Infopython

Midiendo Información de Mass-Media Con Python

(O continuando la charla de Lipe)

Juan B Cabral <jbc.develop@gmail.com>

Pyday Gonzales Catán 2011

¿Quien Soy?

- Soy ingeniero en sistemas de la UTN-FRC.
- Investigo minerias de datos orientado a redes de haplotipos (Biologia).
- Me gustan los juegos de rol (tengo uno hecho en django y otro a medio hacer).
- Trabajo en java (Suicidio en progreso).

¿Qué es Infopython?

- Infopython es un toolkit que se utiliza para la valoración de medios de información utilizando teorías formales.
- Es una prueba de concepto
- Surge como muchos módulos sueltos en mi trabajo durante el año 2010.
- Por que continua la charla de Lipe?

Un Poco de Teoría

- Existen diferentes teorías para determinar la importancia de los medios. sobre la opinión publica.
- La “Teoría de Información” de Shannon es una formalización matemática de una de la “Aguja hipodérmica”.
- Existen teorías más complejas.

Teoría de La "Agenda-Setting"

- La teoría de la agenda-setting postula que los medios de comunicación de masas tienen una gran influencia sobre el público al determinar qué historias poseen interés informativo y cuánto espacio e importancia se les da. (Wikipedia)
- El punto central de esta teoría es asignar una prioridad para obtener mayor audiencia, mayor impacto y una determinada conciencia sobre la noticia.

¿Qué es un Medio Para Nuestro Caso?

"En el dominio es un "coso" al cual quiero medir el valor de su información"

Entonces tiene que:

- Ser homogéneo en su información.
- Tener la sensación de "unidad".
- Ser Medible.

Formalizando

VALOR = F(AUDIENCIA, IMPACTO)

- VALOR: Es la importancia del medio dada la teoría.
- AUDIENCIA: A cuánta gente le llega la información del medio.
- IMPACTO: Qué tanta importancia le da la audiencia al medio
- NOTA: La conciencia NO es medible (o no se me ocurrió como formalizarlo)

Proponiendo F(AUDIENCIA, IMPACTO)

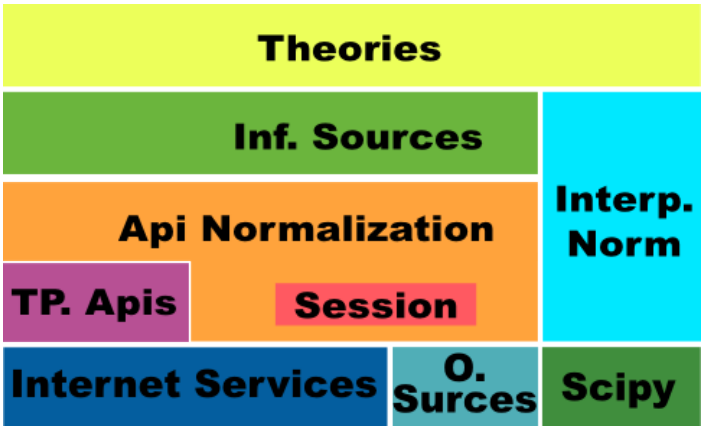
VALOR = AUDIENCIA * IMPACTO

- Funciona bien cuando uno de los valores es 0.
- Refleja mejor la variación de los valores.

Consideraciones finales de diseño

- Es mas fácil empezar por "nuevos medios" (web, twitter, etc).
- Existe una amplia variedad de servicios públicos que miden "cosas" sobre nuevos medios.
- Hay que identificar qué mide la audiencia y qué mide el impacto de estas "cosas".
- El reto no es técnico.

Arquitectura de Infopython



Metodología de Trabajo

1. Configurar la sesión.
2. Crear los medios.
3. Crear "sacadatos". (Opcional)
4. Crear los interpoladores. (Opcional)
5. Crear la/s agenda/s.
6. Evaluar los nodos.

session API

```
from infopython import session

# Listado de todas las llaves OBLIGATORIAS
session.NEEDED_KEYS

# crea una nueva session con las llaves v0, v1, ...
session.set(v0=1, v1=2...)

# retorna el valor de una llave
session.get("v0")

# borra la session
session.clear()
```

IS Webpage

- Representa una página web (PLOP!).
- No importa si es web, un perfil de twitter o un blog.
- **Audiencia:**
    - Compete (http://www.compete.com/).
    - Alexa (http://www.alexa.com/).
- **Impacto:**
    - Page Rank (http://es.wikipedia.org/wiki/PageRank).

WebPage API

```
from infopython.isources import webpages

google = webpages.WebPage("google.com")

print "ID> " + google.id
print "URL> " + google.url
print "HTML>\n" + google.html

print "Compete>"
pprint(google.get_info("compete"))
```

IS TwitterUser

- Representa un usuario de twitter (PLOP²)
- **Audiencia:**
    - Followers.
    - Klout (http://klout.com/).
- **Impacto:**
    - RT.
    - Klout (http://klout.com/).

TwitterUser API

```
from infopython.isources import twitteruser

yo = twitteruser.TwitterUser("leliel12")

print "ID> " + yo.id
print "Username> " + yo.username
print "Tweepy>"
pprint(yo.get_info("tweepy"))
```

Agenda API

```
from infopython import agenda
from infopython.util import interpolator
from infopython.isources import twitteruser

google = webpages.WebPage("google.com")
yahoo = webpages.WebPage("yahoo.com")

aud = lambda w: w.get_info("compete")["metrics"]["uv_count"]
imp = lambda w: w.get_info("pagerank")["pagerank"]
itp = interpolator.PieceWisePolynomial([0,0,1,1,2,45,64], [1,3,1,1,2,4,64])

ag = agenda.AgendaSetting(itype=webpages.WebPage,
                          inf_sources=[google, yahoo],
                          audience_valuator=aud,
                          impact_valuator=imp,
                          audience_interpolator=itp,
                          impact_interpolator=itp)
```

Agenda API 2

```
ag.value_of(google)
ag.impact_of(google)
ag.audience_of(google)
ag.wrap(google)

ag.count(google)
ag.remove(google)
ag.append(google)

ag.for_type
ag.audience_valuator
ag.impact_valuator
ag.audience_interpolator
ag.impact_interpolator
```

Comparando 2 Agendas

```
for i in agenda.rank_isources(ag1, ag2):
    print i
```

Futuro 1

Las que dije que hiba a agregar en Pycon 2010:

- linkedin.
- Integrar más tipos de massmedia (imdb, amazon...).
- y... ¿desde el punto de vista de la audiencia?
- ¿Web semántica?
- nltk.

Futuro 2

- El manejo de sesiones APESTA! (debería hacerlo multi sesión) y con mas configuraciones (tiempo de espera)

```
my_session = sessions.Session(...)
google = webpages.WebPage("google.com", session=my_session)
```

- Análisis de Texto:

  Esta si lo empecé a implementar para Diarios!

- Pedirle a lipe que porte scripts de de scraping a infopython :D

Futuro 3

- Analisis de imagenes:

```
from infopython.isource.images import  Image

img_1 = Image(open("/foto_de_campo.png"))
img_2 = Image(open("/foto_de_ciudad.png"))

impacto = lambda img: contar_pixels_color_verde(img)

.. agenda ..
```

¿Preguntas?

- **Proyecto:**
  - http://bitbucket.org/leliel12/infopython/
- **Esta Charla:**
  - Source: https://bitbucket.org/leliel12/talks/src
  - Pet #2: http://revista.python.org.ar/
- **Contacto:**
  - Juan B Cabral <jbc.develop@gmail.com> / @JuanBCabral