Université de Strasbourg

Recherche d'information
Évaluation

Enzo Doyen

2025 - LGC6KM43 - M2

🕮 Plan

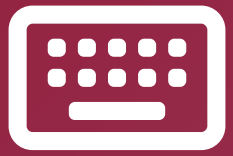
I. Annotation manuelle et QREL

II. Métriques

III. Implémentations en Python

Objectifs de l'évaluation

- Mesurer l'efficacité d'un système de RI : le système correspond-il aux attentes de l'utilisateur ou de l'utilisatrice ? Les résultars retournés sont-ils pertinents vis-à-vis de la requête ?
- Comparaison des systèmes de RI entre eux : quel système obtient les meilleures performances pour telle ou telle tâche de recherche ?
- Amélioration continue des systèmes : connaitre les faiblesses d'un système permet de savoir quels éléments améliorer pour augmenter ses performances.



I. Annotation manuelle et QREL

Jeu de données d'évaluation de référence

En RI, un jeu de données d'évaluation de référence (*ground-truth*) sert de base pour l'évaluation des systèmes de recherche d'information. Il représente les résultats attendus pour un système de RI.

Celui-ci est créé manuellement et contient un ensemble de requêtes (prélablement choisies) et, pour chacune de ces requêtes, un ensemble des documents les plus pertinents.

Format QREL

Pour l'évaluation en RI, on utilise communément le format **QREL** (Query Relevance) : il s'agit un format standardisé pour représenter les jugements de pertinence des documents par rapport à des requêtes spécifiques.

Format QREL

Pour l'évaluation en RI, on utilise communément le format **QREL** (Query Relevance) : il s'agit un format standardisé pour représenter les jugements de pertinence des documents par rapport à des requêtes spécifiques.

Il s'agit d'un fichier texte où chaque ligne représente une **note de pertinence** pour une paire requête-document.

Format QREL

```
1  <id_requete> <iteration> <id_doc> <pertinence>   qrel
```

```
1  101 0 DOC1 1                                     qrel
2  101 0 DOC2 0
3  101 0 DOC3 2
4  102 0 DOC4 0
5  102 0 DOC5 1
```

Format QREL

```
1  <id_requete> <iteration> <id_doc> <pertinence>   qrel
```

<id_requete> : identifiant de la requête.

<iteration> : ancien champ utilisé lors des premières expériences de la TREC (Text Retrieval Conference), qui est à l'initiative de ce format. N'est plus utilisé maintenant, défini sur 0.

<id_doc> : identifiant du document.

<pertinence> : note de pertinence du document par rapport à la requête (0 = non pertinent, 1 = pertinent, 2 = très pertinent). On se limite souvent à 0/1.

Format QREL

En règle générale, un fichier QREL contient uniquement les paires requêtes-documents qui sont jugées pertinentes (avec une note de pertinence supérieure à 0). Les documents non annotés sont considérés comme non pertinents.

Cela permet de gagner du temps lors de l'annotation manuelle en se concentrant uniquement sur les documents jugés pertinents.

QREL et comparaison avec les résultats d'un système de RI

Ce fichier d'annotation manuelle est ensuite comparé aux résultats retournés par un système de recherche d'information (*run file*), ces derniers ayant un format similaire :

```
1  <id_requete> Q0 <id_doc> <rang> <score> <nom_systeme>
```

```
1  101 Q0 DOC3 1 15.7 bm25
2  101 Q0 DOC1 2 12.3 bm25
3  102 Q0 DOC5 1 13.0 bm25
```

II. Métriques

Précision et rappel

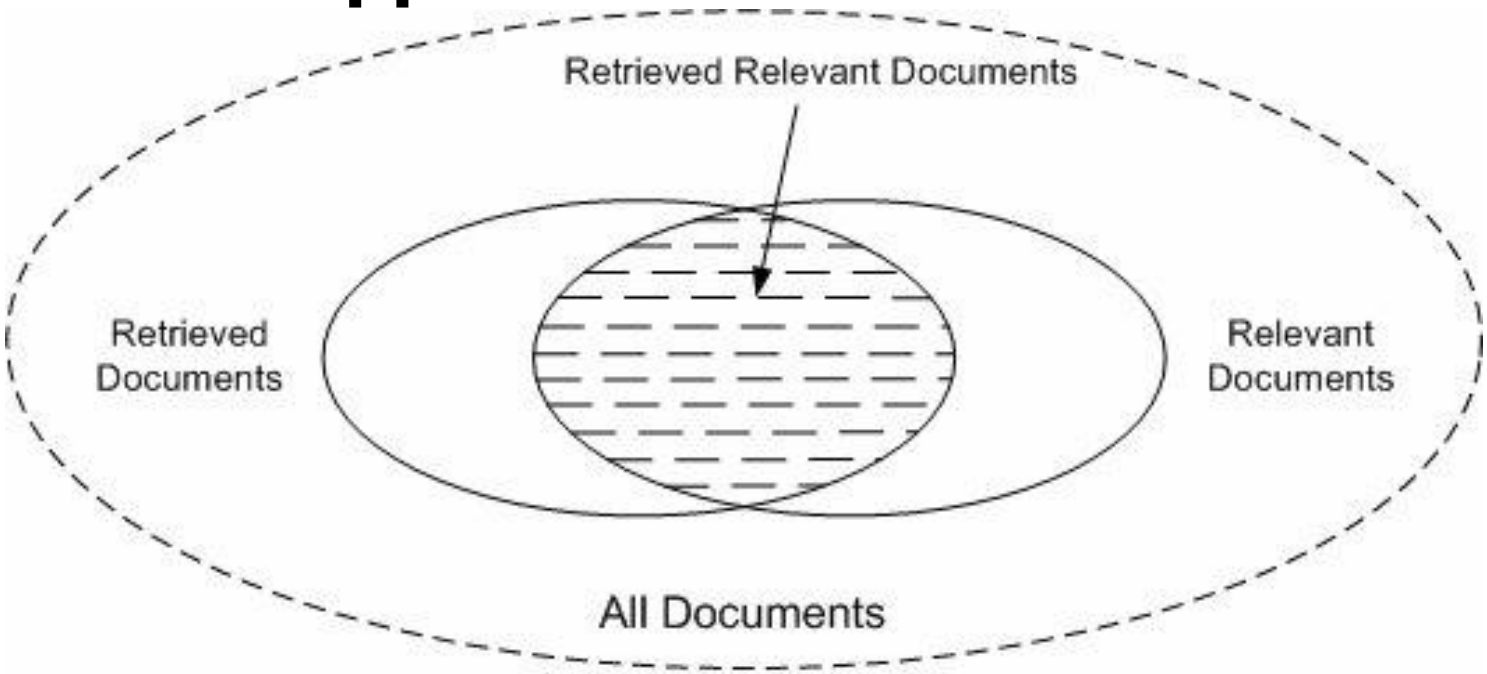
La **précision** correspond au pourcentage de documents récupérés qui sont pertinents.

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Le **rappel** correspond au pourcentage de documents pertinents qui ont été récupérés.

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Précision et rappel



Source : **Soibelman et al. (2006)**

Précision et rappel

	Pertinent	Non pertinent
Récupéré	TP	FP
Non récupéré	FN	TN

Tableau de contingence (adapté de **Manning et al. (2008:155)**)

Classement et métriques à meilleurs k

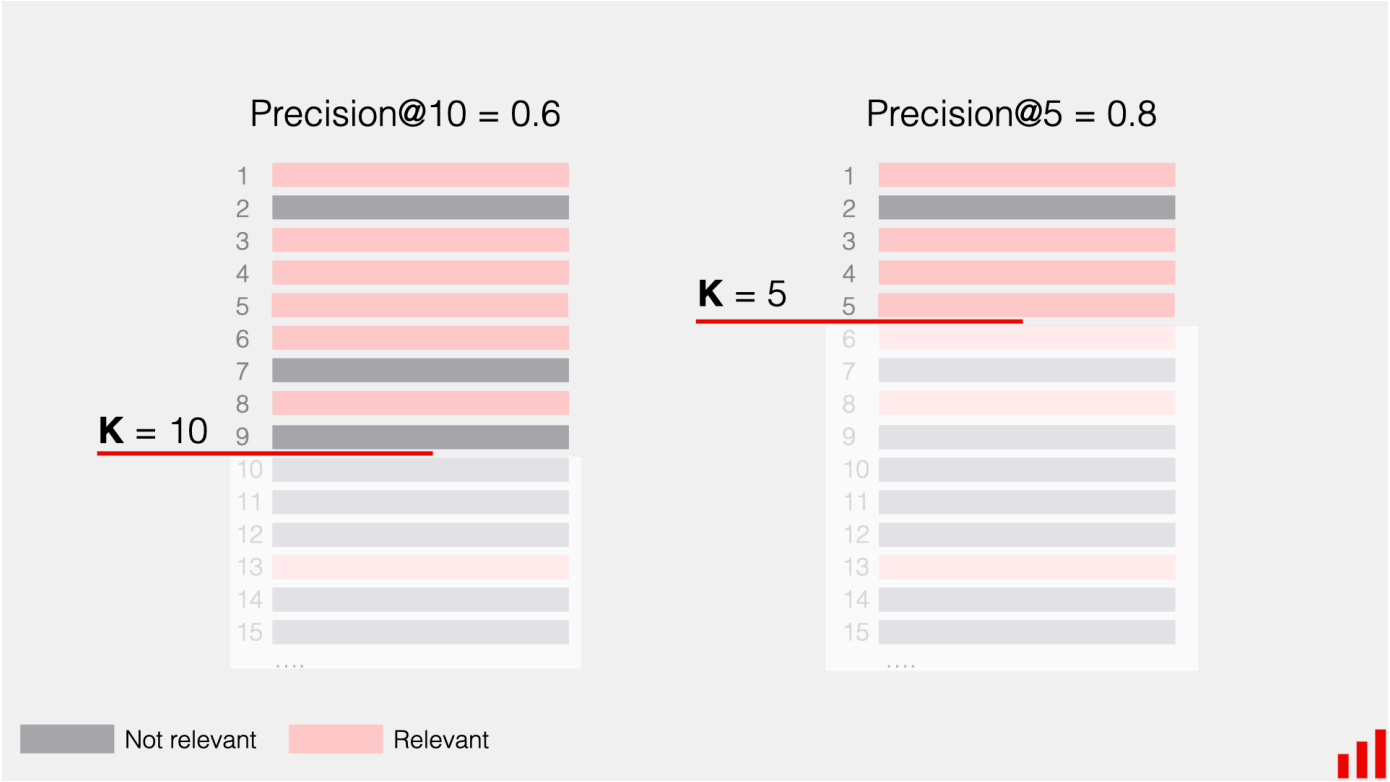
Les systèmes de RI font un classement des documents avant de les renvoyer à l'utilisateur ou à l'utilisatrice.

Ce faisant, les performances des systèmes sont évaluées sur les meilleurs k résultats (*top-k*) : 5 meilleurs, 10 meilleurs, etc.

Précision@k (P@k)

$$\text{Précision@k} = \frac{\#\text{ documents pertinents dans } k}{k}$$

Précision@k (P@k)

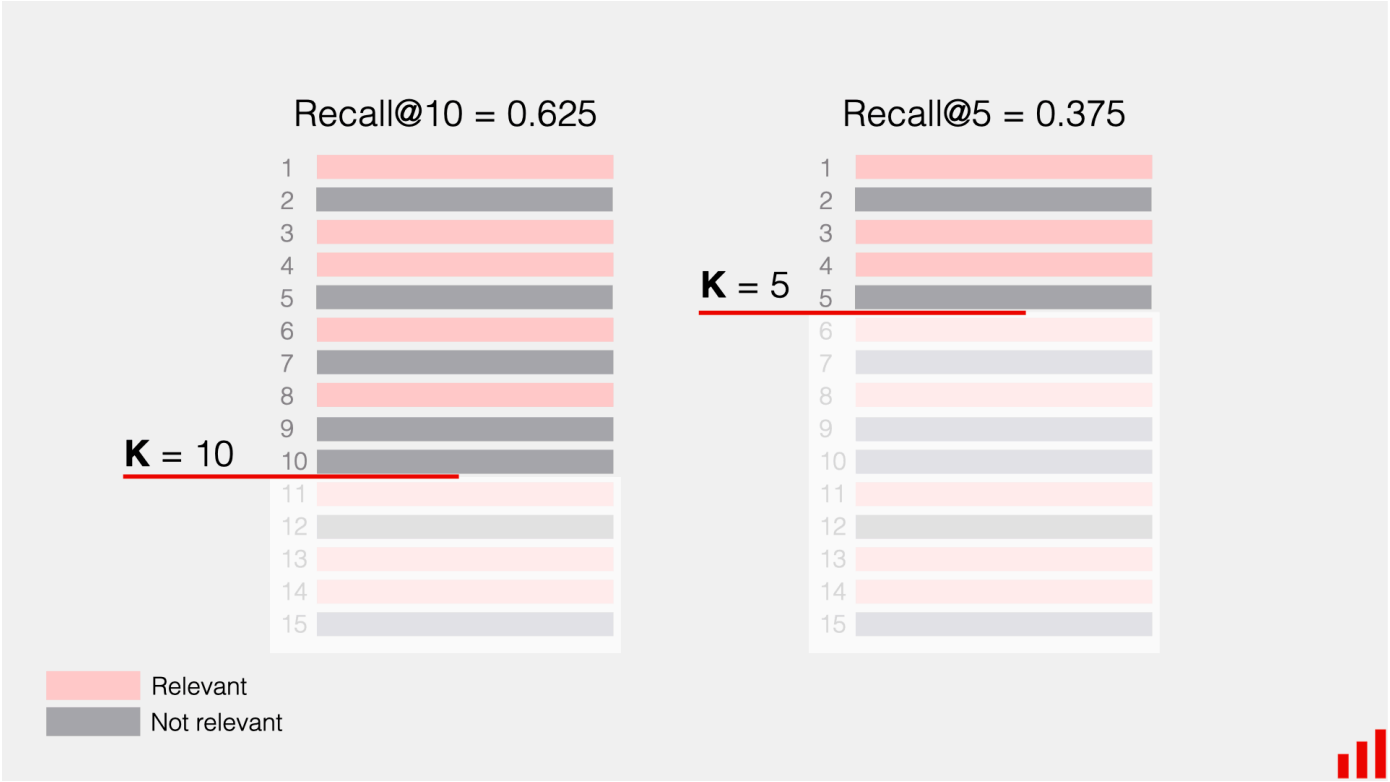


Source : https://www.evidentlyai.com/ranking-metrics/precision-recall-at-k

Rappel@k (R@k)

$$\text{Rappel@k} = \frac{\#\text{ documents pertinents dans } k}{\#\text{ total de documents pertinents}}$$

Rappel@k (R@k)



Source : https://www.evidentlyai.com/ranking-metrics/precision-recall-at-k

Mean Reciprocal Rank (MRR)

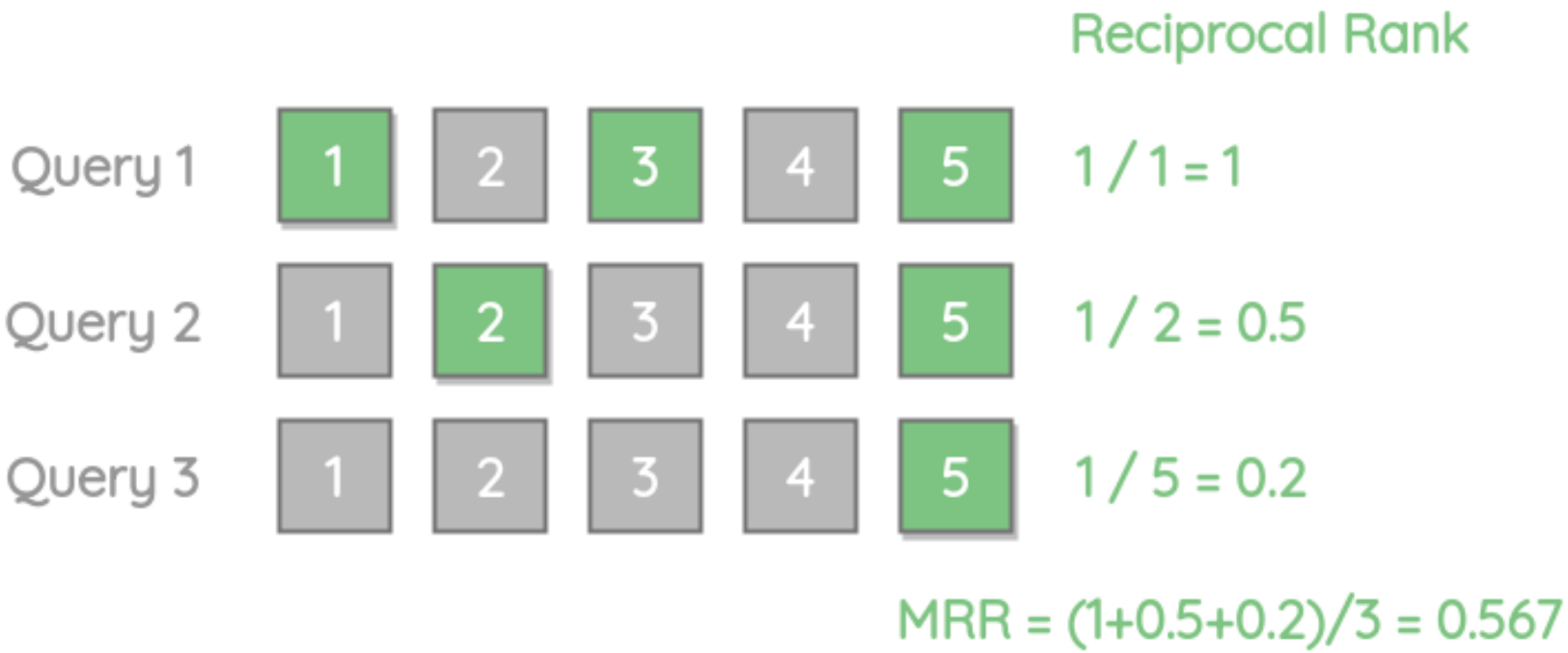
Une autre métrique couramment utilisée pour la RI est le **Mean Reciprocal Rank** (MRR), qui mesure la position du premier document pertinent dans les résultats sur un ensemble de requêtes Q, et retourne la moyenne de ces positions.

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rang}_i}$$

Où rang_i est le rang du premier document pertinent pour la requête q_i.

Renvoie une valeur entre 0 et 1. Plus le MRR est élevé, plus le système est performant.

Mean Reciprocal Rank (MRR)



Source : **https://amitness.com/posts/information-retrieval-evaluation**

Mean Reciprocal Rank (MRR)

La principale limite du MRR est qu'il ne prend en compte que le premier document pertinent. Si un système de RI renvoie plusieurs documents pertinents, le MRR ne permet pas de refléter cette performance.

Mean Average Precision (mAP)

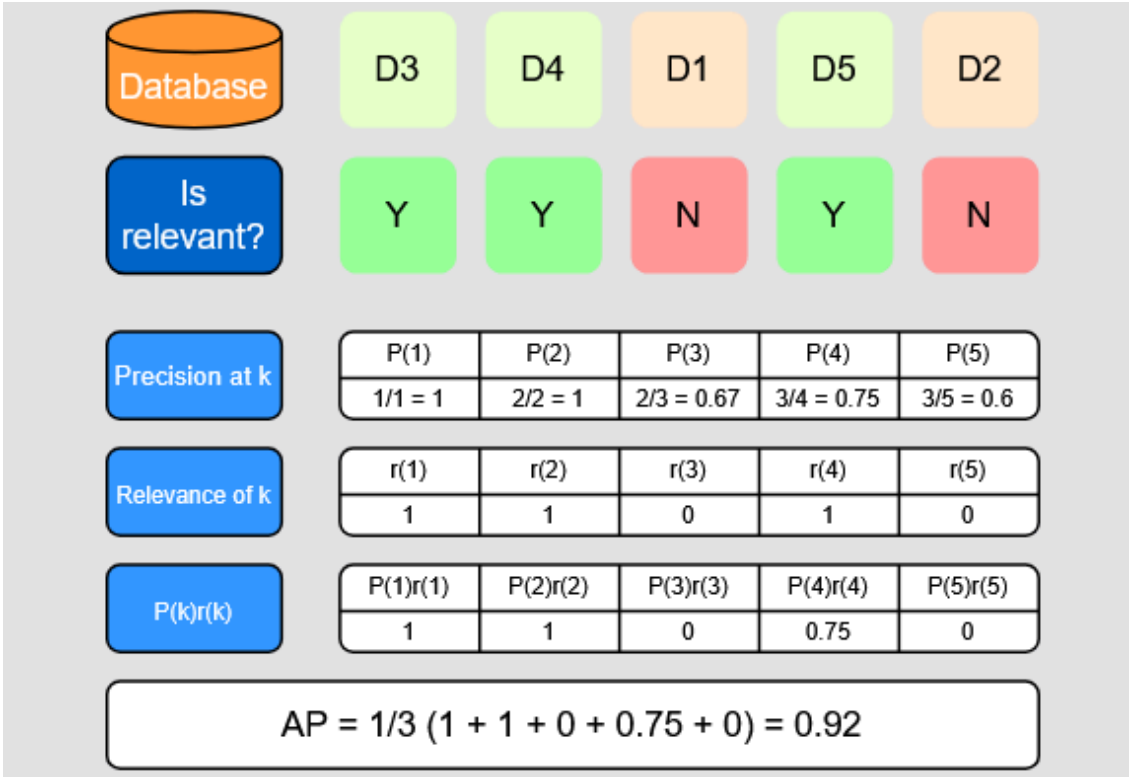
Une autre mesure, mAP (Mean Average Precision), permet de contrebalancer la limite du MRR en prenant en compte tous les documents pertinents dans les résultats.

Il faut d'abord calculer la précision pour chaque requête, puis faire la moyenne de ces précisions.

$$\mathrm{AP} = \frac{1}{\mathrm{RD}} \sum_{k=1}^{n} P(k) \times r(k)$$

Où RD est le nombre total de documents pertinents pour la requête, $P(k)$ la précision à k pour la requête, et $r(k)$ la note de pertinence du k^e document (0, 1 ou plus selon la gradation utilisée pour l'évaluation).

Calcul de la précision moyenne (AP)



Source : https://how.dev/answers/what-is-the-mean-average-precision-in-information-retrieval

Mean Average Precision (mAP)

En faisant la moyenne des précisions calculées précédemment, on obtient un score unifié pour toutes les requêtes à évaluer.

$$\mathrm{mAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathrm{AP}_i$$

III. Implémentations en Python

ir-measures

La bibliothèque Python *ir-measures* rassemble des implémentations de nombreuses métriques d'évaluation pour les systèmes de RI, et permet d'évaluer facilement les performances des modèles.

https://ir-measur.es/

ir-measures

```python
import ir_measures
from ir_measures import R, P, AP, RR
qrels = ir_measures.read_trec_qrels('path/to/qrels')
run = ir_measures.read_trec_run('path/to/run')
ir_measures.calc_aggregate([P@10, R@10, AP, RR], qrels, run) # renvoie un dict
```

ir-measures

Marche aussi en ligne de commande :

```bash
ir_measures path/to/qrels path/to/run P@10 R@10 AP RR
```

ir-measures

Possibilité de définir un seuil de pertinence :

```python
ir_measures.calc_aggregate([P(rel=2)@10], qrels, run)
```

Seuls les documents ayant une note de pertinence supérieure ou égale à 2 seront pris en compte pour le calcul.

ir-measures

Démo : **https://demo.ir-measur.es/**

Notebook d'exemple : **https://github.com/terrierteam/ir_measures/blob/main/examples/demo.ipynb**

Documentation : **https://ir-measur.es/**

Bibliographie

Manning, C. D., Raghavan, P., et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mitra, B. (2018). *An Introduction to Neural Information Retrieval* (Numéro v.41). Now Publishers.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., et Lin, K.-Y. (2006). Data Analysis on Complicated Construction Data Sources: Vision, Research, and Recent Developments. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, et I. F. C. Smith (éds.), *Intelligent Computing in Engineering and Architecture: Vol. 4200. Intelligent Computing in Engineering and Architecture* (p. 637-652). Springer Berlin Heidelberg. **10.1007/11888598_57**

Zhai, C., et Massung, S. (juin 2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool. **10.1145/2915031**