Guidelines : Construction d'un ensemble de données pour l'extraction de *keyphrases*

Florian Boudin

Module X7IT020 - 2013

Les *keyphrases*, ou termes clés en français, sont des termes permettant de caractériser le contenu d'un document. Considéré comme un bref résumé, ils permettent d'organiser et de retrouver plus facilement les documents dans une collection. Cependant, très peu de documents possèdent un ensemble de *keyphrases* associé. De nombreux travaux se sont donc penché sur la problématique de l'extraction automatique *keyphrases*.

La dernière séance de travaux pratiques du module d'introduction au TALN sera consacrée au développement d'un système d'extraction de *keyphrases*. Afin de vous familiariser avec les problèmes d'évaluation, vous allez devoir créer un corpus annoté. L'idée est d'assigner manuellement des *keyphrases* à un ensemble de documents issus de Wikinews [1]. Ces dernières seront par la suite utilisées pour évaluer les performances de votre système.

L'extraction des *keyphrases* est une tâche difficile et subjective. Il s'agit de trouver un ensemble de termes permettant de résumer le contenu d'un document. Chaque document sera annoté par deux personnes différentes afin de pouvoir calculer des indices d'accord inter-annotateurs. Voici la démarche à suivre pour extraire les *keyphrases* à partir d'un document :

1. Lire le document une première fois en entier.
2. Analyser son contenu et en détacher la thématique principale.
3. Générer un ensemble de termes couvrant les aspects principaux du document. La lecture seule de ces derniers doit permettre de se faire une idée du contenu réel du document.

Vous devez extraire au plus 10 termes par document. Dans la plupart des cas, les termes sont des syntagmes nominaux, e.g. Université de Nantes, travail personnel, étudiants fantastiques.

1. `http://fr.wikinews.org/`

Pour vous aider, un exemple de document ainsi que les *keyphrases* associées est donné ci-dessous :

Vatican : l'accès à la chapelle Sixtine pourrait être limité
L'accès à la chapelle Sixtine, l'une des merveilles des palais pontificaux du Vatican, pourrait prochainement être limité dans le but de protéger ses fresques demi-millénaires de dégâts commis par un afflux massif de visiteurs. Les fresques mondialement connues, peintes par Michel-Ange au début du XVIe siècle, sont en effet victimes de leur succès, la chapelle Sixtine étant l'un des lieux les plus visités au monde avec une affluence quotidienne de quelque 20 000 personnes.

- Annotateur 1 : Vatican, chapelle Sixtine, accès limité, affluence, protection des fresques
- Annotateur 2 : chapelle Sixtine ; accès limité ; fresques demi-millénaires ; victimes de leur succès

Les formes fléchies des termes pouront être ramenées à leurs lemme (e.g. limité et limitation). Il est possible que vous souhaitiez assigner un terme qui n'est pas présent dans le document (e.g. accès restreint). Cela n'est pas interdit mais il faut, dans la mesure du possible, l'éviter.

Les corpus et la répartition des documents aux étudiants sont disponibles sur le site web du cours. Vous devez avoir envoyé vos annotations avant la dernière séance de travaux pratiques. Le format du fichier à envoyer est le suivant :

```
43971.html keyphrase1;keyphrase2;keyphrase3
44046.html keyphrase1;keyphrase2;keyphrase3;keyphrase4
44145.html keyphrase1;keyphrase2;keyphrase3;keyphrase4
44231.html keyphrase1;keyphrase2
```