SylNews, un agréfilter multilingue

Olivier Hamon[1] Kévin Espasa[1] Sara Quispe[1]
(1) Syllabs, 35 rue Chanzy, 75011 Paris, France
hamon@syllabs.com, espasa@syllabs.com, quispe@syllabs.com

RÉSUMÉ

Depuis plusieurs années, Syllabs intègre de nombreux composants au sein d'un agréfilter, utilisant des technologies d'extraction d'information développées en interne et dans un contexte multilingue. Originellement conçu pour agréger des contenus issus de la presse, SylNews peut être utilisé à des fins de veille, pour explorer des contenus, ou pour identifier d'une manière plus globale les sujets chauds de l'ensemble ou d'une partie des contenus stockés.

ABSTRACT

SylNews, a multilingual aggrefilter.

For a few years, Syllabs has been integrating a number of components in an aggrefilter, using information extraction technologies developed internally in a multilingual context. Originally designed to aggregate news contents, SylNews may be used to watch news, explore contents, or identify hot elements inside either part or all of the contents to be explored.

MOTS-CLÉS : agréfilter, extraction d'information, clustering
KEYWORDS: aggrefiltering, information extraction, clustering

1 Introduction

En 2015, Syllabs a développé un agréfilter pour le média Les Échos. Après une refonte globale, cet outil de filtrage et d'agrégation (c.-à-d. agréfilter) de contenus n'a cessé d'évoluer, guidé par les besoins de nos clients. Ces évolutions incluent l'apport de nouveaux composants, le passage au multilingue, ou encore le développement d'une interface spécifique en cours de développement.

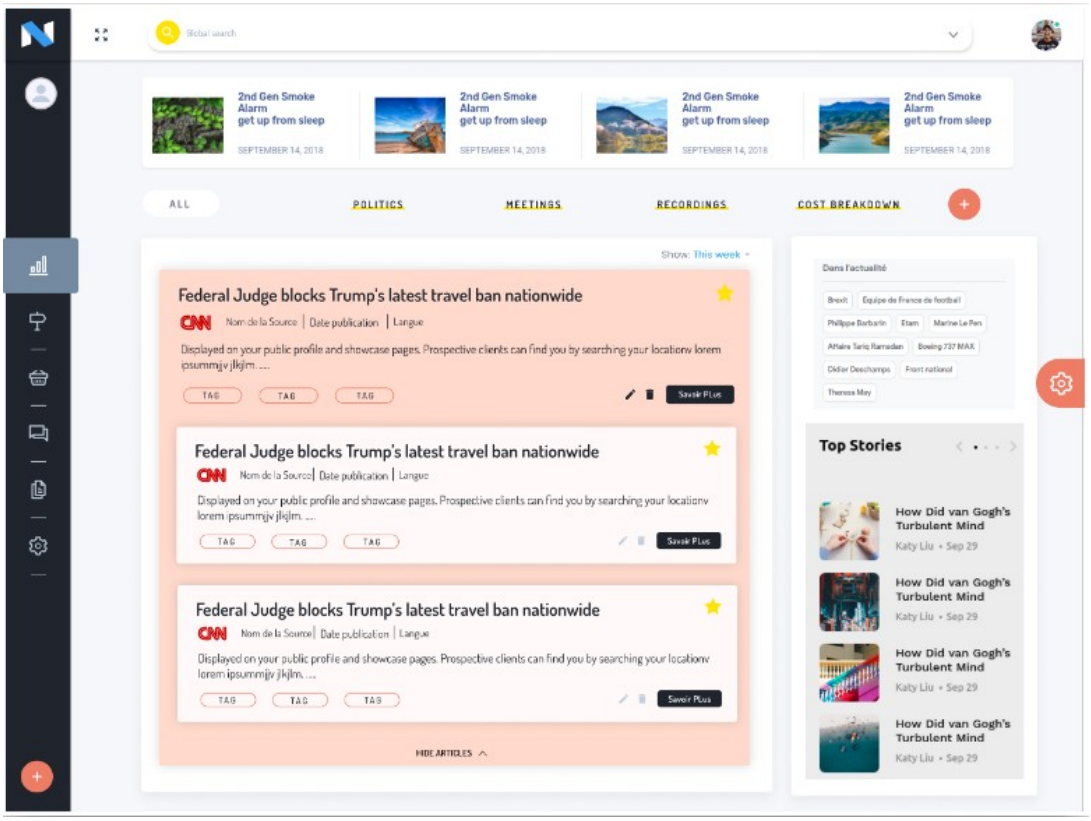

FIGURE 1: Page d'accueil d'un SylNews

SylNews permet principalement de :

- collecter des contenus à partir de sources locales ou en ligne ;
- fournir différents types d'annotations automatiques pour chaque contenu (entités nommées, sujets, thèmes, etc.) ;
- filtrer par thématique ;
- regrouper les contenus par sujet ou par événement ;
- appliquer les technologies précédentes à plusieurs langues.

2 Technologies et approche de SylNews

La première étape lors du déploiement d'un SylNews concerne la collecte des contenus. Cela peut être réalisé *via* des sources en ligne comme des flux RSS, ou bien des fichiers locaux.

Ces contenus sont pré-traités (lemmatisation, tokenisation, statistiques, etc.), puis différentes annotations sont appliquées. Parmi elles, nous pouvons citer la reconnaissance d'entités nommées (Ma, 2011), la catégorisation IPTC[1] utilisant une méthode d'amorçage (Baroni & Bernardini, 2004), le lien vers des articles Wikipédia, etc. Ces technologies d'extraction d'informations sont régulièrement adaptées en fonction des besoins des projets successifs.

Ensuite, les pré-traitements et annotations sont utilisés afin de regrouper des contenus à l'aide de l'algorithme DBSCAN (Ester et al., 1996) qui permet de regrouper des documents présents dans un espace vectoriel.. Nous avons également ajouté le regroupement multilingue des contenus sur cinq

[1] https://iptc.org/standards/photo-metadata/iptc-standard/

langues (De, En, Es, Fr et It) en utilisant comme pivot les entités nommées extraites et leur traduction *via* Wikipédia.

Chaque nouveau contenu est ainsi annoté, et l'ensemble des contenus sont régulièrement regroupés, d'après une fréquence et un volume donné. Les regroupements obtenus étant eux-mêmes annotés par combinaison, il nous est possible de recueillir des caractéristiques notables, telles que des « sujets chauds » sur l'actualité.

Le déploiement d'un SylNews est flexible et automatisé. Il permet ainsi d'effectuer les traitements selon différentes sources, projets ou thématiques.

3 Utilisation

SylNews a été utilisé par le passé et encore aujourd'hui dans le cadre de plusieurs projets de recherche et clients. Son utilisation principale concerne le regroupement de contenus, et surtout l'analyse de ces regroupements pour obtenir les « sujets chauds » de l'actualité. C'est également un bon moyen pour disposer, rapidement, d'annotations sur des volumes de données importants, telle que pourrait le faire une plateforme d'annotation.

Références

BARONI M., BERNARDINI S. (2004). BootCaT: Bootstrapping corpora and terms from the web. Actes de *LREC 2004*.

ESTER M., KRIEGEL H.-P., SANDER J., XU X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Actes de *KDD-96*.

MA J., MOUNIER M., BLANCAFORT H., COUTO J., DE LOUPY C. (2011). LOL: Langage objet dédié à la programmation linguistique. Actes de *TALN 2011*.