Recherche documentaire

Présenté par : Noé De Caestecker et Killian Darras

Stemmer de Porter

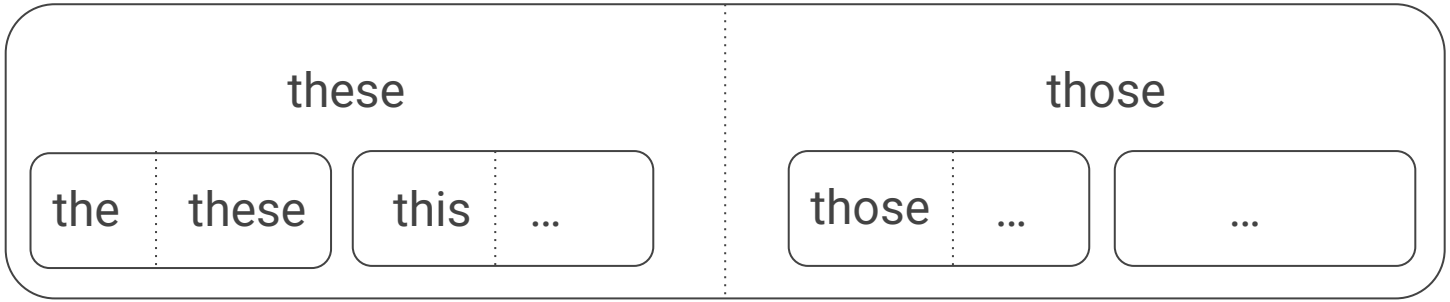
- Lecture du corpus de documents

- Structure sous forme de dictionnaire

  ex : {logical=logic, logically=logic, logics=logic, etc …}

Structure de l’index

- arbre 2-4 :



- rotation des mots : abc -> abc, cab, bc$a

Structure de l’index

- pointeurs pour les rotations

- création des vecteurs
- normalisation (avec les normes stockées)

mot / idf	**query**	**doc1**	**doc2**
new / 0.5	0.5	0.5	
york / 0.75	0.75	0.9	
is / 0.2	0.2		0.4
good / 0.6	0.6		0.6

Correcteur d’orthographe

- Soundex avec le dictionnaire anglais

  ex : {z352 = [zootomical,zootomically,zootomist]}

- Utilisation des Trigrammes

  ex : unfortunately = {un,nf,fo,or,rt,tu,un,na,at,te,el,ly}

- Mesure de Jaccard