Exploration de documents par navigation conceptuelle

Caroline Barrière, chercheure, CRIM
Jean-François Lavallée, agent de recherche, CRIM

41e congrès de l'Association des archivistes du Québec
Le 31 mai 2012

Partenaire financier :
Développement économique, Innovation et Exportation Québec

Plan

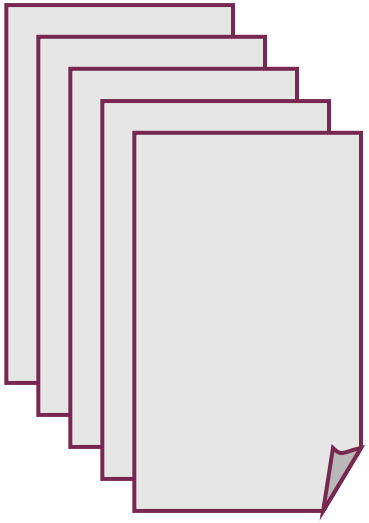
1. Devant une collection de documents, par où débuter?

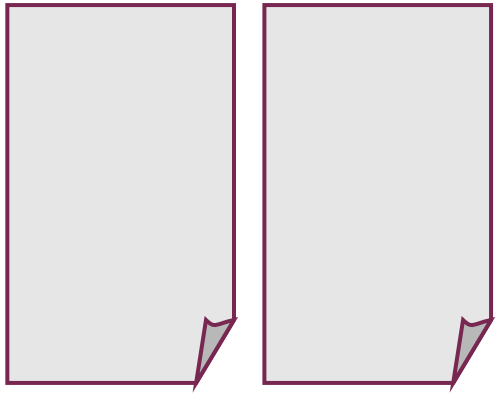
2. Génération automatique d'un pseudo-thésaurus
    1. Extraction de termes
    2. Calcul de similarité distributionnelle

3. Exploration d'une collection de documents à l'aide du pseudo-thésaurus
    1. Visualisation/navigation
    2. Accès à des listes de mots-clés des documents

1. Devant une collection de documents, par où débuter?

Collection de documents





Humain qui regarde un document à la fois pour annoter, ajouter des métadonnées.

Et si on n'a pas ce schéma???

Collection de documents

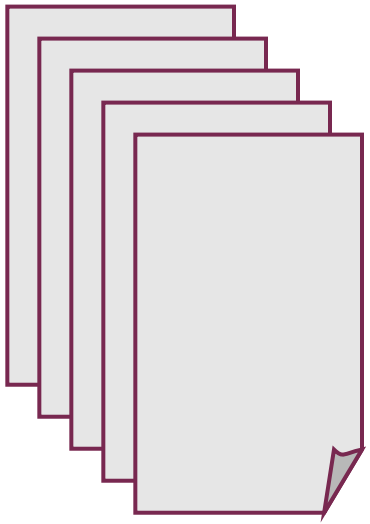
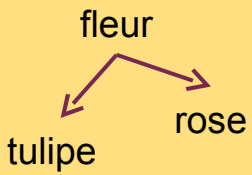


Schéma de classification externe (thésaurus)



Création automatique de l'index (inverted index) par repérage des termes :

Terme	Documents
rose	doc1, doc8, doc25
tulipe	doc7, doc28

Des difficultés se posent sur les variantes des formes (e.g. pluriels, déclinaisons différentes).

Plan

1. Devant une collection de documents, par où débuter?

2. Génération automatique d'un pseudo-thésaurus
    1. Extraction de termes
    2. Calcul de similarité distributionnelle

3. Exploration d'une collection de documents à l'aide du pseudo-thésaurus
    1. Visualisation/navigation
    2. Accès à des listes de mots-clés des documents

Collection de documents

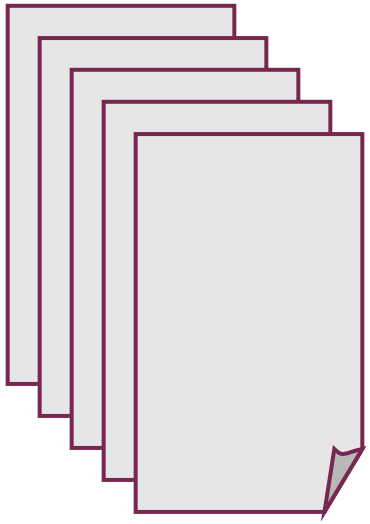
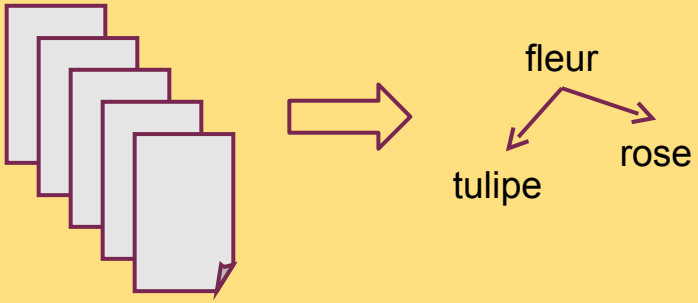


Schéma de classification est inexistant…
Nous pouvons le créer automatiquement.



Création automatique de l'index (inverted index):

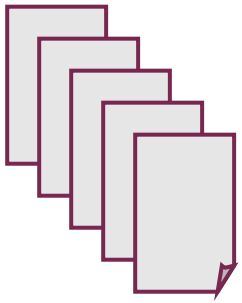
Terme	Documents
rose	doc1, doc8, doc25
tulipe	doc7, doc28

Construction automatique d'un « pseudo-thésaurus »

1.Obtenir un corpus de documents sur un domaine.
2.Extraire les termes pertinents de ce corpus pour la création d'unités d'indexation.
3.Calcul de similarité distributionnelle entre termes – leur tendance à apparaître ensemble dans les textes.

« we know a word by the company it keeps »

1-2. Extraire les termes du domaine à partir d'un corpus ciblé



Corpus sur le jardinage

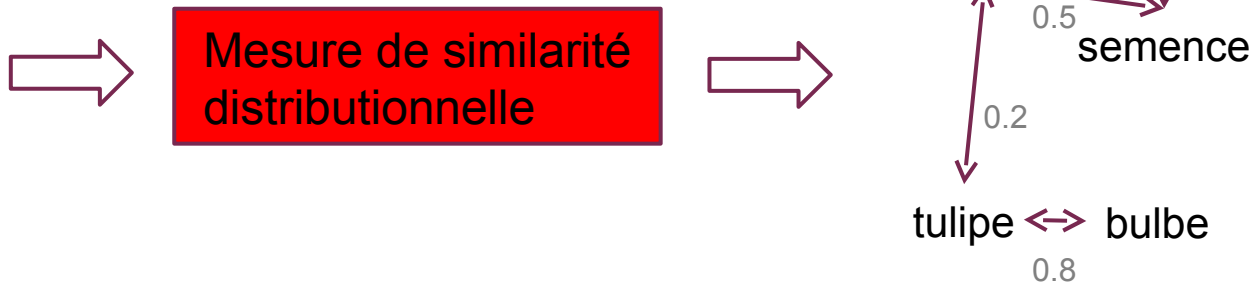
Extraction de termes

échinacée
semence
engrais
tulipe
…

Méthode :
1.Calcul de fréquences comparées : l'état de l'art en extraction de termes utilise des méthodes comparatives.
2.Calcul de cohérence des termes complexes : des modèles statistiques qui évaluent les probabilités de retrouver les termes seuls ou dans leur forme complexe.

3. Calcul de similarité distributionnelle

échinacée
semence
engrais
tulipe
terre
bulbe



Module de calcul de similarité

1.Recherche de co-occurrences des termes dans des fenêtres précises (fenêtre = paragraphe, phrase, document).
2.Utilisation d'une mesure de co-occurrence (Jacquard, Dice, Information Mutuelle).

Plan

1. Devant une collection de documents, par où débuter?

2. Génération automatique d'un pseudo-thésaurus
    1. Extraction de termes
    2. Calcul de similarité distributionnelle

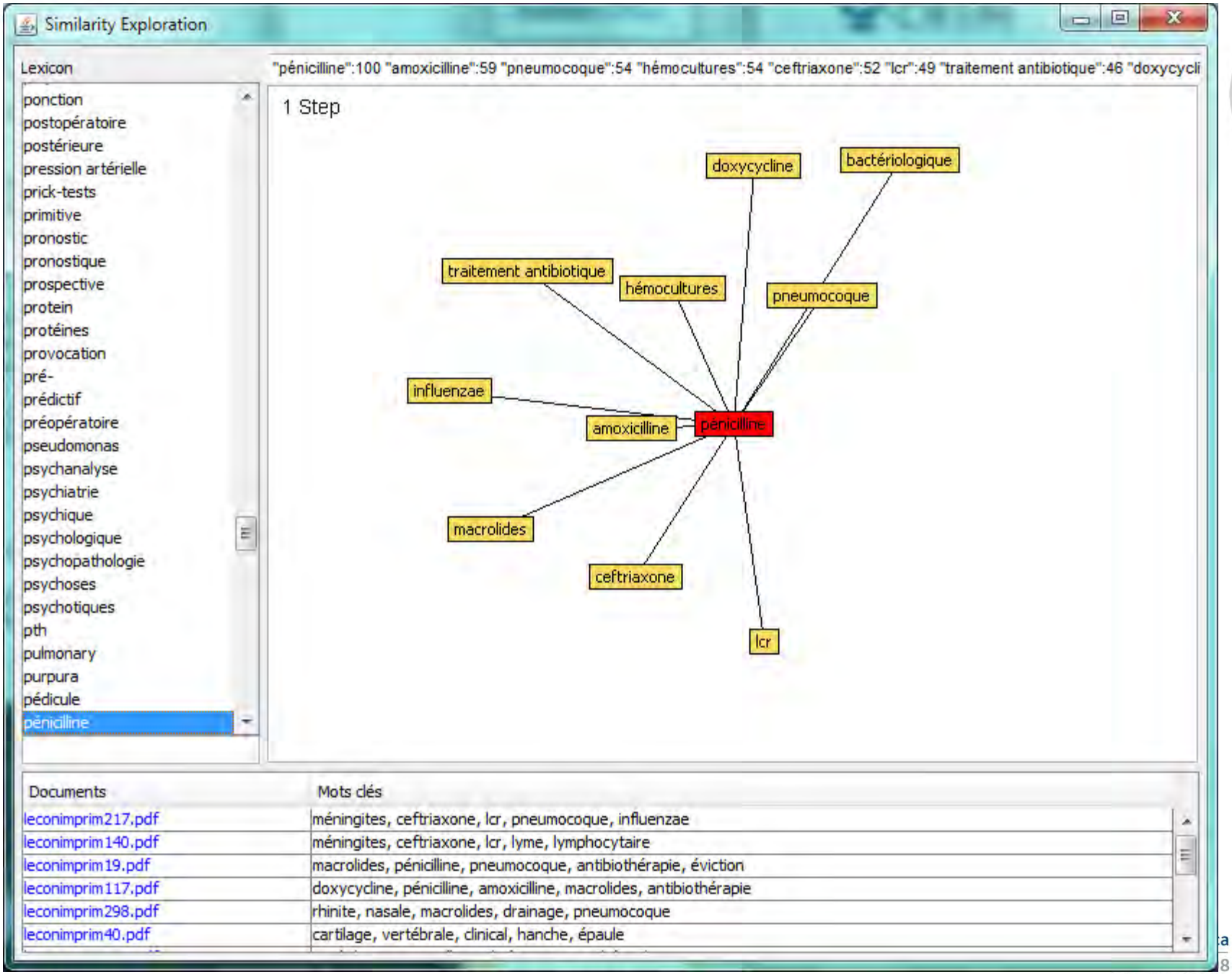
3. Exploration d'une collection de documents à l'aide du pseudo-thésaurus
    1. Visualisation/navigation
    2. Accès à des listes de mots-clés des documents

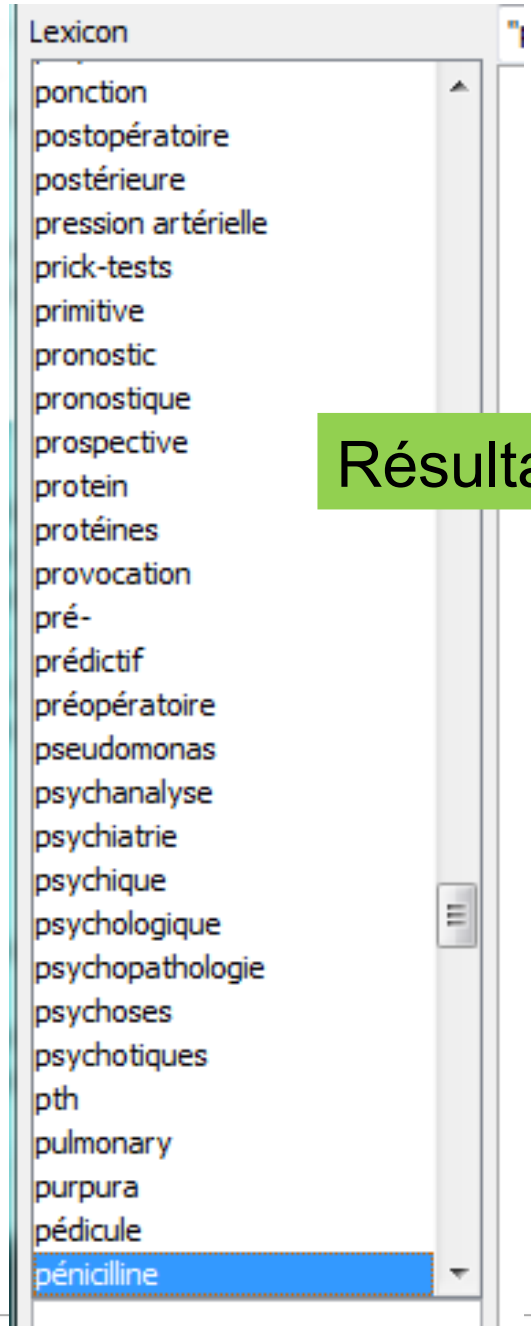
Prototype de navigation conceptuelle

Captures d'écran

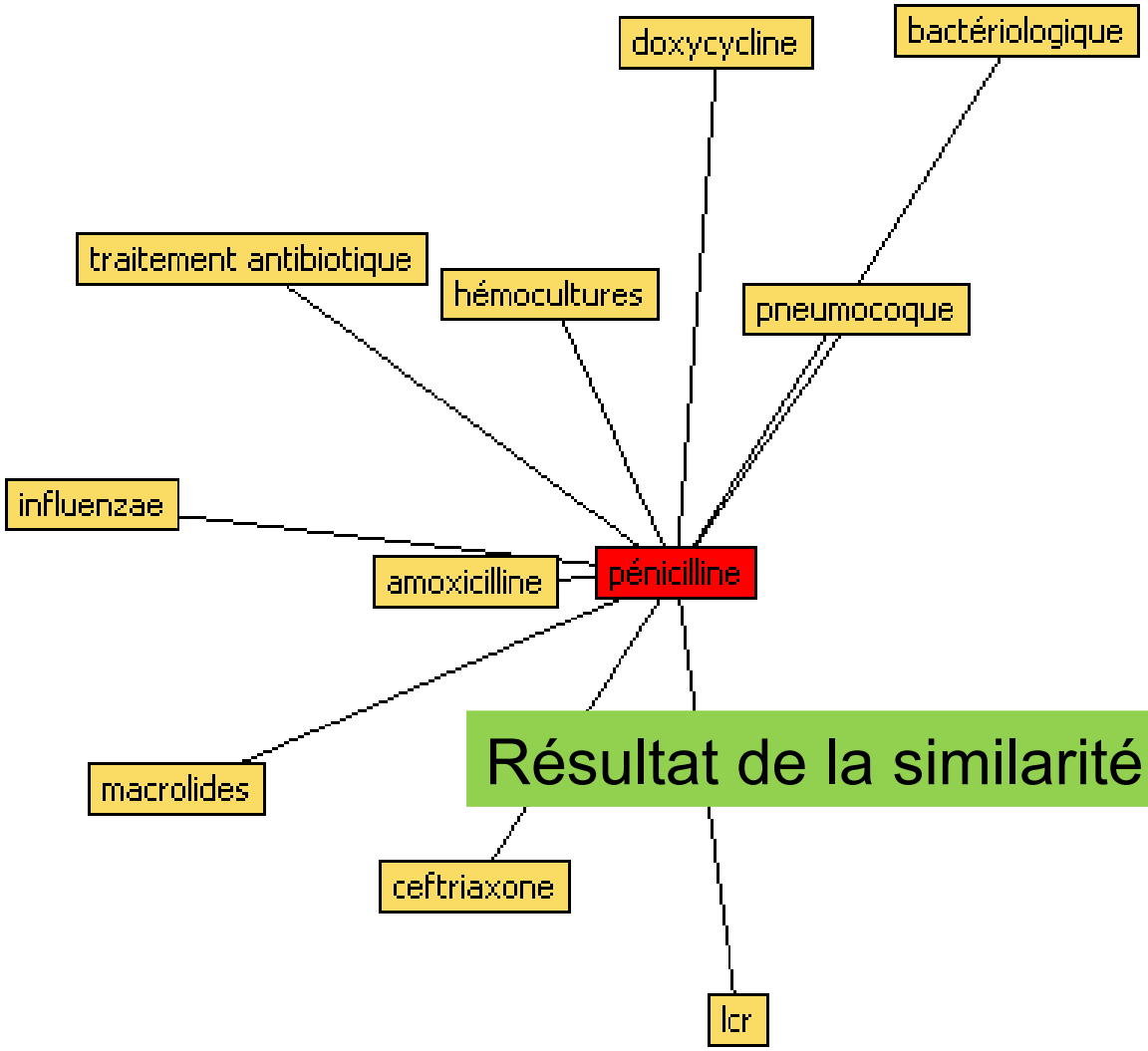
Pseudo-thésaurus:
Corpus médical Français du centre de recherche en terminologie et traduction
http://perso.univ-lyon2.fr/~maniezf/Corpus/Corpus_medical_FR_CRTT.htm

Collection à indexer:
Les documents PDF indexés proviennent de la faculté de médecine de Grenoble
http://www-sante.ujf-grenoble.fr/SANTE/alpesmed/corpus.htm


Similarity Exploration
Lexicon
"pénicilline":100 "amoxicilline":59 "pneumocoque":54 "hémocultures":54 "ceftriaxone":52 "lcr":49 "traitement antibiotique":46 "doxycycli
ponction
postopératoire
postérieure
pression artérielle
prick-tests
primitive
pronostic
pronostique
prospective
protein
protéines
provocation
pré-
prédictif
préopératoire
pseudomonas
psychanalyse
psychiatrie
psychique
psychologique
psychopathologie
psychoses
psychotiques
pth
pulmonary
purpura
pédicule
pénicilline
1 Step
doxycycline
bactériologique
traitement antibiotique
hémocultures
pneumocoque
influenzae
amoxicilline
pénicilline
macrolides
ceftriaxone
lcr
Documents
Mots clés
leconimprim217.pdf
méningites, ceftriaxone, lcr, pneumocoque, influenzae
leconimprim140.pdf
méningites, ceftriaxone, lcr, lyme, lymphocytaire
leconimprim19.pdf
macrolides, pénicilline, pneumocoque, antibiothérapie, éviction
leconimprim117.pdf
doxycycline, pénicilline, amoxicilline, macrolides, antibiothérapie
leconimprim298.pdf
rhinite, nasale, macrolides, drainage, pneumocoque
leconimprim40.pdf
cartilage, vertébrale, clinical, hanche, épaule



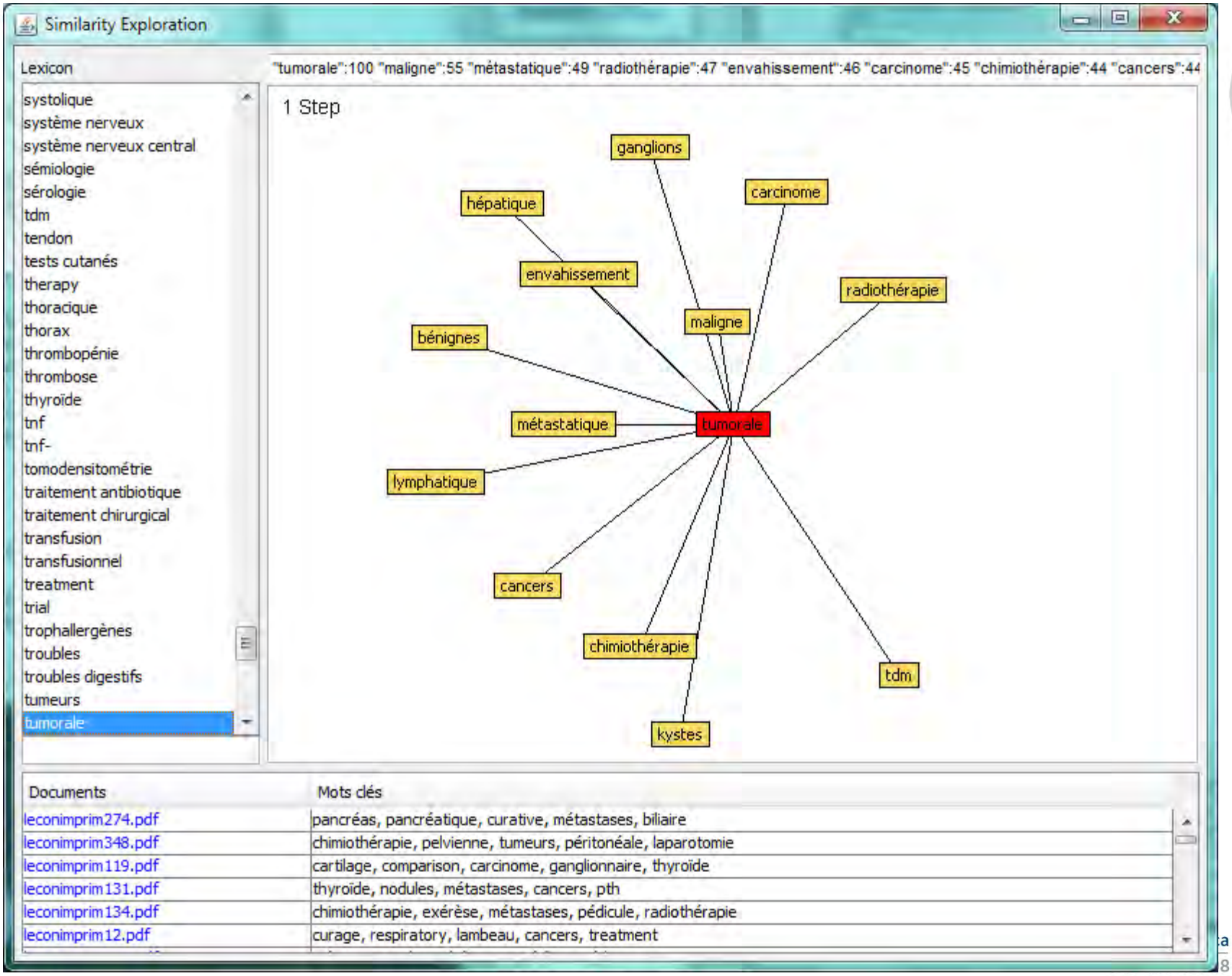
Résultat de l'extraction de termes

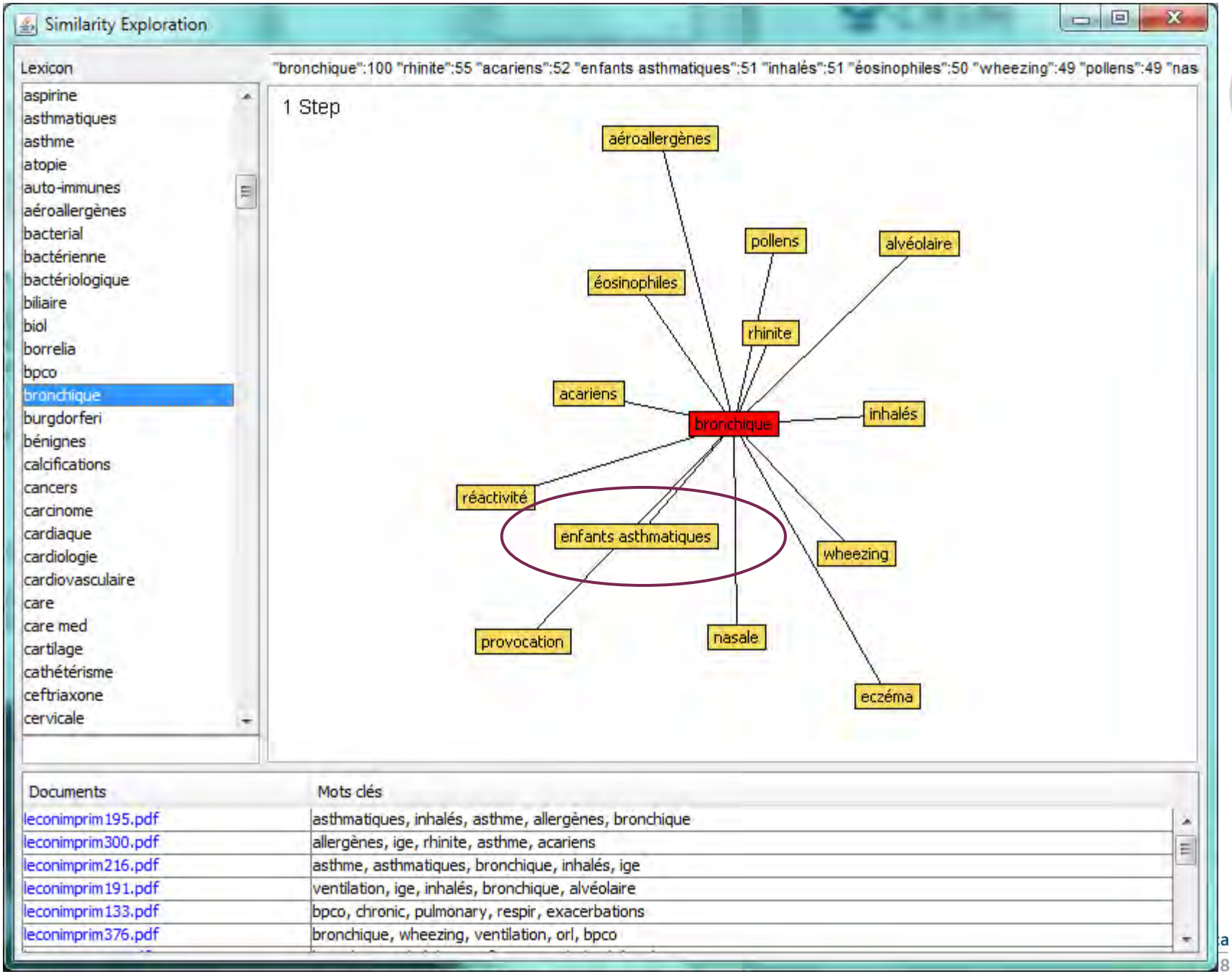


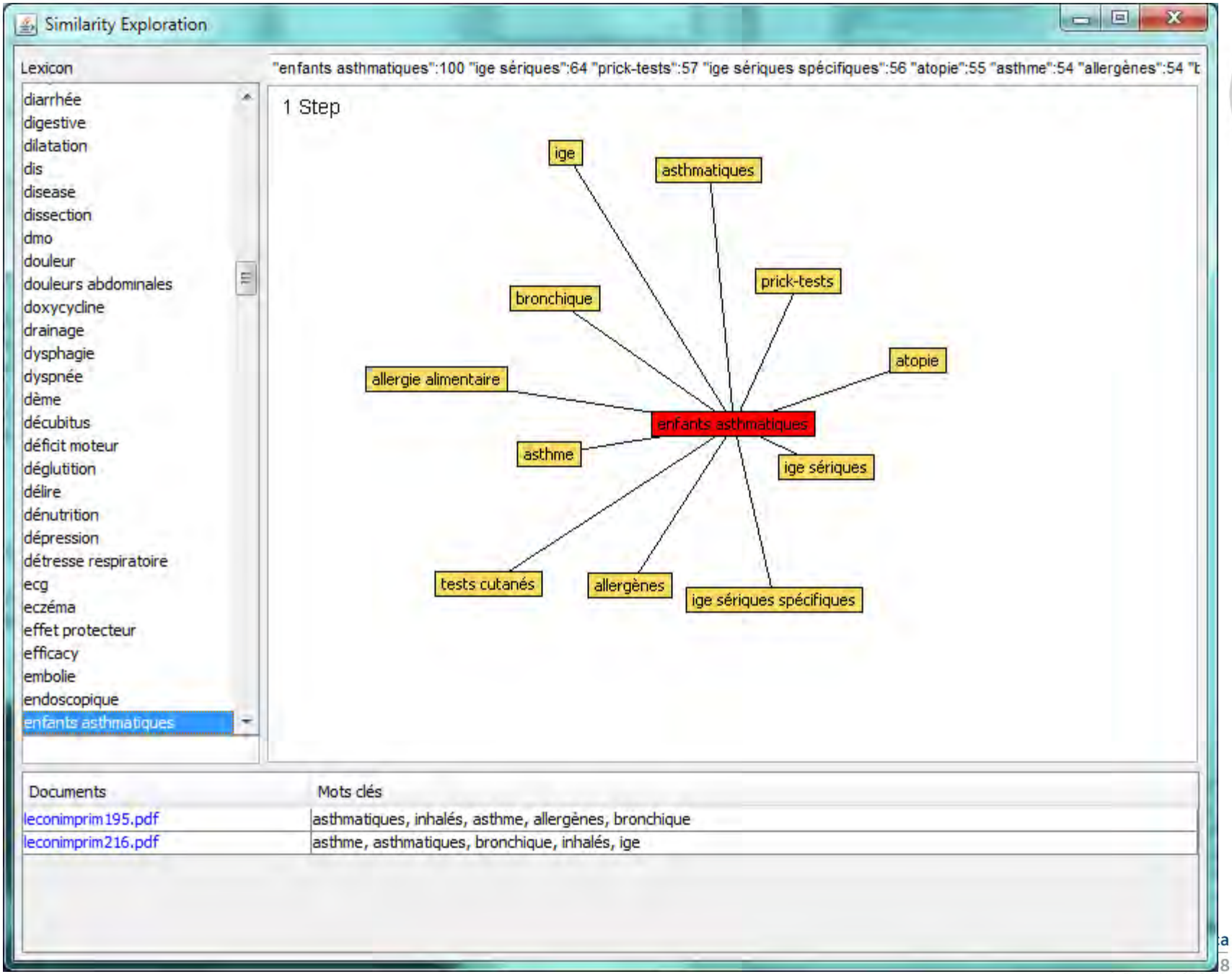
Résultat de la similarité distributionnelle

Documents	Mots clés
leconimprim217.pdf	méningites, ceftriaxone, lcr, pneumocoque, influenzae
leconimprim140.pdf	méningites, ceftriaxone, lcr, lyme, lymphocytaire
leconimprim19.pdf	macrolides, pénicilline, pneumocoque, antibiothérapie, éviction
leconimprim117.pdf	doxycycline, pénicilline, amoxicilline, macrolides, antibiothérapie
leconimprim298.pdf	rhinite, nasale, macrolides, drainage, pneumocoque
leconimprim40.pdf	cartilage, vertébrale, clinical, hanche, épaule

Résultats de la recherche de documents et la visualisation d'un ensemble de mots-clés extraits des documents.


Similarity Exploration
Lexicon
"tumorale":100 "maligne":55 "métastatique":49 "radiothérapie":47 "envahissement":46 "carcinome":45 "chimiothérapie":44 "cancers":44
systolique
système nerveux
système nerveux central
sémiologie
sérologie
tdm
tendon
tests cutanés
therapy
thoracique
thorax
thrombopénie
thrombose
thyroïde
tnf
tnf-
tomodensitométrie
traitement antibiotique
traitement chirurgical
transfusion
transfusionnel
treatment
trial
trophallergènes
troubles
troubles digestifs
tumeurs
tumorale
1 Step
ganglions
carcinome
hépatique
envahissement
radiothérapie
maligne
bénignes
métastatique
tumorale
lymphatique
cancers
chimiothérapie
tdm
kystes
Documents
Mots clés
leconimprim274.pdf
pancréas, pancréatique, curative, métastases, biliaire
leconimprim348.pdf
chimiothérapie, pelvienne, tumeurs, péritonéale, laparotomie
leconimprim119.pdf
cartilage, comparison, carcinome, ganglionnaire, thyroïde
leconimprim131.pdf
thyroïde, nodules, métastases, cancers, pth
leconimprim134.pdf
chimiothérapie, exérèse, métastases, pédicule, radiothérapie
leconimprim12.pdf
curage, respiratory, lambeau, cancers, treatment


Similarity Exploration
Lexicon
aspirine
asthmatiques
asthme
atopie
auto-immunes
aéroallergènes
bacterial
bactérienne
bactériologique
biliaire
biol
borrelia
bpco
bronchique
burgdorferi
bénignes
calcifications
cancers
carcinome
cardiaque
cardiologie
cardiovasculaire
care
care med
cartilage
cathétérisme
ceftriaxone
cervicale
"bronchique":100 "rhinite":55 "acariens":52 "enfants asthmatiques":51 "inhalés":51 "éosinophiles":50 "wheezing":49 "pollens":49 "nas
1 Step
aéroallergènes
pollens
alvéolaire
éosinophiles
rhinite
acariens
bronchique
inhalés
réactivité
enfants asthmatiques
wheezing
provocation
nasale
eczéma
Documents
Mots clés
leconimprim195.pdf
asthmatiques, inhalés, asthme, allergènes, bronchique
leconimprim300.pdf
allergènes, ige, rhinite, asthme, acariens
leconimprim216.pdf
asthme, asthmatiques, bronchique, inhalés, ige
leconimprim191.pdf
ventilation, ige, inhalés, bronchique, alvéolaire
leconimprim133.pdf
bpco, chronic, pulmonary, respir, exacerbations
leconimprim376.pdf
bronchique, wheezing, ventilation, orl, bpco


Similarity Exploration
Lexicon
"enfants asthmatiques":100 "ige sériques":64 "prick-tests":57 "ige sériques spécifiques":56 "atopie":55 "asthme":54 "allergènes":54 "t
diarrhée
digestive
dilatation
dis
disease
dissection
dmo
douleur
douleurs abdominales
doxycycline
drainage
dysphagie
dyspnée
dème
décubitus
déficit moteur
déglutition
délire
dénutrition
dépression
détresse respiratoire
ecg
eczéma
effet protecteur
efficacy
embolie
endoscopique
enfants asthmatiques
1 Step
ige
asthmatiques
prick-tests
bronchique
atopie
allergie alimentaire
enfants asthmatiques
asthme
ige sériques
tests cutanés
allergènes
ige sériques spécifiques
Documents
Mots clés
leconimprim195.pdf
asthmatiques, inhalés, asthme, allergènes, bronchique
leconimprim216.pdf
asthme, asthmatiques, bronchique, inhalés, ige

Conclusions

1. Présentation d'un prototype de navigation conceptuelle, permettant d'indexer et d'explorer une collection de documents.

2. La collection à indexer peut servir pour la construction du pseudo-thésaurus, ou bien une autre collection ciblée (si accessible) peut être utilisée.

3. Les diverses techniques de Traitement Automatique des Langues (TAL) sous-jacentes au prototype démontré peuvent être adaptées pour créer d'autres prototypes répondant à divers besoins des utilisateurs.

4. Plusieurs autres techniques en TAL existent pour la classification automatique de documents, les regroupements automatiques de documents, l'extraction de mots-clés, les résumés automatiques de texte, etc.

Merci!

Questions??