Comment extraire l'information de pages web

Zhentao Li

École Normale Supérieure

2 avril 2014

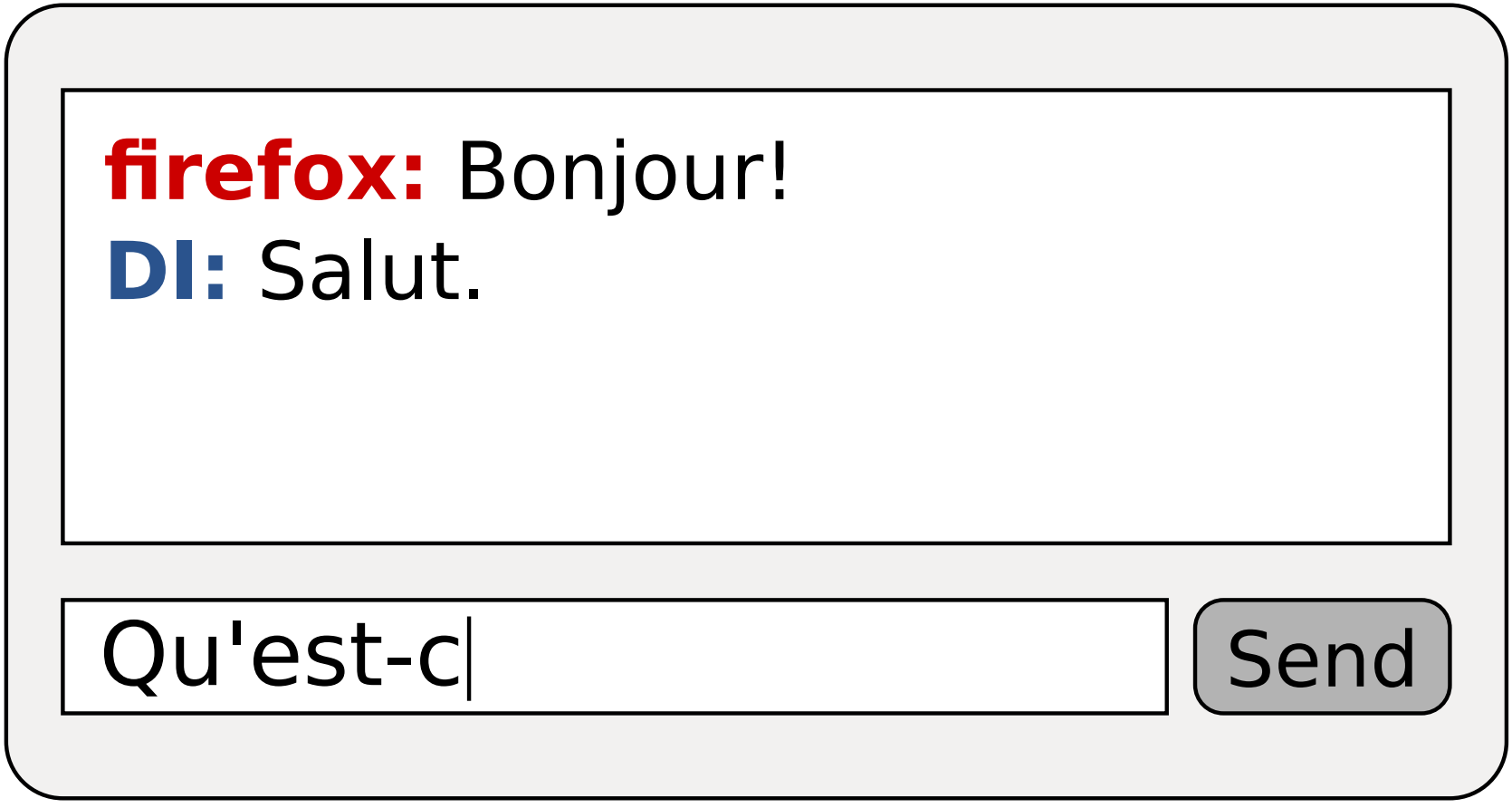
Extraire l'information

s = "Le Département
d'Informatique
de l'ENS (DI ENS)
est à la fois"

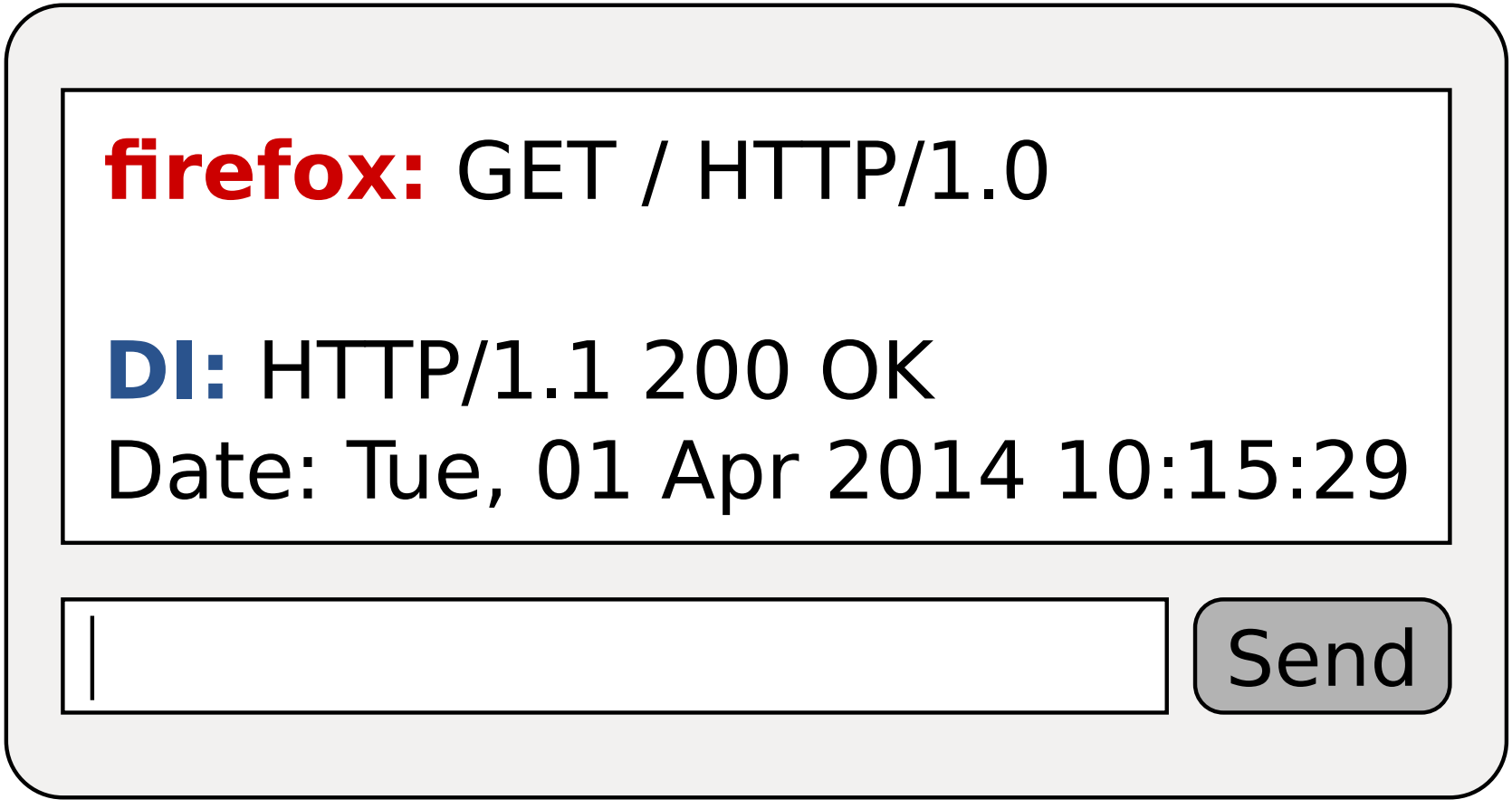
Comment fonctionne un navigateur?*

*jusqu'à un certain point

Comment fonctionne un navigateur?



Comment fonctionne un navigateur?



```
HTTP/1.1 200 OK
Date: Tue, 01 Apr 2014 10:15:29 GMT
Server: Apache/2.2.14 (Ubuntu)
Last-Modified: Mon, 06 Feb 2012 23:10:22 GMT
Accept-Ranges: bytes
Content-Length: 58722
Vary: Accept-Encoding
Connection: close
Content-Type: text/html
```

Entête retour

```
<!-- -*- coding: utf-8; -*- -->
<!-- $Id: index.shtml,v 1.36 2004/09/04 15:19:34 local Exp
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//
<html lang="en">
<head>
<meta http-equiv="Content-Type" content="text/html; charset
<link rel=stylesheet href="astree.css" type="text/css">
<title>The Astr&eacute;e Static Analyzer</title>
<meta name="Keyword" content="validation, verification, sta
debugging, testing, abstract interpretation, abstract, inte
safety-critical, real-time, embedded, static program analys
```

Document

Le document

Écrire en blanc
« Département d'Info »
Puis, en dessous, écrire en plus petit « ENS ».
Ensuite, dessiner l'image de

Interprétation du document

Méthodes de traitement de page web

- Traiter le document comme une chaîne de caractères.
- Utiliser un module spécialisé pour des chaîne de caractères avec ce format (Cours d'aujourd'hui).
- Ce module transforme le document (encodé en HTML) en dictionnaires, listes et chaînes de caractères emboîtés. Il ajoute aussi quelques fonctions utiles.

Le module BeautifulSoup

Site

http://www.crummy.com/software/BeautifulSoup/

- Peut aussi être utilisé sans installation en commençant `python` au bon endroit (dans le premier niveau du répertoire extrait).
- Autre modules aux fonctions similaires:
  - mechanize
  - lxml (BeautifulSoup 4 peut utiliser lxml)
  - html5lib
  - sgmllib (ancien)

Chargement du document

```
>>> from bs4 import BeautifulSoup
>>> soup = BeautifulSoup(html_doc)
```

La variable de base est une variable de type `BeautifulSoup`.

```
>>> soup.title
<title> WebHome - Département d'informatique de l'ENS</title>
>>> soup.title.name
'title'
```

Pour obtenir le document HTML.

```
>>> import urllib
>>> html_doc = urllib.urlopen("http://www.di.ens.fr").read()
```

Fonctions globales simples

Extrait toutes les parties texte.

```
>>> print soup.get_text()
 WebHome - Département d'informatique de l'ENS
        @import url('pub/TWiki/TWikiTemplates/base.css');
        @import url("pub/TWiki/PatternSkin/layout.css");
[...]
Le Département d'Informatique de l'ENS (DI ENS) est à la fois un
département d'enseignement et un laboratoire de recherche affilié au
CNRS et à l'INRIA (Unité Mixte de Recherche 8548).

Du point de vue de l'enseignement, le DI forme ses élèves (admis au
concours de l'ENS comme sur titre) au sein du prédoctorat de l'ENS et
du MPRI. [...]
```

Fonctions globales simples

Afficher le contenu avec indentation

```
>>> print soup.prettify()
[...]
           <ul>
            <li>
             <a class="twikiCurrentTopicLink twikiCurrentWebHomeLink twikiLi
              <span class="di-internal-lang-fr">
               Accueil
              </span>
             </a>
            </li>
[...]
```

Recherche

- Le HTML est un langage de parenthèse. Python utilise `()`, `[]`, `{}`, HTML en utilise beaucoup plus de types de parenthèses appellés ≫ tags ≫.
  - `<span>` ouvre une parenthèse span.
  - `</span>` ferme une parenthèse span.
- De plus, une paire de parenthèses peut avoir un dictionnaire associé à son ouverture.
- Trouver tous les tags de type 'a'

```
>>> soup.find_all('a')
[<a id="PageTop" name="PageTop"></a>, <a href="http://www.ens.fr/" target=
"_top">ENS</a>, <a href="http://www.cnrs.fr/" target="_top">CNRS</a>, <a
href="http://www.inria.fr/" target="_top">INRIA</a>,
[...]
```

Recherche

- Trouver tous les liens (tous les parentheses de type ‘a’ qui ont une cle ‘href’)

```
>>> liens = []
>>> for tag in soup.find_all('a'):
>>>     if tag.has_attr("href"):
>>>         liens.append(tag["href"])
>>> liens
['http://www.ens.fr/', 'http://www.cnrs.fr/', 'http://www.inria.fr/', 'http
[...]
```

- Trouve le premier tag où la clé `id` prends valeur `PageTop`

```
>>> soup.find(id="PageTop")
<a id="PageTop" name="PageTop"></a>
```

Exploration des structures emboîtés

- Sauter au bon endroit: recherche par type de tag et paires (clé,valeur) du dictionnaire associé.

```
>>> noeud = soup.find_all('body')[0]
```

- Se déplacer dans la structure emboîtée.

```
>>> # Se déplacer a la premiere parenthese a l'interieur de la parenthese a
>>> premier_fils = list(noeud.children)[0]
>>> # Se déplacer a la premiere parenthese contenant la parenthese actuelle
>>> parent = noeud.parent
```

- Dans tous les cas, il faut trouver un modèle pour l'information recherchée. À partir de ce modèle, on peut ensuite écrire le programme requis.

Utiliser l'interface du site lui-même

- Pour des sites « modernes », il existe souvent des pages spécialement conçues pour la navigation par machine. Exemple:

```
http://fr.wikipedia.org/w/api.php?format=json&action=query&
titles=Python&prop=revisions&rvprop=content
```

- Normalement sous le nom du API (*Application programming interface* ou *interface de programmation*) web.
- La page retourné resemble à des objets emboîtés, mais ici nous avons en plus l'interprétation du créateur de la page. En fait, la plupart du temps, c'est encodé en JSON (*JavaScript Object Notation*).

Pour transformer une chaîne de caractère en JSON en des dictionnaires python:

```
import json
dict_python = json.loads(doc)
```

Exemple d'extraction de données à partir de l'API web

```
import urllib
import json

adresse_base = "http://fr.wikipedia.org/w/api.php"
parametres = {"format":"json", "action":"query", "titles":"Python",
              "prop":"revisions", "rvprop":"content"}
adresse_url = adresse_base + "?" + urllib.urlencode(parametres)
doc = urllib.urlopen(adresse_url).read()
dict_python = json.loads(doc)
# Examiner dans l'interpreteur pour trouver les elements recherches
# Par exemple le contenu de la page se trouve ici
print dict_python['query']['pages']['2302']['revisions'][0]['*']
```

Sortie:

```
{{Homonymie}}
{{Homophone|Piton}}
{{Autres projets|wiktionary = Python}}

Le mot '''python''' peut désigner :
* [[Python (mythologie)|Python]], un animal monstrueux de Delphes dans la my
* ''[[Python (genre)|Python]]'', genre de serpent qui doit son nom au précéd
```