Noticias en RSS

- El sistema operativo UNIX -

Christian Pareja Jensen

Fundación Vodafone España

MASTERINFTEL

UNIVERSIDAD DE MÁLAGA

MASTERINFTEL

RSS

- **RSS** son las siglas de **Really Simple Syndication**, un formato XML para sindicar o compartir contenido en la web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos. El formato permite distribuir contenidos sin necesidad de un navegador, utilizando un software diseñado para leer estos contenidos RSS.

MASTERINFTEL

Código RSS

{**'summary_detail'**: {**'base'**: u'http://20minutos.feedsportal.com/c/32489/index.rss', **'type'**: u'text/html', **'value'**: u' EFE / V\xcdDEO : ATLAS El tribunal resolver\xe1 los catorce recursos presentados al auto con el que finaliz\xf3 la instrucci\xf3n de esta causa judicial.
Los magistrados tienen en sus manos confirmar la postura del juez Castro.
El instructor ha mantenido la imputaci\xf3n de do\xf1a Cristina por dos delitos fiscales y uno de blanqueo', **'language'**: None}, **'published_parsed'**: *time.struct_time*(tm_year=2014, tm_mon=11, tm_mday=7, tm_hour=5, tm_min=5, tm_sec=15, tm_wday=4, tm_yday=311, tm_isdst=0), **'links'**: [{**'href'**: u'http://20minutos.feedsportal.com/c/story01.htm', **'type'**: u'text/html', **'rel'**: u'alternate'}], **'title'**: u'La Audiencia de Palma decide este viernes sobre la imputaci\xf3n de la infanta Cristina', 'authors': [{}], **'updated'**: u'2014-11-07T05:05:15Z', **'summary'**: u' EFE , **'guidislink'**: False, **'title_detail'**: {'base': u'http://20minutos.feedsportal.com/c/32489/f/478284/index.rss', 'type': u'text/plain', **'value'**: u'La Audiencia de Palma decide este viernes sobre la imputaci\xf3n de la infanta Cristina', 'language': None}, **'link'**: u'http://20minutos.com/story01.htm', **'author'**: u'EFE / V\xcdDEO : ATLAS', **'published'**: u'Fri, 07 Nov 2014 05:05:15 GMT', **'author_detail'**: {**'name'**: u'EFE / V\xcdDEO : ATLAS'}, **'updated_parsed'**: time.struct_time(tm_year=2014, tm_mon=11, tm_mday=7, tm_hour=5, tm_min=5, tm_sec=15, tm_wday=4, tm_yday=311, tm_isdst=0)}

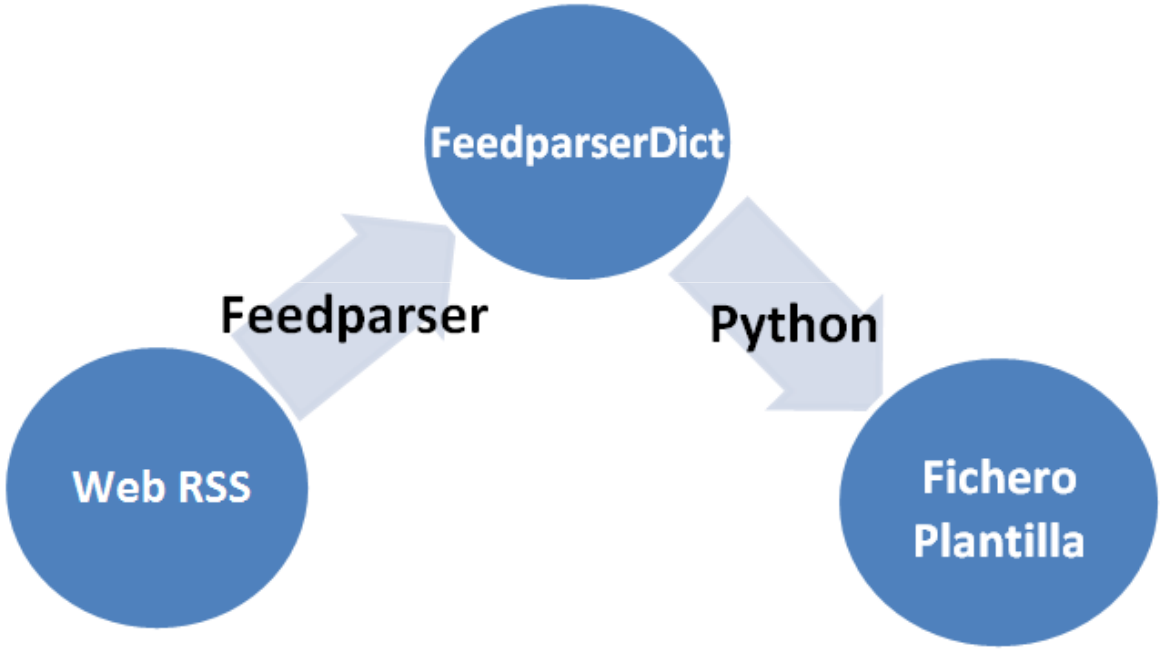
MASTERINFTEL

Plantilla

<ID>
Identificador. Un número entero que identifica de forma unívoca la noticia en el fichero.
</ID>
<TITULAR>
Titular de la noticias
</TITULAR>
<DESCRIPCION>
Texto de la noticia (puede contener varias líneas)
</DESCRIPCION>
<ENLACE>
URL del sitio donde ha sido publicada la noticia
</ENLACE>
<ORIGEN>
Medio que publica la noticias
</ORIGEN>
<FECHA>
Fecha de publicación en formato aaaa-mm-dd
</FECHA>
<RELACIONADAS
Una lista de IDs de noticias separadas por comas. Cuando se añade una nueva noticia esta lista estará vacía.
</RELACIONADAS>

MASTERINFTEL

Diseño



MASTERINFTEL

Características

- Dos ficheros como parámetros
  - ***Fichero con webs de noticias RSS***
  - ***Fichero donde escribir las noticias ya formateadas***
- Evitar añadir noticias duplicadas

MASTERINFTEL

Características

- Librería Feedparser

import feedparser

feedparser.parse(linea)

- Codificar texto entre lectura y escritura

f.write(lista[k][l].encode('utf-8', 'ignore'))

- Convertir fecha a formato especificado

import time

time.strftime("%Y-%m-%d",d['entries'][i].published_parsed)

MASTERINFTEL

Características

- Comprobar que el campo pedido exista

```
if(d['entries'][i].has_key("author")):
```

- Limpiar código de etiquetas (Opcional)

```
import re
TAG_RE = re.compile(r'<[^>]+>')

def remove_tags(text):
    return TAG_RE.sub('', text)
```