**Documentatie**

**Submodulul 1**

**17 decembrie 2017**

1. **Descrierea proiectului**

**1.1 Obiective si scop**

Submodulul 1 are drept scop prelucrarea unui text in limba romana extras din fisiere de diverse tipuri (.pdf, .docx, .doc).

Prelucrarea consta in parsarea textului si curatarea acestuia de orice element irelevant contextului (numerotarea paginilor, footere si headere, imagini, etc.). Textul curatat si reparat va oferi token-uri, propozitii, parti de vorbire, grupuri nominale si rezolutia anaforei. In plus, vor fi puse la punct repararea encoding si fixarea automata a diacriticelor.

**1.2 Echipa**

* Bivol Daniel
* Ghergu Andreea
* Lungeanu Ionut
* Matrana Karina
* Palihovici Andrei
* Prodan Marius
* Tanasa Petru
* Trifan Tamara
* Turcu Nicursor
* Ulinici Cristina
* Vasilache Diana

**1.3 Avantajele proiectului**

Solutie functionala sub forma informatiilor in forma ideala alaturi de un adnotator perfect.

1. **Parser**

**2.1 Tool-uri de downloadat**

**2.1.1** [**NLTK**](http://nltk.org/)

1. Deschideti un command line ca administrator
2. Tastati comanda: **pip install nltk**
3. Dupa ce s-a downloadat, deschideti Python de la linia de comanda: **python**
   * Deschideti un command line daca cel precedent a fost inchis.
4. In Python tastati comenzile:
   * **import nltk**
   * **nltk.download()**
5. Ultima comanda ar trebui sa va deschida o interfata:
   * Din interfata (**colectii**) selectati **all** si apasati **download**.
6. Dupa ce ati downloadat ar trebui sa arate ca in screenshot.

**2.1.2** [**TreeTagger**](http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)

1. Deschideti un command line ca administrator
2. Tastati comanda: **pip install treetaggerwrapper**
3. Copiati directorul **LbisTreeTagger** pe disk-ul C
4. Adaugati o noua variabila de sistem. Numele variabilei va fi "**TAGDIR**" iar valoarea va fi "**C:TreeTagger**"
5. Scriptul "treetaggerwrapper.py" va fi modificat astfel:
   * Localizati script-ul in directorul : "**C:Program FilesPython37Libsite-packages**"
6. Cautati dictionarul "**g\_langsupport**" in acest script.
7. Adaugati cheia si valoarea urmatoare la acest dictionar:

"ro": {  
 "encoding": "utf-8",  
 "tagparfile": "romanian-utf8.par",  
 "abbrevfile": "romanian-abbreviations",  
 "pchar": ALONEMARKS + "'",  
 "fchar": ALONEMARKS + "'",  
 "pclictic": "",  
 "fclictic": "",  
 "number": NUMBER\_EXPRESSION,  
 "dummysentence": "Ana are mere. Câte mere are"  
 "Ana?",  
 "replurlexp": 'sustituir-url>',  
 "replemailexp": 'sustituir-email',  
 "replipexp": 'sustituir-ip',  
 "repldnsexp": 'sustituir-dns'  
}

**3. Extragere de texte**

**3.1 Scop**

Vom extrage textul important din fisiere de tip .doc, .docx si .pdf. Din fiecare fisier vom elimina continutul din header si footer, vom elimina imaginile si informatiile din subsolul imaginilor, vom elimina numarul paginilor si restul informatiile irelevante.

**3.2. Cum functioneaza**

In fisierele *text\_from\_text.py*, *text\_from\_docx.py*, *text\_from\_pdf.py* se afla cate o functie care extrage textul din fisierele *.doc*, .*docx*, *.pdf* si returneaza cate un fisier cu extensia *.txt*.

Fisierul *docx\_from\_doc.py* contine o functie care converteste un fisier *.doc* in *.docx*.

Fisierul *convert.py* inglobeaza toate fisierele de mai sus folosind o functie care primeste calea unui fisier, si in functie de extensia respectivului fisier, apeleaza una din functiile de mai sus.

Fisierele rezultate in urma extragerii textului sunt *testfile.txt* pentru extragerea textului din documente de tip .pdf, si *FisierText.txt* pentru extragerea textului din documente de tip .docx.

**4. Output**

**4.1 Exemple de fisier**

Informatiile vor fi trimise catre submodulul 2, printr-un fisier stabilit in prealabil, care va contine cuvintele extrase, semnele de punctuatie corespunzatoare, precum si alte informatii necesare submodulului 2.

<?xml version="1.0" encoding="UTF-8"?>  
<S id="47" offset="0">  
 <W LEMMA="/" MSD="SLASH" id="47.1" offset="0">/</W>  
 <W LEMMA="/" MSD="SLASH" id="47.2" offset="0">/</W>  
 <W Type="preposition" WordFormation="simple" Case="accusative" LEMMA="până" MSD="Spsa" POS="S" id="47.3" offset="0">Până</W>  
 <W Type="IntRel" Person="third" Case="direct" LEMMA="ce" MSD="Pw3--r" POS="RELR" id="47.4" offset="0">ce</W>  
 <W Type="main" VerbForm="indicative" Tense="present" Person="third" Number="plural" LEMMA="muri" MSD="Vmip3p" POS="V3" id="47.5" offset="0">mor</W>  
 <W LEMMA="," MSD="COMMA" id="47.6" offset="0">,</W>  
 <W LEMMA="/" MSD="SLASH" id="47.7" offset="0">/</W>  
 <W Type="proper" LEMMA="Pleaca~te" MSD="Np" POS="NP" id="47.8" offset="0">Pleaca te</W>  
 <W Type="proper" LEMMA="Îngere" MSD="Np" POS="NP" id="47.9" offset="0">Îngere</W>  
 <W LEMMA="/" MSD="SLASH" id="47.10" offset="0">/</W>  
 <W Type="preposition" WordFormation="simple" Case="accusative" LEMMA="la" MSD="Spsa" POS="S" id="47.11" offset="0">La</W>  
 <W Type="qualificative" Degree="positive" Gender="feminine" Number="singular" Case="direct" Definiteness="yes" LEMMA="trist" MSD="Afpfsry" POS="ASRY" id="47.12" offset="0">trista</W>  
 <W Type="personal" Person="first" Number="singular" Case="dative" Clitic="yes" PronForm="weak" LEMMA="eu" MSD="Pp1-sd--y-----w" POS="PPSD" id="47.13" offset="0">-mi</W>  
 <W Type="common" Gender="feminine" Number="singular" Case="direct" Definiteness="no" LEMMA="plângere" MSD="Ncfsrn" POS="NSRN" id="47.14" offset="0">plângere</W>  
 <W LEMMA="/" MSD="SLASH" id="47.15" offset="0">/</W>  
 <W Type="qualificative" Degree="positive" Gender="feminine" Number="singular" Case="direct" Definiteness="no" LEMMA="plin" MSD="Afpfsrn" POS="ASN" id="47.16" offset="0">Plină</W>  
 <W Type="preposition" WordFormation="simple" Case="accusative" Clitic="yes" LEMMA="de" MSD="Spsay" POS="S" id="47.17" offset="0">de-</W>  
 <W Type="common" Gender="masculine" Number="singular" Definiteness="no" LEMMA="amor" MSD="Ncms-n" POS="NSN" id="47.18" offset="0">amor</W>  
 <W LEMMA="." MSD="PERIOD" id="47.19" offset="0">.</W>  
</S>

Mai jos, pentru descrierea structurii documentului este prezentata Schema XML:

<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified" xmlns:xs="http://www.w3.org/2001/XMLSchema">  
 <xs:element name="S">  
 <xs:complexType>  
 <xs:sequence>  
 <xs:element name="W" maxOccurs="unbounded" minOccurs="0">  
 <xs:complexType>  
 <xs:simpleContent>  
 <xs:extension base="xs:string">  
 <xs:attribute type="xs:string" name="LEMMA" use="optional"/>  
 <xs:attribute type="xs:string" name="MSD" use="optional"/>  
 <xs:attribute type="xs:float" name="id" use="optional"/>  
 <xs:attribute type="xs:byte" name="offset" use="optional"/>  
 <xs:attribute type="xs:string" name="Type" use="optional"/>  
 <xs:attribute type="xs:string" name="WordFormation" use="optional"/>  
 <xs:attribute type="xs:string" name="Case" use="optional"/>  
 <xs:attribute type="xs:string" name="POS" use="optional"/>  
 <xs:attribute type="xs:string" name="Person" use="optional"/>  
 <xs:attribute type="xs:string" name="VerbForm" use="optional"/>  
 <xs:attribute type="xs:string" name="Tense" use="optional"/>  
 <xs:attribute type="xs:string" name="Number" use="optional"/>  
 <xs:attribute type="xs:string" name="Degree" use="optional"/>  
 <xs:attribute type="xs:string" name="Gender" use="optional"/>  
 <xs:attribute type="xs:string" name="Definiteness" use="optional"/>  
 <xs:attribute type="xs:string" name="Clitic" use="optional"/>  
 <xs:attribute type="xs:string" name="PronForm" use="optional"/>  
 </xs:extension>  
 </xs:simpleContent>  
 </xs:complexType>  
 </xs:element>  
 </xs:sequence>  
 <xs:attribute type="xs:byte" name="id"/>  
 <xs:attribute type="xs:byte" name="offset"/>  
 </xs:complexType>  
 </xs:element>  
</xs:schema>

<?xml version="1.0" encoding="UTF-8"?>  
<S id="2" offset="0">  
 <W Type="nominal" LEMMA="I." MSD="Yn" POS="Y" id="2.1" offset="0">I.</W>  
 <W Type="IntRel" LEMMA="cum" MSD="Rw" POS="R" id="2.2" offset="0">Cum</W>  
 <W Type="auxiliary" Person="third" Number="singular" LEMMA="avea" MSD="Va--3s" POS="VA3S" id="2.3" offset="0">a</W>  
 <W Type="main" VerbForm="participle" Number="singular" Gender="masculine" LEMMA="muri" MSD="Vmp--sm" POS="VPSM" id="2.4" offset="0">murit</W>  
 <W Type="proper" LEMMA="Eminescu" MSD="Np" POS="NP" id="2.5" offset="0">Eminescu</W>  
 <W Type="ordinal" Gender="feminine" Number="plural" Case="direct" NumForm="letter" Definiteness="yes" LEMMA="ultimele" MSD="Mofprly" POS="M" id="2.6" offset="0">Ultimele</W>  
 <W Type="personal" Person="third" Gender="masculine" Number="singular" Case="oblique" PronForm="strong" LEMMA="el" MSD="Pp3mso--------s" POS="PPSO" id="2.7" offset="0">lui</W>  
 <W Type="common" Gender="feminine" Number="plural" Definiteness="no" LEMMA="ceas" MSD="Ncfp-n" POS="NPN" id="2.8" offset="0">ceasuri</W>  
 <W LEMMA="," MSD="COMMA" id="2.9" offset="0">,</W>  
 <W Type="main" VerbForm="participle" Number="plural" Gender="feminine" LEMMA="povesti" MSD="Vmp--pf" POS="VPPF" id="2.10" offset="0">povestite</W>  
 <W Type="preposition" WordFormation="simple" Case="accusative" LEMMA="de" MSD="Spsa" POS="S" id="2.11" offset="0">de</W>  
 <W Type="indefinite" Gender="masculine" Number="singular" Case="direct" LEMMA="un" MSD="Timsr" POS="TSR" id="2.12" offset="0">un</W>  
 <W Type="common" Gender="masculine" Number="singular" Definiteness="no" LEMMA="martor" MSD="Ncms-n" POS="NSN" id="2.13" offset="0">martor</W>  
 <W Type="qualificative" Degree="positive" Gender="masculine" Number="singular" Definiteness="no" LEMMA="ocular" MSD="Afpms-n" POS="ASN" id="2.14" offset="0">ocular</W>  
 <W LEMMA="„" MSD="DBLQ" id="2.15" offset="0">„</W>  
 <W Type="indefinite" Gender="masculine" Number="singular" Case="direct" LEMMA="un" MSD="Timsr" POS="TSR" id="2.16" offset="0">Un</W>  
 <W Type="qualificative" Degree="positive" Gender="masculine" Number="singular" Definiteness="no" LEMMA="modest" MSD="Afpms-n" POS="ASN" id="2.17" offset="0">modest</W>  
 <W Type="common" Gender="masculine" Number="singular" Definiteness="no" LEMMA="cetățean" MSD="Ncms-n" POS="NSN" id="2.18" offset="0">cetățean</W>  
 <W LEMMA="," MSD="COMMA" id="2.19" offset="0">,</W>  
 <W Type="qualificative" Degree="positive" Gender="masculine" Number="singular" Definiteness="no" LEMMA="mic" MSD="Afpms-n" POS="ASN" id="2.20" offset="0">mic</W>  
 <W Type="preposition" WordFormation="simple" Case="accusative" LEMMA="de" MSD="Spsa" POS="S" id="2.21" offset="0">de</W>  
 <W Type="common" Gender="masculine" Number="singular" Definiteness="no" LEMMA="stat" MSD="Ncms-n" POS="NSN" id="2.22" offset="0">stat</W>  
 <W Type="portmanteau" WordFormation="simple" CoordType="simple" SubordType="positive" LEMMA="și" MSD="Crssp" POS="CR" id="2.23" offset="0">și</W>  
 <W Type="preposition" WordFormation="simple" Case="accusative" LEMMA="cu" MSD="Spsa" POS="S" id="2.24" offset="0">cu</W>  
 <W Type="indefinite" Gender="feminine" Number="singular" Case="direct" LEMMA="un" MSD="Tifsr" POS="TSR" id="2.25" offset="0">o</W>  
 <W Type="common" Gender="feminine" Number="singular" Case="direct" Definiteness="yes" LEMMA="barbută" MSD="Ncfsry" POS="NSRY" id="2.26" offset="0">barbuta</W>  
 <W Type="qualificative" Degree="positive" Gender="feminine" Number="singular" Case="direct" Definiteness="no" LEMMA="alb" MSD="Afpfsrn" POS="ASN" id="2.27" offset="0">albă</W>  
 <W LEMMA="," MSD="COMMA" id="2.28" offset="0">,</W>  
 <W Type="auxiliary" Person="third" Number="singular" LEMMA="avea" MSD="Va--3s" POS="VA3S" id="2.29" offset="0">a</W>  
 <W Type="main" VerbForm="participle" Number="singular" Gender="masculine" LEMMA="sui" MSD="Vmp--sm" POS="VPSM" id="2.30" offset="0">suit</W>  
 <W Type="general" Degree="positive" LEMMA="alaltăieri" MSD="Rgp" POS="R" id="2.31" offset="0">alaltăieri</W>  
 <W Type="common" Gender="feminine" Number="plural" Case="direct" Definiteness="yes" LEMMA="scară" MSD="Ncfpry" POS="NPRY" id="2.32" offset="0">scările</W>  
 <W Type="common" Gender="feminine" Number="singular" Case="oblique" Definiteness="yes" LEMMA="redacție" MSD="Ncfsoy" POS="NSOY" id="2.33" offset="0">redacției</W>  
 <W Type="possessive" Person="first" Gender="feminine" Number="singular" Case="oblique" OwnNumber="plural" LEMMA="meu" MSD="Ds1fsop" POS="PSS" id="2.34" offset="0">noastre</W>  
 <W LEMMA="," MSD="COMMA" id="2.35" offset="0">,</W>  
 <W Type="main" VerbForm="gerund" LEMMA="voi" MSD="Vmg" POS="VG" id="2.36" offset="0">voind</W>  
 <W Type="subjunctive" LEMMA="să" MSD="Qs" POS="QS" id="2.37" offset="0">să</W>  
 <W Type="personal" Person="first" Number="plural" Case="accusative" PronForm="weak" LEMMA="eu" MSD="Pp1-pa--------w" POS="PPPA" id="2.38" offset="0">ne</W>  
 <W Type="main" VerbForm="subjunctive" Tense="present" Person="third" LEMMA="face" MSD="Vmsp3" POS="V3" id="2.39" offset="0">facă</W>  
 <W Type="indefinite" Gender="feminine" Number="singular" Case="direct" LEMMA="un" MSD="Tifsr" POS="TSR" id="2.40" offset="0">o</W>  
 <W Type="common" Gender="feminine" Number="singular" Case="direct" Definiteness="no" LEMMA="comunicare" MSD="Ncfsrn" POS="NSRN" id="2.41" offset="0">comunicare</W>  
 <W LEMMA="." MSD="PERIOD" id="2.42" offset="0">.</W>  
</S>

Mai jos, pentru descrierea structurii documentului este prezentata Schema XML:

<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified" xmlns:xs="http://www.w3.org/2001/XMLSchema">  
 <xs:element name="S">  
 <xs:complexType>  
 <xs:sequence>  
 <xs:element name="W" maxOccurs="unbounded" minOccurs="0">  
 <xs:complexType>  
 <xs:simpleContent>  
 <xs:extension base="xs:string">  
 <xs:attribute type="xs:string" name="Type" use="optional"/>  
 <xs:attribute type="xs:string" name="LEMMA" use="optional"/>  
 <xs:attribute type="xs:string" name="MSD" use="optional"/>  
 <xs:attribute type="xs:string" name="POS" use="optional"/>  
 <xs:attribute type="xs:float" name="id" use="optional"/>  
 <xs:attribute type="xs:byte" name="offset" use="optional"/>  
 <xs:attribute type="xs:string" name="Person" use="optional"/>  
 <xs:attribute type="xs:string" name="Number" use="optional"/>  
 <xs:attribute type="xs:string" name="VerbForm" use="optional"/>  
 <xs:attribute type="xs:string" name="Gender" use="optional"/>  
 <xs:attribute type="xs:string" name="Case" use="optional"/>  
 <xs:attribute type="xs:string" name="NumForm" use="optional"/>  
 <xs:attribute type="xs:string" name="Definiteness" use="optional"/>  
 <xs:attribute type="xs:string" name="PronForm" use="optional"/>  
 <xs:attribute type="xs:string" name="WordFormation" use="optional"/>  
 <xs:attribute type="xs:string" name="Degree" use="optional"/>  
 <xs:attribute type="xs:string" name="CoordType" use="optional"/>  
 <xs:attribute type="xs:string" name="SubordType" use="optional"/>  
 <xs:attribute type="xs:string" name="OwnNumber" use="optional"/>  
 <xs:attribute type="xs:string" name="Tense" use="optional"/>  
 </xs:extension>  
 </xs:simpleContent>  
 </xs:complexType>  
 </xs:element>  
 </xs:sequence>  
 <xs:attribute type="xs:byte" name="id"/>  
 <xs:attribute type="xs:byte" name="offset"/>  
 </xs:complexType>  
 </xs:element>  
</xs:schema>