TextGain API

textgain	**Service d'analyse de textes**	
	Système d'exploitation :	Multiplateforme
	Développé par :	TextGain
Modèle commercial gratuit et payant	Personne de contact :	Katy.Fokou@Smals.be

Fonctionnalités

TextGain est une spin-off du groupe de recherche de l'université d'Anvers qui développe des services d'analyse de textes en real-time. La société propose des *Application Programming Interface* (API) type REST pour le traitement automatique du langage naturel/ *Natural Language Processing* (NLP) et met à disposition les fonctionnalités suivantes :

- Profilage. Les API permettent de déterminer si l'auteur·e du texte est adolescent·e (-25 ans) ou adulte (+25 ans) ainsi que son genre (masculin ou féminin), sa personnalité (extraverti·e ou introverti·e) et son niveau d'éducation
- Analyse de sentiment
- Identification de la langue et du genre (blog, journal…)
- Extraction de concepts. Cela inclut les termes les plus importants: mots, phrases clés, noms, dates, localisations, organisations…
- Géocodage. Latitude et longitude, population, type (ville, pays,…)
- Analyse lexicale. Lemmatisation, *Part-of-Speech Tagging (POS tagging)* c.-à-d. l'association des mots dans une phrase à une catégorie grammaticale (verbe, pronom, nom,…)
- Lisibilité. Utilisation de la voie passive, nombre de syllabes, utilisation de traits d'union

En termes de confidentialité de données, TextGain ne fait pas appel à des tiers-services (cloud) et les données envoyées sur leurs serveurs sont conservées le temps de leur traitement. De plus, pour le traitement des données sensibles, la société propose ses services sur site.

Conclusions & Recommandations

Il existe de nombreux outils de traitement automatique du langage naturel mais peu d'entre eux supportent le Néerlandais. Les API proposées par TextGain sont très efficaces pour les textes en néerlandais et en français. Ils s'intègrent facilement dans une application et le traitement des données est rapide. Outre les performances techniques, les API TextGain peuvent être déployées sur site pour rencontrer les exigences des clients en termes de confidentialité des données.

TextGain prévoit de déployer fin 2018 une API pour extraire les *Named Entities*.

Tests & Résultats

Les documents utilisés pour notre test des API sont des arrêts de la cour de cassation publiés sur le web ; ces documents juridiques sont longs et rédigés dans un jargon propre au métier. Huit textes (3 en français et 5 en néerlandais) qui constituent des segments de texte d'arrêts de cassation ont été utilisés pour les tests.

Pour les huit documents, la **langue** est correctement identifiée par contre les **genres** détectés, le blog, la revue et l'info, n'ont aucun lien avec les documents utilisés. Ceci est peu surprenant étant donné que les documents juridiques sont un genre particulier pour lequel l'API n'est pas optimisée.

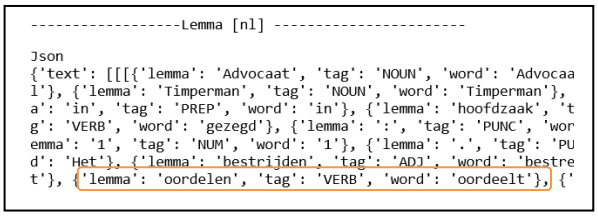

Figure 2. Exemple de lemmatisation en néerlandais. Dans l'encadré en orange, le champ "lemma" contient l'infinitif "oordelen" du verbe "oordeelt".

Les fonctionnalités d'analyse lexicale sont importantes car de nombreux pipelines NLP impliquent de *POS tagger* et/ou lemmatiser les textes à traiter au préalable. Les tests sur les huit textes ont montré que la **lemmatisation** (**Figure 1**) et le **POS tagger** (**Figure 2**) (seulement disponible en néerlandais) de TextGain sont très performants même sur des documents dont l'OCR est de piètre qualité. Le **découpage syllabique** est par contre moins bon en français.

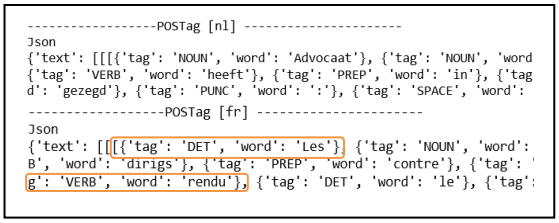

Figure 1. Exemple de POS tagging en néerlandais et en français. Dans l'exemple en français, le mot "Les" est de type "DET" (déterminant) tandis que le mot "rendu" est de type "VERB".

En appliquant l'API de **géocodage** sur le texte brut, les localités ne sont pas toujours détectées. Il est alors nécessaire de réduire le texte aux seuls noms en utilisant le POS tagger avant d'appliquer l'API de géocodage. Cependant, le système détecte toutes les villes homonymes des noms extraits (voir l'exemple *Michel* dans **Figure 3** ci-dessous).

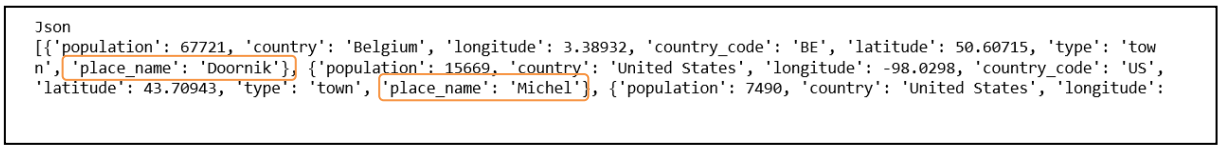

Figure 3. Exemple de géocodage, Doornik est une ville dans le texte tandis que Michel est le nom d'une personne et non d'une ville.

Enfin l'API d'**extraction de concepts** est efficace dans les deux langues pour retrouver les termes les plus pertinents. Il est important de noter que l'extraction de concepts est différent du *Named Entity Recognition*, les noms, dates et localités ne sont détectés que s'ils sont pertinents.

Conditions d'utilisation & Budget

TextGain propose une formule gratuite (300 requêtes/jour sans support) et une formule payante dont les prix varient de 19 euros par mois (1 000 requêtes/jour) à 1 300 euros par mois (100 000 requêtes/jour).