LDA-Based Document Models for Ad-hoc Retrieval

Xing Wei and W. Bruce Croft

SORBONNE UNIVERSITÉ

Amayas SADI &
Ghiles OUHENIA

1. Présentation des algorithmes
2. Ajustement des paramètres
3. Résultats obtenus
4. Conclusion
5. Structure du code

Ad-hoc Retrieval est un processus de recherche d'informations visant à trouver les documents pertinents en réponse à une requête spécifique de l'utilisateur.

1. Présentation des algorithmes

$$P_{model}(q|d) = \prod_{i=1}^{N_q} P_{model}(q_i|d)$$

Query likelihood retrieval (QL) :

$$P_{QL}(q_i|d) = \frac{n_{q_i}}{N_d}$$

Cluster-based retrieval (CBDM) :

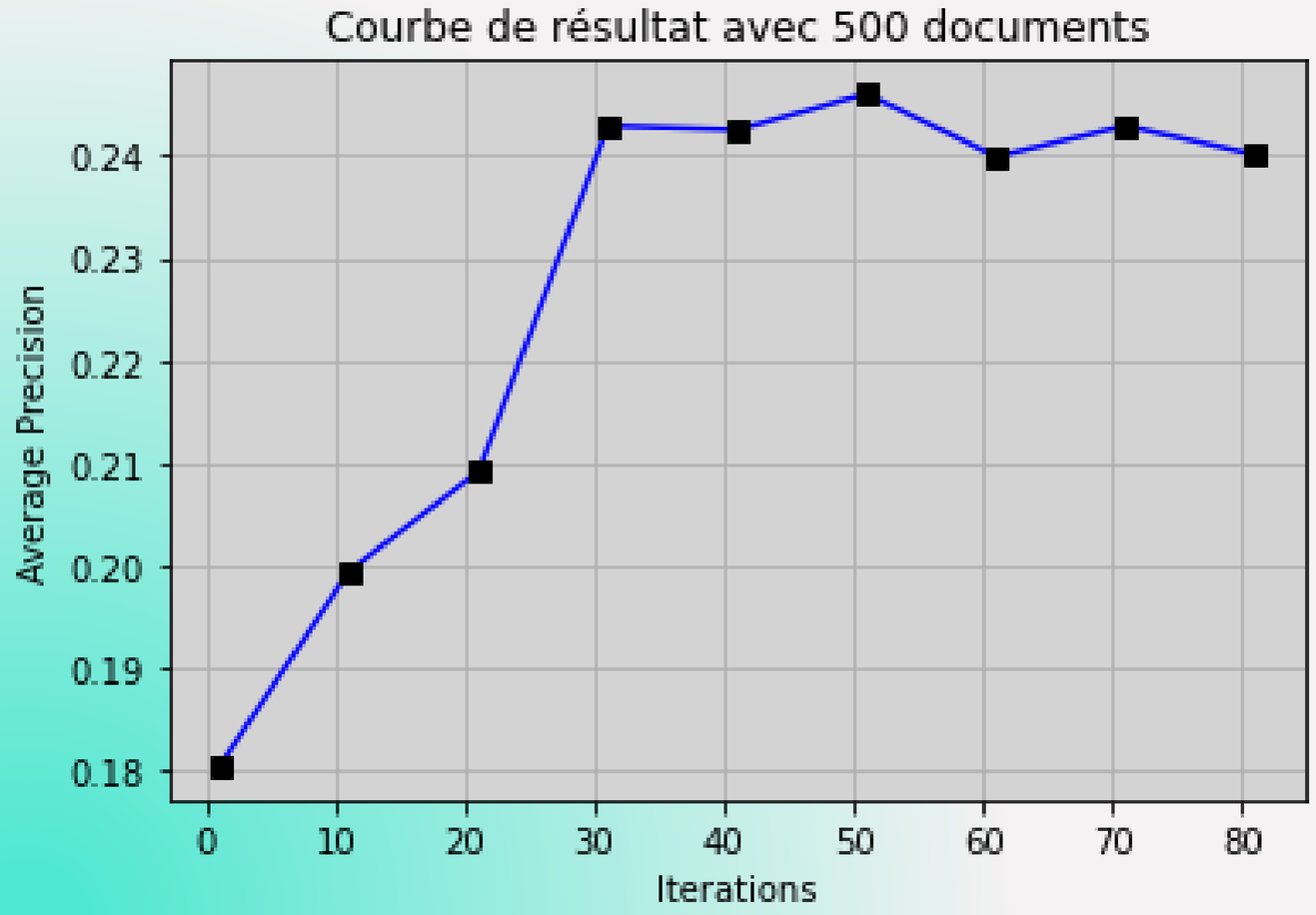
$$P_{CBDM}(q_i|d) = \frac{N_d}{N_d + \mu} P_{QL}(q_i|d) + (1 - \frac{N_d}{N_d + \mu}) P_{QL}(q_i|cluster_d)$$

LDA-based document models (LBDM) :

$$P_{LDA}(q_i|d) = \lambda(\frac{N_d}{N_d + \mu} P_{QL}(q_i|d) + (1 - \frac{N_d}{N_d + \mu}) P_{QL}(q_i|coll)) + (1 - \lambda)(\sum_{i=1}^{K} \frac{n_j^{\alpha_i} + \beta_{q_i}}{\sum_{v=1}^{|V|} n_j^v + \beta_v} \cdot \frac{n_j^d + \alpha_j}{\sum_{t=1}^{K} n_t^d + \alpha_t})$$

2. Ajustement des paramètres

a - Variation du nombre d'itérations

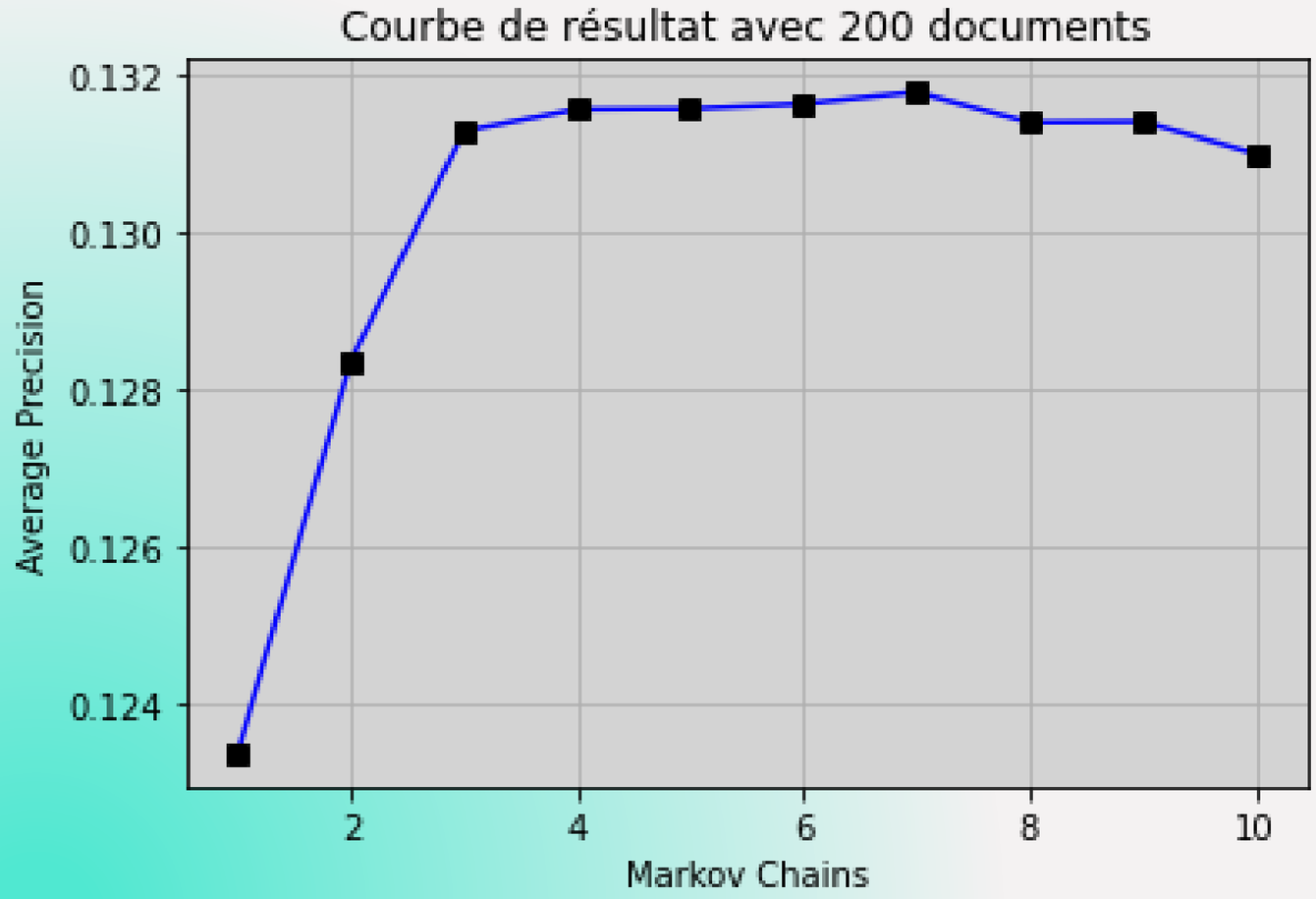


Average Precision pour LDBM avec K = 100, λ = 0.7, 1 MC, pour différents nombres d'itérations.

Optimum : **50 itérations**.

2. Ajustement des paramètres

b - Variation du nombre de chaînes de Markov

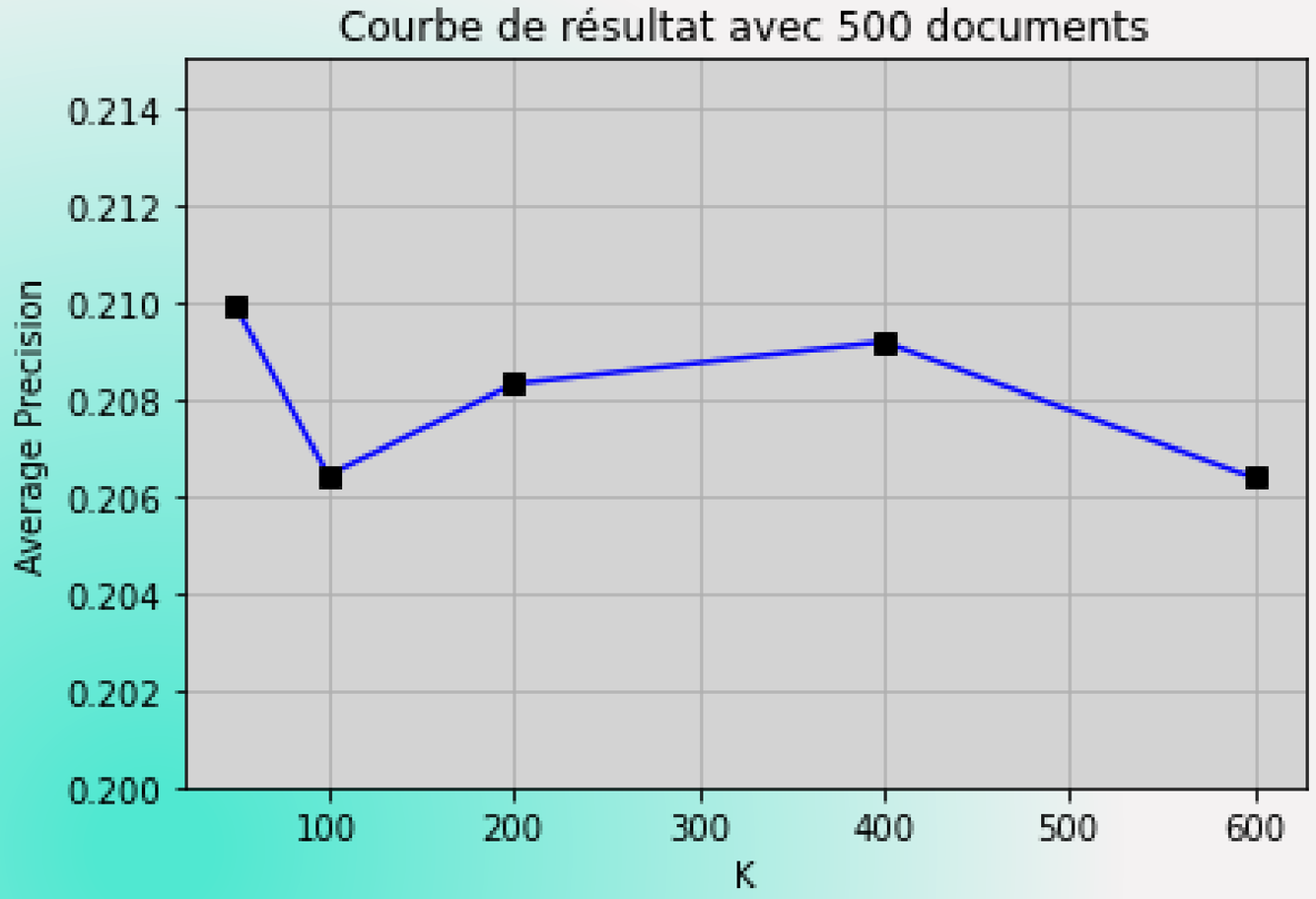


Average Precision pour LDBM avec K = 50, λ = 0.7, 30 itérations, pour différents nombres de chaînes de Markov.

Optimum : **7 chaines de Markov.**

2. Ajustement des paramètres

c - Variation du nombre de topics K



Average Precision pour LDBM avec λ = 0.7, 30 itérations, 2 MC, pour différents nombres de topics K.

Optimum : **50 topics.**

3. Résultats obtenus

Collection utilisée : Scidocs de BEIR.

Collection	QL	CBDM	LBDM
scidocs	0.1866	0.2085	0.2471

Tableau représentant l'Average Precision des modèles sur une collection de 1000 documents et 50 requêtes

Paramètres :

CBDM : 50 topics.

LBDM : K = 50, λ = 0.7, 30 itérations, 2 MC, α = 50 / K, = 0.01 , β = 0.01.

4. Conclusion

- L'approche LDA a démontré son efficacité pour la modélisation des documents dans la recherche ad-hoc.

- Les résultats obtenus avec LDA, en ajustant les paramètres appropriés, ont montré des performances encourageantes.

5. Structure du code

La classe **RetrievalSystem** est la classe de base de nos systèmes de recherche. Elle fournit les méthodes nécessaires pour entraîner et évaluer un système de recherche.

Les deux principales méthodes sont :

- Méthode ***fit(corpus)*** : permet d'entraîner le système de recherche en utilisant un corpus de documents. Elle juste les paramètres internes du système en fonction des caractéristiques des documents.
- Méthode ***predict(queries)*** : permet d'évaluer les requêtes et de prédire les scores de similarité associés à chaque document dans le corpus.

La classe **ClusterBasedRetrieval** hérite de **RetrievalSystem** et implémente l'approche CBDM. Elle utilise le clustering pour regrouper les documents similaires et améliorer la recherche d'informations.

La classe **LikelihoodRetrieval** définit l'algorithme QL. Elle calcule les scores de similarité en se basant sur la probabilité de la requête dans les documents du corpus.

La classe **LdaRetrieval** implémente une approche de recherche d'informations basée sur la LDA et l'estimation par Gibbs Sampling. Cette approche utilise des modèles de sujets pour représenter les documents et les requêtes, permettant ainsi de trouver des documents pertinents en fonction des sujets abordés.

Merci de votre attention !