

---

# **Data Mining**

## **Introduction au web mining**

Thierry Dhorne

2 mars 2015

# Intérêt du web mining

## ❖ Intérêt du web mining

- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- le web est constitué de
  - texte
  - tableaux numériques
  - images
  - vidéos
  - ....
- il contient donc des informations plus ou moins stratégiques
- qu'il est éventuellement judicieux d'analyser

# Comment miner le web ?

❖ Intérêt du web mining

❖ Comment miner le web ?

❖ Exemples d'applications

❖ Exemple 1

❖ Exemple 2

❖ Exemple 3

❖ Vu du browser

❖ XML - HTML

❖ HTML

❖ HTML - Représentation

❖ HTML - Structure

❖ HTML - Structure

❖ Xpath

❖ Navigation

❖ Navigation

❖ R XML

❖ Web mining

- de manière statique et manuelle
  - comme T. Dhorne demande aux étudiants de le faire : copié-collé
- de manière automatique
  - c'est parfois plus rapide mais parfois moins (il faut programmer un peu)
- de manière dynamique
  - il faut programmer (une fois !) mais après ça marche tout seul !

# Exemples d'applications

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- suivre en temps réel les commentaires sur votre site
- disposer d'informations statistiques au jour le jour
- scruter le web pour trouver des informations nouvelles
- surveiller un concurrent (dynamique du site)

# Exemple 1

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

The screenshot displays the 'Premiere' website interface. At the top, there's a navigation bar with 'Premiere' logo, social media icons, and user options like '+ S'inscrire' and '+ S'identifier'. The main content area is titled 'Top des avis spectateurs' (Top of viewer reviews) with the subtitle 'les films les mieux notés par les internautes'. Below this, there are tabs for 'Tout le top cinéma', 'Box office', 'Top spectateurs', 'Top films première', and 'Top DVD'. The 'Top spectateurs' tab is selected, showing a grid of movie posters with their respective review counts and star ratings. The first row includes 'Imitation Game' (27 avis, 4 stars), 'American Sniper' (15 avis, 4 stars), and 'Un Village presque Parfait' (12 avis, 4 stars). The second row includes 'L'Âcher-prise: Cinquante nuances de Grey' (4 avis, 4 stars), 'Jupiter' (5 avis, 4 stars), and 'Papa ou Maman' (6 avis, 4 stars). On the left side, there are sections for 'Leonard Nimoy est mort', 'Le nouveau rôle d'Eddie Redmayne', and 'Avengers : L'Ere d'Ultron'. A 'PRATIQUE' sidebar offers links to 'Séances', 'Programme TV', 'TV-replay', 'Spectacles', and 'Personnalités'. On the right, there's a 'HORAIRES & SALLES' section with a search bar and a list of cinema chains (UGC, mk2, G, eRHEF, KGC). Below that is an 'A VOIR AUSSI' section with a recommendation from Outbrain and a subscription offer for 'ABONNEZ-VOUS à partir de 19€ seulement'. At the bottom right, there's a Facebook social module for 'Première' with 100,694 likes and a 'Gagnez vos places pour' promotion.

## Exemple 2

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- suivre les sites des départements STID et voir ceux qui bougent le plus
- on scane les sites tous les jours et on regarde s'ils ont changé par rapport à la veille

# Exemple 3

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

The screenshot displays the Investing.com website interface. The main content area shows the EUR/RUB exchange rate at 69,082, with a daily change of +0,604 (+0,88%). Below this is a table of daily price movements from February 1, 2015, to February 27, 2015. The table includes columns for Date, Dernier (closing price), Ouv. (opening price), + Haut (high), + Bas (low), and Variation % (percentage change). Summary statistics at the bottom of the table indicate a high of 80,119, a low of 67,956, a difference of 12,163, an average of 73,051, and a variation of -12,06.

On the right side, there is a 'SignalTräder' advertisement with the text 'Continuez vous à faire du trading manuel ?' and 'Automatisez votre trading'. Below the ad is a 'REJOINGNEZ NOUS !' button. A secondary table shows various indices: Dow 30 (-0,45%), DAX (+0,66%), Euro Stoxx 50 (+0,67%), and Indice US Dollar (-0,07%).

At the bottom right, there is a 'Devises' section for EUR/USD, showing a current rate of 1,1195 with a change of -0,0003 (-0,03%). It includes a 'Résumé' section with 'Vente Forte' and a table of indicators for Achat and Vente.

Date	Dernier	Ouv.	+ Haut	+ Bas	Variation %
27/02/2015	69.082	68.515	69.724	68.315	0.83%
26/02/2015	68.515	69.765	69.947	67.956	-1.80%
25/02/2015	69.771	71.555	71.555	69.626	-2.49%
24/02/2015	71.555	72.301	72.579	71.091	-1.03%
23/02/2015	72.301	70.407	73.397	70.407	2.69%
22/02/2015	70.407	70.407	70.407	70.407	-0.00%
20/02/2015	70.407	70.356	71.047	69.604	0.06%
19/02/2015	70.364	70.179	71.339	69.634	0.24%
18/02/2015	70.194	71.318	71.452	69.310	-1.60%
17/02/2015	71.336	71.727	72.413	70.624	-0.57%
16/02/2015	71.748	72.288	72.564	70.667	-0.77%
15/02/2015	72.304	72.304	72.304	72.304	-0.00%
13/02/2015	72.304	74.414	75.028	71.747	-2.86%
12/02/2015	74.433	73.948	76.369	72.954	0.64%
11/02/2015	73.957	74.088	75.619	73.335	-0.20%
10/02/2015	74.107	74.573	75.810	73.635	-0.65%
09/02/2015	74.589	75.923	76.106	73.726	-1.79%
08/02/2015	75.950	75.950	75.950	75.950	-0.00%
06/02/2015	75.950	76.328	77.200	75.000	-0.57%
05/02/2015	76.383	76.962	78.317	75.229	-0.86%
04/02/2015	77.043	74.814	78.499	73.902	2.88%
03/02/2015	74.885	77.431	77.563	74.259	-3.37%
02/02/2015	77.494	77.939	80.119	77.092	-0.85%
01/02/2015	78.155	78.155	78.155	78.155	-0.00%

Le + haut: 80.119    Le + bas: 67.956    Différence: 12.163    Moyenne: 73.051    Variation %: -12.06



# Vu du browser

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
  - ❖ Exemple 1
  - ❖ Exemple 2
  - ❖ Exemple 3
  - ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

```
598     <div class="clear"></div>
599 <div style="text-align:center;padding:10px;">
600   <a href="https://twitter.com/intent/tweet?button_hashtag=SCBFCN" class="twitter-hashtag-button" data-lang="fr" data-size="large">Tweet #SCBFCN</a>
601 <script>!function(d,s,id){var js,fjs=d.getElementsByTagName(s)[0];if(!d.getElementById(id)){js=d.createElement(s);js.id=id;js.src="//platform.twitter.com/widgets.js";fjs.
602 </div>
603 </div>
604 <div class="clear"></div>
605
606 <!-- Tirs au buts -->
607
608 <!-- BUTS -->
609 <br/>
610 <div id="buts">
611   <ul class="club_dom">
612     <ul>
613       <ul class="club_ext">
614     </ul>
615   </ul>
616 </div>
617
618 <!-- CARTONS -->
619 <div id="cartons">
620   <ul class="club_dom clear">
621     <ul>
622       <ul class="club_ext">
623     </ul>
624   </ul>
625 </div>
626 <div class="clear"></div>
627 </div>
628 </div>
629
630 <div class="details">
631   <div class="onglets clear" id="onglets_infos" style="position: relative;">
632     <ul class="nav_onglets">
633       <li id="onglet_infos" class="on"><a href="#bloc_infos"><span class="lien">Infos match</span><span class="fin"></span></a></li>
634       <li id="onglet_statistiques" class=""><a href="#bloc_statistiques"><span class="lien">Statistiques</span><span class="fin"></span></a></li>
635       <li id="onglet_statistiquesJoueurs" class=""><a href="#bloc_statistiquesJoueurs"><span class="lien">Statistiques joueurs</span><span class="fin"></span></a></li>
636       <li id="onglet_videoMatch" class=""><a href="#bloc_videoMatch"><span class="lien">Résumé vidéo</span><span class="fin"></span></a></li>
637     <?php// endif;?>
638   </ul>
639 </div>
640 <div id="bloc_infos" class="onglets_infos" style="display: block;">
641 <div id="bloc_infosMatch_load" class="load"></div>
642 <div id="bloc_infosMatch_data"></div>
643 </div>
```



# XML - HTML

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- XML : eXtensible Markup Language
  - les langages de markup permettent l'annotation de structures distinctes du texte
  - extensible signifie que l'on peut créer ses propres annotations
- en XML, les marqueurs de structures sont < et >
- HTML : HyperText Markup Language est le langage initial du web et reste le plus important aujourd'hui
  - mais plus le plus efficace
- HTML (bien que plus ancien) peut être considéré comme un sous XML
  - moins rigoureux (et moins flexible)

# HTML

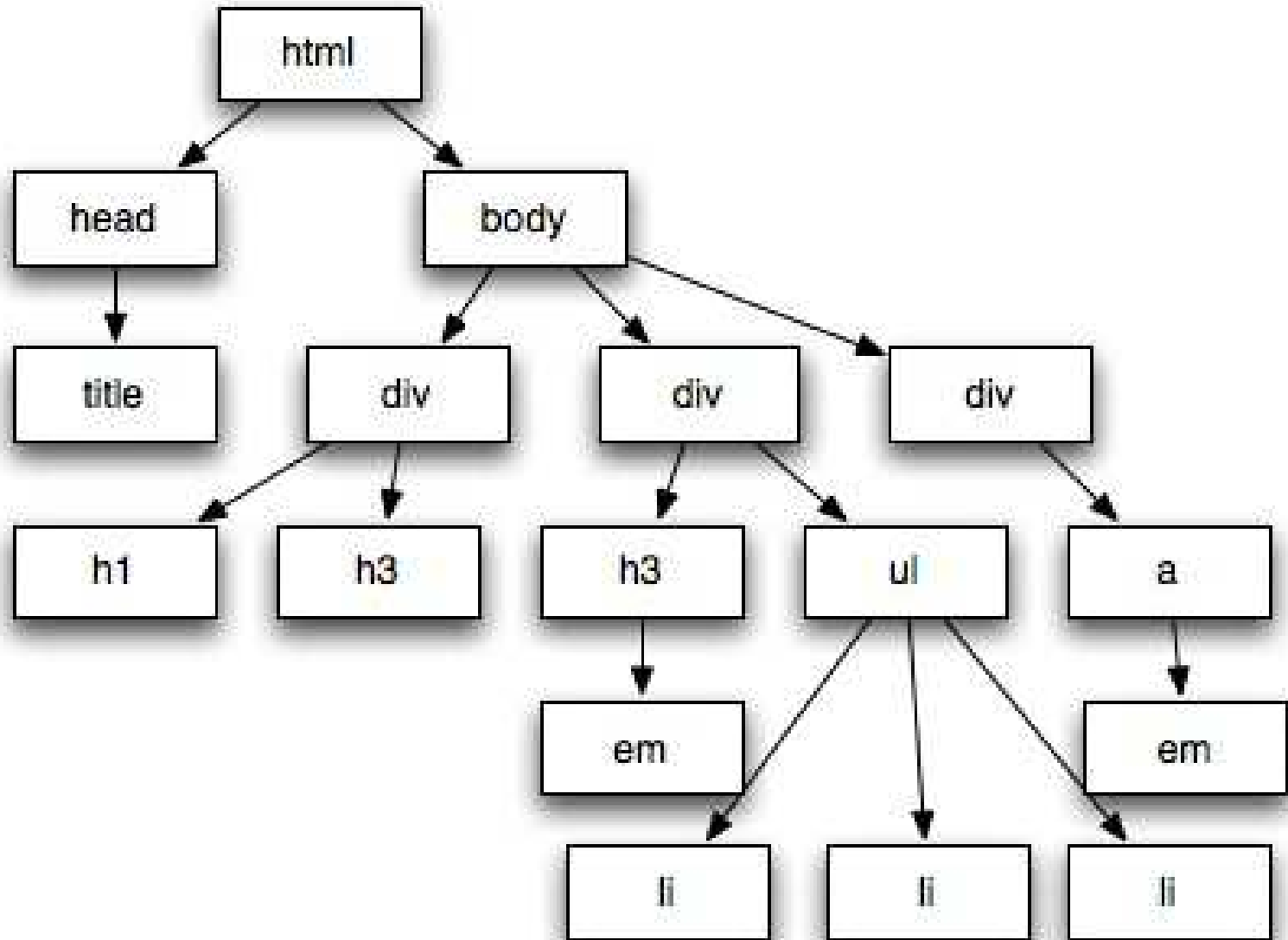
- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ **HTML**
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- HTML est simplement un ensemble de marques

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "h
<html>
<head>
<title>The document title</title>
</head>
<body>
<h1>Main heading</h1>
<p>A paragraph.</p>
<a href = "www.stid-vannes.fr">Statistics Web
</body>
</html>
```

# HTML - Représentation

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining



# HTML - Structure

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- les boîtes s'appellent nœuds
- les lignes s'appellent arêtes
- une arête va de a à b si :

```
<a>
```

```
    <b>
```

```
    </b>
```

```
</a>
```

- ★ il n'y a qu'un unique chemin de la racine à un nœud quelconque

# HTML - Structure

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- à chaque nœud peuvent être associés les éléments suivants :

- un nom (requis mais non nécessairement unique)
- un nombre (quelconque) d'attributs
- un texte (optionnel)

- exemple

```
<a href = "www.stid-vannes.fr">
```

```
Statistiques du site
```

```
</a>
```

- avec

- nom du nœud : a
- attribut : href de valeur "www.stid-vannes.fr"
- texte : Statistiques du site

# Xpath

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

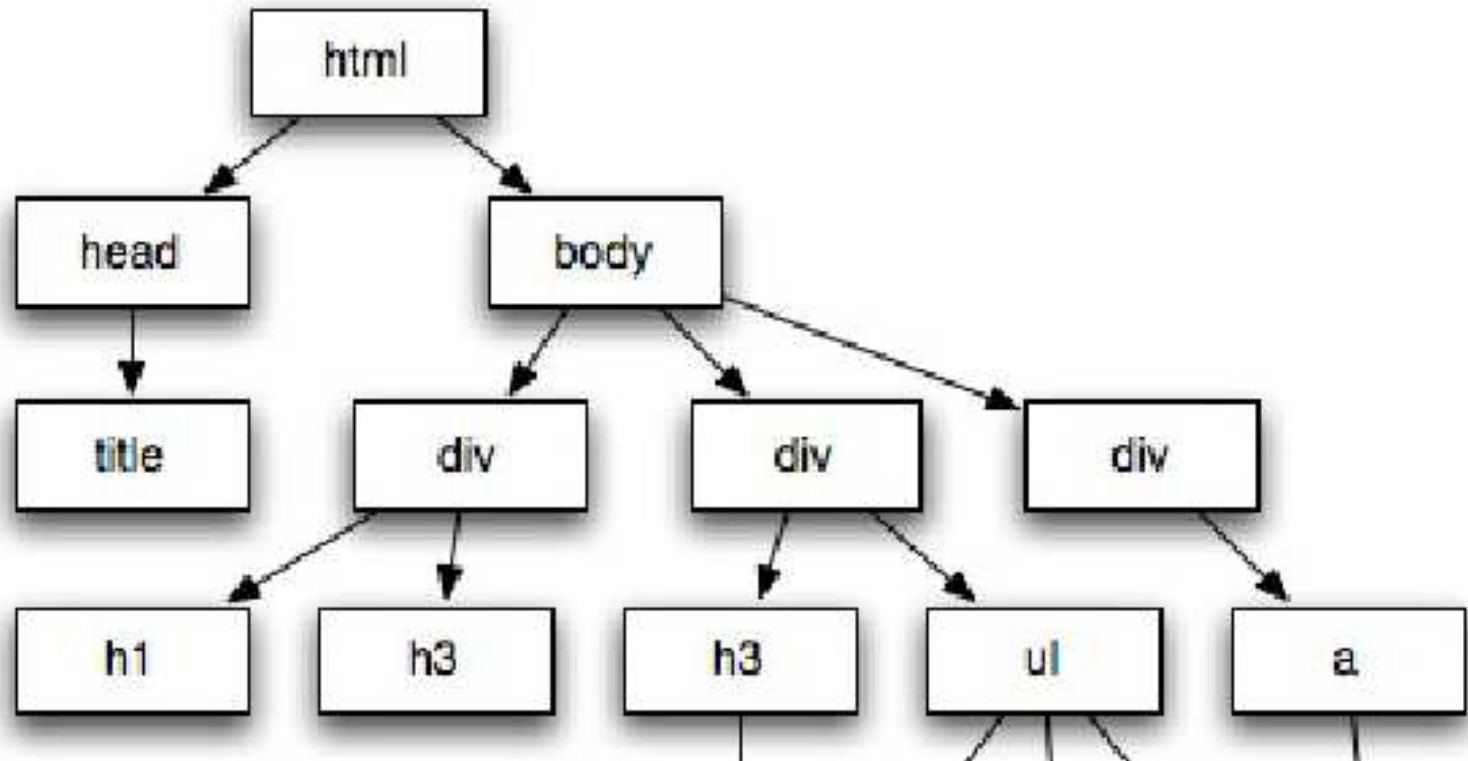
- XPath est un langage de requête qui permet de retrouver des nœuds spécifiés dans un arbre XML/HTML
- quelques symboles utilisables pour représenter un chemin sur l'arbre :
  - / trouve la racine
  - // sélectionne un nœud quelconque sur l'arbre
  - . sélectionne le nœud courant
  - .. sélectionne le parent du nœud courant
  - @ sélectionne des attributs
  - nodename recherche la position du nœud appelé



# Navigation

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- pour aller jusqu'à la balise (ancree) *a* on peut utiliser
  - `"../body/div/a"`
  - `"//a"`



# Navigation

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- si la balise est :

```
<a href = "stid.vannes.fr">stats</a>
```

- on peut chercher avec

```
"//a[@href = 'stid.vannes.fr']"
```

- ou

```
"//a[text() = 'stats']"
```

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- R dispose d'un package qui permet de travailler avec des arbres XML/HTML
- les deux fonctions les plus utiles (pour nous) sont
  - `htmlTreeParse`  
récupère une page HTML et crée une structure d'arbre interne à R
  - on peut alors utiliser XPath pour parcourir l'arbre

```
>url <- "https://fr.wikipedia.org/wiki/Charolaise"
>doc <- htmlTreeParse(url,useInternalNodes = TRUE)
```
  - `getNodeSet` recherche dans le document des nœuds spécifiés :

```
> x1 <- getNodeSet(doc, "//div[@class = 'bd']")
> x2 <- getNodeSet(doc, "//a")
```

# Web mining

- ❖ Intérêt du web mining
- ❖ Comment miner le web ?
- ❖ Exemples d'applications
- ❖ Exemple 1
- ❖ Exemple 2
- ❖ Exemple 3
- ❖ Vu du browser
- ❖ XML - HTML
- ❖ HTML
- ❖ HTML - Représentation
- ❖ HTML - Structure
- ❖ HTML - Structure
- ❖ Xpath
- ❖ Navigation
- ❖ Navigation
- ❖ R XML
- ❖ Web mining

- le web est d'abord un monde automatique
- Google,
- ★ l'effort de recherche est considérable (big data)
- même si on cherche des choses simples (?) il faut automatiser
- d'où la nécessité de réfléchir
  - aux objectifs
  - aux enjeux
  - aux techniques