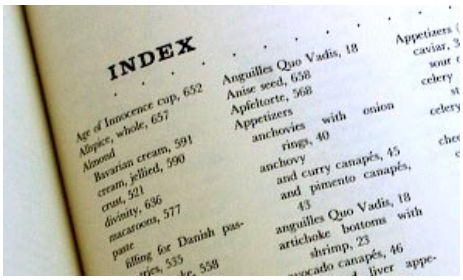
Les moteurs de recherche du Web
Une activité pour primaire, collège, lycée

Comment les moteurs de recherche gèrent-ils les volumes d'information phénoménaux du Web ? Comment aident-ils les utilisateurs à trouver ce qu'ils cherchent dans cette masse ? Serge Abiteboul, auteur de l'article « chercher sur le Web, juste un point fixe et quelques algorithmes » vous propose une activité pour faire sentir à vos élèves la façon dont fonctionne un moteur de recherche du Web.

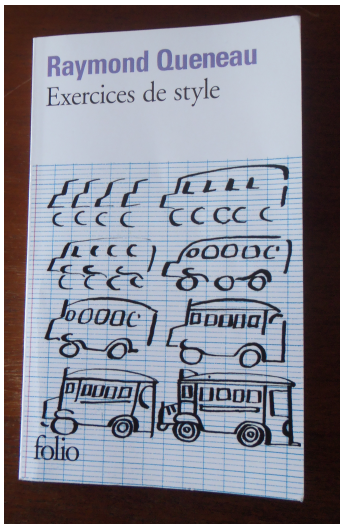
A. Indexation

Le Web c'est d'abord un nombre impressionnant de pages de texte. Nous pouvons voir un moteur de recherche comme un index de ces pages, un peu comme l'index d'un très grand livre. Regardons l'index d'un livre pour nous rappeler comment cela fonctionne.



Pour le Web, le travail est énorme parce qu'il y a des milliards de pages et des millions de questions posées chaque jour à l'index. Pour imprimer un tel index, il faudrait plus de pages de papier que n'en stocke la Bibliothèque Nationale de France. Alors nous allons partager le travail en un très grand nombre de machines. Comment ?

Pour jouer le rôle des pages web dans l'activité proposée ici, nous vous suggérons d'utiliser des pages extraites des Exercices de Style de Raymond Queneau, qui se révèlent particulièrement bien adaptés. En couplant des textes très courts et d'autres un peu moins courts, on répartit le travail également entre les différents ordinateurs.



Vous pouvez par exemple utiliser la répartition ci-dessous:
Ordinateur 1 : 1. Notation 2. Rétrograde
Ordinateur 2 : 3. Récit 4. Tanka
Ordinateur 3 : 5. Surprise 6 : Présent
Ordinateur 4 : 7. Passé simple 8. Lipogramme
Ordinateur 5 : 9 : Imparfait 10 : Alors
Pour toutes précisions sur les *Exercices de Style* ou d'autres suggestions de textes, rendez-vous sur la page :
http://www.math.u-psud.fr/~pansu/explosion_continue_en_classe.html

1) Partageons la classe en cinq groupes que nous appellerons Ordinateur1 à Ordinateur 5. Chaque groupe est responsable de réaliser un index pour deux textes. Pour nous faciliter la tâche, nous n'indexerons que les noms communs. Nous pourrons ignorer les lettres majuscules, le pluriel.
Ainsi l'index d'un groupe qui aura rencontré le nom *autobus* dans les textes 5 et 6 et le nom *chapeau* dans le texte 5 seulement aura un index qui ressemblera à :
autobus 5, 6
chapeau 5
…

2) Confions tous les index à un élève et nous lui demandons de trouver tous les textes contenant (par exemple) le mot *autobus,* puis ceux contenant le mot *chapeau,* puis ceux contenant [*autobus* et *chapeau*]. Il les écrira au tableau.

3) Cherchons les mots *cou*, *bouton* et [*cou* et *bouton*] mais en laissant à chaque groupe la responsabilité de ses pages. Un élève sera responsable d'écrire tous les résultats au tableau. Observer que nous obtenons les réponses beaucoup plus vite.

Pour réaliser un tel index un moteur de recherche utilise de nombreux centres de calculs. Et chaque centre utilise des milliers de machines. Le travail à réaliser est donc partagé entre des dizaines de milliers de machines.

B. Classement des pages

Un moteur de recherche donne les réponses à des questions comme *autobus* et [*autobus* et *chapeau*] en proposant d'abord les textes « les plus populaires ». Pour simplifier, nous allons considérer que les textes les plus populaires sont ceux qui sont le plus souvent choisis par les élèves.

1) Les élèves votent à main levée pour leurs trois textes préférés afin de classer les 10 textes suivant leurs préférences. Chaque groupe introduira cette information dans son index. Par exemple, si 3 élèves ont voté pour le texte 5, un seul pour le texte 6, nous pourrons avoir la ligne d'index :
   autobus (5,3),(6,1)

2) Pour les questions déjà posées (comme *autobus* ou [*autobus* et *chapeau*]*...)*, l'élève chargé d'inscrire les réponses (les numéros des textes qui contiennent ce mot) au tableau devra les reclasser par ordre de popularité.

Les algorithmes de classement des moteurs de recherches utilisent de nombreux critères pour classer les résultats, notamment un critère de popularité. Ces critères sont secrets.

C. Le sens des mots

Supposons que nous cherchons les documents parlant de *chapeau*. Nous aimerions aussi avoir comme résultats les documents parlant de *couvre-chef*, ou *feutre mou* car nous savons que les feutres mous sont une variété de chapeau. Ainsi on tient compte, dans l'indexation, des synonymes (*chapeau* et *couvre-chef*) et des généralisations (*chapeau* et *feutre mou*). Chaque groupe peut ainsi chercher dans son index les noms qu'il pourrait faire figurer avec *citoyen* qui figure dans le texte n°10. En considérant l'index du texte n°8, qui contient des mots peu fréquents, donner d'autres exemples de regroupements que l'on peut effectuer pour diminuer ainsi la taille de l'index.

Le Web est aussi multi-langues. Si nous posons la question « Paris » et si nous parlons italien, peut-être sommes nous aussi intéressés par les pages contenant le mot « Parigi ».

D. Questionnements

Maintenant que nous comprenons les bases de ces systèmes, nous pouvons nous interroger sur d'autres aspects passionnants comme :

- Comment fait un moteur de recherche pour trouver des milliards de pages du Web ?
- Quand vous avez cinq à six mille ordinateurs qui travaillent pour vous, il est certain que plusieurs tomberont en panne chaque jour. Comment un moteur de recherche fait-il pour ne jamais s'arrêter ?
- Toutes ces machines dégagent de la chaleur, comment fonctionnent les systèmes de climatisation super sophistiqués et peu coûteux des moteurs de recherche.
- Pour fabriquer un index, il faut lire toutes les pages qu'il index sur le Web. Le moteur de recherche n'a pas fini de les lire que certaines pages ont déjà changé et qu'il lui faudrait les relire. Et découvrir de nouvelles pages… Comment se tenir au courant des « nouveautés » du Web ?
- On fait des fautes d'orthographe. Comment un moteur de recherche peut-il répondre aux questions quand les textes et les questions contiennent des fautes ?