

LO17 : Indexation et Recherche d'Information

TD1 : Introduction et Recherche Booléenne

Printemps 2025

On veut réaliser un système de recherche booléen simple¹ à partir d'un corpus contenant les titres et résumés de plusieurs fables de Jean de La Fontaine. Considérons tout d'abord les documents suivants :

- **Le Corbeau et le Renard** : Un corbeau tient un fromage dans son bec. Un renard rusé le flatte pour le tromper et lui faire lâcher son repas.
- **Le Renard et le Bouc** : Un renard et un bouc tombent dans un puits. Le renard parvient à tromper le bouc pour s'en sortir, puis l'abandonne au fond.
- **Le Loup et l'Agneau** : Un loup cherche un prétexte pour faire de l'agneau innocent son repas.
- **Le Renard et la Cigogne** : Un renard invite une cigogne à dîner mais lui sert un repas inadapté ; la cigogne se venge de la même manière.
- **Le Loup devenu Berger** : Un loup se déguise en berger pour tromper les moutons, mais il est vite démasqué et puni.

1 Matrice d'incidence

- Construisez une matrice d'incidence pour les termes suivants : "corbeau, renard, cigogne, bouc, loup, berger, fromage, repas, tromper".
- Répondez aux requêtes suivantes :
 1. $\text{renard} \vee \text{loup}$
 2. $\text{renard} \wedge \text{tromper}$
 3. $\text{corbeau} \vee \text{berger}$
 4. $\text{repas} \wedge \neg \text{cigogne}$
- Proposez une requête qui permettrait d'exclure toutes les fables où un personnage est trompé par un autre.

2 Index inversé

- Construisez un index inversé avec tous les mots des trois premiers documents.
- Répondez aux requêtes suivantes :
 1. $\text{loup} \wedge \text{repas}$
 2. $(\text{bouc} \wedge \text{tromper}) \vee \text{fromage}$
 3. $\text{renard} \wedge \text{repas} \wedge \text{tromper}$
- Quel est le besoin en information exprimé par la première requête ?

3 Codage en Python

- Construisez la matrice d'incidence des termes proposés.
- Construisez l'index inversé avec tous les termes du corpus.
- Codez les algorithmes pour répondre aux requêtes proposées avant.

1. Nous nous limiterons à la suppression de la ponctuation et à la mise en minuscules.