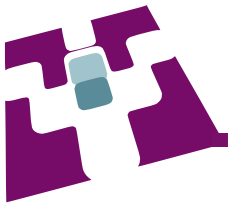
Deft 2011 : Appariement de résumés et d'articles scientifiques

une approche fondée sur des distributions de chaînes de caractères

Gaël Lejeune, Romain Brixtel, Emmanuel Giguet et Nadine Lucas

GREYC, CNRS UMR 6072
6, boulevard du Maréchal Juin
14050 Caen cedex
prénom.nom@unicaen.fr

Juillet 2011



Plan

Réflexion sur le matériau

Extrait d'un article du corpus

```
<?xml version="1.0" encoding="UTF-8" ?>
<corpus>
  <description>
    <titre>Corpus de résumés et d'articles scientifiques DEFT2011</titre>
    <source>~~Consortium Erudit - http ://www.erudit.org/~~</source>
    <disponibilite>Librement réutilisable à des fins non commerciales</disponibilite>
  </description>
  <article>
    <meta>
      <revue>Revue des sciences de l'éducation</revue>
    </meta>
    <texte>
      <titre>Introduction</titre>
      <p>En 1983, Viviane Isambert-Jamati, qui retraçait ... </p>
      ...
    </texte>
  </article>

</corpus>
```

- Pas d'appel à des ressources externes
- Pas d'exploitation des connaissances sur la revue

Réflexion sur le matériau

Extrait d'un article du corpus

```
<?xml version="1.0" encoding="UTF-8" ?>
<corpus>
  <description>
    <titre>Corpus de résumés et d'articles scientifiques DEFT2011</titre>
    <source>~~Consortium Erudit - http ://www.erudit.org/~~</source>
    <disponibilite>Librement réutilisable à des fins non commerciales</disponibilite>
  </description>
  <article>
    <meta>
      <revue>~~Revue des sciences de l'éducation~~</revue>
    </meta>
    <texte>
      <titre>Introduction</titre>
      <p>En 1983, Viviane Isambert-Jamati, qui retraçait ... </p>
      ...
    </texte>
  </article>

</corpus>
```

- Pas d'appel à des ressources externes
- Pas d'exploitation des connaissances sur la revue

Réflexion sur le matériau

Structuration d'un article du corpus

...

<titre>La fermeture d'une ≪ boîte noire ≫**</titre>**

<titre>La Direction de l'Évaluation et de la Prospective et le magistère de Claude Thélot</titre>

** **Jean-Pierre Chevènement, ministre de 1984 à 1986, avait lancé un mot d'ordre, ≪ amener 80 ...** **

...

- Pauvreté des descripteurs de la structure logique du document
- Profondeur et hiérarchie peu exploitables

Réflexion sur le matériau

Structuration d'un article du corpus

...

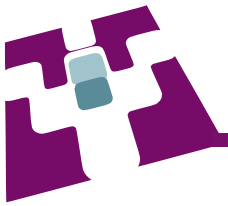
<titre>La fermeture d'une ≪ boîte noire ≫**</titre>**

<titre>La Direction de l'Évaluation et de la Prospective et le magistère de Claude Thélot**</titre>**

Jean-Pierre Chevènement, ministre de 1984 à 1986, avait lancé un mot d'ordre, ≪ amener 80 ...

...

- Pauvreté des descripteurs de la structure logique du document
- Profondeur et hiérarchie peu exploitables



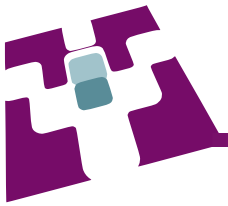
Approche façon Dimeco

Choix des armes

- Un corpus monolingue, une stratégie **multilingue** ;
- Considérer les documents tels qu'ils sont (**avec balisage**).

Dans la même lignée

- Extraction terminologique multilingue (Giguet 2006)
- Détection de plagiat (Brixtel 2011)
- Alignement (Brixtel 2011, Lecluze 2011)
- Veille (Lejeune 2010)



Approche façon DIMECO

Méthodologie

1. Utiliser l'article comme requête sur un corpus de résumés.
   - Chaque article est un **célibataire** auquel on va présenter tous les résumés disponibles, c-à-d ses **prétendants**.
2. Exploiter les phénomènes de recopie entre article et résumés.
   - Pour les relier on va calculer des **affinités** entre le célibataire et ses prétendants, via leurs **chaînes de caractères**.

Sur la base d'un traitement fondé sur les mots

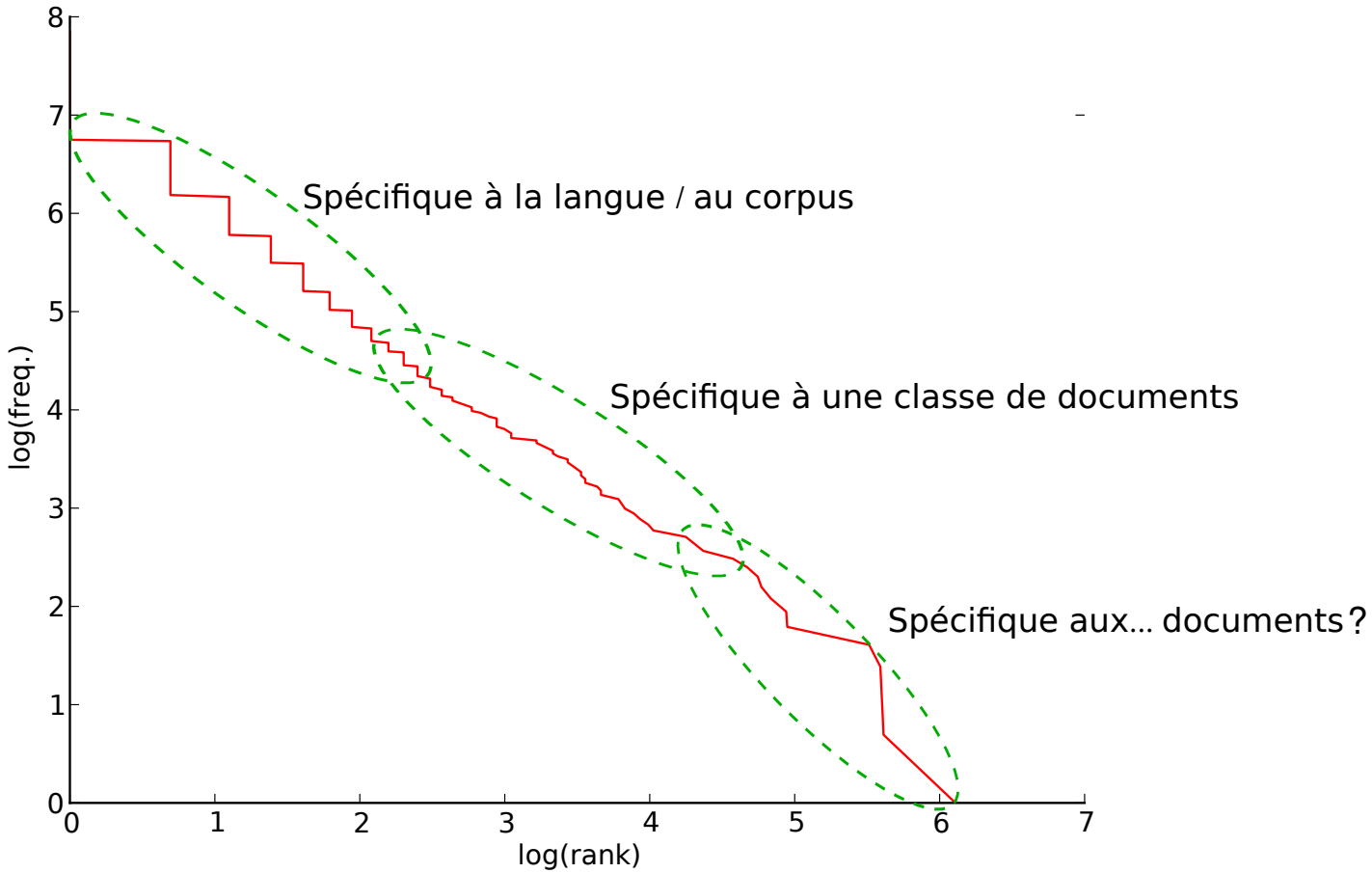


FIGURE: Effectif des mots en fonction de leur rang (échelle logarithmique)

Sur la base d'un traitement multilingue

Exploitation des chaînes de caractères

$rstr_{max}$— chaînes de caractères répétées maximales

1. chaînes de caractères . . .
2. répétées . . .
3. maximales

Algorithmique du texte

- Tableaux de suffixes (KÄRKKÄINEN ET SANDERS, 2003)
- PY-RSTR-MAX — `http://code.google.com/p/py-rstr-max`.

Sur la base d'un traitement
fondé sur les chaînes de caractères

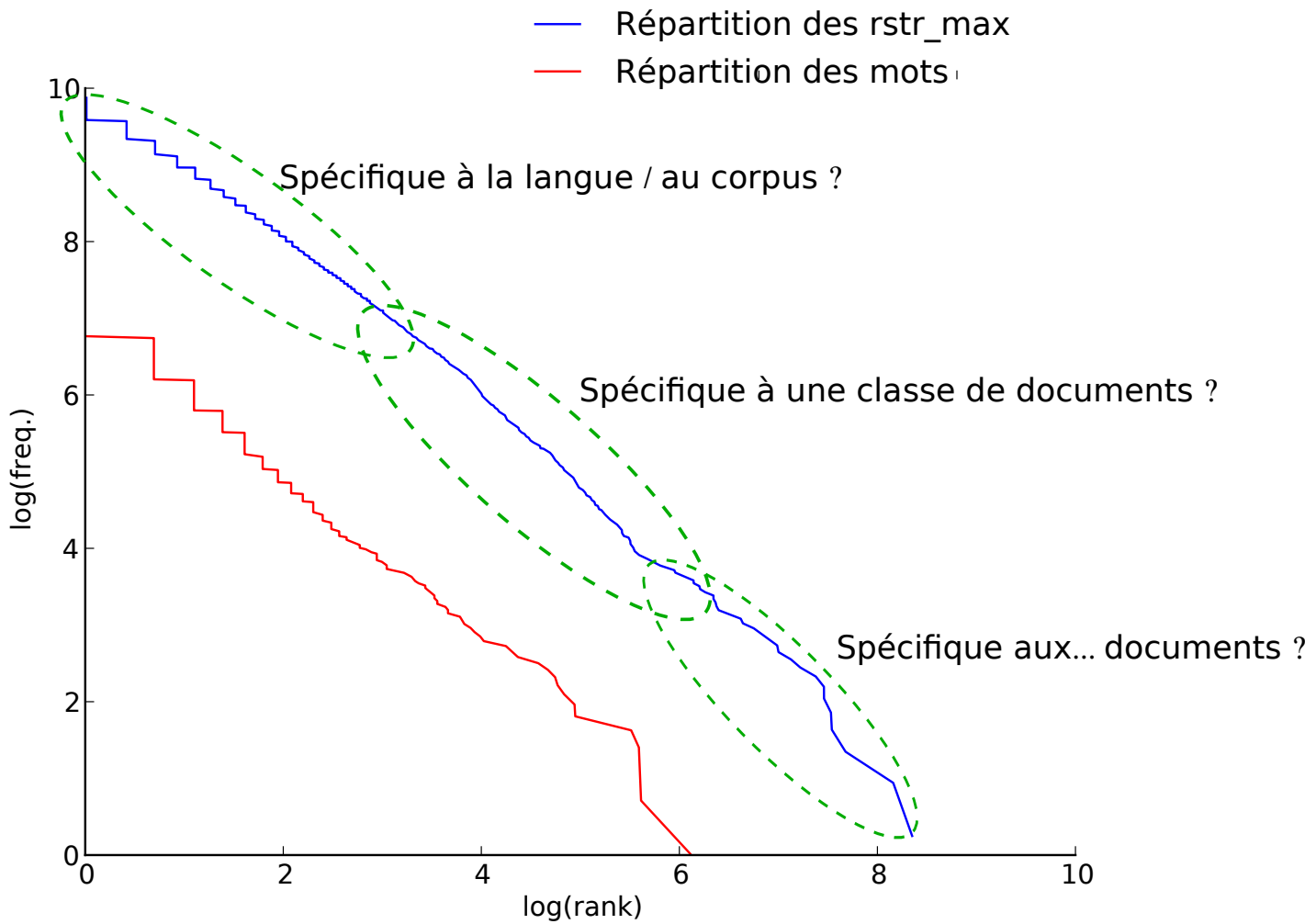


FIGURE: Effectif des $rstr_{max}$ et des mots en fonction de leur rang (échelle logarithmique)

Caractéristiques des $rstr_{max}$ décisives sur les effectifs

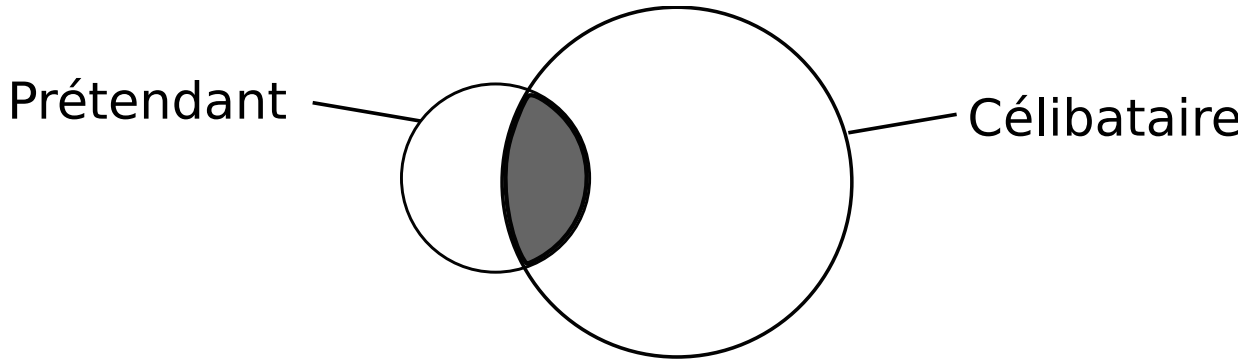


FIGURE: $rstr_{max}$ informative

$rstr_{max}$ informative :

- $rstr_{max}$, donc des chaînes répétées . . .
- . . . mais hapax : à l'intersection d'un célibataire et **d'un seul** prétendant.

Caractéristiques des $rstr_{max}$ décisives sur les effectifs

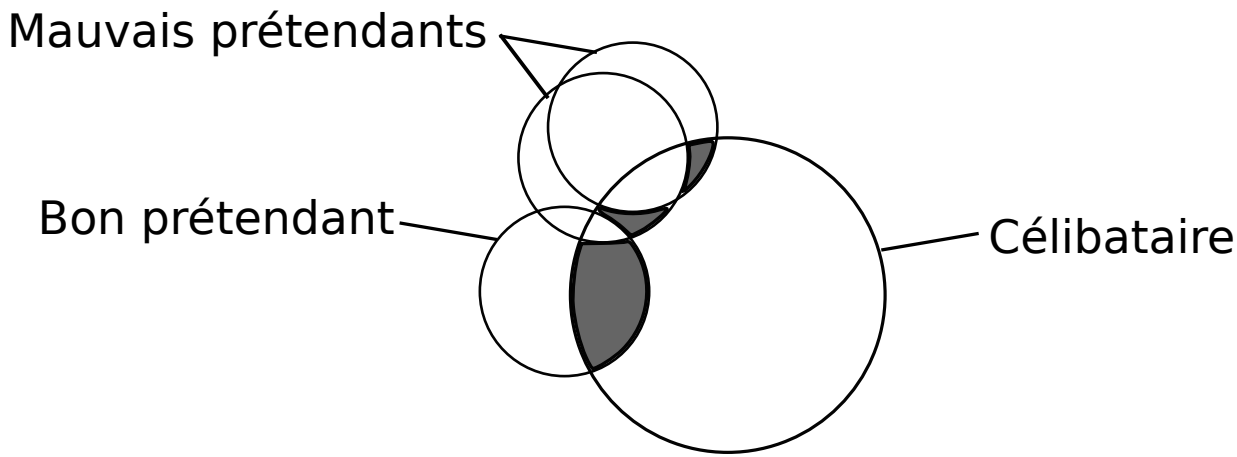


FIGURE: $rstr_{max}$ informative sur plusieurs prétendants

$rstr_{max}$ informative :

- $rstr_{max}$, donc des chaînes répétées ...
- ...mais hapax : à l'intersection d'un célibataire et **d'un seul** prétendant.

Caractéristiques des $rstr_{max}$ décisives
Calcul des affinités

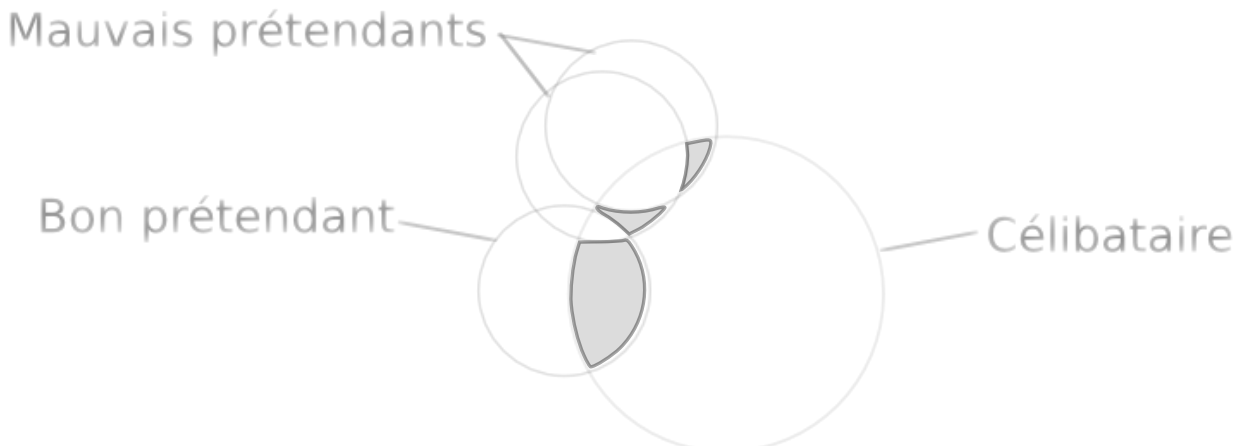


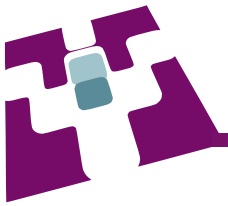
FIGURE: $rstr_{max}$ informative sur plusieurs prétendants

Exploitation des longueurs

- **Calcul de l'affinité entre un célibataire et un prétendant :**

  nombre de $rstr_{max}$ informatives

  \+ taille de la plus longue $rstr_{max}$ informative



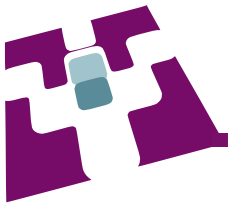
Caractéristiques des $rstr_{max}$ décisives
. . . pour cette tâche !

$rstr_{max}$ informatives

≪ philosophie_politique_d ≫ ≪ s_les_organisations ≫
≪ r_la_reconnaissance_des_ ≫ ≪ des_organisations_internationales ≫
≪ s_les_années_1970 ≫ ≪ _établissements ≫

$rstr_{max}$ non-informatives

≪ resse ≫, ≪ ymbol ≫, ≪ ssib ≫, ≪ est_p ≫, ≪ ns_et ≫, ≪ s_donné ≫, ≪ ntifi ≫,
≪ à_m ≫, ≪ qu'elle ≫, ≪ d'une_co ≫, ≪ e_ap ≫, ≪ les,_ ≫, ≪ s_qua ≫, ≪ ur_l'a ≫



Caractéristiques des $rstr_{max}$ décisives
. . . pour cette tâche !

$rstr_{max}$ informatives

≪ philosophie_politique_d ≫ ≪ s_les_organisations ≫
≪ r_la_reconnaissance_des_ ≫ ≪ des_organisations_internationales ≫
≪ s_les_années_1970 ≫ ≪ _établissements ≫

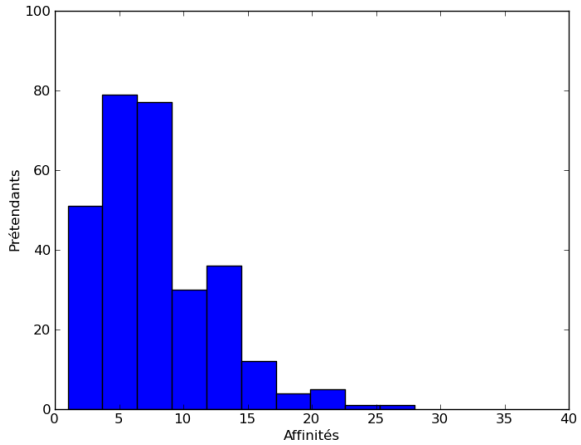
$rstr_{max}$ non-informatives

≪ resse ≫, ≪ ymbol ≫, ≪ ssib ≫, ≪ est_p ≫, ≪ ns_et ≫, ≪ s_donné ≫, ≪ ntifi ≫,
≪ à_m ≫, ≪ qu'elle ≫, ≪ d'une_co ≫, ≪ e_ap ≫, ≪ les,_ ≫, ≪ s_qua ≫, ≪ ur_l'a ≫

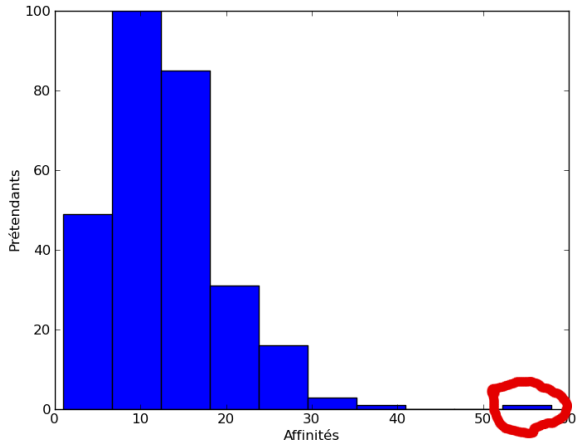
$rstr_{max}$ non-informatives aussi

≪ <?xml version="1.0" encoding="UTF-8" ?>
<corpus>
<description>
<titre>Corpus de résumés et d'articles scientifiques DEFT2011</titre>
<source>Consortium Erudit - http ://www.erudit.org/</source>
<disponibilite>Librement réutilisable à des fins non commerciales</disponibilite>. . . ≫

Sélection du couple (célibataire, prétendant) (article, résumé)



a)



b)

FIGURE: Classement de 300 prétendants selon leur nombre de $rstr_{max}$ informatives avec un célibataire

Une approche différentielle

- a) les prétendants sont regroupés, pas de couple.
- b) un prétendant se détache, un couple existe.
- Si aucune affinité caractéristique ? **Aucun couple de formé**

Optimalité des critères
Longueur

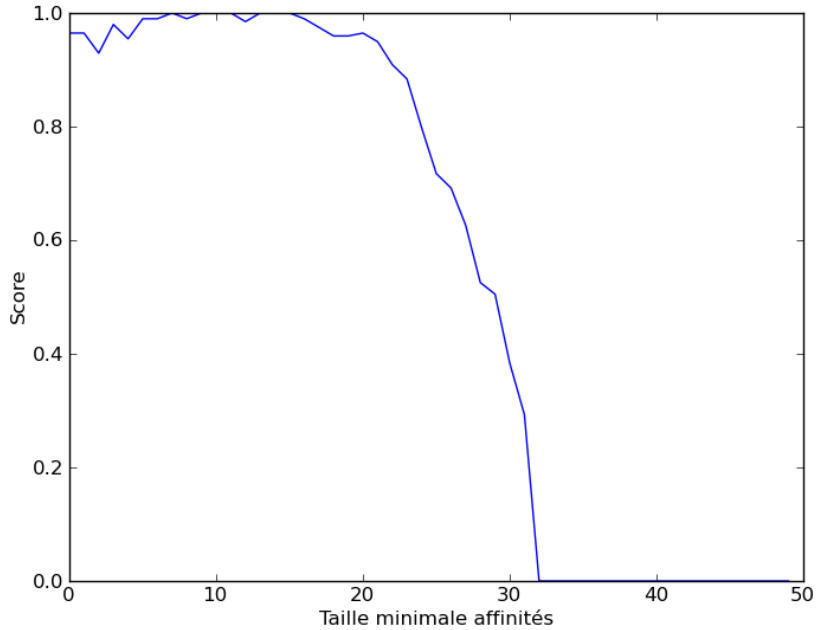


FIGURE: Efficacité du système en fonction d'une taille minimale des $rstr_{max}$

Score optimal : entre 7 et 16 caractères

Optimalité des critères
Effectif

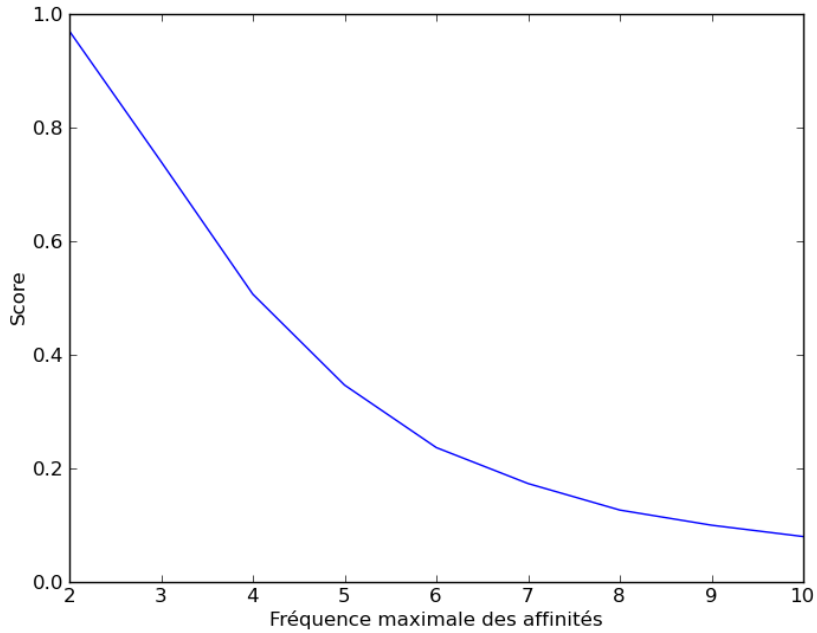


FIGURE: Efficacité du système en fonction de l'effectif des $rstr_{max}$

Score optimal : effectif 2 (l'article et un seul résumé)

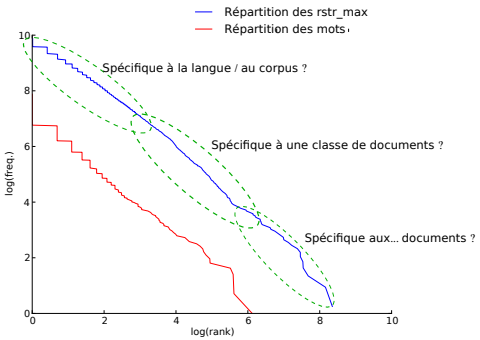
Résultats selon les corpus

	Apprentissage	Test	Apprentissage+test
Piste 1 : Article complet	0.97	1.0	0.98
Piste 2 : Développement	0.96	0.98	0.97

Remarques :

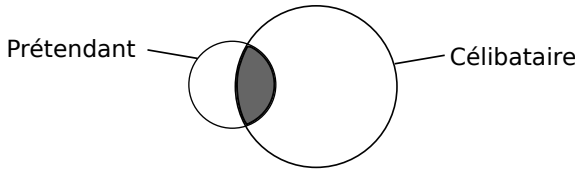
1. La méthode n'est pas sensible à l'ordre de présentation des célibataires
2. La longueur joue un rôle marginal
3. Moins d'affinités sur les articles tronqués

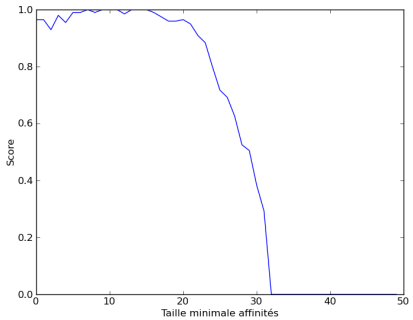
Conclusion



Exploitation de **chaînes de caractères** . . .

. . . avec des critères d'**effectif** et de **longueur** . . .





. . . adaptés à l'**écosystème**.

merci de votre attention