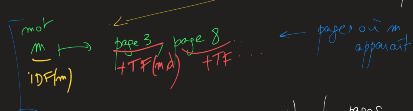


# Aperçu du projet "entru"

TP1 collecteur → f. dictionnaire  
 .index inverse (relation mots-pages)



graphes des pages sous forme CL1

1<sup>er</sup> score des pages selon la fréquence des mots :

\* pour chaque mot  $m$ ,  $IDF(m)$ .

\* pour chaque page  $d$ , chaque mot  $m$

pour la relation mots-pages  
 $TF(m, d)$  normalisée  $(TF/N_d)$

$$= \sum_{m \in d} TF(m, d)^2$$

TP2 (pagerank) :

calcul un score pour

chaque page  $d$

indépendant des mots

contenus dans  $d$

(on regarde le graphe)

→  $pr(d)$

page

trouper les listes

## TP3 traitement de la requête

$$req = m_1 m_2 \dots m_k$$

1. Trouver les pages contenant tous les mots →  $P$

2. Calculer le vecteur normalisé de la requête

(dim.  $k$ , basé  $(IDF(m_1) \dots IDF(m_k))$ )

3. pour chaque page  $d \in P$ , on calcule le

score de la page pour la requête :

produit scalaire du vecteur de la page

(restreint à  $m_1 m_k$ )

$TF(m_1, d) \dots TF(m_k, d)$

avec le vecteur de la requête

Score  $s_1$

4. Pour chaque page  $d \in P$ , calculer  $S_d = \alpha s_1 + \beta pr(d)$

score de  $d$  pour la requête

si  $IDF(m_2) \times TF(m_2, d) \approx 10^{-6}$  (?)  
 très faible  
 → contribue peu au score  
 → on peut supprimer la page  $d$  de la liste de  $m_2$

On embale tout ça dans un site web.

