Εργασία 2 Μεταγλωττιστές

Σίνγκ Αλέξανδρος Π2013101

1 Implementation

Το πρόγραμμα αποτελείται κυρίως απο δυο συναρτήσεις, την `visit`:

```
def visit(urlStr):
    page = urllib.request.urlopen(urlStr)
    text = page.read().decode('utf-8')
    page.close()
    return text
```

Η οποία για δεδομένο url, γυρνάει το κείμενο της ιστοσελίδας, και την `analyse`:

```
def analyse(url):
    text = visit(url)
    links = []
    rexp = re.compile(r'<a[^>]* href="([^"]*)">(.+?)</a>',re.DOTALL)
    matches = rexp.finditer(text)
    for item in matches:
        if not(re.match(r'#.+',item.group(1))): #ignore anchors
            links.append((item.group(2),urllib.parse.urljoin(url,item.group(1))))
    return links
```

Η οποία για δεδομένο url, το επισκέπτεται με την χρήση της `visit`, και στην συνέχεια το φιλτράρει με το regex "<a[^]* href="([^"]*)">(.+?)", επιστρέφοντας τελικά, μια λίστα απο tuples μορφής (ετικέτα συνδέσμου, full url συνδέσμου). Το παραπάνω regex λειτουργεί ως εξής: Αναγνωρίζει strings της μορφής: "<a (οτιδίποτε εκτός του >) href= "(group 1: οτιδίποτε εκτος του ")"> (group 2: optionally οτιδίποτε) ". Τα δυο group αντιστοιχούν στο url και το label αντίστοιχα. Η `analyse` αγνοεί συνδέσμους anchor/local links, αγνοώντας οσα links έχουνε το σύμβολο # στο string μετα το "href=". Εαν θέλουμε να τους συμπεριλάβουμε, αρκεί να αφαιρέσουμε το αντίστοιχο if-statement.
Τέλος καλούμε την `analyse` σε κάποιο url, και εκτυπώνουμε κάθε αποτέλεσμα καθώς και το κείμενο/ιστοσελίδα τους, μέσω της `visit`. πχ:

```
for label,url in analyse('http://di.ionio.gr/~mistral/tp/compilers
                                /lecturedoc/unit3/module1.html'):
    print(label,":",url)
    print(visit(url))
```

2 Παράδειγμα εκτέλεσης

Με αρχικό url το “http://di.ionio.gr/ mistral/tp/compilers/lecturedoc/unit3/module1.html”, έχουμε αντίστοιχα κείμενο εισόδου το:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">


<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />

    <titleE   > 3-1 &mdash; Compiler Lecture Notes 1.0 documentation</title>

    <link rel="stylesheet" href="../_static/sphinxdoc.css" type="text/css" />
    <link rel="stylesheet" href="../_static/pygments.css" type="text/css" />

    <script type="text/javascript">
      var DOCUMENTATION_OPTIONS = {
        URL_ROOT:    '../',
        VERSION:     '1.0',
        COLLAPSE_INDEX: false,
        FILE_SUFFIX: '.html',
        HAS_SOURCE:  true
      };
    </script>
    <script type="text/javascript" src="../_static/jquery.js"></script>
    <script type="text/javascript" src="../_static/underscore.js"></script>
    <script type="text/javascript" src="../_static/doctools.js"></script>
    <link rel="top" title="Compiler Lecture Notes 1.0 documentation" href="../index.ht
  </head>
  ...
```

(Full input text in “input.txt” file)
Στο οποίο η analyse βρίσκει τα εξής links:

```
Compiler Lecture Notes 1.0 documentation : http://di.ionio.gr/~mistral/tp/compilers/le
DFA : http://en.wikipedia.org/wiki/Deterministic_finite_automaton
NFA : http://en.wikipedia.org/wiki/Nondeterministic_finite_state_machine
https://docs.python.org/3/library/re.html : https://docs.python.org/3/library/re.html
Compiler Lecture Notes 1.0 documentation : http://di.ionio.gr/~mistral/tp/compilers/le
Sphinx : http://sphinx-doc.org/
```

Τέλος το πρόγραμμα τα εκτυπώνει μαζί με τα αντίστοιχα κείμενα τους, ενδεικτικά το πρώτο αρχίζει ως εξής:

```
Compiler Lecture Notes 1.0 documentation : http://di.ionio.gr/~mistral/tp/compilers/le
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">


<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />

    <titleM      > - Σ    M     E         &mdash; Compiler Lecture Notes 1.0 documentat

    <link rel="stylesheet" href="_static/sphinxdoc.css" type="text/css" />
    <link rel="stylesheet" href="_static/pygments.css" type="text/css" />

    <script type="text/javascript">
      var DOCUMENTATION_OPTIONS = {
        URL_ROOT:    './',
        VERSION:     '1.0',
        COLLAPSE_INDEX: false,
        FILE_SUFFIX: '.html',
        HAS_SOURCE:  true
      };
    </script>
    <script type="text/javascript" src="_static/jquery.js"></script>
    <script type="text/javascript" src="_static/underscore.js"></script>
    <script type="text/javascript" src="_static/doctools.js"></script>
    <link rel="top" title="Compiler Lecture Notes 1.0 documentation" href="#" />
    <link rel="next" titleB   =" Σ    Python 3" href="unit1/module1.html" />
  </head>
  ...
```

(Full output in results.txt file)