Graphe

Début : . annuaires de sites — . sélection et indexation à la main
. moteurs de recherche ~~basés~~ v1 : documents contenant les mots cherchés
. ——— modernes : — + pagerank + compréhension de la requête + mise à jour en fonction du comportement

Internet : pages, sites, liens
quelques milliards — quelques centaines de millions "achetés"
dépend du comptage
cf. vente en ligne
page créée à la demande


Données : $\simeq 10^{13}$ octets / an

Faire 1 séance sur communautés énergétique

+ 1 séance privacy

À faire : page moodle
~~moodle~~ tester galène
. envoyer lien galène à Carole
. lien dump ~~du~~ wiki + mettre à jour sujets TP

Dans ce cours : * principes moteurs rech.
* ——— sys. de reco.
* climat + vie privée
50 % exam + 50 % projet → mini moteur 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
* Parser les pages * suivre les liens
* Extraire les infos sur chaque page : métadonnées, contenu, liens
* Repérer le spam
algo + lourd
* calculer fréquence, etc.
* si lien mort : on l'ignore ou on réessaye + tard

montrer le code source page + curl

- fichier robots.txt
. fréquence des visites en fonction de la freq. de maj
BFS vs DFS → ~~DFS~~ ou IDDFS.
appeler algo file

Déroulé séance ordi → montrer exemples recherches sur Quant