Comment chercher vite et bien dans un texte ?

Allier rapidité, souplesse et légèreté.

Parcourir le texte

Licence CC By-SA, square_eye, Flickr

	Rapidité	Souplesse	Légèreté
Bilan	●	●	●

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 ...
⋮	

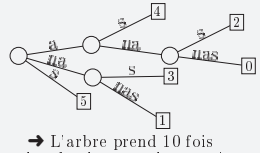
➜ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

	R	S	L
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 !

	R	S	L
Bilan	●	●	●

Indexer et compresser le texte

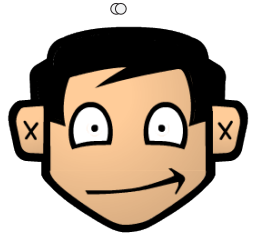
Texte d'origine : a b r a c a d a b r a

Parties communes : a b r a c a d a b r a

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

a b r a c a d 4 , 0

	R	S	L
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 :

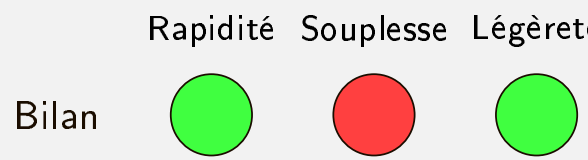
- Recherche sur un mot clé (ou plusieurs)
- Pas de recherche de « sous-mot »

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... » !



Manque de souplesse :

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

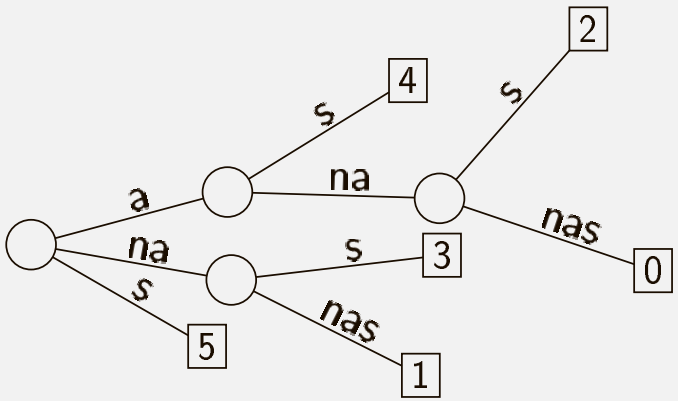
« Indexer » le texte

« Indexer » le texte

Exemple : Arbre des suffixes.

Indexation du texte

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



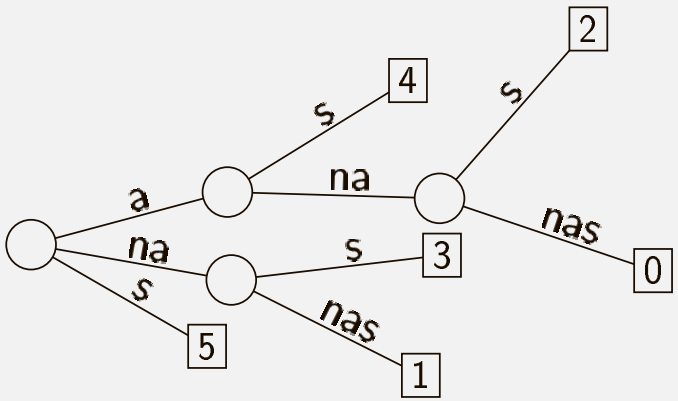
« Indexer » le texte

« Indexer » le texte

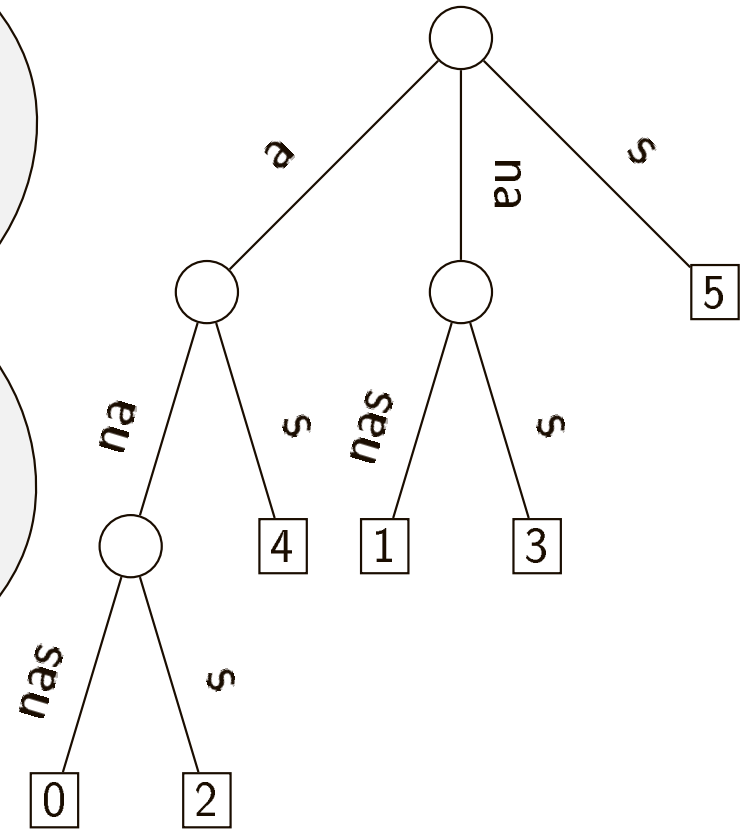
Exemple : Arbre des suffixes.

Indexation du texte

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



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

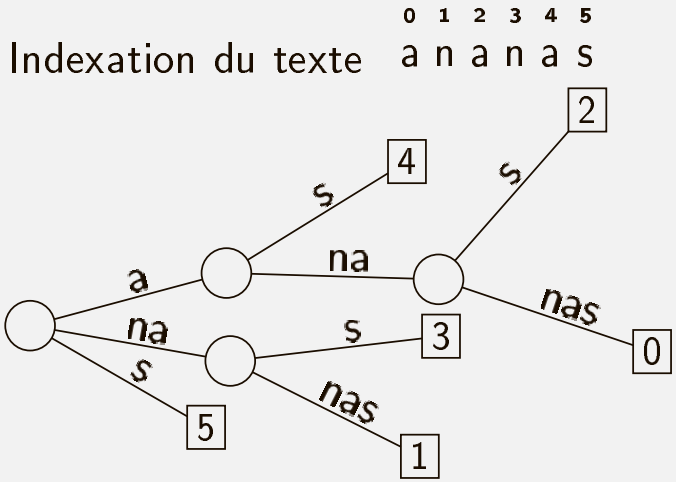


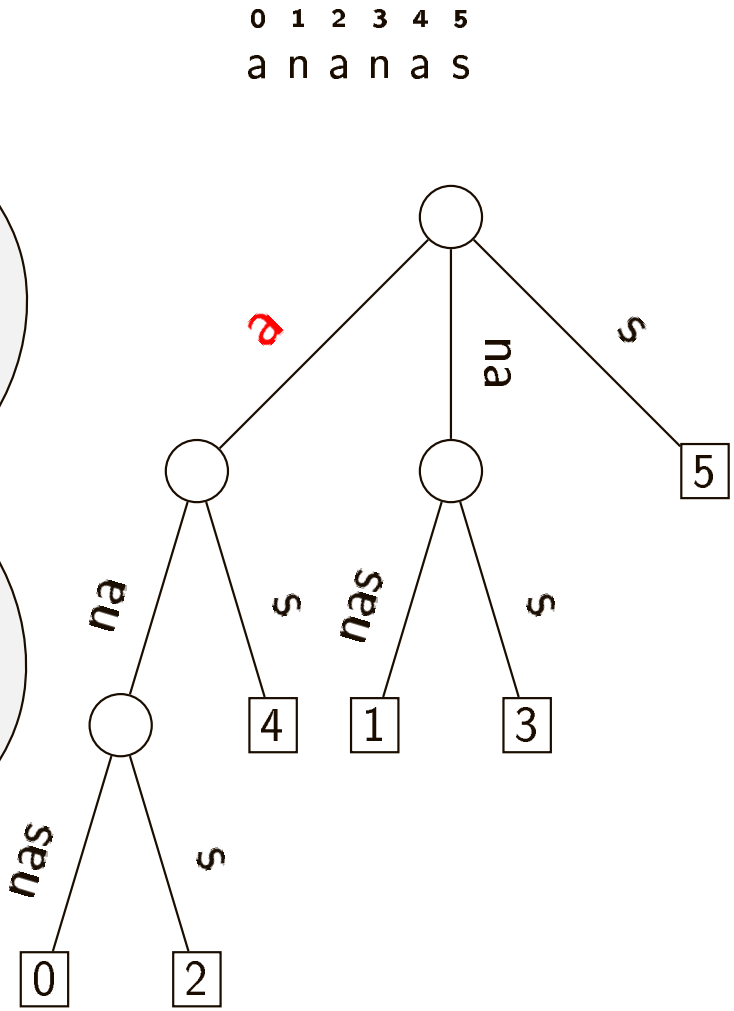
« Indexer » le texte

« Indexer » le texte

Exemple : Arbre des suffixes.

Indexation du texte





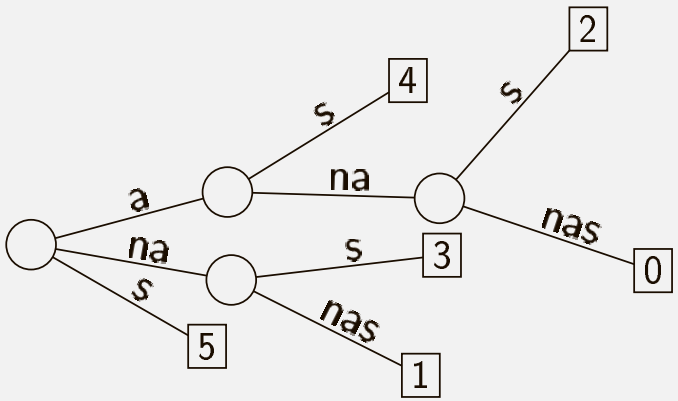
« Indexer » le texte

« Indexer » le texte

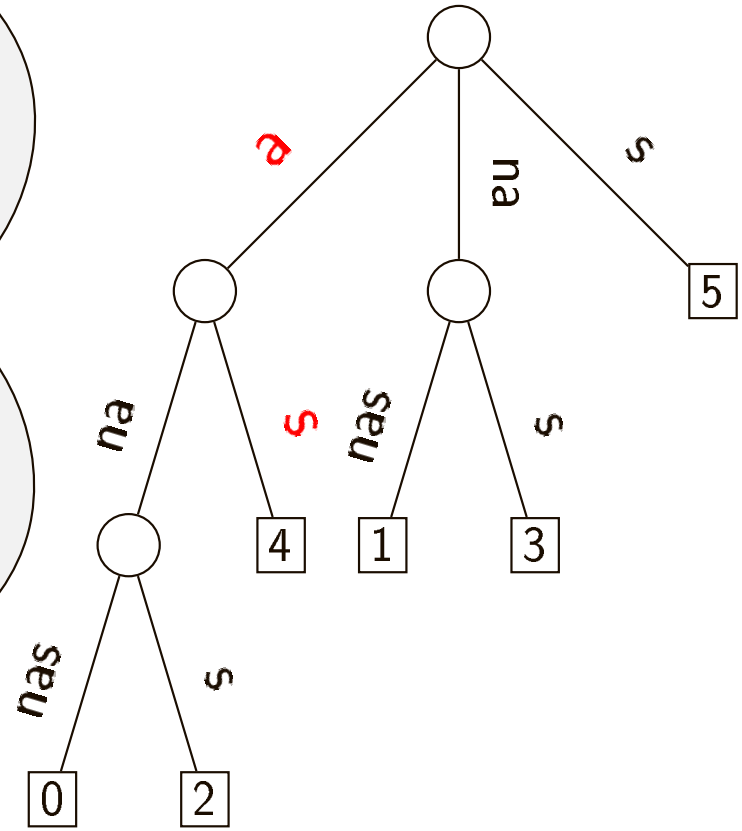
Exemple : Arbre des suffixes.

Indexation du texte

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



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

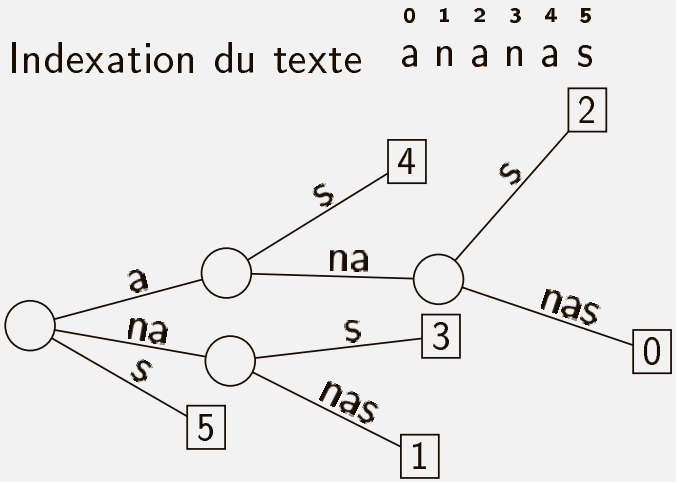


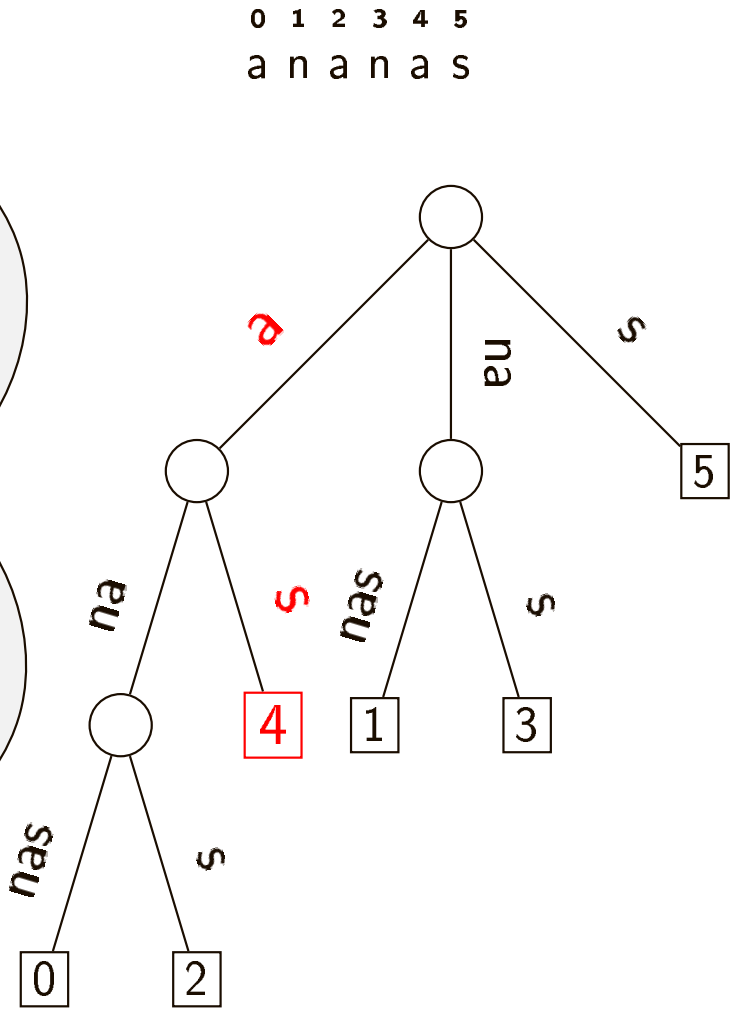
« Indexer » le texte

« Indexer » le texte

Exemple : Arbre des suffixes.

Indexation du texte





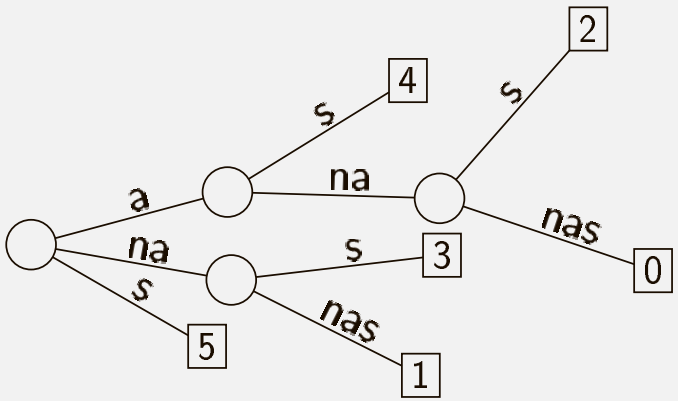
« Indexer » le texte

« Indexer » le texte

Exemple : Arbre des suffixes.

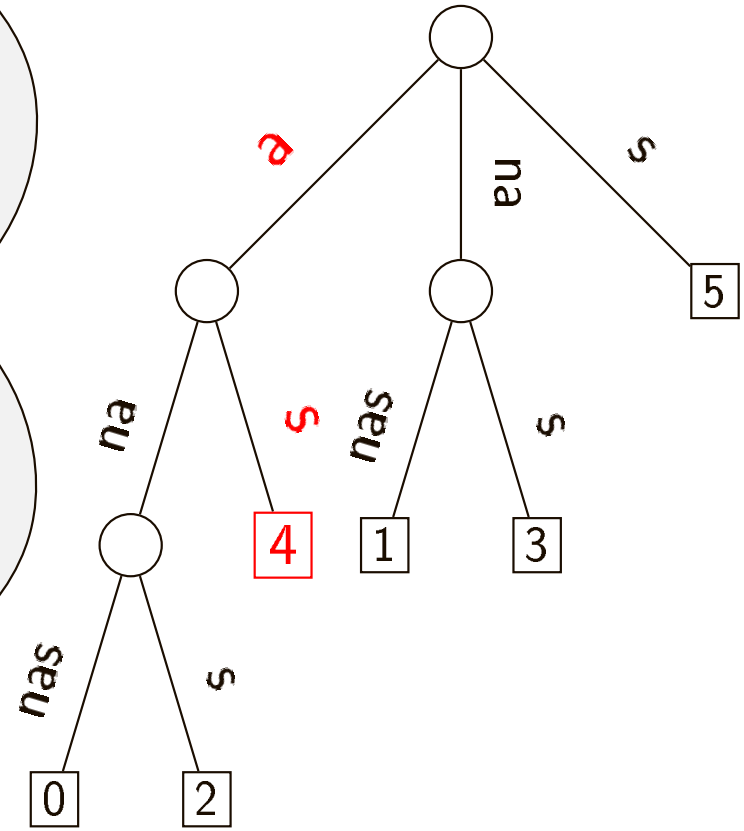
Indexation du texte

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



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

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



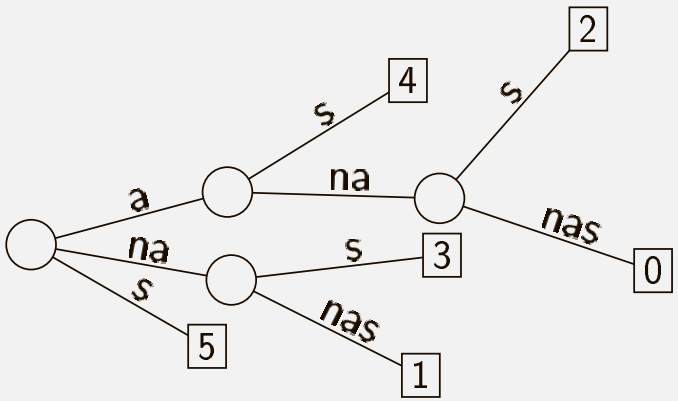
« Indexer » le texte

« Indexer » le texte

Exemple : Arbre des suffixes.

Indexation du texte

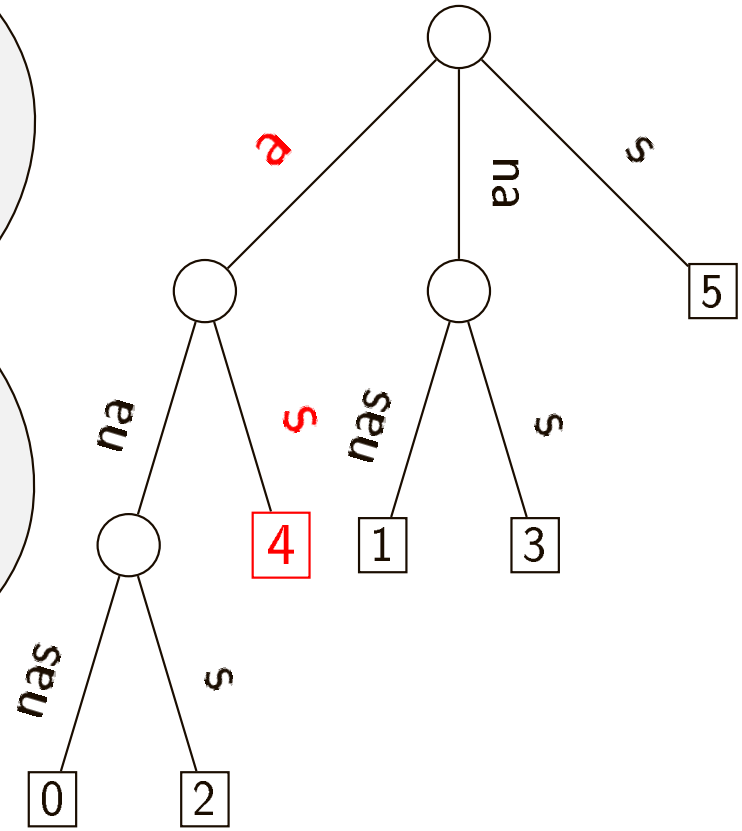
0	1	2	3	4	5
a	n	a	n	a	s



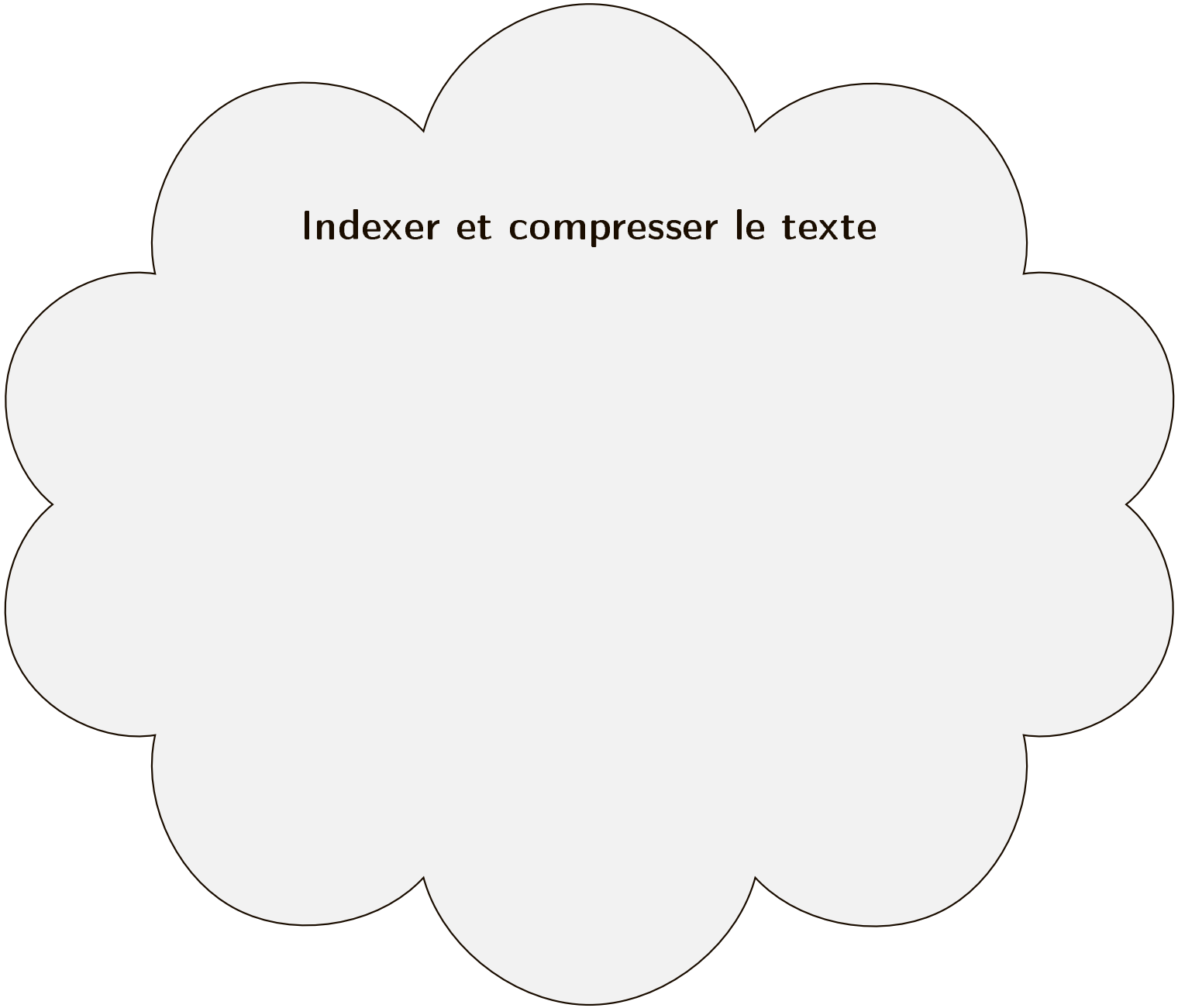
➜ L'arbre prend 10 fois plus de place que le texte !

	Rapidité	Souplesse	Légèreté
Bilan			

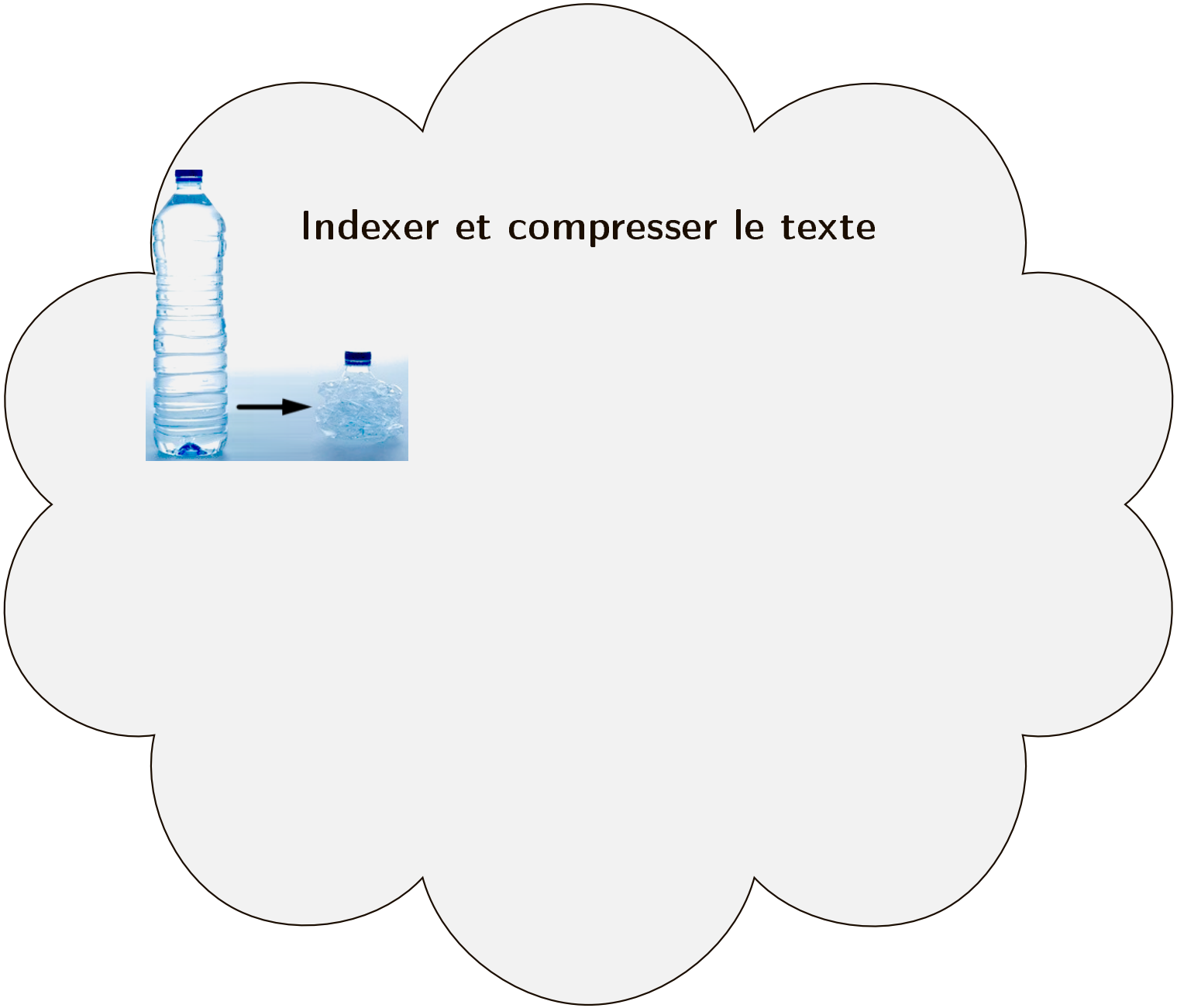
0	1	2	3	4	5
a	n	a	n	a	s



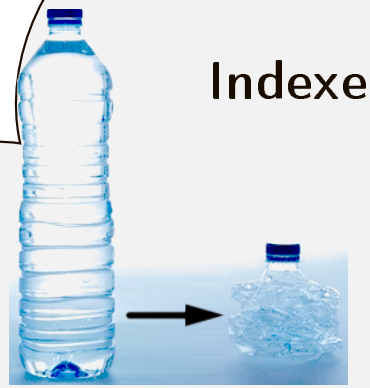
Indexer et compresser le texte



Indexer et compresser le texte



Indexer et compresser le texte



Indexer et compresser le texte

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

↓

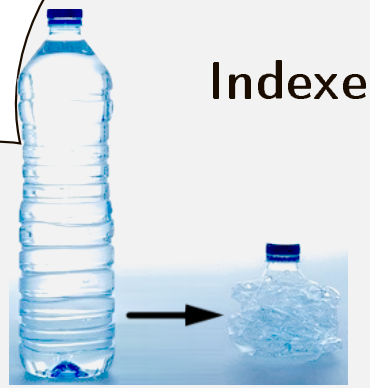
Parties communes a b r a c a d a b r a

↓

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

a b r a c a d 4 , 0

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 a b r a c a d a b r a

↓

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

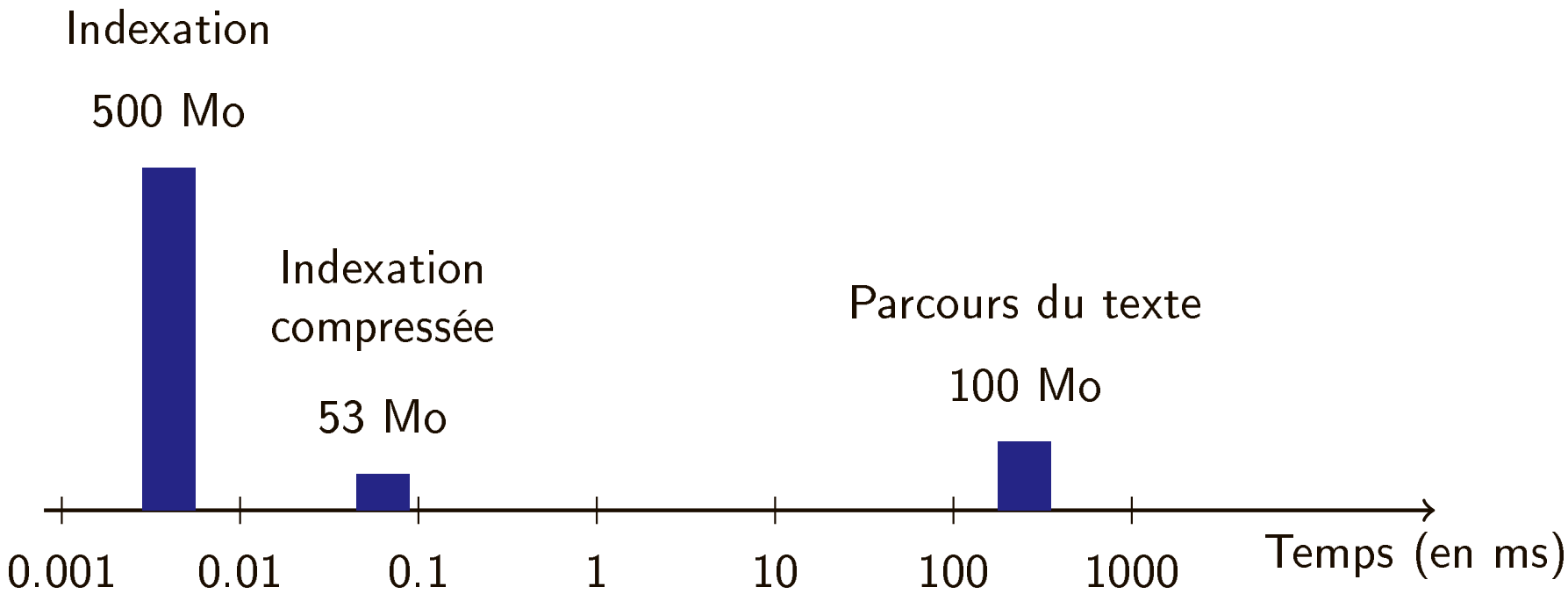
a b r a c a d 4 , 0

Rapidité Souplesse Légèreté

Bilan

Bilan

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



Comment chercher vite et bien dans un texte ?

Allier rapidité, souplesse et légèreté.

Parcourir le texte

Licence CC By-SA, squars_eye, Flickr

	Rapidité	Souplesse	Légèreté
Bilan			

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 ...
⋮	

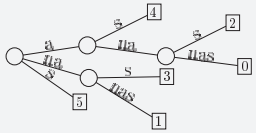
➜ Difficile de rechercher des morceaux de mots, p. ex. : « ...che bo... » !

	R	S	L
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 !

	R	S	L
Bilan			

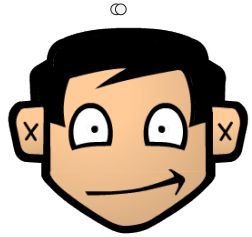
Indexer et compresser le texte

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

Parties communes : a b r a c a d a b r a

La seconde partie est de longueur 4 et vue précédemment en position 0 : a b r a c a d 4 , 0

	R	S	L
Bilan			



Mikaël Salson

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