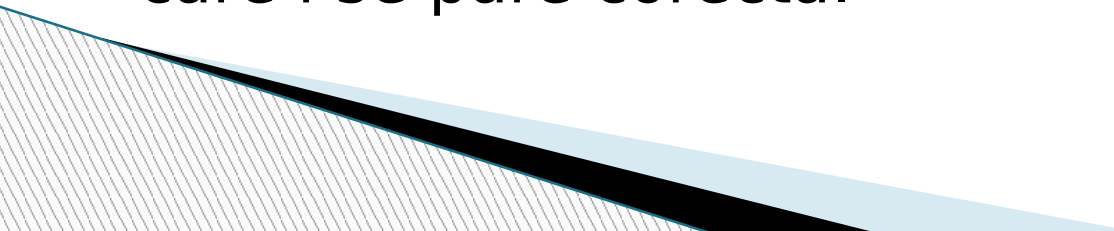
- parsează textele extrase de crawler, căutând cuvintele ce conțin diacritice pentru a identifica formele acestora

- salvează formele găsite în baza de date

► mecanismul de inserare diacritice propriu-zis:

□ este o pagină în care utilizatorul introduce o bucată de text căreia dorește să-i insereze diacritice.

- dacă există mai multe variante ale unui cuvânt atunci va apărea o listă (dropdown select) sortată în ordine descrescătoare a aparițiilor, oferindu-i-se astfel utilizatorului posibilitatea de a alege forma care i se pare corectă.



Bineînțeles că ...

- ▶ În realizarea acestei prezentări am folosit mecanismul pentru inserare diacritice peste textul prezentării scris inițial cu caractere latine.

Tehnologii folosite

- ▶ Crawler

php, php-curl, simple_html_dom, mysql, idiorm si paris

- ▶ Crawler monitor

php, mysql, idiorm si paris, jquery, ajax, html, css

- ▶ Mecanismul de inserare diacritice

php, mysql, idiorm si paris, html si css



Probleme întâmpinate

- ▶ Transformarea linkurilor relative în absolute
- ▶ Canonizarea linkurilor

Un link canonic se poate scrie în mai multe feluri

www.example.com

www.example.com/

www.example.com/index.html

www.example.com////////index.html sau .php, .jsp, aspx, etc (directory index cel mai probabil)

- ▶ identificarea conținutului paginii descărcate

- ▶ Broken HTML

Taguri lăsate deschise, lipsa elementului `<body>`

- ▶ determinarea codului HTTP pe care îl returnează serverul pentru pagina cerută.

- ▶ Lungimea variabilă a caracterelor unicode:
pe utf8-general-ci (pe acesta îl folosește
DexOnline) caracterele latine ocupa 1 octet pe
când cele cu diacritice ocupă 2 octeți, îngreunând
astfel codul și mentenanța acestuia.

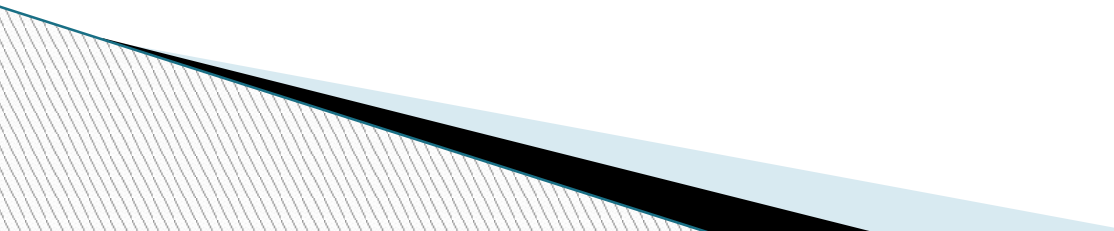
- ▶ diferența dintre null și 0

Valoarea booleană a lor este aceeași, trebuie
verificat tipul variabilei

0 === null returnează false

0 == null returnează true

Ce am învățat?

- ▶ Code review-ul contează foarte mult în dezvoltarea unei aplicații, mai ales dacă cel care implementează aplicația utilizează tehnologii cu care nu este obișnuit, altfel va scrie cod greoi.
 - ▶ N-are rost să reinventez roata, mereu trebuie să caut să folosesc ceva deja implementat.
- 

- ▶ Hardcodările sunt o ‘crimă’ pentru cei care preiau dezvoltarea aplicației, fișierul de configurare este ‘sfânt’.
 - ▶ Separarea tehnologiilor folosite într-o aplicație face modificările ulterioare să fie mult mai simplu de realizat.
- 