

# Comment chercher vite et bien dans un texte ?

Aller rapidement, souplesse et légèreté.

## Parcourir le texte



Ma palette 5 couleurs - Légende

Bilan ● ● ●

## Recenser les mots clés du texte

Mot clé	Lignes et pages	
biche	1.18 p.3	1.53 p.15 ...
bobo	1.13 p.14	1.28 p.17 ...
:		

→ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

Bilan ● ● ●

## « Indexer » le texte

Exemple : Arbre des suffixes.

Indexation du texte : a m a n a s



→ L'arbre prend 30 fois plus de place que le texte !

Bilan ● ● ●

## Indexer et compresser le texte



Texte d'origine : a b r a c a d u b r a

Par les occurrences

a b r a c a d u b r a

La somme de parties est de la gauche à et se présente d'abord en parties 0

a b r a c a d u b r a

Bilan ● ● ●



Mikaël SALSON

Université de Rouen, Laboratoire d'Informatique de Traitement de l'Information et des Systèmes



# Rechercher dans un texte

## Problématique

Rechercher toutes les *occurrences* d'un mot, groupe de mots, sous-mot, dans un texte.

## Rechercher dans un texte

### Problématique

Rechercher toutes les *occurrences* d'un mot, groupe de mots, sous-mot, dans un texte.

### Avec un ordinateur, pourquoi ?

# Rechercher dans un texte

## Problématique

Rechercher toutes les *occurrences* d'un mot, groupe de mots, sous-mot, dans un texte.

## Avec un ordinateur, pourquoi ?

- ▶ C'est moins fatigant !
- ▶ C'est plus rapide.

# Rechercher dans un texte

## Problématique

Rechercher toutes les *occurrences* d'un mot, groupe de mots, sous-mot, dans un texte.

## Avec un ordinateur, pourquoi ?

- ▶ C'est moins fatigant !
- ▶ C'est plus rapide.

## Quels problèmes ?

# Rechercher dans un texte

## Problématique

Rechercher toutes les *occurrences* d'un mot, groupe de mots, sous-mot, dans un texte.

## Avec un ordinateur, pourquoi ?

- ▶ C'est moins fatigant !
- ▶ C'est plus rapide.

## Quels problèmes ?

- ▶ Énormément de données (l'équivalent de centaines de milliards de pages)
- ▶ On est pressés !

## Rechercher avec rapidité, souplesse et légèreté

### Rapidité

La recherche doit prendre peu de temps

### Souplesse

On doit pouvoir effectuer tout type de recherche

### Légèreté

La recherche ne doit pas utiliser trop de mémoire

## Parcourir le texte

### Parcourir le texte



Licence CC By-SA, square\_eye, Flickr

## Parcourir le texte

### Parcourir le texte



Licence CC By-SA, square\_eye, Flickr

- ▶ Parcourir tout un texte prend du temps (même pour un ordinateur !)

## Parcourir le texte

### Parcourir le texte



Licence CC By-SA, square\_eye, Flickr

Rapidité   Souplesse   Légèreté

Bilan



- Parcourir tout un texte prend du temps (même pour un ordinateur !)

## Recenser les mots clés du texte

### Recenser les mots clés du texte

Mot clé	Lignes et pages		
biche	l.18 p.3	l.53 p.15	...
bobo	l.13 p.14	l.28 p.17	...
⋮			

→ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

## Recenser les mots clés du texte

### Recenser les mots clés du texte

Mot clé	Lignes et pages		
biche	l.18 p.3	l.53 p.15	...
bobo	l.13 p.14	l.28 p.17	...
⋮			

→ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

Manque de souplesse :

- Recherche sur un mot clé (ou plusieurs)

## Recenser les mots clés du texte

### Recenser les mots clés du texte

Mot clé	Lignes et pages		
biche	l.18 p.3	l.53 p.15	...
bobo	l.13 p.14	l.28 p.17	...
⋮			

→ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

Manque de souplesse :

- ▶ Recherche sur un mot clé (ou plusieurs)
- ▶ Pas de recherche de « sous-mot »

## Recenser les mots clés du texte

### Recenser les mots clés du texte

Mot clé	Lignes et pages		
biche	l.18 p.3	l.53 p.15	...
bobo	l.13 p.14	l.28 p.17	...
⋮			

→ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

Rapidité   Souplesse   Légèreté

Bilan



Manque de souplesse :

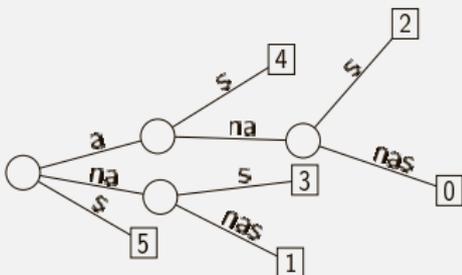
- ▶ Recherche sur un mot clé (ou plusieurs)
- ▶ Pas de recherche de « sous-mot »

## « Indexer » le texte

## « Indexer » le texte

**Exemple** : Arbre des suffixes.

Indexation du texte    a n a n a s  
                                   0 1 2 3 4 5

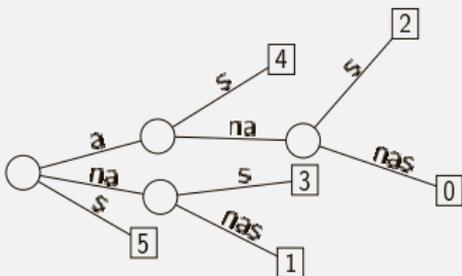


## « Indexer » le texte

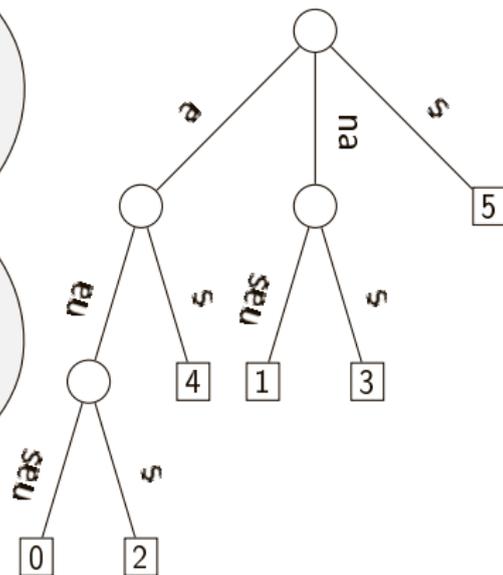
## « Indexer » le texte

**Exemple :** Arbre des suffixes.

Indexation du texte  $\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$



$\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$

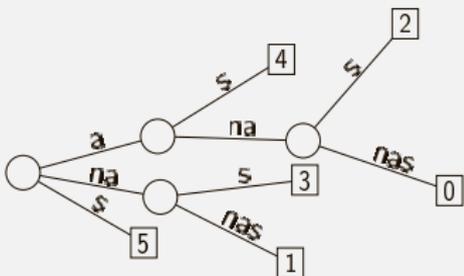


## « Indexer » le texte

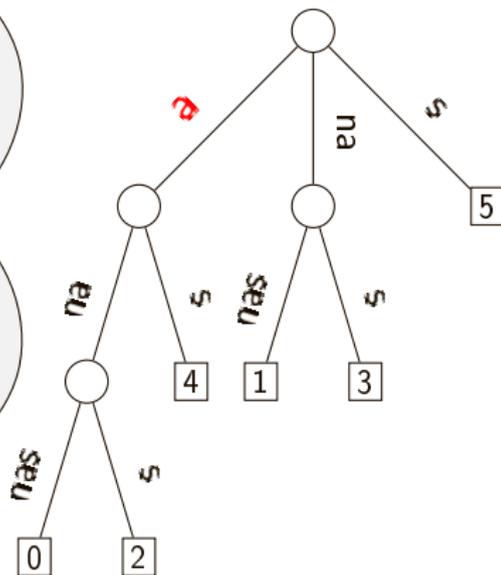
## « Indexer » le texte

**Exemple :** Arbre des suffixes.

Indexation du texte  $\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$



$\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$

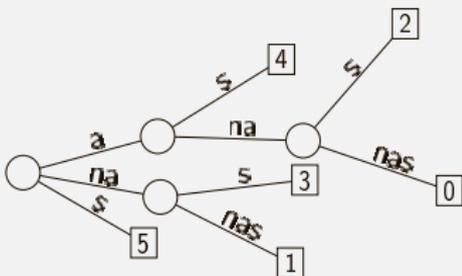


## « Indexer » le texte

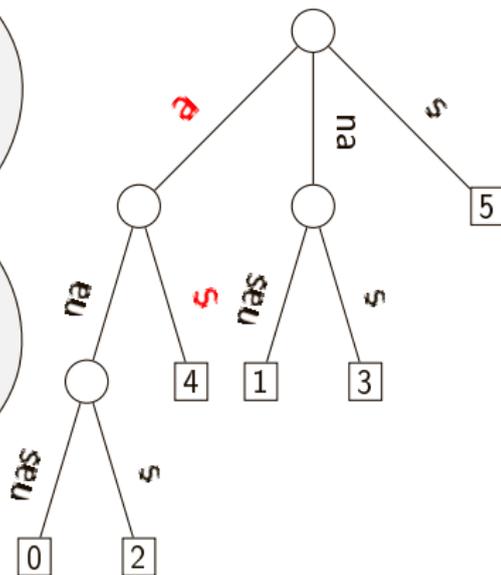
## « Indexer » le texte

**Exemple :** Arbre des suffixes.

Indexation du texte  $\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$



$\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$

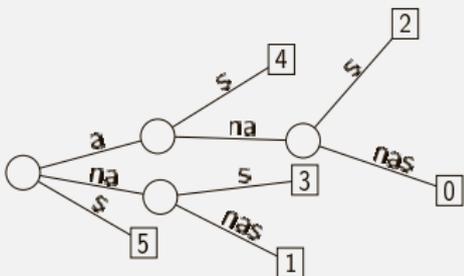


## « Indexer » le texte

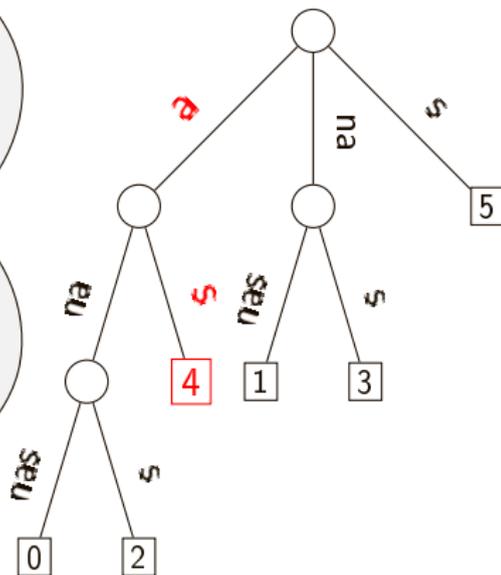
## « Indexer » le texte

**Exemple :** Arbre des suffixes.

Indexation du texte  $\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$



$\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$

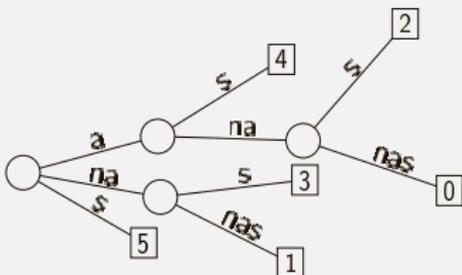


## « Indexer » le texte

## « Indexer » le texte

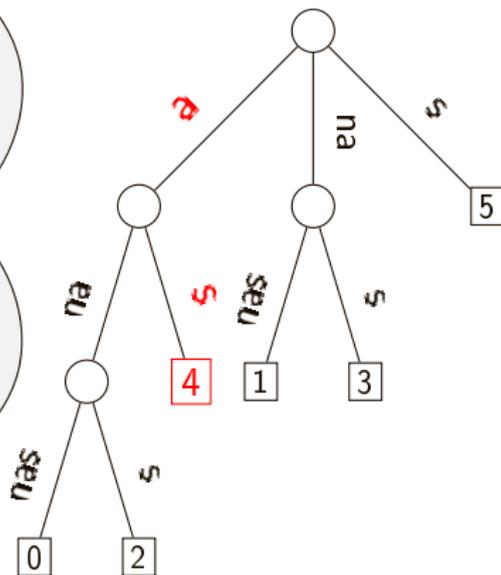
**Exemple :** Arbre des suffixes.

Indexation du texte  $\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$



→ L'arbre prend 10 fois plus de place que le texte !

0 1 2 3 4 5  
a n a n a s

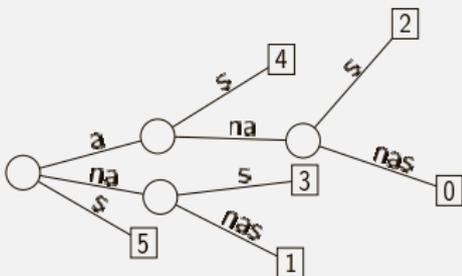


## « Indexer » le texte

## « Indexer » le texte

**Exemple :** Arbre des suffixes.

Indexation du texte  $\begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \\ a & n & a & n & a & s \end{matrix}$



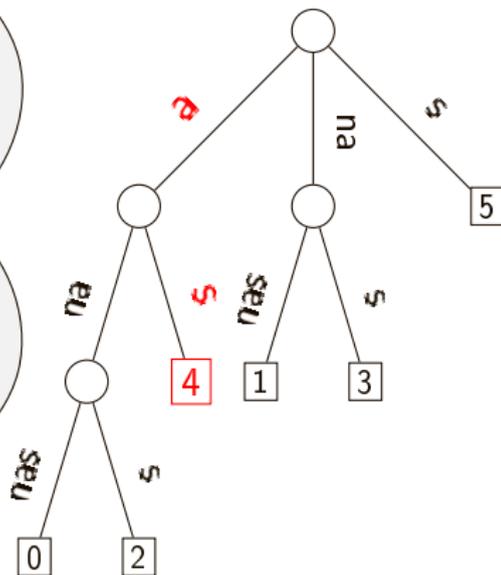
→ L'arbre prend 10 fois plus de place que le texte !

Rapidité    Souplesse    Légèreté

Bilan



0 1 2 3 4 5  
a n a n a s



## Indexer et compresser le texte

**Indexer et compresser le texte**

## Indexer et compresser le texte



**Indexer et compresser le texte**

# Indexer et compresser le texte



## Indexer et compresser le texte

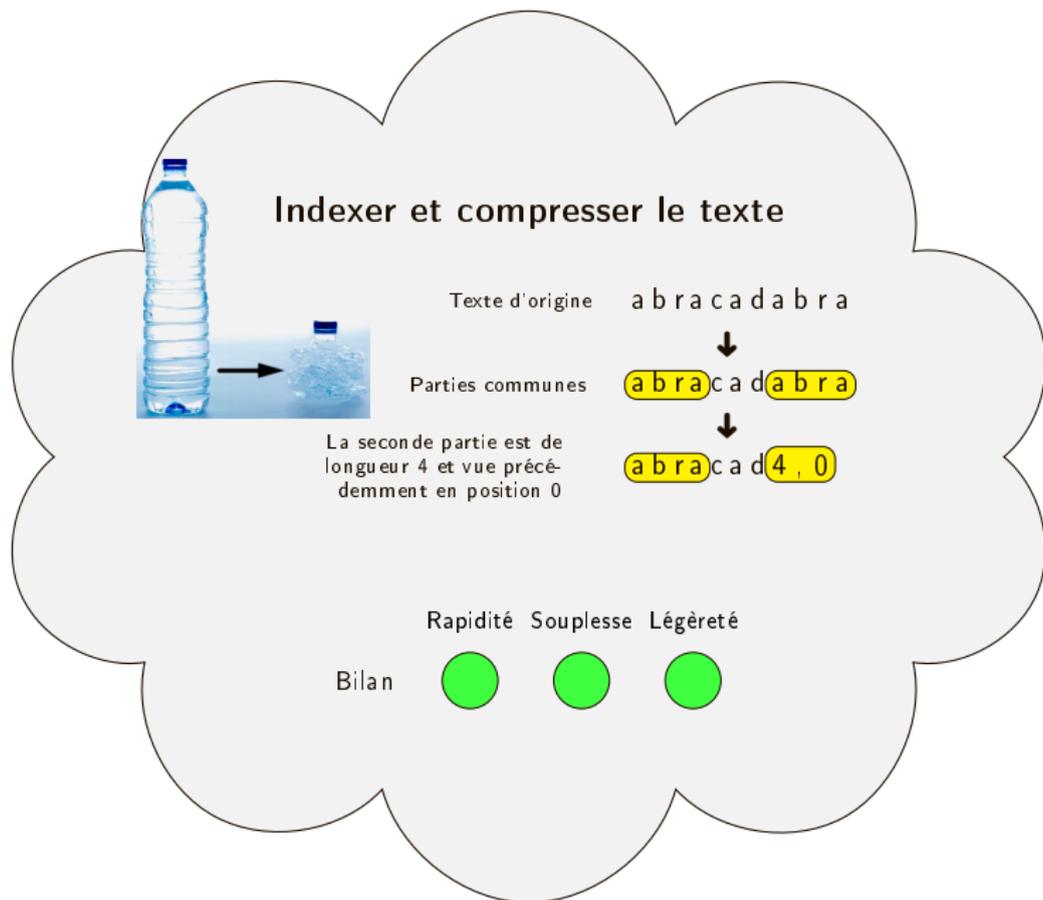
Texte d'origine    a b r a c a d a b r a

Parties communes    abra cad abra

La seconde partie est de longueur 4 et vue précédemment en position 0

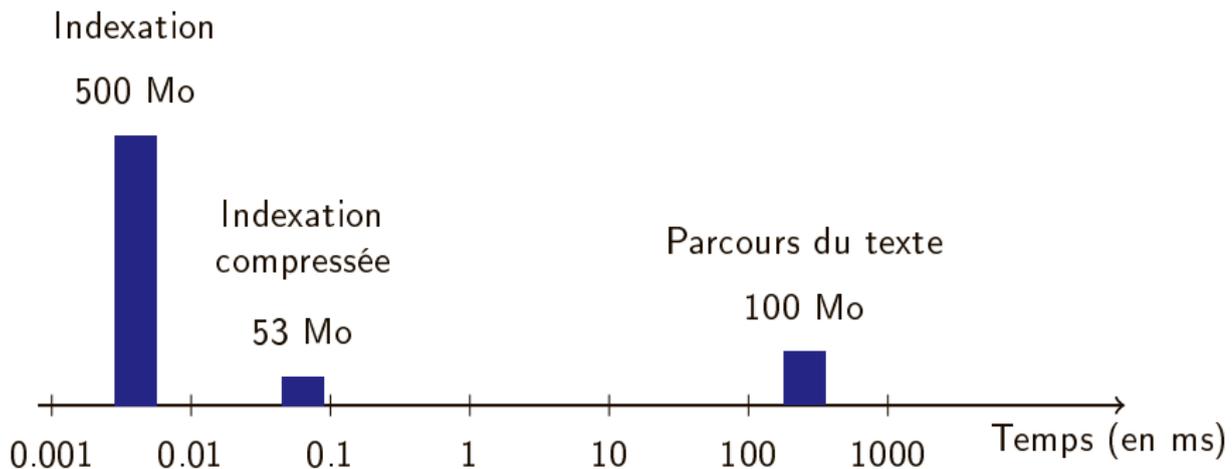
abra cad 4, 0

## Indexer et compresser le texte



# Bilan

Recherche d'un mot de 25 lettres dans un texte de 100 Mo



# Comment chercher vite et bien dans un texte ?

Aller rapidement, souplesse et légèreté.

## Parcourir le texte



Libre de droit, voir page 14

Ma palette Souplesse Légèreté

Bilan ● ● ●

## Recenser les mots clés du texte

Mot clé	Lignes et pages		
biche	1.18 p.3	1.33 p.15	...
bobo	1.13 p.14	1.28 p.17	...
...			

→ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

Bilan ● ● ●

## « Indexer » le texte

Exemple : Arbre des suffixes.  
Indexation du texte a n a n a s



→ L'arbre prend 10 fois plus de place que le texte !

Bilan ● ● ●

## Indexer et compresser le texte



Texte d'exemple : a b r a c a d a b r a

Pour les occurrences

a b r a c a d a b r a

↓

a b r a c a d a b r a

↓

a b r a c a d a b r a

La norme de partie sur de la pointer et si ça grimpe - descendre en position 0

Bilan ● ● ●



Mikaël SALSON

Université de Rouen, Laboratoire d'Informatique de Traitement de l'Information et des Systèmes

