IMT Mines Alès
École Mines-Télécom

DÉFI FOUILLE DE TEXTES 2019
INDEXATION PAR EXTRACTION ET APPARIEMENT TEXTUEL

Jean-Christophe MENSONIDES
Pierre-Antoine JEAN
Andon TCHECHMEDJIEV
Sébastien HARISPE

SOMMAIRE

IMT Mines Alès
École Mines-Télécom

I. INTRODUCTION À LA TÂCHE D'INDEXATION

Indexer des couples de textes cas clinique / discussion: Pour chaque couple, retrouver un ensemble de mots-clés prédéfinis par deux annotateurs.

I. INTRODUCTION À LA TÂCHE D'INDEXATION

Indexer des couples de textes cas clinique / discussion: Pour chaque couple, retrouver un ensemble de mots-clés prédéfinis par deux annotateurs.

Vocabulaire de référence : 1311 mots-clés

2,4-d	virus de l'immunodéficience humaine
2dpmp	virus du papillome humain
4-méthylthioamphétamine	vision des couleurs
4-mta	voie basse
5-azacytidine	voies urinaires
5-méthyl-7-méthoxyisoflavone	vols
7-aminoclonazépam	volume globulaire moyen
abcès	volvulus du sigmoïde
abdomen aigu	vomissement
abl 825 radiometer	von hippel-lindau
accès palustre	voriconazole
acétaldéhyde	vp16
acetaminophen	wolff-parkinson-white syndrome
acétone	wolfram
achalasie	zolpidem
acide gamma hydroxybutyrique	zona
...	zopiclone

I. INTRODUCTION À LA TÂCHE D'INDEXATION

Indexer des couples de textes cas clinique / discussion: Pour chaque couple, retrouver un ensemble de mots-clés prédéfinis par deux annotateurs.

Vocabulaire de référence : 1311 mots-clés

Métrique d'évaluation: Mean Average Precision (MAP)

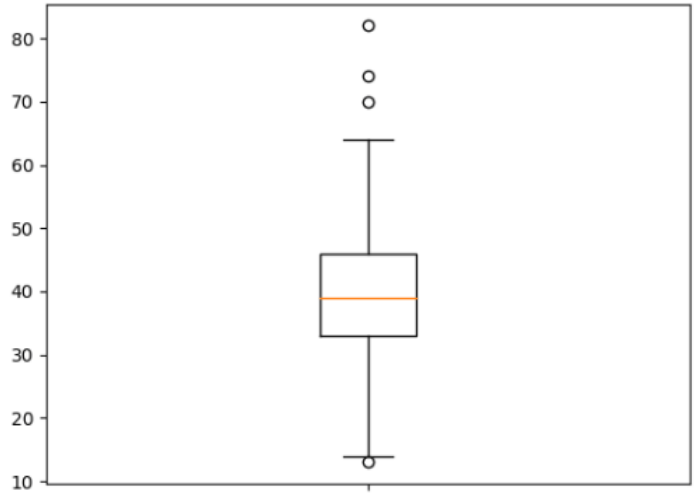
I. INTRODUCTION À LA TÂCHE D'INDEXATION

# couples cas clinique/discussion dans $\mathcal{C}$	290
# moyen de mots dans les cas cliniques de $\mathcal{C}$	332
# moyen de mots dans les discussions de $\mathcal{C}$	764
# de mots-clés dans le vocabulaire contrôlé	1311
# de mots-clés utilisés dans $\mathcal{C}$	1123 (85%)
# de mots-clés avec une correspondance exacte dans les cas cliniques de $\mathcal{C}$	441
# de mots-clés avec une correspondance exacte dans les discussions de $\mathcal{C}$	658
# de mots-clés abstraits au sein d'un couple de $\mathcal{C}$	390

I. INTRODUCTION À LA TÂCHE D'INDEXATION

# couples cas clinique/discussion dans $\mathcal{C}$	290
# moyen de mots dans les cas cliniques de $\mathcal{C}$	332
# moyen de mots dans les discussions de $\mathcal{C}$	764
# de mots-clés dans le vocabulaire contrôlé	1311
# de mots-clés utilisés dans $\mathcal{C}$	1123 (85%)
# de mots-clés avec une correspondance exacte dans les cas cliniques de $\mathcal{C}$	441
# de mots-clés avec une correspondance exacte dans les discussions de $\mathcal{C}$	658
# de mots-clés abstraits au sein d'un couple de $\mathcal{C}$	390

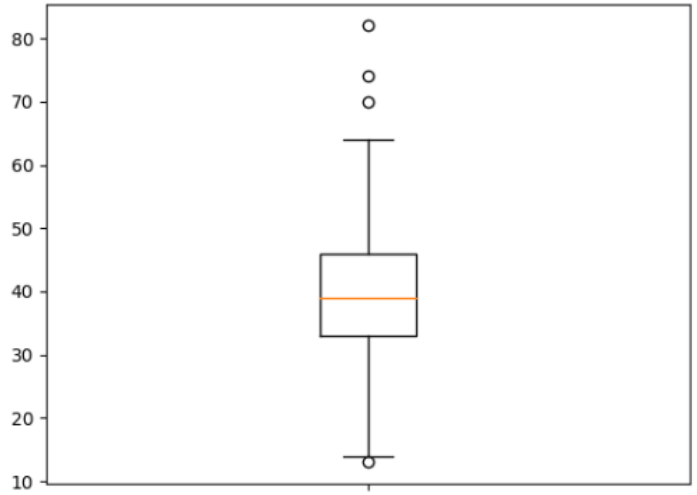
Nombre de mots-clés résultant de l'intersection
entre les textes et la liste de mots-clés de référence



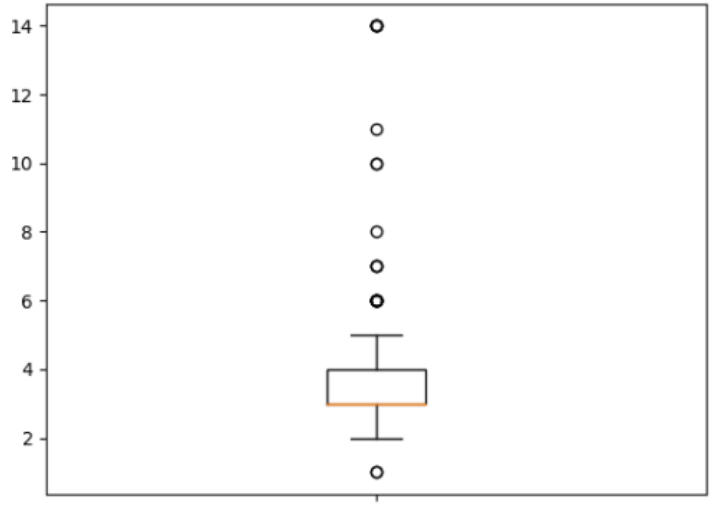
I. INTRODUCTION À LA TÂCHE D'INDEXATION

# couples cas clinique/discussion dans $\mathcal{C}$	290
# moyen de mots dans les cas cliniques de $\mathcal{C}$	332
# moyen de mots dans les discussions de $\mathcal{C}$	764
# de mots-clés dans le vocabulaire contrôlé	1311
# de mots-clés utilisés dans $\mathcal{C}$	1123 (85%)
# de mots-clés avec une correspondance exacte dans les cas cliniques de $\mathcal{C}$	441
# de mots-clés avec une correspondance exacte dans les discussions de $\mathcal{C}$	658
# de mots-clés abstraits au sein d'un couple de $\mathcal{C}$	390

Nombre de mots-clés résultant de l'intersection entre les textes et la liste de mots-clés de référence



Nombre de mots-clés à attribuer à chaque couple cas clinique / discussion

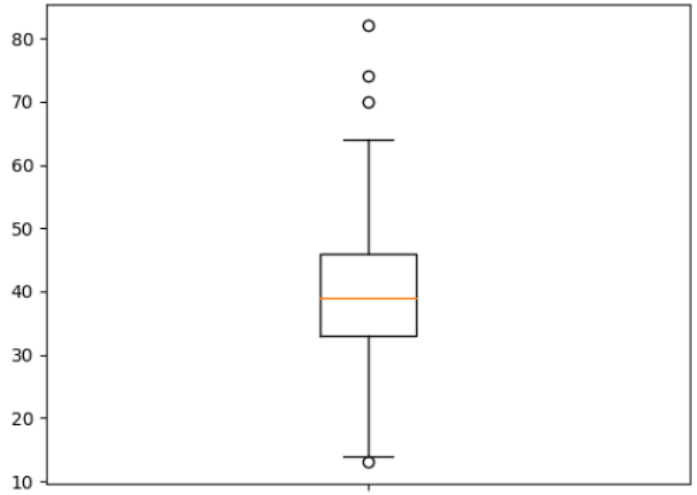


IMT Mines Alès
École Mines-Télécom

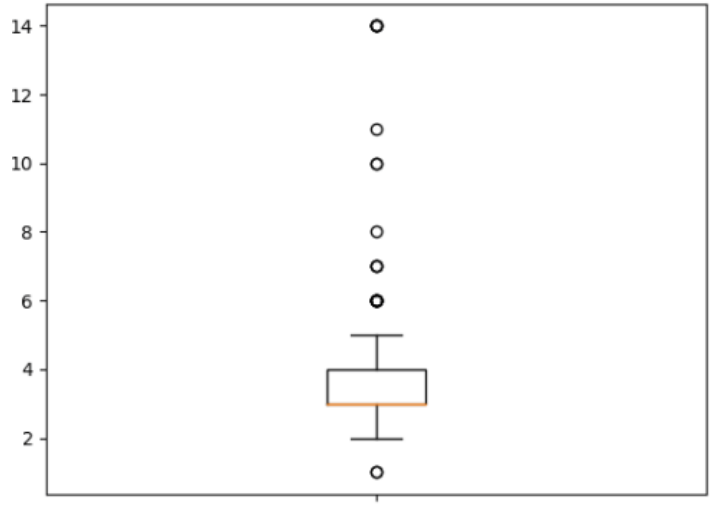
I. INTRODUCTION À LA TÂCHE D'INDEXATION

# couples cas clinique/discussion dans $\mathcal{C}$	290
# moyen de mots dans les cas cliniques de $\mathcal{C}$	332
# moyen de mots dans les discussions de $\mathcal{C}$	764
# de mots-clés dans le vocabulaire contrôlé	1311
# de mots-clés utilisés dans $\mathcal{C}$	1123 (85%)
# de mots-clés avec une correspondance exacte dans les cas cliniques de $\mathcal{C}$	441
# de mots-clés avec une correspondance exacte dans les discussions de $\mathcal{C}$	658
# de mots-clés abstraits au sein d'un couple de $\mathcal{C}$	390

Nombre de mots-clés résultant de l'intersection entre les textes et la liste de mots-clés de référence



Nombre de mots-clés à attribuer à chaque couple cas clinique / discussion



Taux de recouvrement à 72 %: 72% des mots-clés attendus dans le jeu d'entraînement apparaissent dans les couples correspondants

I. INDEXATION PAR EXTRACTION

Méthode par extraction : évaluation de la pertinence des termes d'un document pour son indexation

3 phases séquentielles :

- **Pré-traitement** des mots-clés de référence et du texte des couples à caractériser
- **Evaluation** de la pertinence des termes d'un document
- **Classement** des mots-clés considérés pour l'indexation

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Certains mots-clés sont proches lexicalement mais considérés comme différents

cancer de la prostate – cancer de prostate
urètre – urèthre

Réduction de la variance des mots-clés en limitant la perte / altération de l'information

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Suppression de la ponctuation
Suppression des chiffres
Suppression des stopwords (un, une, le, la, etc.)
Lemmatisation des unigrames, racinisation des lemmes

alcooliques → alcoolique → alcool

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Considérer des mots racinisés proches comme similaires :

uretral – ureteral – urethral – urethr

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Considérer des mots racinisés proches comme similaires :

uretral – ureteral – urethral – urethr

Pour chaque paire d'unigrammes racinisés :

Extraction des radicaux en soustrayant préfixes et suffixes

Evaluation d'une similarité cosinus entre les embeddings des radicaux. Si la similarité est importante, les termes sont considérés comme similaires.

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Préfixes utilisés :

acetyl, acetal, ana, anti, angio, antibiot, ante, ben, meth, eth, prop, but, pent, hex, hept, di, tri,tetra, carboxy, sulf, alca, hyper, hypo, cardio, psych, poly, pneumo, myco, meso, lymph, intra, hydro, immun, homo, endo, dys,chondro, met, micro, osteo, retro, hemangio

Suffixes utilisés :

tom, plast, scop, graph, oid, sarcom, log, om

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Transformations d'unigrammes :

uretero → uret
uretr → uret
uretral → uret
ureteral → uret
ureterocol → uret
urethral → uret
urethr → uret

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Concaténer le texte de chaque cas clinique avec sa discussion

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Concaténer le texte de chaque cas clinique avec sa discussion

Les mots composants le texte subissent un traîtement similaire aux mots-clés

urethro → uret
urethrocel → uret
urethrorrag → uret
urethrorraph → uret
urethrovaginal → uret

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Concaténer le texte de chaque cas clinique avec sa discussion

Les mots composants le texte subissent un traîtement similaire aux mots-clés

urethro → uret
urethrocel → uret
urethrorrag → uret
urethrorraph → uret
urethrovaginal → uret

Seuls les unigrammes correspondants correspondant à des mots-clés (partiels) sont conservés

I. INDEXATION PAR EXTRACTION - PRÉ-TRAITEMENT

Concaténer le texte de chaque cas clinique avec sa discussion

Les mots composants le texte subissent un traîtement similaire aux mots-clés

urethro → uret
urethrocel → uret
urethrorrag → uret
urethrorraph → uret
urethrovaginal → uret

Seuls les unigrammes correspondants correspondant à des mots-clés (partiels) sont conservés

Extension des unigrammes et bi-grammes déterministes

escherichia → escherichia coli

I. INDEXATION PAR EXTRACTION – SCORE PAR TERME

Pour chaque couple, pour chaque mot-clé, calcul d'un score de pondération TF-IDF en tenant compte des ngrams de rang 1 à 5.

$$tfidf(t,c) = idf(t) \times (1 + log\ \ tf(t,c))$$

$$idf(t) = 1 + log\frac{1+|\mathcal{C}|}{1+df(t)}$$

I. INDEXATION PAR EXTRACTION – SCORE PAR TERME

Pour chaque couple, pour chaque mot-clé, calcul d'un score de pondération TF-IDF en tenant compte des ngrams de rang 1 à 5.

$$tfidf(t,c) = idf(t) \times (1 + log\ \ tf(t,c))$$

$$idf(t) = 1 + log\frac{1+|\mathcal{C}|}{1+df(t)}$$

Pondération des scores des ngrams ajoutés par déterminisme

escherichia → escherichia coli

$$tf(t,c) = entier(tf(t,c) \times \lambda_1)$$

I. INDEXATION PAR EXTRACTION – CLASSEMENT

Favoriser les mots-clés fréquents dans le jeu d'entraînement

$$tfidf(t,c) = tfidf(t,c) \times (1 + freq(t) \times \lambda_2)$$

I. INDEXATION PAR EXTRACTION – CLASSEMENT

Favoriser les mots-clés fréquents dans le jeu d'entraînement

$$tfidf(t,c) = tfidf(t,c) \times (1 + freq(t) \times \lambda_2)$$

Privilégier les mots-clés les plus spécifiques lorsque les scores sont proches

tumeur du rein > tumeur
0.65 > 0.67

I. INDEXATION PAR EXTRACTION – CLASSEMENT

Favoriser les mots-clés fréquents dans le jeu d'entraînement

$$tfidf(t,c) = tfidf(t,c) \times (1 + freq(t) \times \lambda_2)$$

Privilégier les mots-clés les plus spécifiques lorsque les scores sont proches

tumeur du rein > tumeur
0.65 > 0.67

Supprimer les mots-clés des candidats à l'indexation s'il existe un autre mot-clé référant à un concept similaire mieux classé

tumeur > tumeur du rein
0.65 > 0.23

I. INDEXATION PAR EXTRACTION – CLASSEMENT

Revenir aux mots-clés non-racinisés

cancer de la prostate → canc prost
cancer de prostate → canc prost

canc prost → ?

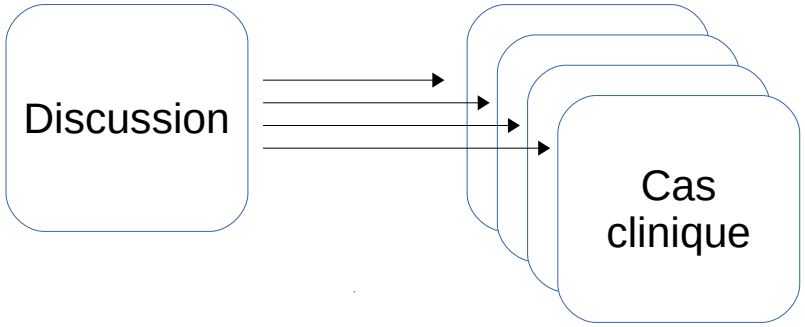
I. INDEXATION PAR EXTRACTION – RÉSULTATS

	Entraînement	Evaluation
MAP	0.42	0.40
MAP max	0.52	0.48

6 participants
MAP : min=0,220 ; max=0,478 ; médiane=0,401 ; moyenne=0,385

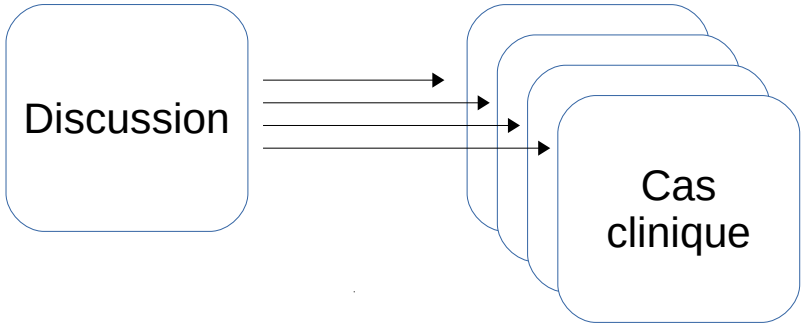
II. INTRODUCTION À LA TÂCHE DE SIMILARITÉ SÉMANTIQUE

Appareiller une discussion à son cas clinique



II. INTRODUCTION À LA TÂCHE DE SIMILARITÉ SÉMANTIQUE

Appareiller une discussion à son cas clinique



Métrique d'évaluation

$$s(a_p(r_i), a_r(r_i)) = \begin{cases} 1 & \text{si } a_p(r_i) = a_r(r_i) \\ 0 & \text{sinon.} \end{cases}$$

$$S(p) = \frac{1}{N} \sum_{i=1}^{N} s(a_p(r_i), a_r(r_i))$$

II. SIMILARITÉ SÉMANTIQUE - MÉTHODOLOGIE

Baseline – Adaptation de la similarité de Lin

$$sim(s_1, s_2) = \frac{2 \times \sum_{w \in s_1 \cap s_2} \log P(w)}{\sum_{w \in s_1} \log P(w) + \sum_{w \in s_2} \log P(w)}$$

Score obtenu : 0.64

II. SIMILARITÉ SÉMANTIQUE - MÉTHODOLOGIE

Baseline – Adaptation de la similarité de Lin

$$sim(s_1, s_2) = \frac{2 \times \sum_{w \in s_1 \cap s_2} \log P(w)}{\sum_{w \in s_1} \log P(w) + \sum_{w \in s_2} \log P(w)}$$

Score obtenu : 0.64

Méthode KNN – Représentation vectorielle TF-IDF

Distances	Score
Euclidienne	**0,748**
Manhattan	0,141
Chebyshev	0,352
Hamming	0,010
Canberra	0,045
Braycurtis	0,741

II. SIMILARITÉ SÉMANTIQUE - MÉTHODOLOGIE

Baseline – Adaptation de la similarité de Lin

$$sim(s_1, s_2) = \frac{2 \times \sum_{w \in s_1 \cap s_2} \log P(w)}{\sum_{w \in s_1} \log P(w) + \sum_{w \in s_2} \log P(w)}$$

Score obtenu : 0.64

Méthode KNN – Représentation vectorielle TF-IDF

Distances	Score
Euclidienne	**0,748**
Manhattan	0,141
Chebyshev	0,352
Hamming	0,010
Canberra	0,045
Braycurtis	0,741

Méthode LSA – Représentation vectorielle TF-IDF

Dimensions	Score
200	0,707
300	0,745
400	0,728
500	0,734
1000	**0,755**

II. SIMILARITÉ SÉMANTIQUE - RÉSULTATS

Méthode KNN – précision : 0.91

Méthode LSA – précision : 0.90

6 participants
Précisions : min=0,617 ; max=0,953 ; médiane=0,862 ; moyenne=0,803

IMT Mines Alès
École Mines-Télécom

DÉFI FOUILLE DE TEXTES 2019
INDEXATION PAR EXTRACTION ET APPARIEMENT TEXTUEL

Jean-Christophe MENSONIDES
Pierre-Antoine JEAN
Andon TCHECHMEDJIEV
Sébastien HARISPE