Seminarski rad iz Informatičkog praktikuma 4

Ukoliko student nije na spisku **http://www.fil.bg.ac.rs/branislava/seminarski/ip4/**, potrebno je da se prijavi slanjem mejla na **milica.ikonic.nesic@fil.bg.ac.rs**. Rok za prijavljivanje je **25.03.2021.** godine. Zaduženje se sastoji u tome da student pretvori tačno određene članke ili vesti u XML dokument u skladu sa uputstvom. Svaki student treba da obradi po tri proizvoljne vesti objavljene određenog datuma u sledećim dnevnim novinama, odnosno TV stanicama:

Politika (latinično izdanje) - http://www.politika.rs/
Danas - http://www.danas.rs/
B92 - http://www.b92.net/
Večernje novosti - http://www.novosti.rs/
Blic - http://www.blic.rs/

Priprema vesti (I deo)

Pošto se naslovne stranice navedenih medija menjaju svakodnevno, umesto adresa medija treba posetiti adrese njihovih arhiva. Pošto pretraga arhiva zahteva vreme, najjednostavnije je konsultovati stranicu **http://www.fil.bg.ac.rs/branislava/seminarski/ip4/** na kojoj je pored imena svakog studenta njegovo zaduženje (datum) i linkovi na odgovarajuće arhive vesti. Ako iz nekog razloga ova stranica ne bude dostupna, rezervno rešenje je stranica **http://www.naslovi.net**. Direktni linkovi na arhivu vesti za, na primer, 14. februar 2021. godine su:

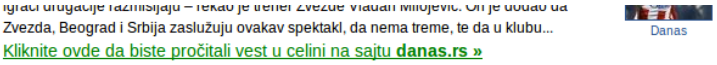
http://www.naslovi.net/**2021-02-14**/izvor/politika
http://www.naslovi.net/**2021-02-14**/izvor/danas
http://www.naslovi.net/**2021-02-14**/izvor/b92
http://www.naslovi.net/**2021-02-14**/izvor/vecernje-novosti
http://www.naslovi.net/**2021-02-14**/izvor/blic

U daljem tekstu **gggg**, **mm** i **dd** označavaju **godinu**, **mesec** i **dan** dobijenog datuma (zaduženja).

Prebacivanje vesti u mašinski čitljiv oblik se vrši u nekoliko koraka

1. Koristeći isključivo Notepad(++)[1], kopirati celokupan sadržaj vesti u tekstuelne datoteke (po jednu za svaku od gore navedenih novina/stranica). Poželjno je da vesti budu raznovrsne, odnosno da pokrivaju različite teme.

OPREZ: stranica **www.naslovi.net** prikazuje samo prvi pasus vesti, za kompletnu vest potrebno je otići na samu stranicu vesti, na sajtu originalnog izvora



Vesti istih novina/stranica se snimaju u jednoj datoteci koristeći Unicode (dakle, sačuvati kao tip **.txt**, polje **Save as type: All Files**, polje **Encoding: UTF-8**). Ime datoteke **ne sme** da bude proizvoljno. Ako **gggg**, **mm** i **dd** označavaju **godinu**, **mesec** i **dan** dobijenog datuma imena datoteka moraju biti:

politika-gggg-mm-dd-lat.txt (npr. **politika-2021-02-14-lat.txt**)
danas-gggg-mm-dd-lat.txt (npr. **danas-2021-02-14-lat.txt)**
b92-gggg-mm-dd-lat.txt (npr. **b92-2021-02-14-lat.txt)**
novosti-gggg-mm-dd-lat.txt (npr. **novosti-2021-02-14-lat.txt)**
blic-gggg-mm-dd-lat.txt (npr. **blic-2021-02-14-lat.txt)**

2. Dobijene datoteke treba iskopirati u datoteke **politika-gggg-mm-dd-ascii.txt** itd. i sve dalje obrade raditi na kopijama (u slučaju fatalne greške pri obradi uvek možete de se poslužite originalima za ponovno kopiranje – **originali se čuvaju netaknuti**).

[1] Student može da koristi i druge programe za obradu teksta dokle god svaku od datoteka snima u odgovarajućem kodnom rasporedu (UTF-8, ANSI i sl.), pri čemu rezultat može da se otvori u Notepad-u sa istovetnim sadržajem (**WordPad i Word to onemogućavaju i zato ih treba izbegavati**). Pogodna rešenja su besplatni program **PSPad** (http://www.pspad.com) i Notepad++ (http://notepad-plus-plus.org). Objašnjenja u tekstu se odnose na Notepad.

U datotekama ***-ascii.txt** zameniti **karaktere karakteristične za srpski jezik: Š, š, Ž, ž, Ć, ć, Č, č, Đ, đ, Dž, dž, Nj, nj, Lj, lj** na sledeći način (tj. u skladu sa kodnom shemom **aurora**):

Veliko slovo	Kôd aurora	Malo slovo	Kôd aurora	Primer teksta kodiranog pomoću sheme aurora
Š	**Sx**	**š**	**sx**	**Šuškati = Sxusxkati**
Ć	**Cx**	**ć**	**cx**	**Ćićevac = Cxicxevac**
Č	**Cy**	**č**	**cy**	**Čačak = Cyacyak**
Đ	**Dx**	**đ**	**dx**	**Đorđe = Dxordxe**
Dž	**Dy**	**dž**	**dy**	**Džordž = Dyordy**
Ž	**Zx**	**ž**	**zx**	**Žižak = Zxizxak**
Nj	**Nx**	**nj**	**nx**	**Njegoš = Nxegosx**
Lj	**Lx**	**lj**	**lx**	**Ljubljana = Lxublxana**

Prilikom zamene koristiti **Find** & **Replace** pri čemu treba **strogo voditi računa da je izabrana opcija Match case**, kako se ne bi velika slova zamenjivala malim i obrnuto. Npr. veliko **Š** se predstavlja isključivo kao **Sx** (**kombinacije SX i sx nisu dozvoljene**). Posebnu pažnju treba obratiti na **Ć** i **Č** (i veliko i malo) kao najčešći izvor grešaka. Greške su moguće i kod **đ i dž**: leksemu *odjednom* ne treba zameniti u *ođednom*, kao ni *nadživeti* u *nadyiveti* (dakle, ne treba koristiti *Replace All* za ova slova). Takođe, ne treba menjati **nj** u leksemama *konjuktura*, *injekcija*, *Tanjug* i sl. Naposletku, **Dž** **ne treba kodirati** kao Dzx (ovo se dobija pri zameni **ž** sa **zx**) ili Dyx (što se dobija zamenom Dz u Dzx sa Dy), već **ISKLJUČIVO** sa **Dy**. Stoga je najbolje **PRVO** zameniti sva pojavljivanja slova **Dž** i **dž**, pa tek onda slova **Ž** i **ž**.

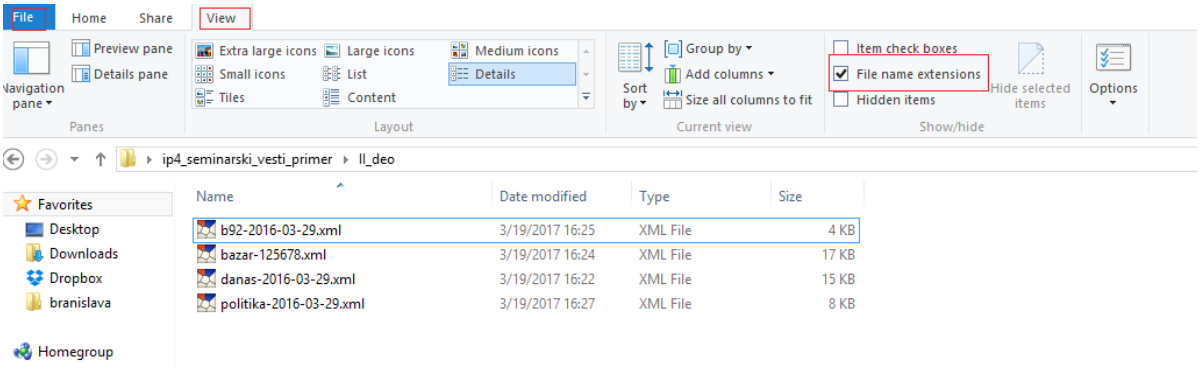
Sve ostale ne-ASCII karaktere treba u Notepad-u zameniti odgovarajućim ASCII-karakterima. Uglavnom će biti u pitanju crtice, navodnici, apostrofi koje treba iskopirati u Find, a potom u Replace otkucati odgovarajući ASCII karakter i primeniti Replace All. Potom treba ponoviti kontrolu. Paziti na sledeće:

- Prilikom kucanja crtica, apostrofa i navodnika koristiti ASCII karaktere **'** i **"**. Navodnik " je jedan karakter i ne treba ga kucati kao dva apostrofa! Znači, navodnike tipa **»** ili **«** ili **″** ili **“** ili **”** ili **„** treba otkucati kao **" (jedan karakter!)**. Takođe, apostrofe tipa ´ ili ` ili ´ ili ’ ili ‘ treba otkucati kao '.
- Akcentovana slova treba kucati kao neakcentovana: npr, ako se u tekstu pojavi neki od karaktera **ō**, **ò**, **ó**, **ô** ili **ő**, svaki od njih treba otkucati kao **o**. Diftonge **æ**, **œ** treba kucati kao **ae**, **oe**, nemačka slova **ä**,**ö**, **ü, ß** se redom zamenjuju sa **ae**, **oe**, **ue, ss**. U slučajevima kada nije jasno kako treba otkucati deo teksta na stranom jeziku, konsultovati se sa asistentom.
- **Kad god je moguće karakterske entitete treba zameniti odgovarajućim ASCII-karakterima.** To se pre svega odnosi na karakterske entitete **"** (navodnik) i **'** (apostrof), ali i na karakterske navodnike kojima su predstavljene crtice ili slova. Jedini karakterski entiteti koji se ne smeju menjati su **<**, **>** i **&**.

3. Po obavljenoj zameni potrebno je izvršiti kontrolu ne-ASCII karaktera u tekstu. Na adresi **http://www.fil.bg.ac.rs/branislava/seminarski/ip4/provera.html** je dostupan formular sa tekstuelnim poljem *Ulaz*. U to polje treba kopirati celokupan tekst dobijen u prethodnom koraku i aktivirati dugme **ne-ASCII?**, a u donjem tekstuelnom polju pojaviće se poruka ili da ne-ASCII karaktera nema, ili gde su se u tekstu pojavili.

4. Tekst u kome više nema ne-ASCII karaktera treba snimiti u Notepad-u komandom **File/Save as** pod istim imenom (npr. **politika-gggg-mm-dd-ascii.txt**), ali **kao ASCII tekst** (dakle, polje **Save as type: All Files**, polje **Encoding: ANSI**).

Napomena: Korisno je uključiti opciju „File name extensions“, jer je moguće da operativni sistem sakriva pravu ekstenziju (na primer, pri čuvanju, kao deo imena se eksplicitno navede i .txt, ali i Notepad automatski doda .txt na kraju naziva, pa se datoteka ustvari zove politika-gggg-mm-dd-ascii.txt.txt)



Priprema vesti (II deo)

U tekstu koji sledi sva objašnjenja biće ilustrovana na primeru datoteke **politika-gggg-mm-dd-ascii.txt**, ali potpuno istu obradu treba obaviti nad preostalim daotekama:

1. Ispravnu verziju (onu koja je takvog sadržaja da je prošla sve kontrole online sistema za proveru) datoteke **politika-gggg-mm-dd-ascii.txt** treba kopirati u datoteku **politika-gggg-mm-dd.xml** i sve dalje obrade vršiti na toj kopiji.

2. Tekst datoteke **politika-gggg-mm-dd.xml** treba strukturno etiketirati, tj. transformisati datoteku u dobro formiran i validan XML dokument. Na početku datoteke su obavezne linije sa XML deklaracijom i deklaracijom tipa dokumenta (DTD), kao i komentar sa podacima o studentu (ime, prezime i broj indeksa):

<?xml version="1.0" encoding="us-ascii" standalone="no"?>
<!DOCTYPE text SYSTEM "vesti.dtd">

<!-- Dragana Aleksicx, 161234 -->

Dokument treba da zadovolji DTD sa adrese:

http://www.fil.bg.ac.rs/branislava/seminarski/ip4/resursi/vesti.dtd

S obzirom na to da ovaj DTD ne određuje precizno strukturu XML dokumenta, sledi prirodnojezičko objašnjenje, a na kraju uputstva i primer XML dokumenta koji sadrži dva članka jednog broja ***Politike***. Pretpostavimo da je u pitanju broj od 14. februara 2021. godine.

Etiketa <text> (koreni element) se koristi za označavanje početka i kraja dokumenta. U okviru elementa text se nalazi jedan element sa obaveznim atributima type i date (ne zaboraviti zatvarajuću etiketu!).

 ...

Ovaj element predstavlja jedan broj (izdanje) ***Politike***, na šta ukazuje vrednost atributa type (issue), a datum izdanja (14. februar 2021. godine) je vrednost atributa date u formatu **ddmmgggg**, gde je **dd** dan, **mm** mesec i **gggg** godina. Ako su dan ili mesec jednocifreni brojevi, ispred se obavezno navodi nula.

Dakle, ovaj XML dokument je tekst (<text> ... </text>) koji se sastoji iz jednog broja ***Politike***, izdatog 14. februara 2021. godine. Sam broj ***Politike*** se sastoji iz jednog ili više članaka. Za označavanje početka i kraja svakog članka koristi se takođe element sa obaveznim atributom type (čija vrednost article ukazuje da se radi o članku) i obaveznim atributom n (čija vrednost predstavlja redni broj vesti u članku)

 ...

Tekstuelni sadržaj članka (ako zanemarimo podatke o autoru, datum objavljivanja i sl.) predstavlja niz odeljaka. Odeljak je deo članka koji počinje podnaslovom i prostire se do idućeg podnaslova (izuzetak je prvi odeljak koji obuhvata tekst sa početka članka do prvog podnaslova).

U okviru jednog članka potrebno je etiketirati više strukturnih delova:

- podatke o izvoru (autoru) navedene odmah nakon glavnog naslova članka (<byline>);
- glavni naslov članka i podnaslove koji stoje neposredno posle glavnog, a pre elementa byline (<head>);
- datum kada je autor objavio članak (<docDate>);
- svaki pojedinačni odeljak u članku ( );
- potpis autora (<signed>). Ukoliko se ovaj deo teksta nalazi na početku članka, treba ga prebaciti na kraj da bi dokument bio validan u odnosu na zadati DTD.

U okviru jednog odeljka potrebno je etiketirati podnaslov tog odeljka (<head> ... </head>) i svaki pojedinačni pasus odeljka (...).

Vešti student može da se oproba u primeni regularnih izraza prilikom obeležavanja pasusa. Regularni izraz za prepoznavanje pasusa je: **(^[^ <].+)** Pasuse automatski obeležiti odgovarajućom XML etiketom primenom sledećeg regularnog izraza na prepoznate pasuse: ** \1 ** Inače, pasuse obeležiti ručno.

Etikete za naslov (<head> ... </head>), članak (...) i pasuse (...) su obavezne, a ostale već prema onome što se nalazi u samom članku (<byline> ... </byline>, <docDate> ... </docDate>, ... i <signed> ... </signed>).

```
<div type="article" n="2">
<head>Holandxani nastavlxaju dominaciju u brzom klizanxu</head>
<byline>B92, Beta</byline>
<docDate>utorak, 14.02.2021. | 15:35</docDate>
<div type="section">
<p>Holandski takmicyar Kjeld Nejs osvojio je zlatnu medalxu u brzom klizanxu na Zimskim olimpijskim igrama u Pjongcyangu. </p>
<p>Nejs je cyetvrtu zlatnu medalxu u brzom klizanxu za Holandiju u Pjongcyangu osvojio rezultatom 1:44,01.</p>
<p>Srebrnu medalxu osvojio je Patrik Rust sa 0,85 sekundi zaostatka, a bronzu je osvojio Kim Min-seok sa 0,92 sekunde zaostatka. </p>
</div>
<signed>N. Dx.</signed>
</div>
```

Ako u samom tekstu ima evidentnih grešaka treba ih ispraviti i označiti da je greška ispravljena, na sledeći način. Primer:

G. Gerajinov, atasxe ruskog poslanstva, doputovao je iz Vranxe.

Etiketiranje:

G. Gerajinov, atasxe ruskog poslanstva, doputovao je iz **<choice><sic>**Vranxe**</sic><corr>**Vranxa**</corr></choice>**.

Ako u tekstu ima reči koje nisu greške u kucanju već je u pitanju stari pravopis i njih treba označiti, na sledeći način. Primer:

<head>Japanski gubitci - Japanska akcija</head>

Etiketiranje:

<head>Japanski **<choice><reg>**gubici**</reg><orig>**gubitci**</orig></choice>** - Japanska akcija</head>

Strane reči u tekstu se obrađuju pomoću etikete <foreign>:

Jovan jede **<foreign xml:lang="fr">**croissant**</foreign>** svakog jutra.

Atribut je standardna dvoslovna skraćenica za odgovarajući jezik (videti ISO 639-1).

Prilikom etiketiranja, student treba da pročita vest i da popravi sledeće:

- Loše otkucane interpunkcijske znakove (tačke, zapete, dve tačke, tačke zapete, zagrade i sl.), tj. potrebno je poštovati uobičajene konvencije da **posle** njih **obavezno** sledi razmak, a **nikako pre** njih.
- **Tekst jednog pasusa mora biti otkucan u jednoj liniji**! Takođe, **ne treba vršiti rastavljanje reči na slogove na kraju reda**, čak i kad je to slučaj u originalnom tekstu.

Slanje seminarskog rada

Seminarski rad mora biti predat do **11.04.2021.** u 23:59h kao dobro formiran i validan dokument. Pre slanja student je dužan da proveri da li je dokument dobro formiran i validan u odnosu na dati DTD (na primer, korišćenjem programa XML Copy Editor[2]), da li su adekvatno zamenjeni svi ne-ASCII karakteri i karakterski entiteti, da li ima grešaka u primeni koda aurora, kao i da li je tekst svakog pasusa otkucan u jednoj liniji. U tu svrhu student može da koristi formular na stranici **http://www.fil.bg.ac.rs/branislava/seminarski/ip4/provera.html** (aktiviranjem odgovarajućih dugmadi formulara mogu se obaviti pojedinačne provere, dok dugme *Izveštaj* obavlja sve provere odjednom).

Ako seminarski rad nije poslat u predviđenom roku ili je poslat kao dokument koji nije dobro formiran ili validan, student nije odbranio seminarski rad.

Ako je student zadužen za 14. februar 2021. godine, treba da pošalje elektronskom poštom datoteke

politika-gggg-mm-dd.xml (npr. **politika-2021-02-14.xml**)
danas-gggg-mm-dd.xml (npr. **danas-2021-02-14.xml)**
b92-gggg-mm-dd.xml (npr. **b92-2021-02-14.xml)**
novosti-gggg-mm-dd.xml (npr. **novosti-2021-02-14.xml)**
blic-gggg-mm-dd.xml (npr. **blic-2021-02-14.xml)**

Seminarski rad slati na adresu **ip4.seminarski.rad@gmail.com**. Takođe treba sačuvati datoteke *-lat.txt i *-ascii.txt jer u nekom trenutku mogu i one biti zatražene. Da bi seminarski bio uopšte uzet u obzir, u polju **Subject** mora da stoji isključivo **Vesti gggg-mm-dd Ime Prezime** (npr, ako je zaduženje Dragane Aleksić datum 14. februar 2021. godine,

Vesti 2021-02-14 Dragana Aleksicx

).

Svako kašnjenje, kao vraćanje, povlači negativne poene. Seminarski radovi za školsku 2019/20. godinu se ne primaju posle završetka nastave, tj. posle **31.05.2021**. godine.

Sve naknadne poruke (pitanja, nedoumice) vezane seminarski rad moraju u polju **Subject** imati isključivo **Vesti gggg-mm-dd Ime Prezime** da bi bile uzete u obzir.

[2] http://xml-copy-editor.sourceforge.net