

18/01/2019

# Notepad++

Expressions régulières

M. PERES LEBLANC

# Notepad++



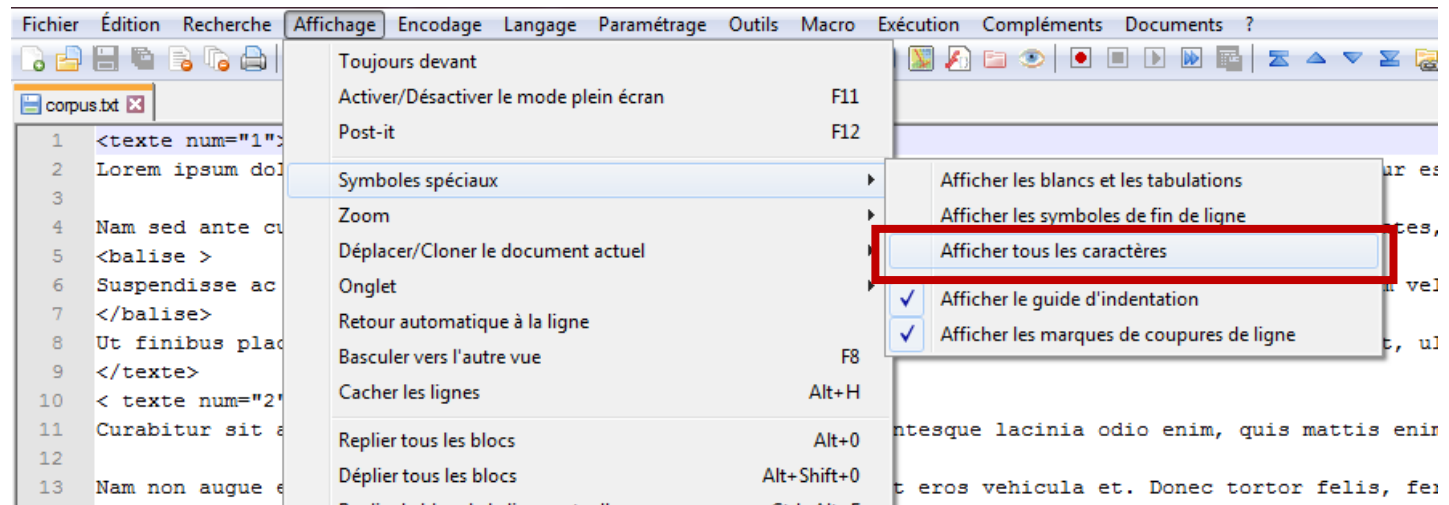
Bien démarrer

# Afficher tous les caractères

3

- Show symbol > Show All Characters ou version française ci-dessous.
- Fait apparaître, entre autre, les sauts de lignes et retours chariots

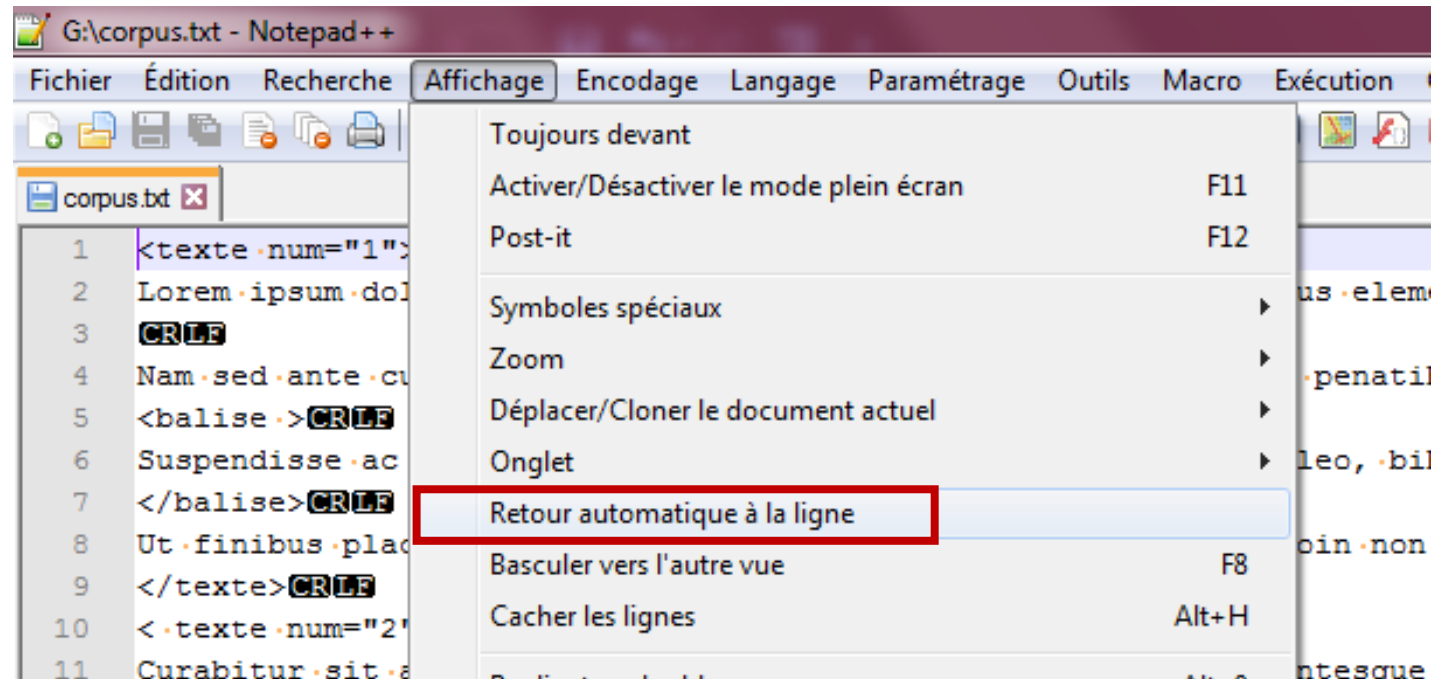
- CR => \r
- LF => \n



# Retour automatique à la ligne

4

- Word Warp
- => permet de replier les lignes dans les limites de la largeur de l'écran

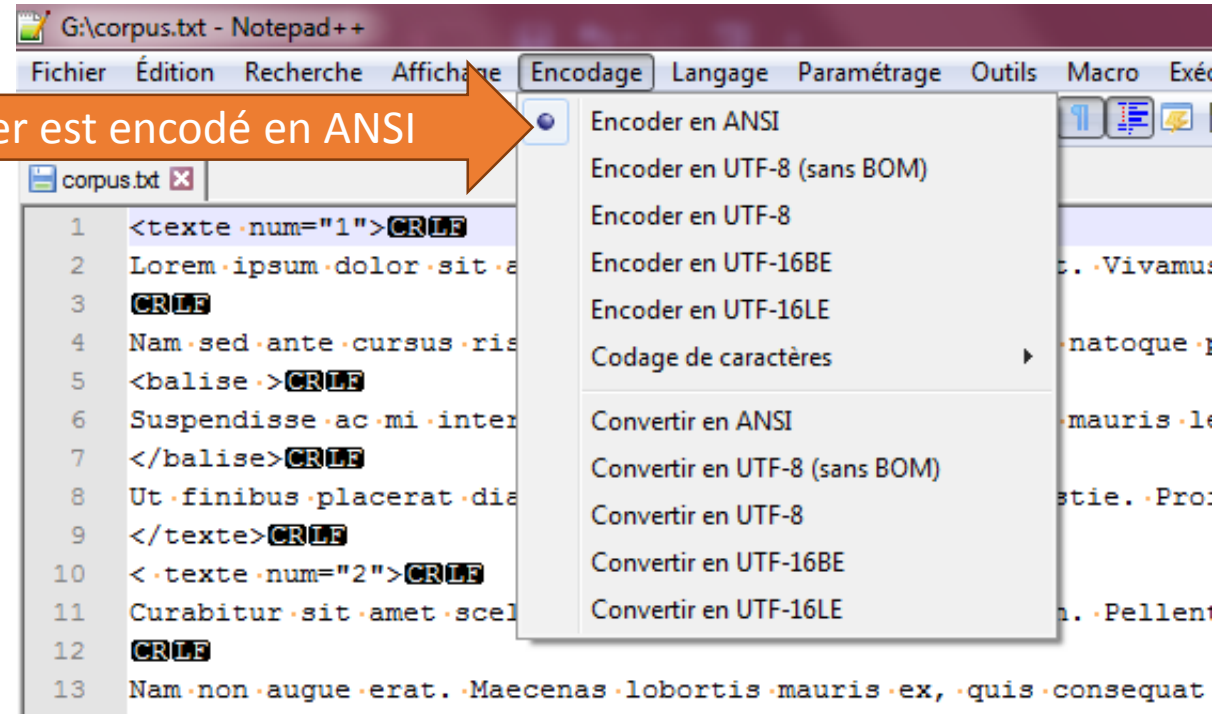


# Encodage

5

- Vérifier si le corpus est encodé en ANSI ou en UTF-8

Ici le fichier est encodé en ANSI



# Langage

6

- Choisir le langage xml permet entre autre d'avoir une aide visuelle pour la fabrication des balises

```
1 <texte num="1">CRLE
2 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Viv
3 CRLE
4 Nam sed ante cursus risus commodo feugiat. Orci varius natoq
5 CRLE
6 Suspendisse ac mi interdum elit blandit aliquet. Donec mauri
7 CRLE
8 Ut finibus placerat diam, a interdum sem suscipit molestie.
9 </texte>CRLE
10 <.texte num="2">CRLE
11 Curabitur sit amet scelerisque neque, la turpis sapien. Pel
12 CRLE
13 Nam non augue erat. Maecenas lobortis mauris ex, quis conseq
14 CRLE
15 Mauris vitae tortor non risus pretium sodales. Phasellus lac
16 </texte>CRLE
17 <texte num="3" loc="machin">CRLE
```

Balise mal rédigée

```
G:\corpus.txt - Notepad++
Fichier  Édition  Recherche  Affichage  Encodage  Langage  Paramétrage  Outils  Macro  Exécut
corpus.txt x
1 <texte num="1">CRLE
2 Lorem ipsum dolor sit amet, con
3 CRLE
4 Nam sed ante cursus risus commo
5 <balise>CRLE
6 Suspendisse ac mi interdum elit
7 </balise>CRLE
8 Ut finibus placerat diam, a int
9 </texte>CRLE
10 <.texte num="2">CRLE
11 Curabitur sit amet scelerisque
12 CRLE
13 Nam non augue erat. Maecenas lo
14 <balises>CRLE
15 Mauris vitae tortor non risus p
16 </texte>CRLE
17 <texte num="3" loc="machin">CR
18 Aliquam in feugiat neque, id ve
19 CRLE
20 Duis lobortis iaculis tincidunt
21 CRLE
22 >.Sed a
23 CRLE
24 Morbi s
25 CRLE
26 Suspendisse tincidunt tortor ip
27 </texte>CRLE
28 <texte num="4">CRLE
29 Donec sed tristique justo. Sed pretium nulla sed libero tempor se
```

Choisir xml

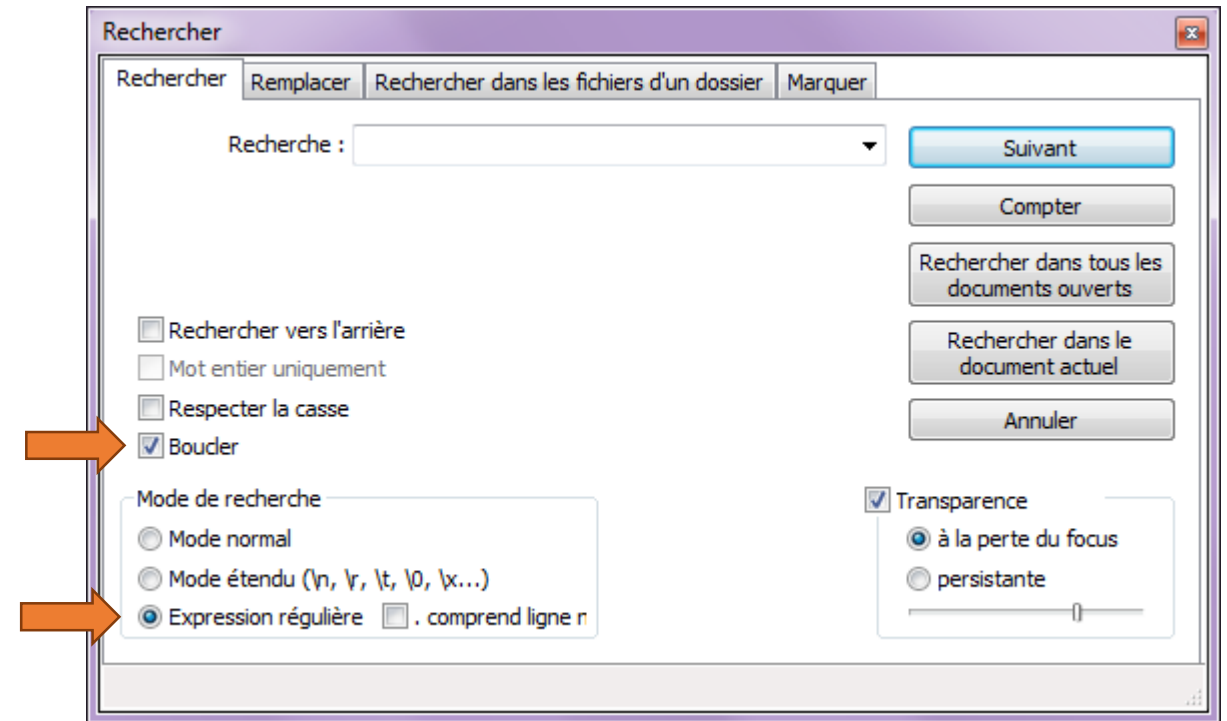
# REGEX

A thick orange horizontal bar spans the width of the slide. In the middle-right section, the bar is interrupted by a diagonal segment that rises from the baseline to a higher level, then continues as a horizontal segment at that higher level.

# Fenêtre rechercher

8

- en mode expression régulière
- Permet de chercher ou de compter

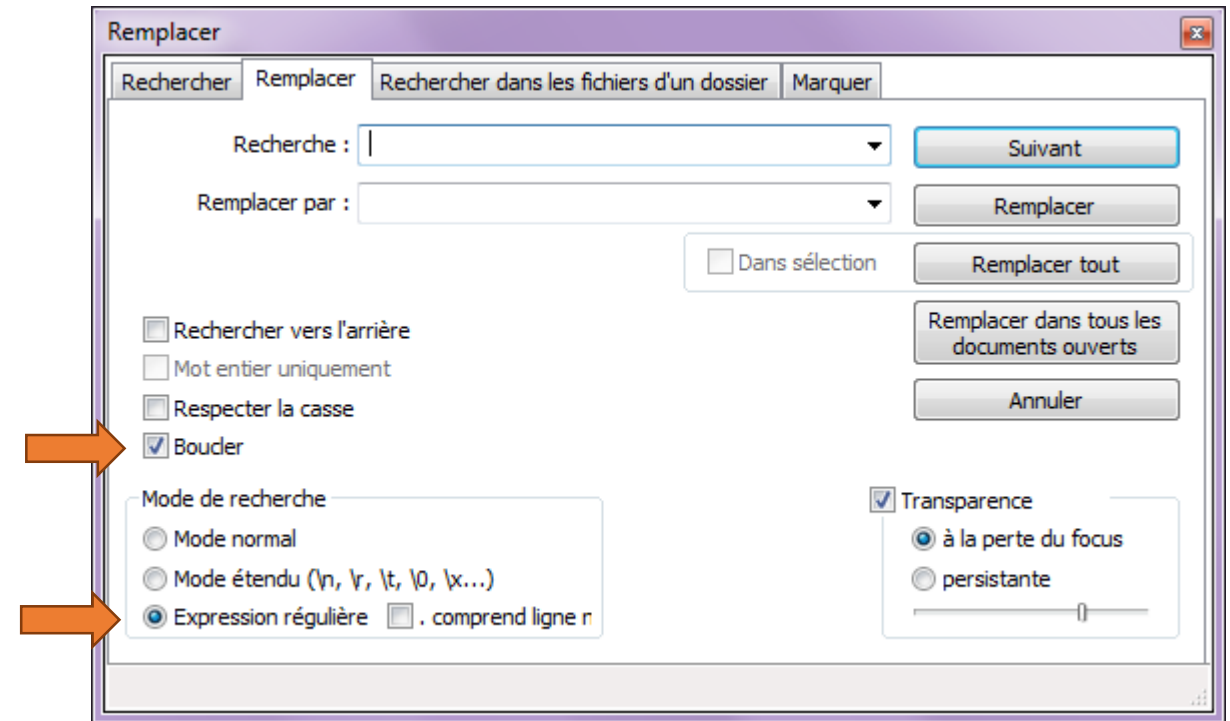




# Fenêtre de remplacer

9

- en mode expression régulière
- Permet de chercher et de remplacer dans le document actif ou dans les documents ouverts



# Rechercher des caractères

10

- `.` => n'importe quel caractère : `sa.le` recherche « sable », « salle » et « salles »... mais aussi « brisa les »
- `.*` => n'importe quel caractère en nombre indéfini (dont espace et ponctuation) : `a.*r` recherche « affiner », « accoudoir »... mais « sable, elle est d'une manipulation rapide »

# Rechercher des caractères

11

- `[]` => l'un des caractères entre les crochets : `[ae]` trouve le ou la ou les...
- `\n` => recherche les fin de ligne
- `\r` => recherche les retour chariot
- `\r\n` => le plus courant sur Windows
- `\t` => les tabulations

# Les classes de caractères

12

- intervalle 0-9 représente l'ensemble des chiffres {0,1,2,3,4,5,6,7,8,9}
- Intervalle a-z représente l'ensemble des lettres {a,b,c...,x,y,z} en minuscule
- A-Za-z représente l'ensemble des lettres en majuscule et en minuscule
- [-;.,] cherche : un tiret, un point-virgule, un point ou une virgule

# Rechercher des caractères

13

- `l[ea]` => recherche la lettre « l » suivie de « e » ou de « a »
- `m[a-z]` => recherche la lettre « m » suivie de toutes les lettres entre a et z dans l'alphabet
- `l[^a]` => recherche la lettre « l » suivie de tout sauf de « a »
- `m{2}` => trouve « mm »
- `10{1,3}` => trouve « 10 », « 100 » et « 1000 »
- `cal+e` => trouve « cale » et « calle »

# Début et fin de mot



14

- \`<inter` : trouve « internationale » ou « interlocuteurs » et pas « réinterprétation »
- in\`>` : trouve « lendemain » ou « vain » et pas « intéressant »

# Début et fin de paragraphe



15

- $\wedge \Rightarrow$  début de paragraphe
- $\$ \Rightarrow$  fin de paragraphe

# Sensible à la casse



16

- `\<[Nn]ous\>` => cherche Nous et nous
- `\<[Vv]ous\> aurez` => vous ou Vous suivis de « aurez »



# Rechercher un mot

17

- Pourquoi pas...

`([a-zA-zéèàùïôûçëâ]*)[^\s!.,?;]`

- Mais on peut faire mieux (exclu aussi la ponctuation) :

`\w+`

# Formatage de dates

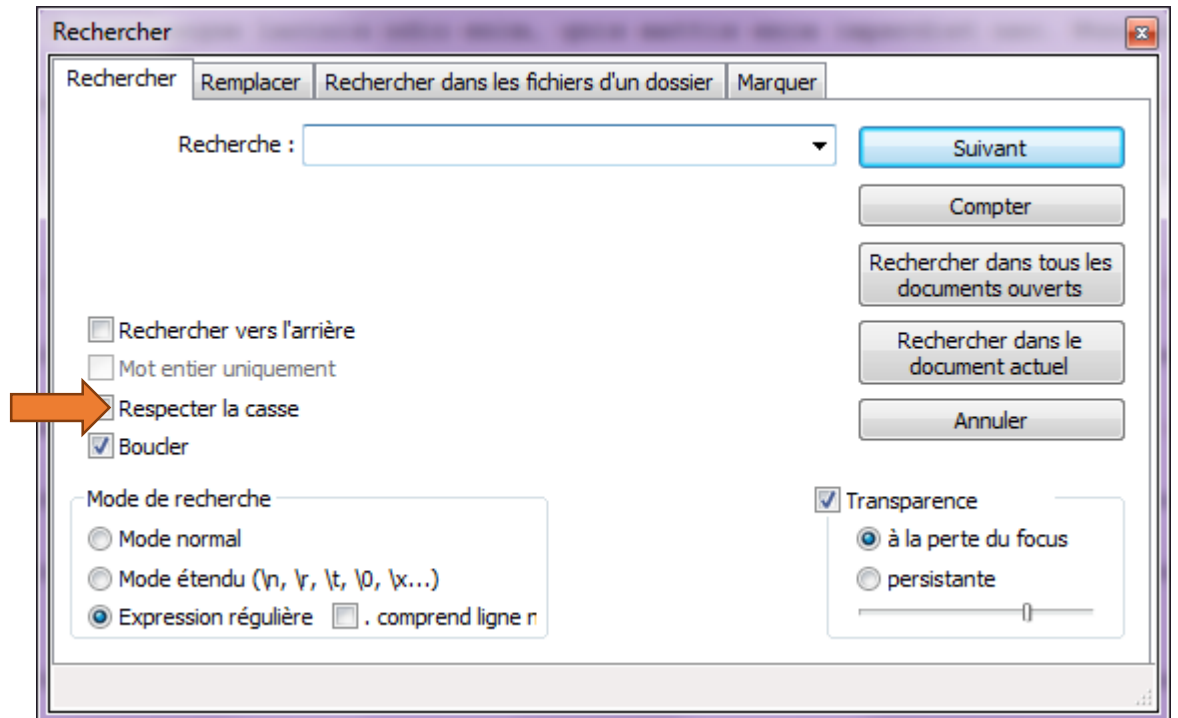
18

- 01/12/2017 => 2017\_12\_01
- Rechercher : ([0-9]{1,2})/([0-9]{1,2})/([0-9]{4})
- Remplacer : \3\_\2\_\1

# Je ou j'

19

- Je, je, J' ou j'
  - Avec respecter la casse  
[Jj][e']
  - Sans respecter la casse  
j[e']



# Transformer un corpus Alceste en Lexico...



et autres déclinaisons

# Alceste => Lexico

- \*\*\*\* \*num\_15 \*loc\_Bernard



- <num=15>

- <loc=Bernard>

4 étoiles qui se suivent

Au moins 1 espace

\*num\_

Une suite de caractères

Au moins 1 espace

\*loc\_

Une suite de caractères

fin de ligne et retour chariot

- Rechercher : `[*]{4}{1,}[*]num_(.*){1,}[*]loc_(.*)\r\n`

- Remplacer : `<num=\1>\r\n<loc=\2>\r\n`

# Et Lexico => Alceste...

22

- Rechercher : `<num=(.*)>\r\n<loc=(.*)>\r\n`
- Remplacer : `**** *num_\1 *loc_\2\r\n`

# Traiter un corpus



Exemple du corpus exemple à modifier (sur Textopol)

- Trouver les > seuls en début de paragraphe

^>

> Sed aliquet, risus nec posuere tempus, nulla felis bibendum ligula, in efficitur velit eros nec magna. Proin imperdiet dolor sed dapibus faucibus. Quisque lectus ligula, consequat efficitur fringilla quis, porttitor sit amet ipsum. Sed eros ipsum, consequat quis tellus at, consequat dignissim justo. Nunc ultrices placerat ipsum, quis tristique nisi ornare gravida. Phasellus volutpat efficitur risus at sollicitudin. Aliquam malesuada nibh enim, sed lacinia magna molestie sit amet.



- Trouver les > en milieu de paragraphe sans < avant (donc pas des fins de balises)

$$\wedge[\wedge<]^*>.*\$$$

varius scelerisque. Curabitur > vestibulum odio eget magna convallis, quis gravida metus bibendum. Curabitur congue

# Trouver des balises

26

- Dans l'extrait ci-dessous sélectionner juste ce qui ne commence pas par <texte ou </texte (en début de paragraphe)

```
<texte num="1">
```

```
Lorem ipsum dolor sit amet,[...]
```

```
<balise >
```

```
[...]
```

```
</balise>
```

```
Ut finibus placerat diam, a interdum[...]
```

```
</texte>
```

$\wedge <(?!(\text{texte}|\text{/texte}))$

- Nettoyer les balise non prévues (donc toutes sauf <texte > ou </texte>)

`<(?!(texte|/texte)).[^\>]*>`

- Balise avec un espace ou plusieurs entre le < et le nom (erreur xml)

<[ ]{1,}texte

# Repérer le 1<sup>er</sup> mot d'une partie

29

- Chercher la balise de début un premier paragraphe suivit d'autant de paragraphes qu'on veut et on s'arrête au début de la balise fermante ?

```
(<petition idpet=".*[>]>\r\n)(.*\r\n)^[^<]*
```

- Remplacer par
- La balise ouvrante et le titre précédé de debut

```
\1debut\2
```

# Webographie



- <http://nppmanuel.nliautaud.fr/expressions-regulieres>
- [http://docs.notepad-plus-plus.org/index.php/Regular\\_Expressions](http://docs.notepad-plus-plus.org/index.php/Regular_Expressions)
- <http://astuces.jeanviet.info/bureautique/rechercher-et-remplacer-du-texte-avec-notepad-et-quelques-regex.htm>