INSTRUKCJA DO PROJEKTU

PRZETWARZANIE JĘZYKA NATURALNEGO

Ćwiczenie 1: Wstępne przetwarzanie tekstu

Jan Daciuk

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
KPRM Kancelaria Prezesa Rady Ministrów
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014–2020
Oś Priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa” Działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”
Tytuł projektu: „Akademia Innowacyjnych Zastosowań technologii Cyfrowych (AI Tech)”

1 Cel

Zanim zacznie się właściwe przetwarzanie tekstu związane bezpośrednio z celem tego przetwarzania, należy tekst odpowiednio przygotować. Tekst może występować w różnych formatach, z których najpopularniejsze to:

- czysty tekst
- HTML
- XML
- PDF

Nawet czysty tekst może wymagać np. zmiany kodowania liter. Wydobycie czystego tekstu z innych formatów wymaga pewnego wysiłku. Celem tego ćwiczenia jest zapoznanie studentów z problemami związanymi z pozyskiwaniem tekstu i możliwymi ich rozwiązaniami. Spodziewanym wynikiem ma być nabycie umiejętności wydobycia tekstu z plików nie tylko tekstowych, ale także w formatach takich jak HTML, XML i PDF. Nabycie tych umiejętności nastąpić ma przez stopniowe rozwijanie przykładowych programów.

2 Używane narzędzia

Językiem świetnie nadającym się do przetwarzania tekstu jest Python. Książka „*Natural Language Processing with Python*" Stevena Birda, Ewana Kleina i Edwarda Lopera dostępna jest pod adresem: `http://www.nltk.org/book`. Opisuje ona bardzo przydatną bibliotekę `nltk`.

Do przetwarzania tekstów w HTML-u przydatna jest biblioteka BeautifulSoup.

Do wydobywania tekstu z formatu PDF najlepiej skorzystać z bibliotek pdfplumber lub pdfminer. Do odmiany słów można użyć programu `morfeusz2` (z interfejsem w Pythonie) dostępnego na stronie `http://morfeusz.sgjp.pl/`. Do wyszukiwania wzorców w tekście służy biblioteka re, chociaż można użyć bardziej rozbudowanej biblioteki pyparsing.

3 Źródła danych

W ćwiczeniu będziemy używać następujących źródeł danych:

1. podręczniki systemowe Linuksa dostępne za pomocą polecenia `man`
2. Portal:Current Events dostępny pod adresem `https://en.wikipedia.org/wiki/Portal:Current_events` (część Wikipedii). Jego polskojęzyczny odpowiednik Portal:Aktualności `https://pl.wikipedia.org/wiki/Portal:Aktualności` niestety (jeszcze) nie zawiera wiele materiału.
3. Narodowy Korpus Języka Polskiego dostępny pod adresem `http://nkjp.pl/index.php?page=14&lang=0`, a konkretnie jego milionowy podkorpus
4. Akty prawne z Internetowego Systemu Aktów Prawnych dostępnego pod adresem `http://isap.sejm.gov.pl/`

4 Przetwarzanie czystych plików tekstowych

Na zajęciach wprowadzających wykonywane ćwiczenia zmierzały do wytworzenia narzędzia do pokazywania użycia słów w kontekście. Zadanie oceniane: należy napisać program realizujący KWIC: *key word in context*. Program powinien akceptować jako parametry słowo oraz listę plików tekstowych i drukować w kolejnych wierszach szukane słowo w środku wiersza począwszy od tej samej kolumny w każdym wierszu, a po jego lewej i prawej stronie lewy i prawy kontekst — słowa, cyfry i znaki przestankowe — w jakim znajduje się kolejne wystąpienie szukanego słowa w danym pliku. Przykład wyszukiwania w Kodeksie Postępowania Cywilnego podany jest poniżej.

./kwic.py uczestnik D19640296Lj.txt

```
lub 3, wskazany jako wytwórca lub uczestnik  procesu wprowadzenia do obrotu ta
niowym: 1) od dnia czynności, gdy uczestnik  był przy niej obecny lub był o je
ów prawnych. Art. 520. § 1. Każdy uczestnik  ponosi koszty postępowania związa
y stosuje się odpowiednio, jeżeli uczestnik  postępował niesumiennie lub oczyw
ienia o zniesieniu współwłasności uczestnik  nie może dochodzić roszczeń przew
iu wniosku wnioskodawca bądź inny uczestnik  postępowania zmarł lub został poz
zenie akt uprawniona jest strona, uczestnik  postępowania lub interwenient. §
zenia może żądać każda strona lub uczestnik  postępowania, jeżeli uprawdopodob
ci egzekucyjnych. § 4. Strona lub uczestnik  postępowania zawiadamiają sądowy
zeniu lub zawieszeniu albo jeżeli uczestnik  nie otrzymał zawiadomienia o licy
 od dnia jego doręczenia albo gdy uczestnik , który takiego żądania nie zgłosi
eprocesowego, jeżeli pozwany albo uczestnik  postępowania nieprocesowego podni
```

W wersji rozszerzonej program powinien mieć opcję włączającą poszukiwanie form odmienionych podanej formy hasłowej słowa. Np. po podaniu formy hasłowej „ołówek”, program znajdzie także formy odmienione „ołówka”, „ołówkowi”, „ołówkiem”, „ołówku”, „ołówki”, „ołówków”, „ołówkom”, „ołówkami” i „ołówkach”, każdą pokazując w kontekście użycia w tekście w podanych plikach.

./kwic.py -m uczestnik D19640296Lj.txt

```
 Sejmu s. 2/524 Art. 3. Strony i uczestnicy    postępowania obowiązani są dokon
rzepisach postępowania stronom i uczestnikom   postępowania nie wolno czynić uż
zeby sąd może udzielić stronom i uczestnikom   postępowania występującym w spra
yjaśnienia sprawy. § 2. Strony i uczestnicy    postępowania obowiązani są przyt
ególny stanowi inaczej. Strony i uczestnicy    postępowania mają prawo przegląd
otokołów lub pism. § 2. Strony i uczestnicy    postępowania mają prawo do otrzy
zy drzwiach zamkniętych strony i uczestnicy    postępowania mają prawo do otrzy
u na utrwalanie przez strony lub uczestników   postępowania przebiegu posiedzeń
strującego dźwięk. § 2. Strony i uczestnicy    postępowania uprzedzają sąd o za
k. § 3. Sąd zakazuje stronie lub uczestnikowi  postępowania utrwalenia przebieg
unku zaś do przeciwnika i innych uczestników   - z chwilą doręczenia im tego za
zacji gwarancji procesowych jego uczestników   , ochronę praw osób, którym pisma
 na odległość. W takim przypadku uczestnicy    postępowania mogą brać udział w
o postępowanie do miejsca pobytu uczestników   postępowania oraz 2022-08-24 ©Ka
Sejmu s. 56/524 z miejsca pobytu uczestników   postępowania do sali sądowej sąd
pełnomocników, świadków i innych uczestników   postępowania, wymaga przedstawie
słuchania stron w sprawie między uczestnikami  tej czynności na fakt jej dokona
awną może być dopuszczony między uczestnikami  tej czynności tylko w wypadkach,
ązki są dla nich wspólne. Współ- uczestników   tych należy zawiadomić o rozpraw
 kopiowane przez strony i innych uczestników   postępowania jedynie, w przypadk
ub 3, wskazany jako wytwórca lub uczestnik     procesu wprowadzenia do obrotu t
 Jeżeli weźmie udział, staje się uczestnikiem . Na odmowę dopuszczenia do wzięc
 się, że zainteresowany nie jest uczestnikiem , sąd wezwie go do udziału w spra
iału w sprawie wezwany staje się uczestnikiem . W razie potrzeby wyznaczenia ku
 złożeniu przez któregokolwiek z uczestników   oświadczenia na piśmie cofnięcie
 skuteczne tylko wtedy, gdy inni uczestnicy    nie sprzeciwili się temu w termi
zędu. Art. 513. Niestawiennictwo uczestników   nie tamuje rozpoznania sprawy. P
zygnięciem sprawy może wysłuchać uczestników   na posiedzeniu sądowym lub zażąd
rzyrzeczenia oraz w nieobecności uczestników   , może również zażądać od osób, k
eż zażądać od osób, które nie są uczestnikami , złożenia wyjaśnień na piśmie. A
iowym: 1) od dnia czynności, gdy uczestnik     był przy niej obecny lub był o j
od dnia doręczenia zawiadomienia uczestnika    o dokonaniu czynności; 3) w przy
 Krajowego Rejestru Sądowego dla uczestników   postępowania, którym postanowien
w prawnych. Art. 520. § 1. Każdy uczestnik     ponosi koszty postępowania związ
em w sprawie. § 2. Jeżeli jednak uczestnicy    są w różnym stopniu zainteresowa
sztów lub włożyć go na jednego z uczestników   w całości. To samo dotyczy zwrot
ów postępowania wyłożonych przez uczestników   . § 3. Jeżeli interesy uczestnikó
czestników. § 3. Jeżeli interesy uczestników   są sprzeczne, sąd może włożyć na
są sprzeczne, sąd może włożyć na uczestnika    , którego wnioski zostały oddalon
owania poniesionych przez innego uczestnika    . Przepis powyższy stosuje się od
 stosuje się odpowiednio, jeżeli uczestnik     postępował niesumiennie lub oczy
2. Zainteresowany, który nie był uczestnikiem  postępowania zakończonego prawom
525. Akta sprawy dostępne są dla uczestników   postępowania oraz za zezwoleniem
. Sąd wzywa do udziału w sprawie uczestników   postępowania, w którym zapadło p
ełnoletniej, sąd może na wniosek uczestnika    postępowania lub z urzędu, przy
aniem postanowienia sąd wysłucha uczestników   postępowania. § 3. Na postanowie
zędu doręcza odpis postanowienia uczestnikom   postępowania, prokuratorowi, Pol
dopuszczalne, sąd może skierować uczestników   do mediacji. Przedmiotem mediacj
ia władzy rodzicielskiej. Jeżeli uczestnicy    postępowania nie uzgodnili osoby
 osobistego stawiennictwa innych uczestników   postępowania stosuje się w spraw
 z 1980 r. obowiązuje zastępstwo uczestników   postępowania przez adwokatów lub
 lub radcy prawnego, a także gdy uczestnikiem  postępowania, jego organem, jego
wezwaniem na rozprawę poucza się uczestnika    o treści przepisów art. 211a § 2
oraz o tym, że w razie działania uczestnika    w celu osiągnięcia korzyści mają
asadnieniem doręcza się z urzędu uczestnikom   postępowania oraz prokuratorowi.
mowa w art. 5182 § 1, zawiadamia uczestników   postępowania, prokuratora, podmi
ddziale niniejszym, sąd wysłucha uczestników   postępowania. Art. 59819. § 1. D
enia o zniesieniu współwłasności uczestnik     nie może dochodzić roszczeń prze
ść rzeczy własność przechodzi na uczestników   wskazanych w postanowieniu. Jeże
```

```
u wniosku wnioskodawca bądź inny uczestnik     postępowania zmarł lub został po
 dokonanym wpisie sąd zawiadamia uczestników   postępowania. Nie zawiadamia się
postępowania. Nie zawiadamia się uczestnika   , który na piśmie zrzekł się zawi
wiąże się wpis. § 12. Na wniosek uczestnika    postępowania zawarty w akcie not
enia zawiadomienia o wpisie. Dla uczestnika   , który zrzekł się zawiadomienia,
nik zawiadamia wnioskodawcę oraz uczestników   postępowania o sporządzenie spis
enie, na które wezwie wszystkich uczestników   sprawy. Jeżeli przed tym posiedz
e lub wykaz, sąd zawiadomi o tym uczestników   . Art. 659. Spadkobierca, który s
onym przez spadkobiercę sąd oraz uczestnicy    mogą zadawać spadkobiercy pytani
nne osoby niż te, które wskazali uczestnicy   . W postanowieniu o stwierdzeniu
rozdziału. Jednakże ten, kto był uczestnikiem  postępowania o stwierdzenie naby
 cywilnego. Art. 683. Na żądanie uczestnika    działu, zgłoszone nie później ni
go na podstawie zgodnego wniosku uczestników   , dział spadku będzie rozpoznany
Krajowego Rejestru Sądowego jest uczestnikiem  postępowania, chociażby nie był
której wnioskodawca jest jedynym uczestnikiem  postępowania, przepis § 1 stosuj
enie akt uprawniona jest strona, uczestnik     postępowania lub interwenient. §
enia może żądać każda strona lub uczestnik     postępowania, jeżeli uprawdopodo
sąd uwzględni interesy stron lub uczestników   postępowania w takiej mierze, ab
ować prawa i obowiązki stron lub uczestników   postępowania na czas trwania pos
enia o udzieleniu zabezpieczenia uczestnikom   postępowania, wraz z pouczeniem
 Organ egzekucyjny może żądać od uczestników   postępowania złożenia wyjaśnień.
Przepis § 1 stosuje się także do uczestnika    postępowania, którego dotyczy cz
. 7631. Komornik poucza strony i uczestników   postępowania niezastępowanych pr
ancję bezpieczeństwa komornika i uczestników   postępowania, poszanowania godno
ości egzekucyjnej oraz stronom i uczestnikom   obecnym podczas czynności dokony
i egzekucyjnych. § 4. Strona lub uczestnik     postępowania zawiadamiają sądowy
§ 1. Komornik zawiadamia strony, uczestników   postępowania i administracyjny o
§ 1. Komornik zawiadamia strony, uczestników   postępowania, z wyjątkiem dłużni
ienie możliwości obrony praw jej uczestników   . § 3. Postanowienie o nadaniu kl
 5. Komornik umożliwia stronom i uczestnikom   postępowania zapoznanie się z za
celarii komorniczej. Stronom ani uczestnikom   postępowania nie wydaje się zapi
 2) wskazać miejsce zamieszkania uczestników   postępowania. § 2. Jeżeli nieruc
a komornik zawiadamia znanych mu uczestników   . § 2. Komornik wzywa ponadto prz
etowej Krajowej Rady Komorniczej uczestników   , o których nie ma wiadomości, or
oszacowania zawiadomi znanych mu uczestników    oraz dokona obwieszczenia stosow
nia liczy się od dnia doręczenia uczestnikowi  zawiadomienia, o którym mowa w a
tórym mowa w art. 945 § 4, a dla uczestników   , którym nie doręczono zawiadomie
ancelaria Sejmu s. 421/524 1) 2) uczestnikom   postępowania; organowi gminy, ur
zaofiarował najwyższą cenę, oraz uczestników   przetargu dokonuje się jedynie w
uchaniu tak jego, jak i obecnych uczestników   . Art. 988. § 1. Postanowienie o
po wysłuchaniu wnioskodawcy oraz uczestników   , jeżeli stawią się na posiedzeni
eniu lub zawieszeniu albo jeżeli uczestnik     nie otrzymał zawiadomienia o lic
dzającym, na wniosek nabywcy lub uczestnika    postępowania. Nabywca może być n
 O sprzedaży komornik zawiadamia uczestników   stosownie do art. 954. § 3. Jeże
ie licytacji komornik zawiadamia uczestników   postępowania stosownie do art. 9
4) sumę, jaka przypada każdemu z uczestników   podziału; sumy, które mają być w
cego sprzedaż zarządca zawiadomi uczestników   stosownie do art. 954. Art. 1064
haskiej z 1996 r. i wyznacza dla uczestników   postępowania termin sześciu tygo
w terminie, o którym mowa w § 3, uczestnicy    postępowania nie wystąpią z wnio
nnego organu państwa obcego albo uczestników   postępowania o stwierdzenie jury
nnego organu państwa obcego albo uczestników   postępowania o stwierdzenie jury
nnego organu państwa obcego albo uczestników   postępowania o stwierdzenie jury
y (WE) nr 2201/2003; 4) wezwania uczestników   postępowania do złożenia wniosku
nnego organu państwa obcego albo uczestników   postępowania. Art. 1107. § 1. Do
a sporządza się tylko na żądanie uczestnika    postępowania zgłoszone w termini
od dnia jego doręczenia albo gdy uczestnik     , który takiego żądania nie zgłos
 uzasadnienie na wniosek strony, uczestnika    postępowania lub osoby ubiegając
procesowego, jeżeli pozwany albo uczestnik     postępowania nieprocesowego podn
```

5 Przetwarzanie stron w HTML-u.

Skrypt `getnews.py` pobiera wiadomości z portalu Current Events Wikipedii i drukuje znalezione odnośniki do Wikipedii z podziałem na dni i kategorie wydarzeń. Korzysta z biblioteki BeautifulSoup. Należy rozbudować go tak, aby w przypadku, gdy zdarzenia zaszły w konkretnych, wymienionych bezpośrednio w tekście państwach, drukował informacje o polu powierzchni tych państw. Informacje te można znaleźć na stronach Wikipedii korzystając z wyłuskanych przez skrypt odnośników. Warto wcześniej obejrzeć takie strony i skorzystać z funkcji „Inspect element". W wersji rozszerzonej powinien także drukować liczbę mieszkańców wymienionych miast. Uwaga: istnieje wiele bibliotek Pythona, które ułatwiają to zadanie, np. `pycountry`, `countrylist`, `geography3`, `flashgeotext`.

getnews.py

```
#!/usr/bin/python
from bs4 import BeautifulSoup
import re
import urllib.request
```

```
def parse_ul_tree(el, top_level = True):
    """
    Parse events of a specific theme at a specific date.
    :param el: unordered list containing themes or individual events.
    :return: a list of all links in that element.
    """
    all_links = []
    for ll in el.find_all("li", recursive=False):
        # Each line contains also subordinate ULs, but up to \n is the current line text.
        tt = ll.text.split("\n")[0]
        tt = re.sub("\([^()]*\)$", "", tt, count=1)  # remove news agency in parentheses if present (.*? didn't work)
        links = [[lnks["href"], lnks.text] for lnks in ll.find_all("a", recursive=False) if lnks["href"][:6] == "/wiki/"]
        all_links.extend(links)
        for uu in ll.find_all("ul", recursive=False):
            all_links.extend(parse_ul_tree(uu, False))
    return all_links


def parse_month():
    """
    Parse a description of events in one month of Wikipedia events.
    It can also be used to parse the current events.
    :return: nothing
    """
    base = "https://en.wikipedia.org/wiki/Portal:Current_events"
    with urllib.request.urlopen(base) as fp:
        soup = BeautifulSoup(fp, "html.parser")
        # The div with region role is deep in the body hierarchy. It represents one day.
        for d in soup.find_all("div", role="region"):
            date = []
            category = ""
            if "id" in d.attrs:
                year, month, day = d["id"].split("_")
                date = [day, month, year]
                print("{0} {1} {2}".format(day, month, year))
            for t1 in d.find_all(recursive=False):
                if t1.name == "div" and "class" in t1.attrs and "description" in t1["class"]:
                    for t in t1.find_all(recursive=False):
                        if t.name == "p":
                            # An opening paragraph contains category name
                            category = t.b.contents[0]
                            print(category)
                        if t.name == "ul":
                            # The following unordered list contains events grouped in themes.
                            # Each theme is one line in the top-level list
                            # Parse those events
                            events = parse_ul_tree(t)
                            print(events)



if (__name__ == "__main__"):
    parse_month()
```

6 Przetwarzanie plików XML

Narodowy Korpus Języka Polskiego jest zbiorem oznaczonych tekstów w języku polskim. Jego stumilionowy podzbiór jest publicznie dostępny. Każdy tekst opisany jest w katalogu zawierającym pliki:

- `ann_groups.xml` — grupy składniowe
- `ann_morphosyntax.xml` — interpretacje morfoskładniowe poszczególnych segmentów
- `ann_named` — jednostki nazewnicze
- `ann_segmentation` — reprezentacja segmentacji tekstu na zdania i na segmenty (w przybliżeniu: słowa ortograficzne)
- `ann_senses` — informacja o znaczeniach wybranych słów wieloznacznych
- `ann_words` — słowa składniowe i ich interpretacje morfoskładniowe
- `header.xml` — nagówek, zawiera m.in. informacje o pochodzeniu tekstu
- `text.xml` — właściwy tekst

Skrypt `proc_nkjp.py` wydobywa i drukuje z katalogu opisu tekstu o danej nazwie informacje z pliku `ann_morphosyntax.xml` o słowie i jego możliwych interpretacjach morfoskładniowych.

proc_nkjp.py

```
#!/usr/bin/python3
import xml.etree.ElementTree as ET


def tag_uri_and_name(elem):
    """
    Divide element name with a prefix being namespace
    into namespace uri and tag.
    :param elem: element in an xml tree
    :return: a 2-tuple: namespace uri and tag
    """
    if elem.tag[0] == "{":
        uri, ignore, tag = elem.tag[1:].partition("}")
    else:
        uri = None
        tag = elem.tag
    return uri, tag


def get_next_morph(filename):
    """
    Get the next word with all its annotations.
    :param filename: file named ann_morphosyntax.xml in NKJP
    :return: a 3-tuple with the word, its possible interpretations,
    and the correct interpretation
    """
    events = ["start-ns", "start", "end"]
    xmlns = ""
    path = []
    ctags = []
    msds = []
    interps = []
    disamb = []
    base = ""
    orth = ""
    for (event, elem) in ET.iterparse(filename, events=events):
        # React to opening and closing of tags
        tag = ""
        if event == "start-ns":
            if elem[0] == "":
                xmlns = elem[1]
        elif event == "start":
            # For tag openings, construct a path to the current tag.
            # The path is a list.
            # Use either tags, or tags concatenated with some attribute
            # values (e.g. "type", "name", or "value") as path items.
            # Store path current leaves in variable `branch'
            ns, tag = tag_uri_and_name(elem)
            branch = tag
            if tag == "fs":
                branch = tag + ":" + elem.attrib["type"]
            elif tag == "f":
                branch = tag + ":" + elem.attrib["name"]
            elif tag == "symbol":
                if path[-1] == "f:ctag":
                    # The part-of-speech tag for the word
                    ctags.append(elem.attrib["value"])
                elif (path[-1] == "f:msd"
                      or path[-1] == "vAlt" and path[-2] == "f:msd"):
                    # Possible morphosyntactic descriptions of the word,
                    # as obtained from a morphological analyzer
                    msd = elem.attrib["value"]
                    # They are appended to the part-of-speech
                    msd_suffix = msd
                    if msd != "":
                        msd_suffix = ":" + msd
                    for c in ctags:
                        # Append msds to each part-of-speech unless empty
                        if c == "ign":
                            interps.append("ign")
                        else:
                            interps.append(base + ":" + c + msd_suffix)
                    msds.append(elem.attrib["value"])
            path.append(branch)
        elif event == "end":
            branch = path.pop()
            if branch == "f:interps":
```

```
                # The end of possible interpretations
                # Clear the list
                ctags = []
                msds = []
            elif branch == "string":
                # Assign extracted values
                if path[-1] == "f:orth":
                    # Word in the text
                    orth = elem.text
                elif path[-1] == "f:base":
                    # Lemma
                    base = elem.text
            elif branch == "seg":
                yield orth, interps, disamb
                orth = ""
                interps = []
                disamb = ""


if (__name__ == "__main__"):
    import sys
    for setname in sys.argv[1:]:
        print(setname)
        mm = get_next_morph(setname + "/ann_morphosyntax.xml")
        for m in mm:
            orth, interps, disamb = m
            print("orth=", orth, ", interps=", interps, ", disamb=", disamb,
                  sep='')
```

Dostarczony skrypt należy rozszerzyć w ten sposób, aby drukowana była też informacja o właściwym oznaczeniu (w skrypcie przypisywana do zmiennej `disamb`). Aby była właściwie wydrukowana, należy oczywiście wpisać do niej odpowiednią wartość. Nazwa zmiennej sugeruje, gdzie w drzewie XML znaleźć pożądaną informację. Poniżej przedstawiono początek wzorcowego wyjścia skryptu dla pierwszego podkatalogu NKJP.

`./proc_nkjp-base.py $NKJP/010-2-000000001`

```
orth=Zatrzasnął, interps=['zatrzasnąć:praet:sg:m1:perf', 'zatrzasnąć:praet:sg:m2:perf', 'zatrzasnąć:praet:sg:m3:perf'], disamb=zat
rzasnąć:praet:sg:m1:perf
orth=drzwi, interps=['drzwi:subst:pl:nom:n', 'drzwi:subst:pl:gen:n', 'drzwi:subst:pl:acc:n', 'drzwi:subst:pl:voc:n'], disamb=drzwi
:subst:pl:acc:n
orth=od, interps=['Oda:subst:pl:gen:f', 'od:subst:gen:nwok', 'od:prep:gen:nwok'], disamb=od:prep:gen:nwok
orth=mieszkania, interps=['mieszkanie:subst:sg:gen:n', 'mieszkanie:subst:pl:nom:n', 'mieszkanie:subst:pl:acc:n', 'mieszkanie:subst
:pl:voc:n', 'mieszkać:subst:sg:gen:n:imperf:aff', 'mieszkać:ger:sg:gen:n:imperf:aff'], disamb=mieszkanie:subst:sg:gen:n
orth=,, interps=[',:interp'], disamb=,:interp
orth=dwa, interps=['dwa:num:pl:nom:m2:congr', 'dwa:num:pl:nom:m3:congr', 'dwa:num:pl:nom:n:congr', 'dwa:num:pl:acc:m2:congr', 'dwa
:num:pl:acc:m3:congr', 'dwa:num:pl:acc:n:congr', 'dwa:num:pl:voc:m2:congr', 'dwa:num:pl:voc:m3:congr', 'dwa:num:pl:voc:n:congr'],
disamb=dwa:num:pl:acc:m3:congr
orth=razy, interps=['raz:subst:pl:nom:m3', 'raz:subst:pl:gen:m3', 'raz:subst:pl:acc:m3', 'raz:subst:pl:voc:m3'], disamb=raz:subst:
pl:acc:m3
orth=przekręcił, interps=['przekręcić:praet:sg:m1:perf', 'przekręcić:praet:sg:m2:perf', 'przekręcić:praet:sg:m3:perf'], disamb=prz
ekręcić:praet:sg:m1:perf
orth=klucz, interps=['klucz:subst:sg:nom:m3', 'klucz:subst:sg:acc:m3', 'kluczyć:subst:sg:sec:imperf', 'kluczyć:impt:sg:sec:imperf'
], disamb=klucz:subst:sg:acc:m3
orth=,, interps=[',:interp'], disamb=,:interp
orth=nacisnął, interps=['nacisnąć:praet:sg:m1:perf', 'nacisnąć:praet:sg:m2:perf', 'nacisnąć:praet:sg:m3:perf'], disamb=nacisnąć:pr
aet:sg:m1:perf
orth=klamkę, interps=['klamka:subst:sg:acc:f'], disamb=klamka:subst:sg:acc:f
orth=,, interps=[',:interp'], disamb=,:interp
orth=by, interps=['by:comp', 'by:comp', 'by:qub'], disamb=by:comp
orth=sprawdzić, interps=['sprawdzić:inf:perf'], disamb=sprawdzić:inf:perf
orth=,, interps=[',:interp'], disamb=,:interp
orth=czy, interps=['czy:conj', 'czy:conj', 'czy:qub'], disamb=czy:qub
orth=dobrze, interps=['dobrze:adv:pos'], disamb=dobrze:adv:pos
orth=zamknięte, interps=['zamknięty:adj:sg:nom:n:pos', 'zamknięty:adj:sg:acc:n:pos', 'zamknięty:adj:sg:voc:n:pos', 'zamknięty:adj:
pl:nom:m2:pos', 'zamknięty:adj:pl:nom:m3:pos', 'zamknięty:adj:pl:nom:f:pos', 'zamknięty:adj:pl:nom:n:pos', 'zamknięty:adj:pl:acc:m
2:pos', 'zamknięty:adj:pl:acc:m3:pos', 'zamknięty:adj:pl:acc:f:pos', 'zamknięty:adj:pl:acc:n:pos', 'zamknięty:adj:pl:voc:m2:pos',
'zamknięty:adj:pl:voc:m3:pos', 'zamknięty:adj:pl:voc:f:pos', 'zamknięty:adj:pl:voc:n:pos', 'zamknąć:adj:sg:nom:n:perf:aff', 'zamkn
ąć:ppas:sg:nom:n:perf:aff', 'zamknąć:adj:sg:acc:n:perf:aff', 'zamknąć:ppas:sg:acc:n:perf:aff', 'zamknąć:adj:pl:nom:m2:perf:aff', '
zamknąć:ppas:pl:nom:m2:perf:aff', 'zamknąć:adj:pl:nom:m3:perf:aff', 'zamknąć:ppas:pl:nom:m3:perf:aff', 'zamknąć:adj:pl:nom:f:perf:
aff', 'zamknąć:ppas:pl:nom:f:perf:aff', 'zamknąć:adj:pl:nom:n:perf:aff', 'zamknąć:ppas:pl:nom:n:perf:aff', 'zamknąć:adj:pl:acc:m2:
perf:aff', 'zamknąć:ppas:pl:acc:m2:perf:aff', 'zamknąć:adj:pl:acc:m3:perf:aff', 'zamknąć:ppas:pl:acc:m3:perf:aff', 'zamknąć:adj:pl
:acc:f:perf:aff', 'zamknąć:ppas:pl:acc:f:perf:aff', 'zamknąć:adj:pl:acc:n:perf:aff', 'zamknąć:ppas:pl:acc:n:perf:aff'], disamb=zam
knąć:ppas:pl:nom:n:perf:aff
orth=,, interps=[',:interp'], disamb=,:interp
orth=zbiegł, interps=['zbiec:praet:sg:m1:perf', 'zbiec:praet:sg:m2:perf', 'zbiec:praet:sg:m3:perf'], disamb=zbiec:praet:sg:m1:perf
orth=po, interps=['PO:subst:sg:nom:f', 'PO:subst:sg:nom:n', 'PO:subst:sg:gen:f', 'PO:subst:sg:gen:n', 'PO:subst:sg:dat:f', 'PO:sub
st:sg:dat:n', 'PO:subst:sg:acc:f', 'PO:subst:sg:acc:n', 'PO:subst:sg:inst:f', 'PO:subst:sg:inst:n', 'PO:subst:sg:loc:f', 'PO:subst
:sg:loc:n', 'PO:subst:sg:voc:f', 'PO:subst:sg:voc:n', 'PO:subst:pl:nom:f', 'PO:subst:pl:nom:n', 'PO:subst:pl:gen:f', 'PO:subst:pl:
gen:n', 'PO:subst:pl:dat:f', 'PO:subst:pl:dat:n', 'PO:subst:pl:acc:f', 'PO:subst:pl:acc:n', 'PO:subst:pl:inst:f', 'PO:subst:pl:ins
t:n', 'PO:subst:pl:loc:f', 'PO:subst:pl:loc:n', 'PO:subst:pl:voc:f', 'PO:subst:pl:voc:n', 'po:subst:acc', 'po:prep:acc', 'po:subst
```

```
:loc', 'po:prep:loc'], disamb=po:prep:loc
orth=schodach, interps=['schody:subst:pl:loc:n', 'schód:subst:pl:loc:m3', 'schód:subst:pl:loc:m3'], disamb=schody:subst:pl:loc:n
orth=,, interps=[',:interp'], disamb=,:interp
orth=minął, interps=['minąć:praet:sg:m1:perf', 'minąć:praet:sg:m2:perf', 'minąć:praet:sg:m3:perf'], disamb=minąć:praet:sg:m1:perf
orth=furtkę, interps=['furtka:subst:sg:acc:f'], disamb=furtka:subst:sg:acc:f
orth=,, interps=[',:interp'], disamb=,:interp
orth=także, interps=['także:qub'], disamb=także:qub
orth=ją, interps=['on:ppron3:sg:acc:f:ter:akc:npraep'], disamb=on:ppron3:sg:acc:f:ter:akc:npraep
orth=zamknął, interps=['zamknąć:praet:sg:m1:perf', 'zamknąć:praet:sg:m2:perf', 'zamknąć:praet:sg:m3:perf'], disamb=zamknąć:praet:s
g:m1:perf
orth=,, interps=[',:interp'], disamb=,:interp
orth=i, interps=['i:conj', 'i:conj', 'i:interj', 'i:conj', 'i:interj', 'i:qub'], disamb=i:conj
orth=znalazł, interps=['znaleźć:praet:sg:m1:perf', 'znaleźć:praet:sg:m2:perf', 'znaleźć:praet:sg:m3:perf'], disamb=znaleźć:praet:s
g:m1:perf
orth=się, interps=['się:qub'], disamb=się:qub
orth=na, interps=['na:interj', 'na:interj:acc', 'na:prep:acc', 'na:interj:loc', 'na:prep:loc'], disamb=na:prep:loc
```

W wersji rozszerzonej skrypt powinien jako pierwszy argument przyjąć słowo, dla którego informacje (jak w powyższym przykładzie) mają być wypisane, razem z analogiczną informacją dla słowa poprzedzającego i następującego. Przykład poniżej („<=” oznacza poprzednie słowo, „==” — bieżące, a „”>=” — następne słowo):

./proc_nkjp-ext.py o $NKJP/010-2-000000001

```
<=orth=stanowi, interps=['stan:subst:sg:dat:m3', 'stanowić:subst:sg:ter:imperf', 'stanowić:fin:sg:ter:imperf', 'stanowy:subst:pl:n
om:m1:pos', 'stanowy:fin:pl:nom:m1:pos', 'stanowy:adj:pl:nom:m1:pos', 'stanowy:subst:pl:voc:m1:pos', 'stanowy:fin:pl:voc:m1:pos',
'stanowy:adj:pl:voc:m1:pos'], disamb=stanowić:fin:sg:ter:imperf
==orth=o, interps=['o:interj', 'o:interj:acc', 'o:prep:acc', 'o:interj:loc', 'o:prep:loc', 'ojciec:interj:pun', 'ojciec:prep:pun',
 'ojciec:brev:pun'], disamb=o:prep:loc
>=orth=wymiarze, interps=['wymiar:subst:sg:loc:m3', 'wymiar:subst:sg:voc:m3'], disamb=wymiar:subst:sg:loc:m3
<=orth=decydującą, interps=['decydować:pact:sg:acc:f:imperf:aff', 'decydować:pact:sg:inst:f:imperf:aff', 'decydujący:pact:sg:acc:f
:pos', 'decydujący:adj:sg:acc:f:pos', 'decydujący:pact:sg:inst:f:pos', 'decydujący:adj:sg:inst:f:pos'], disamb=decydować:pact:sg:i
nst:f:imperf:aff
==orth=o, interps=['o:interj', 'o:interj:acc', 'o:prep:acc', 'o:interj:loc', 'o:prep:loc', 'ojciec:interj:pun', 'ojciec:prep:pun',
 'ojciec:brev:pun'], disamb=o:prep:loc
>=orth=losie, interps=['los:subst:sg:loc:m3', 'los:subst:sg:voc:m3'], disamb=los:subst:sg:loc:m3
<=orth=chodzi, interps=['chodzić:fin:sg:ter:imperf'], disamb=chodzić:fin:sg:ter:imperf
==orth=o, interps=['o:interj', 'o:interj:acc', 'o:prep:acc', 'o:interj:loc', 'o:prep:loc', 'ojciec:interj:pun', 'ojciec:prep:pun',
 'ojciec:brev:pun'], disamb=o:prep:acc
>=orth=sensowność, interps=['sensowność:subst:sg:nom:f', 'sensowność:subst:sg:acc:f'], disamb=sensowność:subst:sg:acc:f
<=orth=już, interps=['już:qub'], disamb=już:qub
==orth=o, interps=['o:interj', 'o:interj:acc', 'o:prep:acc', 'o:interj:loc', 'o:prep:loc', 'ojciec:interj:pun', 'ojciec:prep:pun',
 'ojciec:brev:pun'], disamb=o:prep:acc
>=orth=religię, interps=['religia:subst:sg:acc:f'], disamb=religia:subst:sg:acc:f
<=orth=,, interps=[',:interp'], disamb=,:interp
==orth=o, interps=['o:interj', 'o:interj:acc', 'o:prep:acc', 'o:interj:loc', 'o:prep:loc', 'ojciec:interj:pun', 'ojciec:prep:pun',
 'ojciec:brev:pun'], disamb=o:prep:acc
>=orth=tajemnicę, interps=['tajemnica:subst:sg:acc:f'], disamb=tajemnica:subst:sg:acc:f
<=orth=ale, interps=['Al:depr:pl:nom:m2', 'Al:depr:pl:voc:m2', 'Ala:depr:pl:nom:f', 'Ala:subst:pl:nom:f', 'Ala:depr:pl:acc:f', 'Al
a:subst:pl:acc:f', 'Ala:depr:pl:voc:f', 'Ala:subst:pl:voc:f', 'Ali:depr:pl:nom:m2', 'Ali:subst:pl:nom:m2', 'Ali:depr:pl:nom:m2', '
Ali:depr:pl:voc:m2', 'Ali:subst:pl:voc:m2', 'Ali:depr:pl:voc:m2', 'Alo:depr:pl:nom:m2', 'Alo:subst:pl:nom:m2', 'Alo:depr:pl:nom:m2
', 'Alo:depr:pl:nom:m2', 'Alo:depr:pl:voc:m2', 'Alo:subst:pl:voc:m2', 'Alo:depr:pl:voc:m2', 'Alo:depr:pl:voc:m2', 'Alo:depr:pl:nom
:m1', 'Alo:subst:pl:nom:m1', 'Alo:depr:pl:nom:m1', 'Alo:depr:pl:nom:m1', 'Alo:subst:pl:nom:m1', 'Alo:depr:pl:voc:m1', 'Alo:subst:p
l:voc:m1', 'Alo:depr:pl:voc:m1', 'Alo:depr:pl:voc:m1', 'Alo:subst:pl:voc:m1', 'ale:depr', 'ale:subst', 'ale:depr', 'ale:depr', 'al
e:subst', 'ale:conj', 'ale:depr', 'ale:subst', 'ale:depr', 'ale:depr', 'ale:subst', 'ale:conj', 'ale:qub'], disamb=ale:conj
==orth=o, interps=['o:interj', 'o:interj:acc', 'o:prep:acc', 'o:interj:loc', 'o:prep:loc', 'ojciec:interj:pun', 'ojciec:prep:pun',
 'ojciec:brev:pun'], disamb=o:prep:acc
>=orth=normę, interps=['norma:subst:sg:acc:f'], disamb=norma:subst:sg:acc:f
```

7 Przetwarzanie plików PDF

Prawie żadne akty prawne publikowane przez Sejm RP i rząd nie zawierają spisów treści w metadanych. Nie zawierają ich też, ze względu na charakter aktów prawnych, w treści. Czasami okazują się nawet skanami bez bezpośredniego dostępu do treści jako ciągu znaków. Skrypt `ustawa.py` drukuje bardzo uproszczony spis treści podając numery rozdziałów i strony, na których się zaczynają, o ile w treści dokumentu są takowe wyróżnione i dokument nie jest skanem.

ustawa.py

```
#!/usr/bin/python3
import pdfplumber
import re


def print_chapters(filename):
```

```
    """
    Print chapters in a PDF file (in Polish).
    :param filename: PDF file name
    :return: nothing
    Processing is simplified. The file may contain no chapters.
    """
    with pdfplumber.open(filename) as pdf:
        page_number = 1
        for p in pdf.pages:
            t = p.extract_text()
            for r in re.findall(r'^(Rozdział)\W(\w*)', t,
                                re.M | re.I):
                print(r[0], r[1], 'strona', page_number)
            page_number = page_number + 1


if (__name__ == "__main__"):
    import sys
    for filename in sys.argv[1:]:
        print(filename)
        print_chapters(filename)
```

Sam numer rozdziału pozwala się szybciej poruszać po dokumencie, ale nie informuje o znaczeniu poszczególnych części dokumentu. Należy podany skrypt rozszerzyć, tak aby drukował tytuły części, ksiąg, działów, rozdziałów i numery artykułów, o ile takie występują. Poniżej podany jest przykład początkowych wierszy wyjścia dla kodeksu postępowania cywilnego.

./ustawa-base.py D20211805Lj.pdf

```
D20211805Lj.pdf
TYTUŁ WSTĘPNY strona 1
CZĘŚĆ PIERWSZA strona 4
KSIĘGA PIERWSZA strona 4
TYTUŁ I strona 4
DZIAŁ I strona 4
Rozdział 1 strona 4
Oddział 1 strona 4
Oddział 2 strona 5
Rozdział 2 strona 7
Oddział 1 strona 7
Oddział 2 strona 7
Oddział 3 strona 9
Oddział 4 strona 10
DZIAŁ II strona 11
DZIAŁ III strona 13
TYTUŁ II strona 15
TYTUŁ III strona 16
TYTUŁ IIIa strona 17
TYTUŁ IIIb strona 18
TYTUŁ IIIc strona 18
TYTUŁ IV strona 18
DZIAŁ I strona 18
...
```

W wersji rozszerzonej należy także drukować także tytuł aktu prawnego. Można to osiągnąć na różne sposoby, np. inspirując się rozwiązaniem podanym na stronie:
`https://pdfminersix.readthedocs.io/en/latest/howto/character_properties.html`. Przykładowe wyjście podane jest poniżej. Swoboda prawodawcy w ustalaniu formatu jednostek redakcyjnych aktu czyni dokładne wykonanie tego zadania dość karkołomnym; w przykładzie widać pewne uproszczenia.

./ustawa-ext.py D20211805Lj.pdf

```
D20211805Lj.pdf
Ustawa z dnia 17 listopada 1964 r. Kodeks postępowania cywilnego
TYTUŁ WSTĘPNY strona 1
CZĘŚĆ PIERWSZA strona 4
KSIĘGA PIERWSZA strona 4
TYTUŁ I strona 4
DZIAŁ I strona 4
Rozdział 1 strona 4
Oddział 1 strona 4
Oddział 2 strona 5
Rozdział 2 strona 7
Oddział 1 strona 7
Oddział 2 strona 7
Oddział 3 strona 9
Oddział 4 strona 10
```

8 Kryteria oceny

Za poszczególne zadania można dostać 2 punkty w wersji podstawowej. Rozszerzenie pozwoli zdobyć dodatkowe pół punktu.

- KWIC: wersja podstawowa — bez odmiany, wersja rozszerzona — możliwość odmiany
- informacje o miejscach wydarzeń: wersja podstawowa — kraje z podanym polem powierzchni, wersja rozszerzona — także miejscowości z liczbą mieszkańców
- właściwe oznaczenie segmentu w NKJP: wersja podstawowa — wypisanie realizowane w programie przykładowym, wersja rozszerzona — umożliwienie wyszukiwanie określonego słowa i wypisanie informacji jak w wersji podstawowej, ale tylko o poprzednim, danym i następnym słowie
- spis treści aktu prawnego: wersja podstawowa — tylko spis treści, wersja rozszerzona — także tytuł ustawy.

Dodatek — Instalacja oprogramowania

Przykładowe programy wymagają Pythona i jego bibliotek. Najlepiej używać nowszych ich wersji, w szczególności Pythona w wersji 3. W najnowszych dystrybucjach Linuksa jest on dostępny, ale w innych, aby domyślnie Python był używany wersji 3, konieczna może być instalacja pakietu `python-is-python3`. Biblioteki można instalować korzystając z polecenia `pip3` (na Fedorze itp. zapewne po prostu `pip`) lub `conda`. To ostatnie wymaga instalacji minicondy (lub pełnej anacondy). Można to zrobić w przestrzeni użytkownika. Opis na stronie `https://docs.conda.io/en/latest/miniconda.html`. Jeżeli okaże się, że na danym komputerze w laboratorium `conda` nie jest zainstalowana, można ją doinstalować w przestrzeni użytkownika korzystając z opisu na podanej stronie, ale należy skorzystać z archiwum ze strony `https://repo.anaconda.com/miniconda/` do ściągnięcia wersji działającej z Pythonem 2.7; potem można zainstalować Python 3 i biblioteki dla Pythona 3.

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
KPRM Kancelaria Prezesa Rady Ministrów
Unia Europejska Europejski Fundusz Rozwoju Regionalnego