Techniques Informatiques & Web

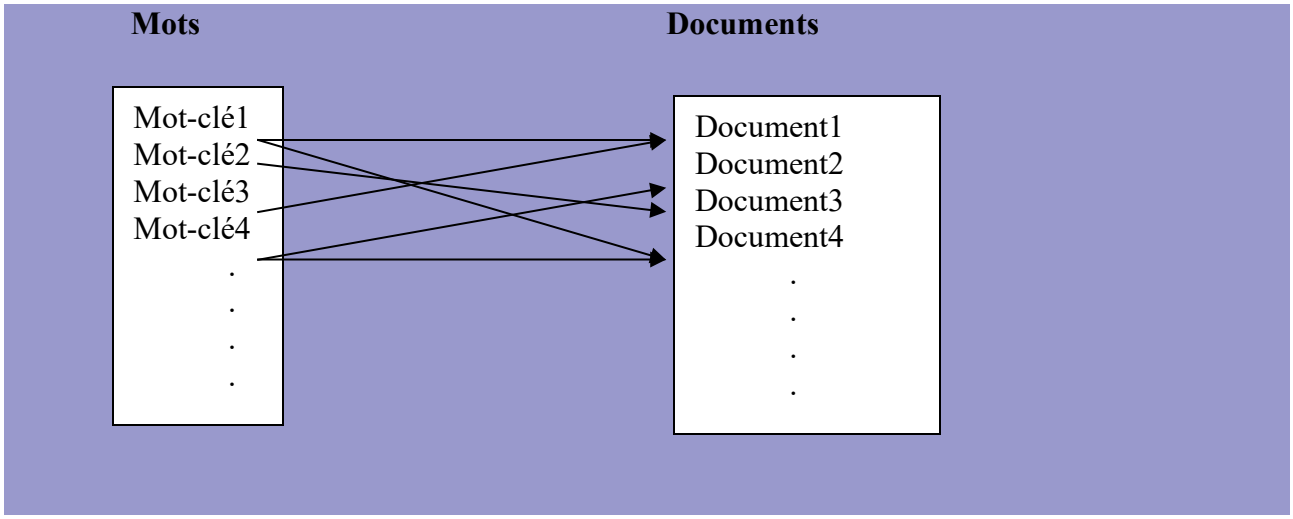
Nasreddine Bouhai
nasreddine.bouhai@univ-paris8.fr

Introduction à l'indexation de documents textuels (non structurés)

1. Définitions et principes

L'indexation est la technique qui consiste à caractériser le contenu d'un document par un certain nombre de mots-clés ou termes convenus. Il s'agit d'établir une relation :

Où chaque mot-clé est associé au moins à un document (ou liste de documents).



2. Segmenter un texte en mot : la tokenisation

Dans le processus du traitement automatique de la langue, la notion de mot se trouve très avant. La première difficulté est donc de rendre la machine capable de reconnaître et de représenter les mots. Ce traitement se fait souvent très en amont car il est à la base des différents processus comme : analyse du style d'un texte, catégorisation des mots pour analyse syntaxique, etc. Cette opération de découpage d'un texte en mot s'appelle la **tokenisation**.