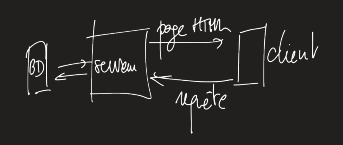


Graphes

Début : annuaire de sites
moteurs de recherche v1 : documents contenant les mots cherchés
modernes : + pagerank + compréhension de la réputation + mise à jour en fonction du comportement

Internet : pages, sites, liens

quelques milliards
dépend du comptage
cf. vente en ligne
page créée à la demande
quelques centaines de millions "achats"



Données : $\approx 10^{15}$ octets/an.

Faire 1 séance sur consommation énergétique

+ 1 séance privacy

A faire :
- page modèle
- ~~envoyer~~ tester galine
- envoyer une galine à Carole
- lien dump wiki + mettre à jour sujets TP

Dans ce cours :
* principes moteurs rech.
* sys. de reco.
* climat + vie privée
50% exam + 50% projet → mot. rech. sur wiki
→ expliqué demain

- Principes de base
- Prétraitement
1. Crawler
 2. Indexation
 3. Traitement de la requête
 4. Calcul du résultat

1. CRAWLER Parcours du graphe du web.

- * Parcourir les pages → suivre les liens
- * Extraire les infos sur chaque page :
 - métadonnées
 - contenu
 - liens
- * Repérer si spam
- * calculer fréquence, etc.
- * si lien mort : on l'ignore ou on renvoie 404
- * montrer p. code source page + curl

algo + fast

Déroulé séance :
excl → montrer exemples recherches sur Quora

· fichier robots.txt
· fréquence des visites en fonction de la freq. de maj
· BFS vs DFS
→ ou IDDFS.
↓
rappeler algo file.