Mise en place d'un système d'acquisition semi-automatique d'un corpus de données hétérogènes (images et textes) : application à la problématique de la sécurité alimentaire en Afrique de l'Ouest

Camille Schaeffer

Mise en place d'un système d'acquisition semi-automatique d'un corpus de données hétérogènes (Images et Textes) – Application à la problématique de la sécurité alimentaire en Afrique de l'Ouest

SCHAEFFER
Camille

Sous la direction de Olivier KRAIF

Réalisé au sein de l'organisme CIRAD

Sous la direction de Roberto INTERDONATO et Mathieu ROCHE

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 Sciences du Langage et FLE – Orientation Professionnelle - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2018-2019

Mise en place d'un système d'acquisition semi-automatique d'un corpus de données hétérogènes (Images et Textes) – Application à la problématique de la sécurité alimentaire en Afrique de l'Ouest

SCHAEFFER
Camille

Sous la direction de Olivier KRAIF

Réalisé au sein de l'organisme CIRAD

Sous la direction de Roberto INTERDONATO et Mathieu ROCHE

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 Sciences du Langage et FLE – Orientation Professionnelle - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2018-2019

Remerciements

Je tiens tout d’abord à remercier mon tuteur universitaire Olivier Kraif pour son amabilité, ses remarques et critiques lors la réalisation de ce mémoire.

Je remercie fortement Roberto Interdonato et Mathieu Roche, mes deux tuteurs CIRAD, pour avoir partagé leurs connaissances et expériences, et de m’avoir permis de participer à un tel projet.

Je remercie également Maguelonne Teisseire et Elodie Maître d'Hôtel pour leur expertise sur le domaine de la sécurité alimentaire, ainsi qu’à Hugo Deléglise et à Jacques Fize pour l’aide et les conseils apportés.

Et enfin un grand merci aux stagiaires de la salle 262 pour tous les bons moments passés à Montpellier.

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : SCHAEFFER

PRENOM : Camille

DATE : 05/09/19

Sommaire

Introduction

La famine en Afrique est l'urgence la plus grave depuis la Seconde Guerre Mondiale selon les Nations Unies. Au cœur de l'Afrique de l'Ouest, le Burkina Faso est un pays à faible revenu, où l'agriculture représente 32% du produit intérieur brut et emploie 80% de la population. Bien qu'il y ait eu une amélioration dans ce secteur au cours des dernières années, ce pays est affecté par la sécheresse, le changement climatique, le terrorisme et une démographie croissante. Une partie de la population n'arrive pas à satisfaire ses besoins alimentaires, nous parlons alors d'insécurité alimentaire.

Des systèmes d'alerte précoce sur la gestion des risques liés à la sécurité alimentaire sont mis en place pour informer les acteurs afin d'adapter et améliorer le développement de l'agriculture. Les systèmes suivent et analysent notamment la situation agricole et la situation pluviométrique grâce aux images satellitaires, ainsi que la situation économique (par exemple le prix du marché alimentaire) grâce aux données quantitatives. Cependant, ces systèmes d'alerte précoce intègrent peu de données textuelles dans leur traitement. Avec la révolution du numérique, ces données, présentes sur des sites d'actualités ou sur les réseaux sociaux, sont de plus en plus abondantes et accessibles sur Internet. Elles peuvent être utilisées afin de rechercher des informations spécifiques et apporter certaines observations sur des événements. Le domaine qui se concentre sur l'extraction des connaissances à partir du texte est généralement appelée fouille de texte.

Le présent travail est d'appliquer des traitements de fouille de texte consacrés à la problématique de la sécurité alimentaire au Burkina Faso, en analysant des articles traitants de cette thématique, extraits de sites d'actualités de ce pays. Les informations à identifier dans les articles sont notamment le lieu et la temporalité (informations spatio-temporelles). Ces informations présentes dans les journaux ne décrivent pas parfaitement la situation géographique, mais elles apportent une information thématique complémentaire par rapport aux données présentes dans les systèmes d'alerte précoce, afin de mieux anticiper et informer en temps quasi réel les potentiels problèmes de sécurité alimentaire au Burkina Faso.

Ce mémoire sera structuré en trois parties : la première consistant en l'acquisition d'un corpus d'articles de sites d'actualités, la deuxième développant l'identification d'articles pertinents au thème de la sécurité alimentaire au Burkina Faso et la troisième présentant l'identification d'entités spatio-temporelles et leur mise en lien avec les données des systèmes.

Environnement de travail

Ce mémoire présente le travail effectué au Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), dans le cadre du master Industries de la Langue, à l'Université Grenoble Alpes.

Le CIRAD mène des activités relevant des sciences du vivant, des sciences sociales et des sciences de l'ingénieur appliquées à l'agriculture, à l'alimentation, à l'environnement et à la gestion des territoires et comprend 33 unités, dont l'UMR TETIS (Territoires, environnement, télédétection et information spatiale), hébergée à la maison de la télédétection à Montpellier, où le stage a été effectué.

L'UMR TETIS est composée de plusieurs équipes, dont l'équipe SISO (Système d'Information Spatialisée : modélisation, extraction et diffusion des dOnnées et connaissances), équipe qui a abrité le stage, et dont les tuteurs sont Roberto Interdonato et Mathieu Roche.

Le stage s'insère dans le cadre d'une thèse réalisée par Hugo Deléglise, également à l'UMR TETIS, concernant le renforcement des systèmes de sécurité alimentaire en Afrique de l'Ouest, grâce à la mise en relation de données hétérogènes (données quantitatives, données satellitaires, données textuelles, etc.). La sécurité alimentaire, a été définie comme telle au Sommet mondial de l'alimentation à Rome, en 1996 : *"La sécurité alimentaire existe lorsque tous les êtres humains ont, à tout moment, un accès physique et économique à une nourriture suffisante, saine et nutritive leur permettant de satisfaire leurs besoins énergétiques et leurs préférences alimentaires pour mener une vie saine et active".*

Le projet de la thèse est de développer un outil de prédiction des pénuries alimentaires rapide, fiable et gratuit, en concevant un algorithme de *Machine Learning* supervisé. Avec des données antérieures (données quantitatives et satellitaires relatives à la sécurité alimentaire) et des indices de sécurité alimentaire pour classifier ces données (par exemple ; indice 1 : sécurité alimentaire faible, indice 5 : sécurité alimentaire élevée...), l'algorithme apprend un modèle sur ces données d'entraînement. Puis des données actuelles sont soumises (données de prédiction) au modèle afin d'obtenir une prédiction sur l'état de la situation alimentaire actuel.

L’objectif du stage est de traiter la partie des données textuelles afin de fournir ces données antérieures au système pour renforcer l’apprentissage du modèle. Nous pourrons ainsi analyser si les données textuelles apportent des informations complémentaires aux systèmes d’information de sécurité alimentaire.

Pour cela, les missions du stage étaient de produire un corpus via un processus de récolte des données en ligne, d’adapter des techniques de fouille de textes pour identifier les thèmes de la sécurité alimentaire dans les documents du Burkina Faso et de lier les informations spatio-temporelles aux données présentes dans les systèmes d’alerte précoce.

Chapitre 1

-

Acquisition du corpus

Le premier chapitre présente l'acquisition du corpus de données textuelles. Les données à extraire sont des articles de sites d'actualités en ligne. Nous automatisons la collecte de ces données afin de les exporter en format csv. Le corpus sera alors exploitable pour les traitements des deux prochains chapitres.

Contexte

1. Définition corpus

Le dictionnaire Le Trésor de la Langue Française, définit un corpus comme un "ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique".

Les corpus existent depuis longtemps ; nous pouvons mentionner le corpus Frantext composé de textes littéraires du 19ème et 20ème siècle (254 millions de mots), ou encore le corpus 88milSMS constitué de plus de 88 000 SMS en 2011, afin de mieux comprendre comment la langue française évolue avec les usages du numérique.

Nous définissons un corpus comme un ensemble de textes regroupés en fonction d'objectifs précis, et pouvant être étudié et exploité avec des outils et méthodes de traitement du texte.

2. Données à disposition

Il existe sur le web plusieurs site d'actualités du Burkina Faso, dont les principaux sont *l'Observateur Paalga, Fasozine, Lefaso.net, Sidwata, Le Pays, Mutations, Aujourd'hui au Faso, Ouaga24, Burkina24, Faso amazone, Laborpresse, Le Journal du jeudi.*

Nous allons traiter trois de ces sites d'actualités en français ayant une bonne fréquence d'articles et de sujets variés :

l'observateur paalga[1], un quotidien d'informations générales (politique, société, arts & culture, sports, éditorial, technologie, annonces, opinions). Le site web a été créé en 2002, mais les articles datant d'avant 2012 ne sont pas accessibles sur le site. Nous n'avons accès qu'aux articles du 24 septembre 2012 à aujourd'hui, ce qui fait un corpus de 3 657 articles (peu d'articles sont publiés au début mais à partir de 2014 le nombre de parution d'articles augmente).

[1] http://www.lobservateur.bf/

Burkina24[2], un quotidien d'informations générales (monde, politique, économie, sport, culture, hi-tech, Burkina). Ce journal étant assez général, nous avons seulement extrait les articles apparaissant dans la sous-rubrique "Société" de la rubrique "Burkina". Le site web a été créé en 2011, et nous avons 7 170 articles datant du 01 juin 2011 à aujourd'hui.

Lefaso.net[3], un quotidien d'informations générales (politique, société, économie, coopération, culture, portraits, multimédia, sport, international). Ce journal étant très volumineux, nous avons extrait les articles présents dans la rubrique "Société". Le site web a été créé en 2003 et nous avons 23 229 articles datant du 26 octobre 2003 à aujourd'hui.

3. *Extraction des données textuelles*

Afin d'analyser les articles, nous devons les collecter via le web, en vue de constituer un corpus exploitable. Les articles sont par la suite convertis en format texte ou csv, pour être traités par des scripts informatiques, de façon à extraire les informations spatio-temporelles. Pour cela, nous devons parcourir toutes les pages d'un site d'actualités contenant un article, et collecter les informations de ces pages. Ce procédé est possible grâce au système *WordWideWeb* (la toile) et à Internet.

Les sites web sont écrits en HTML (*HyperText Markup Language*), langage de balisage spécifique pour la création de pages web (Figure 1). Ce langage permet de mettre en forme le contenu des pages web, tels que du texte, des liens, des images etc. Ces pages HTML constituent le code source (Figure 2) des pages web, et sont interprétables par les logiciels de navigation pour les afficher aux utilisateurs. Lorsque nous naviguons sur le web, nous visualisons la page web telle que l'auteur l'a écrite et veut que l'utilisateur la voit (Pour apercevoir le code source d'une page, clic droit sur la page et cliquer sur *code source de la page* sur *Firefox* ou *Google Chrome* ou Ctrl+U sous *Windows*.).

Les codes sources des pages web permettent d'avoir des données semi-structurées grâce au balises contenant les informations de l'article comme le titre, la date de parution et le texte de l'article. Nous pouvons donc extraire le contenu des balises <titre><date> et <texte> des articles présents dans le code sources des pages web, pour récupérer les données textuelles voulues afin de pouvoir appliquer des programmes informatique dessus.

[2] https://www.burkina24.com/
[3] https://lefaso.net/

Figure 2 : Exemple page web d'un article

Figure 1 : extrait code source d'une page web

Outils et méthodes de l'état de l'art

Pour parcourir toutes les pages d'un site web et en extraire automatiquement des données spécifiques du code source, deux techniques sont nécessaires : le *crawler* (robot d'indexation) et le *scraper* (extraction de contenus).

Ces techniques aspirent des données textuelles présentes sur Internet, écrites par des auteurs. Dans le cadre de projets de recherche scientifique, l'exploration de textes est possible pour des organismes **p**ublics de recherche et institutions engagés :

Code de la propriété intellectuelle, article L122-5[4] :

> *"10° Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. Un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en œuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites ; ces fichiers constituent des données de la recherche."*

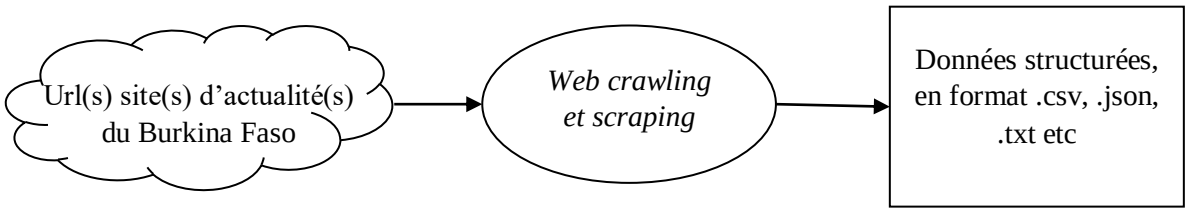


Figure 3 : Acquisition des données en ligne

[4] https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=0245D3957F8A4D07CD24F3B433858AC5.tplgfr44s_2?idArticle=LEGIARTI000037388886&cidTexte=LEGITEXT000006069414&categorieLien=id&dateTexte

1. *Web Crawling*

Le *crawler* est une technique permettant d'explorer le web (comme une toile d'araignée) en parcourant toutes les pages. L'algorithme a en entrée une ou plusieurs url(s), et consulte ces pages pour notamment retrouver des liens d'autres pages.

Avec l'expansion des données sur Internet, de nombreux logiciels pour *crawler* (Xenu, Botify…) ont été créés. Ils permettent de parcourir un site sans coder, en renseignant au logiciel l'adresse du site web à *crawler*. Cependant, ces logiciels sont généralement utiles pour connaître la structure du site et extraire des informations, comme le temps de réponse du serveur, des problèmes d'arborescence etc. Nous voulons parcourir des pages web afin d'extraire du contenu assez précis grâce au code source de celles-ci, nous créerons donc notre propre crawler avec Python et des librairies Python spécifiques aux techniques de *crawling* et de *scraping*, comme Pyspider, MechanicalSoup ou encore Scrapy.

A noter que d'une manière générale, les sites ne permettent pas d'extraire la totalité de leurs pages avec un programme informatique, si nous envoyons trop de requêtes, nous pouvons nous faire blacklister du site.

2. *Web Scraping*

Le scraper est une technique permettant d'extraire du contenu spécifique du code source de pages web. Le contenu à extraire sur les pages est seulement l'article en lui-même et non toute la page de l'article (qui contient l'article mais également d'autres informations non pertinentes pour notre analyse comme des commentaires ou les onglets du site). Pour sélectionner le contenu spécifique à extraire, nous utilisons les informations présentes dans les codes sources qui permettent d'avoir une structuration : soit les sélecteurs css ou récupérer le contenu des balises html, comme ci-dessous.

```
<titre>Ceci est le titre</titre>
<date>Voici la date<date/>
<texte>Et pour finir le texte !<texte/>
```

Figure 4 : exemple page html

Avec un outil de *scraper*, nous devons indiquer le nom de la balise pour en extraire son contenu Unicode, par exemple pour obtenir le titre *Ceci est le titre*, la ligne d'en code sera : outilscraper.titre (des exemples plus précis sont mentionnés dans *Notre approche,2.1 BeautifulSoup*).

Comme les outils du *crawler*, des logiciels permettent de *scraper* des sites web sans utiliser des scripts (Import.io, Web Scraper, Data Scraper), mais nous utiliserons également un programme plus spécifique et précis sous Python avec des librairies spécifiques au processus du scraper, comme BeautifulSoup, Scrapy ou encore Selenium.

Les outils du scraper peuvent être confrontés à des problèmes si le code HTML est mal formaté et/ou est différent sur le même site.

Approche

Nous devions parcourir beaucoup de pages des sites d'actualités du Burkina Faso, nous adopterons une méthode pouvant éviter les limites des techniques de parcours (*crawler*) et d'extraction de contenus (*scraper*). Nous n'utilisons au final aucun outil de *crawler* listé ci-dessus : nous emploierons des *headers* pour parcourir les sites d'actualités, la librairie *Requests* permettant d'obtenir le code source de page http. Nous utiliserons la librairie *BeautifulSoup* pour la technique de *Scraper*. La démarche est décrite ci-dessous, en trois phases.

1. Parcours du web

Headers

Nous devions visiter beaucoup de pages web avec un programme Python (4 500 pages pour l'*Observateur paalga*, 250 pages pour *Burkina24*, 11 000 pages pour *Lefaso.net*). Pour ne pas avoir un nombre de page limité afin de ne pas encombrer le site, et se faire *blacklister* par celui-ci, nous avons utilisé des *headers*[5]. Ceux-ci permettent de *crawler* le site en tant que requête navigateur (Windows, Linux, Mac) et non en tant que programme informatique Python. En ajoutant ces *headers* au script de base du *crawler*, le script exploite les fichiers de données (fichier .dat) présents sur le github mentionnés dans la note de page n°5., permettant d'émuler le défilement infini des pages et de ne pas être expulsé par le site.

Obtention des codes sources des pages

Le code source des pages http est obtenu grâce à la librairie HTTP Python appelée Requests, avec le code *requests.get("url").text*. (.text : retourne le contenu en unicode) La librairie Requests permet d'envoyer des requêtes http et d'accéder aux pages web via Python.

[5] https://github.com/keitakurita/twitter-past-crawler/tree/master/src/twitterpastcrawler

Parcours des urls des sites web

Pour parcourir toutes les pages web des sites d'actualités, nous devons d'abord examiner comment les urls des pages sont connectées entre elles.

Tableau 1 : Parcours des urls des sites L'Observateur Paalga, Burkina24, Lefaso.net

Journal	**Url principale**	**Exemple url article**
Observateur paalga	*http://www.lobservateur.bf/index.php?*	*http://www.lobservateur.bf/index.php?option=com_k2&view=item&id=21*
Burkina24	*https://www.burkina24.com/category/actualite-au-burkina-faso/societe/*	*https://www.burkina24.com/2019/08/12/journee-internationale-de-la-jeunesse-500-plants-mis-en-terre-dans-la-region-du-centre/*
Lefaso.net	*https://lefaso.net/spip.php?rubrique4*	*https://lefaso.net/spip.php?article91488*

Nous cherchons à obtenir l'url de chaque article pour extraire le contenu voulu du code source. Après analyse des urls des journaux (cf tableau 1), nous définissons que le *crawling* des articles se fera de différentes manières selon le site d'actualités :

- *Observateur paalga*

Pour parcourir les urls des articles de ce site, le script a en entrée l'url principale du site (*http://www.lobservateur.bf/index.php?*). Afin d'accéder aux articles, nous devons envoyer des informations à la page php. Pour cela nous écrivons à la suite de l'url les paramètres et les valeurs de paramètres de l'url pour composer l'url complète des articles (*option="com_k2", view="item",id=page,* soit :

*http://www.lobservateur.bf/index.php?option=com_k2&view=item&**id=0***).

Puis nous parcourons les codes sources des pages grâce aux headers et à la librairie *Requests* mentionnés ci-dessus avec une boucle pour incrémenter la variable page de 0 à 4 500, afin de générer les urls des articles. (Lors de la création du script, nous avons constaté qu'il y avait un peu moins de 4 000 articles sur le site. Nous avons mis 4 500 pour avoir une marge).

- *Burkina24*

Pour ce journal, les id des articles n'étant pas écrits dans l'url, nous ne pouvons pas procéder à la même méthode que l'Observateur paalga.

Pour parcourir les urls des articles de ce site, le script a en entrée l'url principale du site (*https://www.burkina24.com/category/actualite-au-burkina-faso/societe/*). Les urls de articles n'ont pas de paramètres comme le site de l'Observateur paalga, nous n'avons donc pas de paramètres à ajouter à la suite de l'url, et nous ne pouvons parcourir les articles avec leur id respectif à partir de l'url.

A partir de l'url de base, nous devons parcourir toutes les pages d'indexation du site en ajoutant la variable page pour incrémenter cette variable de 0 à 250 :

https://www.burkina24.com/category/actualite-au-burkina-faso/societe/**page/{0}**.

Pour chaque page parcourue, le script recherche les urls des articles qui, après analyse du code source des pages, sont toujours situées dans une balise nommée <article>. Ces urls seront collectées dans un fichier json, pour que le code puisse enregistrer les codes sources ciblées des pages de ces articles. Puis, nous procédons à un nettoyage de ce fichier afin de ne pas avoir des urls en doublons.

- *Lefaso.net*

Les id des articles de Lefaso.net sont écrits dans l'url, comme l'Observateur paalga, mais contrairement à ce site, nous ne voulions pas extraire tous les articles de ce site. Les articles qui nous intéressaient étaient dans la rubrique Société, soit la rubrique4 présente dans l'url (le site contenant énormément d'articles, le traitement de parcours et d'extraction des articles de toutes les rubriques aurait pris trop de temps). Après l'analyse des parcours des urls sur le site, nous avons trouvé que comme pour le site Burkina24, nous pouvions parcourir les codes sources des pages 0 à Xmax de la rubrique 4 ; en envoyant à la page Php le paramètre debut_articles et la valeur de ce paramètre qui permettra le parcours des pages (index) :

https://lefaso.net/spip.php?rubrique4&debut_articles={0}

Cette url sera donc l'url que le script prendra en entrée. Pour ce site, la variable d'incrémentation sera de 0 à 23 250.

Comme pour le site Burkina24, pour chaque page parcourue, le script recherche les urls des articles qui, après analyse du code source des pages, ont toujours une syntaxe précise : *"http://lefaso.net/spip.php\?article[0-9]+"*

Ces urls seront également collectées dans un fichier json, pour que le code puisse parcourir les codes sources des pages de ces articles. Puis, nous procédons à un nettoyage de ce fichier afin de ne pas avoir des urls en doublons.

2. Récupération des données textuelles

Pendant que le script Python parcourt les codes sources des articles des sites d'actualités (*crawler*), le script *parse* ces codes sources Html pour rechercher le contenu texte de balises qui constitue l'article, soit la balise contenant le titre de l'article, la balise contenant la date de parution de l'article et la balise contenant l'ensemble du texte de l'article.

La librairie Python *BeautifulSoup* permet de *parser* du code Html et d'en extraire des données. Pour analyser le contenu du code Html et le rendre interrogeable pour cette librairie, nous utilisons le parseur *lxml*, qui permet d'analyser le code Html et xml.

```
Soup = BeautifulSoup(contenu.text, features= "lxml")
```

Nous pouvons ensuite accéder au contenu d'éléments html ;

- Par balise

```
Balise_title=soup.find("title")
```

- Par class

```
Balise_div_class_specifique=soup.find("div",{"class" :"nom_class"})
```

- Par id

```
Balise_div_id_specifique=soup.find("div",{"id" :"nom_id"})
```

- Pour trouver plusieurs éléments

```
Balise_title=soup.find_all("title")
```

3. Filtrage et écriture des articles

Une fois les données textes de tous les articles du site récoltées par le script, nous pouvons les écrire dans un fichier csv, en procédant à un petit filtrage pour obtenir de meilleurs résultats : nous supprimons de façon automatique les articles qui sont écrits dans une langue autre que le français (certains articles étaient écrits en *Lorem ipsum,* des morceaux de faux latin souvent utilisés pour remplir un document), grâce à la librairie python TextBlob[6] (détection de la langue alimentée par Google Translate), ainsi que les articles ayant une longueur inférieure à 15 caractères (commentaires qui sont passés dans le traitement informatique ou articles erronées).

Nous faisons attention à l'écriture des articles : nous concaténons le titre, un point, la date, un point, puis le texte et pour finir le motif *xxxxx*.

Soit : Un article = Titre+.+Date+.+texte+xxxxx

Cette annotation permet de structurer le texte afin d'avoir une séparation entre le titre, la date, le texte grâce au point, et notamment d'avoir une séparation entre chaque article grâce au motif *xxxxx,* pour créer une sorte de frontière pour que le script puisse distinguer les différents éléments de l'article.

Nous avons ainsi obtenu un corpus d'articles bruts (non classifiés et non annotés).

[6] https://textblob.readthedocs.io/en/dev/

Résultats

1. *Format des données*

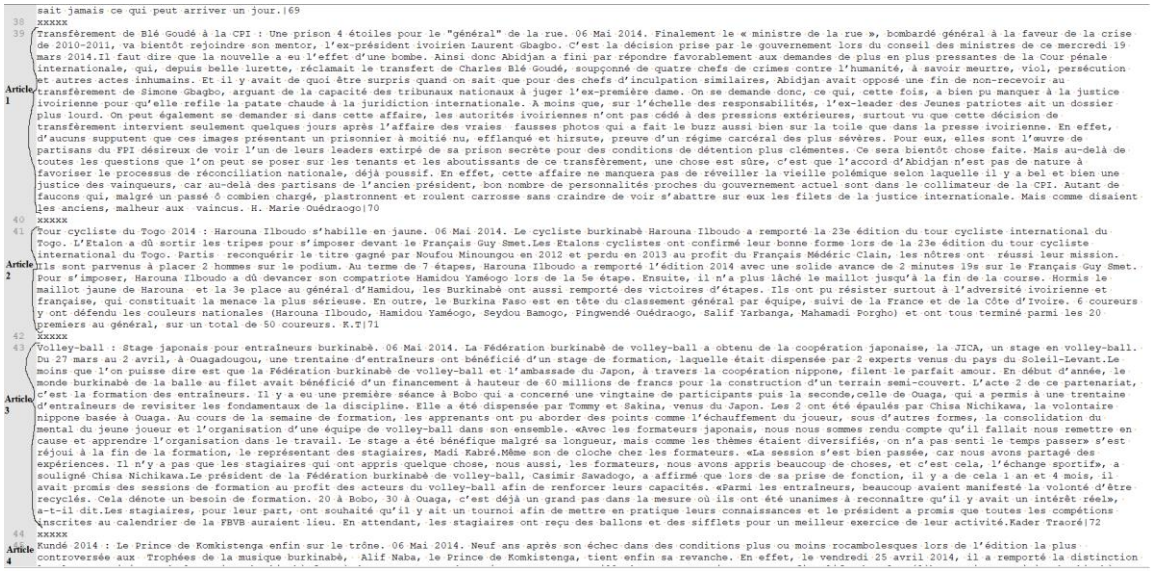


Figure 5 : Extrait du fichier csv contenant le corpus bruts de l'Observateur paalga

2. *Synthèse*

Tableau 2 : Synthèse des informations des journaux

Nom journal	**Date parution articles**	**Langue**	**Nombre articles**
Observateur paalga	24 septembre 2012 à aujourd'hui	Français	2019
Burkina24	01 juin 2011 à aujourd'hui	Français	6365
Lefaso.net	26 octobre 2003 à aujourd'hui	Français	23229
			31 613

Chapitre 2

-

Identification de textes pertinents

Ce chapitre expose les étapes d'identification d'un/de plusieurs texte(s) spécifique(s) dans un ensemble de textes. La problématique de la sécurité alimentaire en Afrique de l'Ouest étant notre domaine de recherche, nous cherchons à identifier quel(s) article(s) sont pertinents à cette thématique dans le corpus acquis dans le chapitre précédent.

Contexte

Le corpus d'articles de sites d'actualités étant récolté en format csv, nous pouvons procéder à l'identification d'articles pertinents pour le thème voulu, c'est-à-dire la problématique de la sécurité alimentaire en Afrique de l'Ouest. Les articles pertinents peuvent traiter de tout ce qui se rapproche de la situation agricole, pluviométrique, de famine. Pour identifier une telle thématique, nous utiliserons le *Topic Modeling*, un algorithme d'apprentissage non-supervisé.

1. Définition Fouille de texte

Grâce aux progrès de capacité de calcul et de mémorisation des ordinateurs dans les années 90, de grandes bases de données ont pu être créées, comme des données tabulaires pour le domaine de la vente ou pour les banques. Pour pouvoir traiter ces bases de données, le domaine de la fouille de données (*data mining)* est apparu. La fouille de texte, *Text mining*, est un processus descendant de la fouille de données : les données sont alors des textes non structurés (textes non catégorisés dans des balises, non organisés en base de données, exemple : fichier pdf, email etc) ou semi-structurés (textes contrôlés par des balises ou mis dans des cases permettant leur interprétation et leur traitement, exemple : page xml, html, base de données, etc).

La fouille de texte permet le traitement de données textuelles volumineuses pour extraire et indexer des informations, et d'en générer de nouvelles.

Les étapes de ce processus sont de collecter l'information de données non ou semi structurées, de convertir cette information en données structurées, d'identifier les motifs des données structurées, d'analyser les motifs, d'extraire les informations utiles et les stocker dans une base de données.

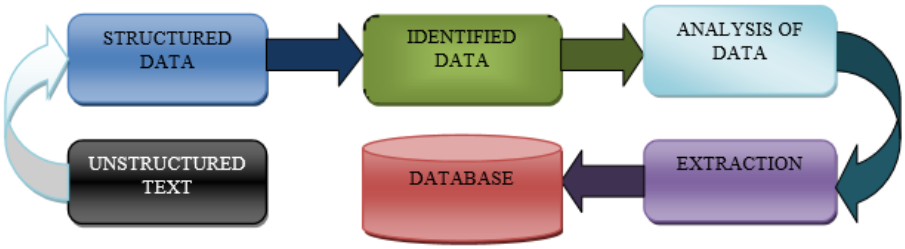


Figure 6 : Etapes générales de la fouille de texte[1]

Ce domaine utilise des techniques telles que l'apprentissage automatique, la recherche d'information ou encore le traitement automatique du langage, et peut effectuer plusieurs tâches dites élémentaires comme (réf. Teiller [2]) :

La Recherche d'Information (RI) ou *Information Retrieval (IR)*) a pour objectif de retrouver un ou plusieurs document(s) pertinent(s) dans une base documentaire, à l'aide d'une requête plus ou moins informelle. L'exemple le plus connu est le moteur de recherche servant à s'orienter et naviguer sur l'ensemble du web.

La classification, permettant d'associer une "classe" à chaque donnée d'entrée et à la ranger dans un domaine existant. Un exemple d'application de cette tâche est la reconnaissance automatique des spams dans les emails. La classification peut se faire avec des approches de *machine learning*, de réseaux de neurones, des modèles probabilistes, des méthodes d'induction de règles (ex : arbre de décision), etc.

L'annotation associe les unités (mots ou segments de mots) d'un ou plusieurs texte(s) à une étiquette. Par exemple, l'annotation *Part Of Speech* caractérise la nature morpho-syntaxique des mots d'un texte : *Le → Det ; chat → NC etc.*

L'extraction d'information (EI) ou *Information Extraction (IE)*) permet d'extraire automatiquement de documents textuels des informations factuelles servant à remplir les champs d'un formulaire prédéfini. La reconnaissance des entités nommées est une application permettant d'illustrer cette tâche (mots ou groupe de mots qui identifient des entités nommées, telles que des noms de lieux, de personnes, d'organisation, etc).

2. *Définition Traitement Automatique du Langage*

Le traitement automatique du langage (TAL, *Natural Language Processing (NLP)*) est un ensemble de méthodes et d'outils qui a comme objectif de modéliser le langage écrit et parlé afin que l'ordinateur puisse manipuler le langage humain, grâce à l'association de l'informatique, de l'intelligence artificielle et de l'analyse linguistique (analyse morpho-lexicale, syntaxique, sémantique et pragmatique).

Le TAL est apparu dans les années 50, en même temps que l'informatique, afin de traduire les messages du camp adverse (URSS vs USA) durant la guerre froide. Nous pouvons citer l'expérience Georgetown-IBM (1954) qui fut une des premières démonstrations de traduction automatique du russe vers l'anglais. Les résultats de la traduction automatique ne furent pas concluants, mais cela a permis d'ouvrir la voie, dans les décennies qui suivirent, à d'autres champs d'applications. En effet, Le TAL permet la mise au point d'application telles que l'analyse de sentiments, les agents virtuels (*chatbots*), la reconnaissance de la parole, la traduction automatique, la correction orthographique, la recherche de mots-clés, ou encore l'extraction d'informations.

3. *Le TAL et la Fouille de texte*

Les différentes étapes d'une chaine de traitement standard en TAL appliqués à la fouille de texte sont :

a. **L'analyse morpho-lexicale** permet d'extraire les mots ou groupes de mots d'un texte, et de les structurer :

   i. La tokenisation : découper un texte en composants, c'est-à-dire en mots ou en groupes de mots, plus précisément tokens.

      La fouille de textes est une spécialisation de la fouille de données.
      La|fouille|de|textes|est|une|spécialisation|de|la|fouille|de|données|.

      Cette étape peut couper des segments de noms de lieux comme *Ganzaga de Sangha, Vallée du Kou*…, alors qu'ils constituent un seul motif.

   ii. La normalisation : éliminer les erreurs, la ponctuation, les chiffres, les *stopwords*[7] ("mots vides", redondants et inutiles pour certains traitements informatiques). Pour la reconnaissance des entités nommées, ces mots sont

[7] https://www.ranks.nl/stopwords/french

importants, par exemple un nom de ville peut être contenir un mot vide, par exemple : *Ganzaga de Sangha*, *Vallée du Kou*…

le, avec, du, ce, alors, été, ...

iii. La racinisation (*stemmatization*) : garder la "racine" du mot (i.e. la forme commune à tous les mots d'une même famille morphologique), en supprimant les préfixes, les suffixes, les formes dérivées et conjuguées, etc.

La racine du mot *chercher* et *recherche* est *cherch*.

iv. La lemmatisation : retrouver la forme canonique du mot, à partir d'une de ces formes dérivées.

Le lemme de *était* est *être*.

v. Le *Part Of Speech Tagging* (POS) : étiqueter les mots selon leur catégorie grammaticale (nom, verbe, déterminant...).Cependant, certains mots peuvent avoir plusieurs catégories syntaxiques selon son sens, par exemple *été* : Nom commun ou participe passé du verbe être.

Texte original :
La fouille de textes est une spécialisation de la fouille de données.
Texte étiqueté :
La/DET fouille/NOM de/PRE textes/NOM est/VER une/DET spécialisation/NOM de/PRE la/DET fouille/NOM de/PRE données/NOM ./SENT

b. **L'analyse syntaxique** (ou *parsing*): consiste à analyser la structure hiérarchisée des phrases et des relations syntaxiques entre les mots qui les unissent (arbres syntaxiques). Cette étape utilise un lexique (vocabulaire) et soit un ensemble de règles syntaxiques (grammaire), soit une grammaire probabiliste (probabilité qu'un mot apparait après un autre selon leur catégorie grammaticale).

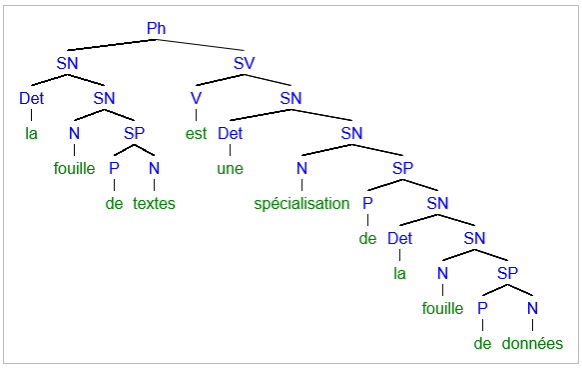


Figure 7 : Exemple d'une phrase analysée syntaxiquement

c. **L'analyse sémantique** quant à elle s'intéresse au sens des mots et des phrases individuellement, généralement hors contexte. Pour cela, l'analyse sémantique se préoccupe du sens de chaque mot d'une phrase, puis le sens de la phrase peut être déduit grâce aux relations existantes entre les mots.

A noter qu'un texte n'est pas obligatoirement soumis à la mise en œuvre successive de ces différentes analyses, et ne peut être traité que par une seule analyse, car celles-ci ne sont pas toujours interdépendantes et essentielles à l'analyse automatique du texte.

Le TAL est parfois difficile à traiter informatiquement, en effet la langue peut être très ambiguë à cause des divers sens des mots, de la position des mots dans la phrase, de la position des phrases dans le texte, du fait que l'interprétation peut dépendre du contexte, etc.

4. Apprentissage automatique

Pour comprendre ce qu'est le *Topic Modeling*, nous devons d'abord définir ce qu'est l'apprentissage automatique (*machine learning*).

L'apprentissage automatique (*machine learning*) est une technologie issue de l'intelligence artificielle, permettant d'élaborer et d'entraîner des algorithmes pour qu'ils puissent faire des prédictions sur un grand volume de données. La popularité de l'apprentissage automatique a augmenté dans les dernières années grâce à la disponibilité du Big Data (large ensemble de données). Il existe deux grands types d'apprentissage : l'apprentissage supervisé et non-supervisé.

Les algorithmes d'apprentissage supervisé sont utilisés pour classifier des données futures. Ils ont en entrée des données connues et étiquetées, appelées données

d'apprentissage, qui sont généralement élaborées et vérifiées à la main, puis servant à entraîner les modèles. Ces données étiquetées nécessitent un travail à la main (étiquetage) et les modèles qui en sont issus donnent des résultats assez précis. La tâche de classification est une application des algorithmes d'apprentissage supervisé assez répandue.

Les données d'entrée des algorithmes d'apprentissage non supervisé ne sont pas quant à elle pré-annotées à la main, et ces algorithmes tentent de comprendre et d'explorer ces données, sans autre information que ces données elles-mêmes. Le *clustering* et le *topic modeling* sont les deux algorithmes d'apprentissages non supervisés généralement utilisés pour les données textuelles. Le *clustering* segmente des documents en partitions, où les documents d'un même groupe (cluster) sont plus proches les uns des autres. Le *topic modeling* est un type spécifique de *clustering* utilisé pour identifier des topics latents dans une collection textuelle.

5. Topic modeling

Le *topic modeling* (modèle thématique) permet donc de découvrir et d'extraire des thématiques dans un grand ensemble de documents textuels, et d'associer une ou plusieurs thématiques à ces documents. Ce n'est pas de la classification de document, nous ne pourrons avoir un thème appelé "politique", un autre "sport" (pas d'étiquettes), mais nous pourrons savoir si un article traite d'une thématique particulière qui a une collection de mots similairement sémantiques.

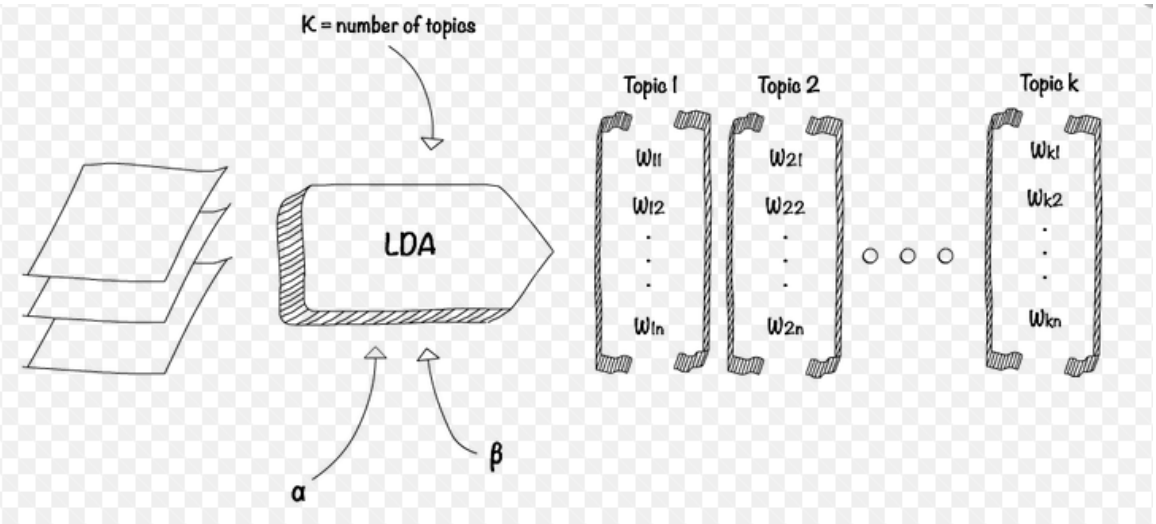


Figure 8 : Schéma général de l'algorithme Topic Modeling avec le modèle LDA[3]

Outils et méthodes de l'état de l'art

1. *Méthodes de Topic Modeling*

Les méthodes actuelles pour effectuer du *topic modeling* sont l'analyse sémantique latente, l'analyse sémantique latente probabiliste et l'allocation de Dirichlet latente, décrites dans les paragraphes suivants :

Analyse sémantique latente (LSA), par Deerwester et al. en 1990 : modèle statistique permettant de découvrir la relation sémantique des mots dans un ensemble de documents. Cette méthode s'appuie sur les données d'une matrice d'occurrences des mots dans les documents (term-document matrix), et d'une réduction de dimensions de cette représentation vectorielle des textes, afin de faire émerger des liaisons sémantiques entre des termes dans un grand corpus.

Analyse sémantique latente probabiliste (PLSA), par Thomas Hofmann en 1999, est un modèle qui a le même principe que LSA, mais utilise une méthode probabiliste au lieu d'une matrice d'occurrences dans les documents.

Allocation de Dirichlet Latente (LDA), par Blei et al en 2003, est un modèle génératif probabiliste qui est une amélioration de la LSA et PLSA. Ce modèle est basé sur le théorème de Bayes et est génératif, c'est-à-dire qu'il caractérise la distribution statistique de chaque classe en apprenant la distribution de probabilité commune P(x,y) (contrairement au modèle discriminant, qui apprend la distribution de probabilité conditionnelle).

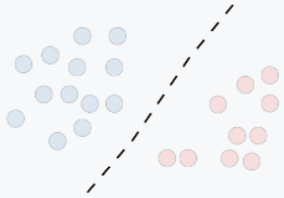
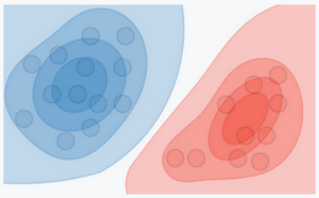
	Modèle discriminant	**Modèle génératif**
But	Estimer directement $P(y\|x)$	Estimer $P(x\|y)$ puis déduire $P(y\|x)$
Ce qui est appris	Frontière de décision	Distribution de probabilité des données
Illustration		
Exemples	Régressions, SVMs	GDA, Naive Bayes

Figure 9 : Schéma de modèle génératif et discriminant[4]

2. *Outils de Topic Modeling*

Plusieurs librairies Python permettent d'effectuer l'algorithme de *Topic Modeling* sur un ensemble de données textuelles ;

Scikit-Learn[8] (Cournapeau, 2010) est une librairie Python open source d'apprentissage statistique, contenant également l'algorithme LDA.

Stanford topic modeling toolbox[9] (Ramage et Rosen 2009) est une boîte à outils écrit en langage Scala, ayant plusieurs algorithmes y compris LDA, Labeled LDA et PLDA.

[8] https://scikit-learn.org/stable/
[9] https://nlp.stanford.edu/software/tmt/tmt-0.4/

Approche

Après avoir recueilli les articles, nous devrons déterminer lesquels traitent de sécurité alimentaire. En effet, en analysant les articles récoltés, nous constatons que les articles traitent de sujets divers. Nous voulons donc procéder à une identification thématique dans ce corpus. N'ayant pas de données étiquetées pour l'analyse thématique des sujets (par exemple l'étiquette *politique, sports* etc.), mais assez de données pour effectuer un traitement d'identification du thème, nous appliquerons un algorithme non supervisé aux articles en utilisant le modèle LDA (*Latent Dirichlet Allocation*), car il permet une catégorisation plus précise des articles, et ce modèle a été utilisé dans des problèmes similaires liés aux catastrophes naturelles [5] [6].

Nous aurons en sortie des thématiques composées de groupe de mots qui auront une probabilité d'occurrence pour le thème donné, et les articles ayant une probabilité d'appartenance à une ou plusieurs thématique(s).

1. Apprentissage non supervisé avec le modèle LDA

1.1. Choix de la librairie Machine Learning

Pour effectuer l'algorithme LDA, nous utilisons la librairie *Scikit-learn*. Cette bibliothèque peut être utilisée pour des tâches de classification (identifier à quelle catégorie un objet appartient), des tâches de régression (prédire une valeur associée à un objet) ou des tâches de *clustering* (regroupement automatique d'objets similaires dans des ensembles). Cette librairie est notamment très connue et utilisée par la communauté, nous avons donc à disposition une documentation assez complète.

1.2. Prétraitement

Pour ne pas avoir de thèmes contenants des mots non informatifs pour notre recherche, par exemple ;

- thème 1 = *le, du, avoir, cette, mange ;*
- thème 2 = *0, faire, l', ce, remettre, vendredi ; etc.*

Nous devons supprimer les *stopwords* dans les articles, mots vides en français (mots communs et très fréquents dans le texte, comme les déterminants, les pronoms etc.., par

exemple *le, alors, du, être...*). Les stopwords sont utiles notamment dans les segments et ne sont généralement pas à supprimer. Mais pour le traitement d'identification thématique, ils peuvent être éliminer, afin d'avoir moins de bruit et d'avoir plus de mots significatifs dans les thèmes. Nous avons également ajouté à cette liste d'autres mots non-pertinents qui apparaissent beaucoup dans notre corpus et donc dans les thématiques, notamment des prénoms "jean", des verbes sans information "mettre", "faire", et des nombres comme "2" etc. Nous n'avons pas lemmatisé le texte car nous voulons en sortie les textes originaux afin de pouvoir les traiter (reconnaissance des entités) par la suite.

1.3. Vectorisation du vocabulaire des textes

Une fois les articles normalisés, la bibliothèque *Scikit-learn* crée le vocabulaire des articles avec la fonction *CountVectorizer* (comme l'algorithme n'interprète pas les lettres mais les chiffres, il faut transformer les textes en vecteurs). Cette fonction compte le nombre de mots dans un texte, en convertissant un ensemble de textes en une matrice de nombre d'occurrences de chaque mot dans le texte. Cela permet d'évaluer l'importance des mots dans les articles.

Cette fonction a certains paramètres, comme :

max_df, qui permet de supprimer les mots apparaissant trop souvent. Par exemple, *max_df* = 0.8 ignore les termes apparaissant dans plus de 80 % des fichiers;

min_df supprime les mots qui apparaissent trop peu dans le corpus. Par exemple *min_df* = 5 ignore les mots apparaissant dans moins de 5 articles.

1.4. Modèle LDA

Puis, le script crée le modèle LDA avec plusieurs paramètres, notamment le nombre de thématiques ;

LDA = LatentDirichletAllocation(n_components=X),

et apprend le modèle LDA sur le vocabulaire transformé en vecteurs pour créer des thématiques composées de mots similaires, grâce aux probabilités des vecteurs des mots et des thématiques ;

LDA.fit()

Nous aurons en sortie une liste de thématiques comportant les mots les plus probables pour chaque thématique, cf. annexe 1 p.67 (sur le corpus d'articles de l'observateur paalga, avec

: n_components = 100, nombre de mots pour chaque thématique = 10, max_df=0,7, min_df=4)

Pour finir, le script transforme les données selon le modèle (LDA) ajusté

Topic_values=LDA.transform()

Sortie_csv=Topic_values.argmax(axis=1) ;

Nous aurons en sortie le csv d'entrée avec une nouvelle colonne, où sera affiché le numéro du topic le plus probable de chaque article, de telle sorte :

Tableau 3 : Exemple de la sortie du script utilisant le Topic Modeling avec LDA

Article	Nom_fichier_article	Numéro du topic correspondant
Titre, date, texte de l'article1	Nom du fichier correspondant à l'article 1 : 1.txt	19
Titre, date, texte de l'article2	Nom du fichier correspondant à l'article 2 : 2.txt	3
…	…	…

1.5. Résultats

Les résultats du *topic modeling* avec le modèle LDA ne mettaient pas forcément en évidence le thème de la sécurité alimentaire, car les articles traitant de ce thème ne sont pas majoritaires dans les journaux. Comme nous pouvons le voir dans l'annexe 1 (p.67), les thématiques étaient trop généralistes et peu pertinentes pour la sécurité alimentaire, seulement trois thématiques (8 ; 61 et 85) correspondaient au thème recherché (céréales, plantes, aliments, repas, nutrition, alimentation, eau), mais ces mots étaient associés au thème de la santé/médecine.

2. Apprentissage non-supervisé avec le modèle LDA et Word2Vec

Les résultats précédents n'étant pas concluants, nous avons procédé d'une autre manière, en utilisant toujours l'algorithme d'apprentissage non-supervisée avec le modèle LDA, et en ajoutant également le modèle Word2vec et un lexique spécifique à la sécurité alimentaire afin d'avoir une identification d'articles spécifique au thème voulue [7].

2.1. Modèle LDA

En entrée, nous aurons le fichier csv contenant les articles bruts (obtenu dans le chapitre précédent). Le script fera le même processus explicité dans la partie ci-dessus (1.

Apprentissage non supervisé avec le modèle LDA). Les thématiques seront ainsi créées (nous spécifions au script de rechercher 100 thématiques dans le corpus d'articles). Puis le script concatène le ou les articles dans le fichier de la thématique correspondante (les articles sont séparés par le retour chariot \n), par exemple si les articles 1, 13 et 83 correspondent au thème 10, alors les articles seront concaténés les uns à la suite des autres dans le fichier correspondant au thème 10 ; 10_lda_topic_subsets.txt.

2.2. Lexique

Puis, afin de trouver les thématiques (générées par LDA) les plus proches du thème de la sécurité alimentaire, nous utilisons un lexique ayant 90 entrées de noms communs, comportant des termes de sécurité alimentaire comme "alimentaire", "agriculture", "élevage", "légumineuse", "céréales", "malnutrition", etc. Le lexique provient du site *Cultivoo*, contenant des informations historiques et scientifiques, dont des lexiques spécifiques. Le script classe les thématiques (qui ont été générées par LDA) en fonction de leur similarité sémantique globale avec le lexique de la sécurité alimentaire (en considérant tous les articles de la thématique comme un seul texte), avec le modèle *Word2Vec* (plongement lexical).

Pour finir, nous prenons les thématiques les mieux classées (au-dessus de 70 %) et nous classons ensuite les articles présents dans ces meilleures thématiques selon leur similarité individuelle avec le lexique, également avec *Word2Vec*.

2.3. Word2vec

Word2Vec est un ensemble de modèles de réseau neuronal qui traite des données textuelles. La méthode de *Word2vec* est notamment implémenté dans la bibliothèque python *Gensim*. Word2vec permet de transformer les mots en vecteurs, et de comparer ces vecteurs entre eux. Cette comparaison permet d'avoir une représentation sémantique des mots : par exemple, les mots *chat* et *chien* sont modélisés par des vecteurs peu distants, donc leur lien sémantique est relativement proche. Nous pourrons donc avoir les mots sémantiquement proches d'un mot, les synonymes et les abréviations des mots, ou encore les différentes formes des mots (fautes de frappe ou d'orthographe).

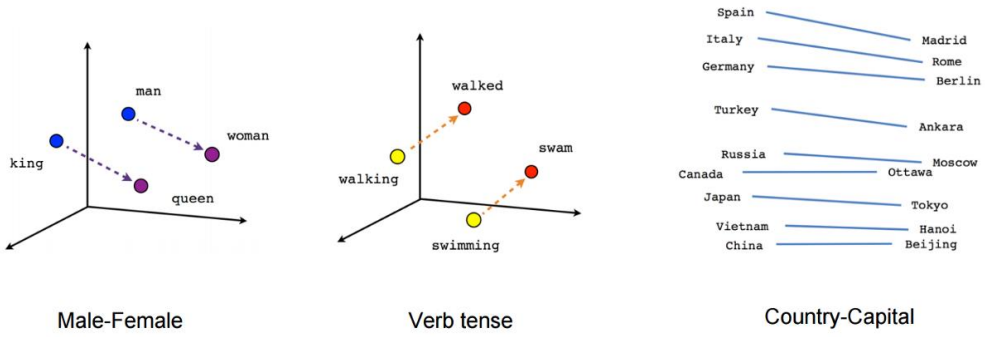


Figure 10 : Relations d'analogie entre les mots avec Word2vec[8]

Word2Vec a besoin d'avoir un modèle de vecteurs pré-entraîné sur des données volumineuses. Nous utilisons un modèle pré-entraîné sur le Wikipédia français.

Wikipédia étant publié dans un langage standard (orthographe, grammaire, etc. unique et stable), le modèle est donc adapté et spécifique à nos données, car les articles de quotidien utilisent également un langage standard.
De même, le modèle Word2vec a été utilisé dans des problèmes similaires liés aux catastrophes naturelles, nous devrions avoir des résultats significatifs [9].

Résultats

1. Format des données

En sortie, nous aurons un fichier csv avec les articles les plus cohérents par rapport au lexique de la sécurité alimentaire (articles classés avec la probabilité la plus haute à la moins haute), et un dossier de fichiers .txt, contenant pour chaque fichier .txt un article, afin de pouvoir effectuer plus facilement les traitements d'annotations et d'avoir une meilleure visualisation de chaque article.

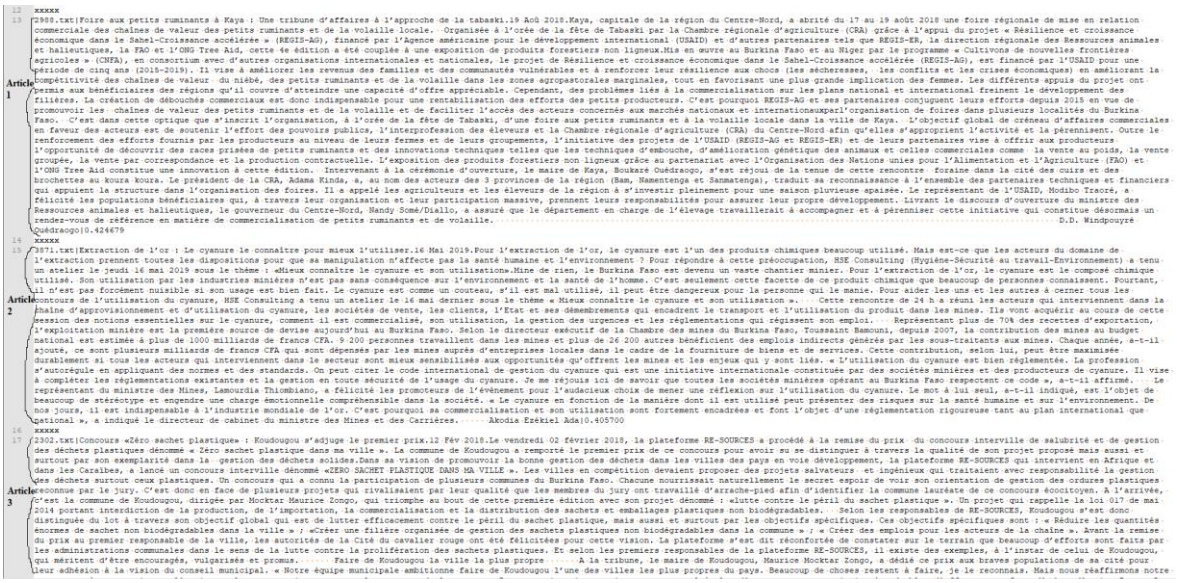


Figure 11 :Extrait csv corpus articles classés de l'Observateurpaalga

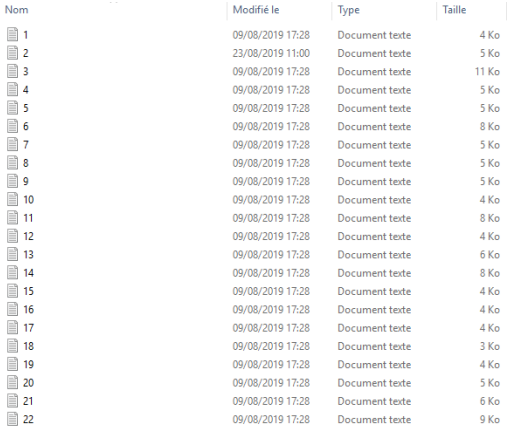


Figure 12 : Extrait du dossier de fichiers txt des articles classés de l'Observateur paalga

2. Evaluation

Pour déterminer si l'identification d'articles par mots clés a été fructueuse, nous avons dû analyser les résultats manuellement.

Nous avons donc établi un protocole d'évaluation pour analyser l'identification des thèmes, en se posant pour chaque article :

- Porte-t-il sur la sécurité alimentaire au Burkina Faso ?

  Oui tout à fait, Oui partiellement, Non.

  A noter que nous définissons un article traitant de la problématique de la sécurité comme tout article mentionnant ou ayant des liens avec les termes de la famine, climat, des prix alimentaires, de l'agriculture.

- Contient-il des informations complémentaires et d'intérêt sur la sécurité alimentaire ?

  Oui, Non.

Deux expertes de la sécurité alimentaire, Elodie MAITRE D'HOTEL, chercheuse Cirad, et Maguelonne TEISSEIRE, directrice de recherche Irstea, ont évalué les articles les plus pertinents par rapport au thème pour le journal l'Observateur paalga et Burkina24 (les articles de Lefaso.net n'ont pas été évalué par manque de temps) :

Observateur paalga

Tableau 4 : Résultats évaluation thématiques des premiers articles de l'Observateur paalga

	Elodie MAITRE D'HOTEL		*Maguelonne TEISSEIRE*
Article classé	Le texte porte-t-il sur le sujet de la sécurité alimentaire ?	Le texte contient-il des informations complémentaires et d'intérêt sur la sécurité alimentaire ?	Le texte porte-t-il sur le sujet de la sécurité alimentaire ?
1	Non (recrutement)	Oui	Non
2	Oui	Oui	Oui
3	Non (pub projet)	Oui	Oui
4	Non (pub projet)	Oui	Oui
5	Oui partiellement (sécurité sanitaire)	Oui	Oui partiellement
6	Non	Oui	Oui
7	Non (pub projet)	Oui	Oui
8	Non	Non	Non
9	Non	Non	Non
10	Non	Non	Non
11			Non
12			Oui partiellement
13			Non
14			Non
15			Non
20			Oui partiellement
25			Non
30			Non
35			Non
40			Non

Sur les 10 premiers articles de l'observateur paalga, les deux expertes n'ont pas eu le même avis pour 4 articles (articles 3 ; 4 ; 6 ; 7).

Burkina24

Tableau 5 : Résultats évaluation thématiques des premiers articles de Burkina24

	Elodie MAITRE D'HOTEL		*Maguelonne TEISSEIRE*
Article classé	Le texte porte-t-il sur le sujet de la sécurité alimentaire ?	Le texte contient-il des informations complémentaires et d'intérêt sur la sécurité alimentaire ?	Le texte porte-t-il sur le sujet de la sécurité alimentaire ?
1	Oui	Oui	Oui
2	Non (pub atelier)	Oui	Oui
3	Oui	Oui	Oui
4	Oui	Oui	Oui
5	Oui (pub atelier)	Oui	Oui
6	Oui partiellement (pub atelier)	Oui	Oui
7	Oui	Oui	Oui (production de viande pertinent même si la notion de sécurité alimentaire n'est pas explicitement discutée)
8	Oui	Oui	Non (c'est sur le coton)
9	Oui	Oui	Oui partiellement (hésitation entre non pertinent - visite d'une exploitation chinoise sur la production agricole mais dans la thématique)
10	Non (pub)	Oui	Oui partiellement
11			Oui
12			Non
13			Non
14			Oui partiellement
15			Non
20			Oui
25			Non
30			Non
35			Non
40			Oui partiellement

Sur les 10 premiers articles de Burkina24, les deux expertes n'ont pas eu le même avis pour 3 articles (articles 2 ; 8 ; 10).

Les expertes ont notamment indiqué que certains articles des journaux, notamment ceux du *Burkina24*, étaient des publicités par des institutions organisant des ateliers, des journées thématiques etc., et non des articles de presse présentant un événement "réel", une situation particulière. Cela peut biaiser les résultats car ces articles peuvent parfois ne pas avoir de rapport avec la situation réelle de la sécurité alimentaire au Burkina Faso.

Produits forestiers non ligneux : Une alternative de lutte contre la faim.
Publié le 12 septembre 2018 at 13 septembre 2018.
Le Comité Permanent Inter-Etats de lutte contre la sécheresse dans le sahel (CILSS) commémore sa 33e journée ce mercredi 12 septembre 2018. Le rapport conjoint publié par la FAO sur l'état de la sécurité alimentaire et de la nutrition dans le monde 2018, paru le 11 septembre 2018, pointe les dérèglements climatiques tels les sécheresses et les inondations comme facteurs clés expliquant la hausse de la faim dans le monde. Près de 821 millions de personnes ont eu faim dans le monde en 2017 contre 804 millions en 2016. L'Afrique est la plus touchée avec un taux de 21% de sa population mal nourrie suivie de l'Asie avec 11,5%. Le rapport indique que la faim a progressé ces trois années consécutives rendant l'objectif d'éradiquer la faim en 2030, problématique. L'utilisation des produits forestiers non ligneux fait partie des stratégies de diversification d'alimentation et de revenus pour de nombreuses communautés des pays sahéliens et de l'Afrique de l'ouest confrontées à une variabilité climatique. Le CILSS, cet instrument de lutte contre la sécheresse en Afrique subsaharien, marque chaque année un arrêt pour « évaluer les résultats atteints afin de dégager les gaps à compenser pour les années à venir ». En cette journée commémorative, il tourne la réflexion sur la promotion des produits forestiers non ligneux, comme une alternative à la lutte contre l'insécurité alimentaire et nutritionnelle et source de revenus des populations. Selon le secrétaire exécutif du CILSS, Adoum Djimé, 40% de la population en Afrique de l'ouest souffre d'une malnutrition. La science et la technologie ont bien démontré que les produits forestiers non ligneux possèdent des propriétés importantes en nutriments. (De droite à gauche) Adoum Djimé, secrétaire exécutif du CILSS et Alassane Guiré, Secrétaire général du ministère de l'agriculture Pour lui, il faut réapprendre à vivre avec ce qu'offre l'environnement. D'où l'importance de la protection et la gestion durable de ces ressources naturelles indispensables. Cependant, une meilleure contribution de ces produits non ligneux à la sécurité alimentaire nécessite un engagement politique, le concours des chercheurs et du secteur privé dans la transformation et la création d'emploi. Il ne pensait pas si bien dire. Dans l'espace d'exposition des savoirs-faire avec les produits non ligneux dressé pour l'occasion, les acteurs ne finissent pas d'égrener les difficultés rencontrées pour la disponibilité, la conservation et la distribution de ces produits finis. Pauline Zoungrana Birba, présente de l'association « Beog Neeré », Pauline Zoungrana Birba, présente de l'association « Beog Neeré », transformation de produits alimentaires, présente entre autres produits, des biscuits de petit mil et de maïs et de manioc. Elle explique que «ce sont des produits bio, aucun ne contient un produit chimique. Ça nous ae beaucoup mais ce sont les moyens pour améliorer les produits qui font défaut. Les matières premières sont chères, les emballages pour la conservation et la protection ». Et Allasane Guiré, secrétaire général du ministère de l'agriculture de rassurer que la réflexion portera sur la contribution de ces produits non ligneux dans la sécurité alimentaire et les difficultés rencontrées dans leur promotion. « Le CILSS se porte très bien », rassure son secrétaire exécutif, et a entre autres grands projets pour les 6 pays membres (Mali, Burkina Faso, Niger, Mauritanie, Tchad, Sénégal), le pastoralisme financé par la Banque mondiale en raison de 250 millions de dollars, un projet de résilience financé par la banque africaine de développement, 120 milliards de F CFA pour les 6 pays plus la Gambie, un projet d'irrigation financé par la Banque mondiale en raison 180 millions de dollars, et bien d'autres à venir.
Revelyn SOME Burkina24|0.525168

Figure 13 : Exemple d'un article présentant une journée thématique

Chapitre 3

-

Identification d’informations spatio-temporelles et mise en lien

Ce chapitre permet l'analyse spatio-temporelle des articles les plus pertinents pour la problématique de la sécurité alimentaire en Afrique de l'Ouest. Nous identifions de façon automatique les lieux et la temporalité présents dans les textes. Nous aurons pour chaque article un index d'entités spatiales et l'année correspondante à l'article, afin de lier ces informations spatio-temporelles aux images satellites.

Contexte

1. Définition Entités nommées

Les entités nommées (EN) ont été définies au MUC (*Message Understanding Conferences*) dans les années 90 comme "tous les éléments du langage qui font référence à une entité unique et concrète, appartenant à un domaine spécifique (i.e. Humain, économique, géographique, etc.)".

2. Reconnaissance d'entités nommées

La reconnaissance d'entités nommées signifie identifier l'entité et déterminer sa catégorie à laquelle elle est référée : Nom, Lieu, Organisation etc. Elle a également été évoqué dans les conférences MUC et trois modèles de représentation d'entités nommées ont ainsi été adoptés : *Enamex* pour les noms de personnes, d'organisation et de lieu, *Timex* pour les expressions temporelles, *et Numex* pour les expressions numériques.

Il existe deux approches pour la reconnaissance d'entités nommées :

- Basée sur des règles ou *rule-based* (p. ex avec des expressions régulières) : elle implique de définir des listes de noms (gazetiers) et des motifs correspondants à des entités nommées. Cette approche est précise mais longue et fastidieuse à mettre en place.
- Basée sur du *machine learning* : Un modèle est entraîné sur des données textuelles déjà annotés en entités nommées, afin de reconnaître et d'annoter automatiquement des entités nommées dans de nouvelles données textuelles. L'apprentissage est alimenté grâce à des indices présents dans le texte tels que la forme des mots (majuscules, nombres, mots entiers), le contexte des mots, les étiquettes grammaticales, etc.

Outils et méthodes de l'état de l'art

1. *Outils de reconnaissance d'entités spatiales*

NLTK[10] est la première librairie Python spécifique au TAL (2001), mais les traitements sont relativement lents et la librairie ne supporte pas toutes les langues. De plus elle n'utilise ni de modèle de réseau neuronal ni d'intégration de vectorisation de mots dans ces traitements.

CORENLP[11] (Stanford) est un outil développé en Java assez robuste et rapide pour le TAL, mais n'intègre pas le plongement lexical de mot (*word embedding*) dans ces traitements.

Spacy[12], créé en 2015, est une librairie Python spécifique au TAL, basé sur du *machine learning*. Elle est simple d'utilisation, possède une communauté assez présente sur Internet, est robuste et rapide. Contrairement aux deux dernières outils présentés, *Spacy* utilise les modèles de réseaux neuronaux pour entraîner ces modèles et intègre une vectorisation des mots dans ces traitements.

2. *Outils de reconnaissance d'entités temporelles*

Pour rappel *Timex* est le modèle de représentation des entités nommées temporelles le plus répandu et usité depuis les MUC (*Message Understanding Conferences*). Nous utiliserons donc un outil de reconnaissance d'entités temporelles suivant ce modèle parmi les outils existants :

SUTime[13], faisant partie de *Stanford CORENLP*. *SUTime* est un tagueur d'expressions temporelles, basé sur un système de règles. L'inconvénient de cet outil est de ne traiter que des textes en anglais.

TRIPS/TRIOS[14], développé par UzZaman et Allen en 2011. Ce tagger d'expressions temporelles utilise un modèle probabiliste (*conditional random filed CRF)* et

[10] https://www.nltk.org/
[11] https://stanfordnlp.github.io/CoreNLP/
[12] https://spacy.io/
[13] https://nlp.stanford.edu/software/sutime.shtml
[14] http://wing.comp.nus.edu.sg/~antho/S/S10/S10-1062.pdf

également un système de règles pour la normalisation d'expressions temporelles. Ce tagueur utilise *Wordnet* et une ontologie en anglais, et traite alors les textes en anglais.

Heideltime[15] développé par Strötgen et Gertz en 2010, est un tagueur temporel multilingue basé sur un système de règles et supportant plusieurs langues (anglais, allemand, français, espagnol, italien, vietnamien, arabe, chinois, russe, croate, portugais, estonien, néerlandais). *Heideltime* est écrit en java, mais un wrapper Python[16] a été développé afin de pouvoir l'utiliser avec ce langage.

[15] https://github.com/HeidelTime/heideltime
[16] https://github.com/amineabdaoui/python-heideltime

Approche

Une fois les articles classés par rapport au lexique de la sécurité alimentaire, nous pouvons analyser les informations spatio-temporelles de ces textes, grâce à des scripts python et des librairies python spécifiques au traitement de texte.

1. Reconnaissance d'entités spatiales

Afin de reconnaître les entités spatiales, nous utilisons *Spacy*, une bibliothèque Python de Traitement Automatique du Langage (TAL).

Spacy permet de traiter du texte, comme pour taguer et *parser* le texte, ou encore reconnaître des entités nommées. Nous utilisons cette librairie pour cette dernière tâche, afin de repérer les entités spatiales dans les articles. Pour rappel, une entité nommée est une expression désignant un nom de personne, un nom de lieu (entité spatiale), ou un nom d'organisation, une date, une abréviation.

Spacy reconnaît ces types d'entités nommées suivants pour le français :

- PER (nom d'une personne, nom de famille),
- LOC (nom d'une localisation : villes, provinces, pays, régions, fleuve, lac, montagne etc.)
- ORG (nom d'une organisation, gouvernemental, etc.)
- MISC (nom d'entités diverses, p.ex. des événements, nationalité, produits, etc.)

Nous utiliserons le modèle 'fr' de *Spacy*, afin que l'outil puisse appliquer des traitements de TAL à des textes français. Ce modèle est entraîné sur deux corpus :

- Le corpus français Sequoia (Dépendances Universelles) pour le parseur et le tagueur. Ce corpus contient des phrases en français, provenant d'Europarl (parlement européen), du journal Est Républicain (journal quotidien), de Wikipédia français (encyclopédie collective en ligne), et des documents de l'Agence Européenne du Médicament.
- Le corpus WikiNER pour la reconnaissance d'entités nommées (données textuelles en français provenant de Wikipédia français).

Ces corpus contiennent un langage standard et sont donc spécifiques à notre corpus d'articles de journaux, les résultats de reconnaissance d'entités devraient être satisfaisants.

Lorsque la bibliothèque rencontre une entité spatiale dans un article, nous ajoutons des balises autour, que nous avons prédéfinie dans le script. Par exemple :

texte ['Ouagadougou', 'LOC'] texte**

texte ['Burkina Faso', 'LOC'] texte**5

A noter que nous ne prenons en compte que l'information spatiale. Par exemple dans l'entité *marché de Bamako…"* seulement *"Bamako"* sera annoté.

Projet «Un ménage vulnérable une vache» : Un bol… de lait pour des familles burkinabè.
25 Jui 2014.
Dans les mois à venir, des familles burkinabè connaîtront une amélioration de leur alimentation à travers le projet «un ménage vulnérable, une vache» du ministère des Ressource animales et halieutiques. Le lancement de ce projet, qui veut augmenter la consommation de lait au ['Burkina', 'LOC'] , a eu lieu le lundi 23 juin 2014 à ['Kaya', 'LOC'] dans la région du ['Centre-Nord', 'LOC'] , concomitamment avec celui de la campagne de production fourragère et la pose de la première pierre de l'abattoir frigorifique de la ville. 5 000 vaches à la disposition de ménages vulnérables, augmentation de la production laitière de 840, 1080 ou 1440 litres par an, accroissement du revenu par ménage, amélioration de la situation nutritionnelle des bénéficiaires… le projet «un ménage vulnérable, une vache» va donc changer les habitudes des populations après sa mise en œuvre. Raison pour laquelle son initiateur, le ministère des Ressources animales et halieutiques, a pris toutes les dispositions, notamment en mobilisant 22 208 018 161 F CFA pour son exécution. «Celui-ci procurera à la famille du lait pour améliorer son bol alimentaire quotidien, mais aussi contribuera à améliorer ses revenus à partir de la vente du surplus de lait», affirme d'ailleurs le chef du département,

Figure 14 Exemple article annoté par Spacy

2. *Reconnaissance d'entités temporelles*

Pour reconnaître les entités temporelles, nous utilisons *HeidelTime*, un programme java, développé par le *Database Systems Research Group de l'Université de Heidelberg*, permettant d'annoter les entités temporelles multilingues de documents selon la norme d'annotation TIMEX3. Nous voulons continuer à utiliser le langage Python, nous nous servons donc d'un wrapper python développé par Amine Abdaoui afin d'appeler *Heideltime* dans python.

HeidelTime nécessite en entrée le texte avec les phrases découpées, les tokens (mots) du texte et leur catégorie syntaxique, appelée *POS-Part Of Speech* (nom, verbe, etc.). Ces traitements sont effectués avec TreeTagger, un étiqueteur qui annote les catégories syntaxiques et les lemmes des tokens du texte. Savoir les catégories syntaxiques peut être utile pour analyser le contexte des phrases, notamment pour le temps des verbes (passé, présent, futur...) pouvant apporter une indication temporelle.

HeidelTime extrait et normalise ensuite les entités temporelles par des règles d'extraction (expression régulières) et des lexiques. Il existe deux types principales d'expressions temporelles : les expressions temporelles absolues (Date du calendrier *mercredi 12 juin 2019*, *2011*, etc.) et relatives (Expressions plus indéfinis dans le temps : *cette semaine, 3 heures,11 h 30*, *ensuite*, etc.)

Les entités temporelles seront annotées selon la typologie de *HeidelTime* :

- *<TIMEX3>* : cette balise permet de marquer les heures (type=TIME), les dates (type=DATE), les intervalles et les durées.

  Par exemple ;

  - *Le 24 novembre 2018 à 17h36* sera annoté :

    <TIMEX3 tid="t2" type="DATE" value="2018-11-24">
    le 24 novembre 2018
    </TIMEX3>
    à
    <TIMEX3 tid="t9" type="TIME" value="2018-11-24T17:36">
    17 h 36
    </TIMEX3> min.

  - *72 heures* sera annoté :

    <TIMEX3 tid="t39"type="DURATION" value="P3D">
    72 heures
    </TIMEX3>(P3D : 3 jours).

- <*SIGNAL*> : cette balise permet d'annoter les mots indiquant les informations temporelles qui ont une relation temporelle *(*Exemple *: avant, après, durant, etc.)*

2.1. Format des données

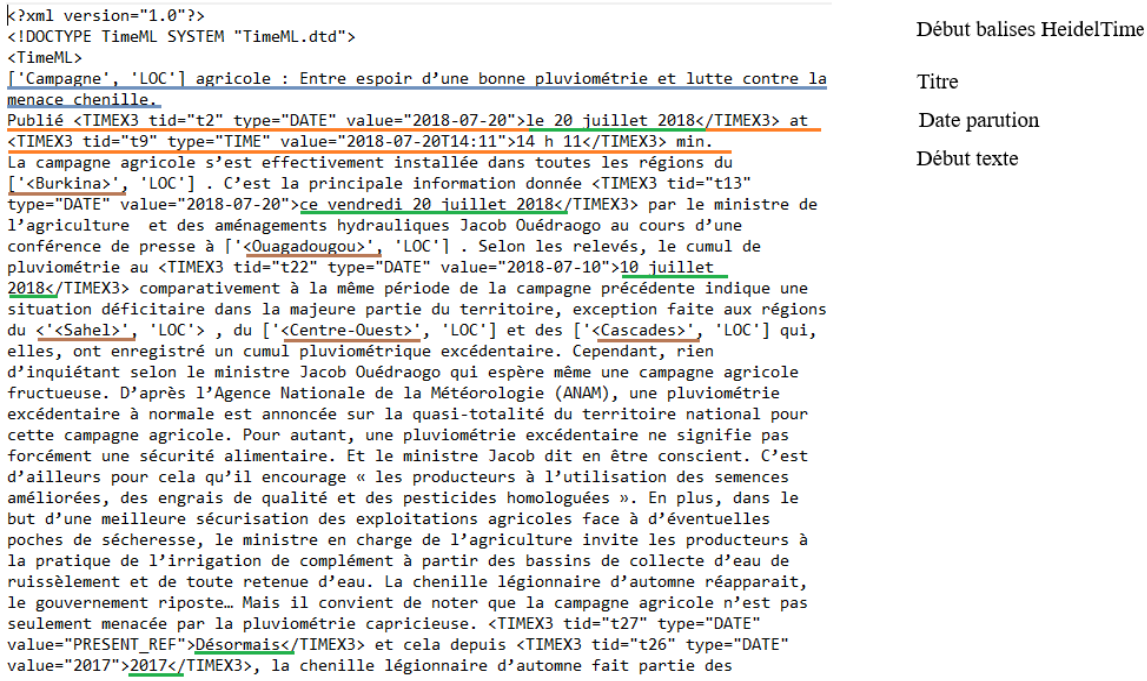


Figure 15 : Exemple article du Burkina24 traitant de sécurité alimentaire et tagué par Spacy et Heideltime

Résultats

1. Entités spatiales

1.1. Analyse

La bibliothèque *Spacy* reconnaît assez bien les entités spatiales, les noms de ville sont majoritairement bien reconnus *("Burkina", "Kaya", etc.)*, mais qu'il y a également des erreurs dans l'annotation des lieux (cf. Annexe 2 p.70).

- *Noms de lieux peu connus*

Par exemple, il y a certains oublis de noms de villages, ou de régions, comme le diminutif de certains noms de villes :

Bobo ; faisant référence à Bobo-Dioulasso,
Ouaga ; faisant référence à Ouagadougou.

Certains quartiers peu connus ou nouveaux ne sont également pas reconnus par la bibliothèque comme *Sikassa Cira, Kosyam, Accart-ville, Dapoya.*

- *Informations graphiques*

Certains mots ayant une capitale sont parfois reconnus en tant qu'entités spatiales, par exemple les mots *Monsieur, Ayant, 23 Mars, J'...*

De nombreux mots sont également annotés en entités spatiales, alors que ce sont des mots communs et n'ont pas de lien avec un lieu, par exemple les mots *Programme, ministère des Ressources, Jan 2019, Quelles, SO.GE.A.O*. Ces erreurs doivent être dues à la majuscule présente dans ces mots. Le système de reconnaissance d'entités nommées s'appuie sur indices dans le texte tels que des informations graphiques et de la position du mot dans la phrase (si le mot présente une majuscule et n'est pas au début de la phrase, alors le mot a une forte probabilité d'être annoté en tant qu'entité.

- *Informations morpho-syntaxiques*

Si un mot est étiqueté en tant que nom propre alors il est possible que le système de reconnaissance d'entités l'annote en tant qu'entités, comme par exemple le mot *programme*, Des oublis d'annotations sont également présentes lorsque l'outil rencontre des noms de lieux ayant des mots communs : *Hauts-Bassins, Boucle du Mouhoun.*

- *Ponctuation*

L'outil de reconnaissance d'entités nommées peut prendre en indice la ponctuation, en effet, il peut y avoir souvent des erreurs dans les acronymes, comme *SO.GE.A.O.* étant annoté en tant qu'entité spatiale, alors que c'est une organisation.

- *Ambiguïté sémantique*

Également, les entités nommées peuvent avoir plusieurs interprétations, selon le contexte de la phrase, par exemple : *Il est 8h et demie lorsque notre fourgonnette stationnait à la mairie de Koubri ;* la mairie de Koubri est ici interprétée comme un lieu. En revanche, la phrase *C'est autour de 15 milliards de francs CFA que la mairie de Ouaga a investi dans la réalisation d'infrastructures de voirie* ; la mairie de Ouaga dans ce cas doit être interprétée comme une organisation, ce qui n'est pas toujours le cas dans les systèmes de reconnaissance d'entités nommées.

1.2. Amélioration du système de reconnaissance d'entités spatiales

Pour pallier certaines erreurs faites par *Spacy*, nous allons apporter des informations issues de lexiques au système de reconnaissance d'entités nommées. *Spacy* étant basé principalement sur un système d'apprentissage automatique, nous allons ajouter à ce système un gazetier : une liste de noms des lieux (provinces, régions, communes, villes et villages) du Burkina Faso. Cette liste a été produite et utilisée par l'état Burkinabé pour leur enquête permanente agricole (EPA). Cela nous permettra l'annotation de villes et notamment de villages du Burkina Faso.

Nous ajouterons également deux autres lexiques : l'ajout d'un dictionnaire d'adverbes[17] obtenu sur un site d'aide en grammaire et l'ajout d'un lexique avec des mots génériques qui étaient souvent annotés comme LOC par le système de reconnaissance d'entités spatiales, alors qu'il n'y a aucun lien entre ces mots et des entités de lieux : *"S'", "L'", "Ndlr", "Lundi", "«", "Ministre de la Culture", "Mademoiselle", "Commerce", "Directeur", "Agriculture", "Prévenir", "Santé"* etc..

Nous spécifions au script que s'il rencontre des mots présents dans les dictionnaires des deux post-traitements (adverbes et mots génériques), alors *Spacy* ne doit pas les annoter en entités spatiales.

[17] http://www.aidenet.eu/grammaire20b.htm

Ajouter des lexiques et post-traitements à un système de reconnaissance d'entités permet d'avoir une performance plus élevée, mais cela nécessite de tenir à jour les lexiques.

1.3. Evaluation de la reconnaissance d'entités spatiales

- *Méthode d'évaluation*

Pour évaluer la performance de la reconnaissance d'entités spatiales faites par *Spacy*, nous procédons à une méthode d'évaluation. Pour cela il faudra comparer deux corpus : un ensemble de textes avec les entités spatiales annotées à la main, et un ensemble de textes avec les entités spatiales annotées automatiquement par *Spacy*. Nous n'évaluons pas la reconnaissance d'entités temporelles, car la typologie d'*Heideltime* est assez complexe et l'annotation manuelle aurait été compliquée et sûrement erronée.

La comparaison sera synthétisée par les mesures de Précision, de Rappel et de F-mesure. Nous pourrons constater les résultats (erreurs, ajout, suppression, etc.) de l'annotation automatique.

Précision (P) : Proportion de solutions trouvées qui sont attendues.
P = Éléments retrouvés et pertinents (VP) / Éléments retrouvés (VP + FP)

Rappel (R) : Proportion de solutions attendues qui sont trouvées.
R = Éléments retrouvés et pertinents (VP) / Éléments pertinents (VP + FN)

F-mesure : Moyenne harmonique de la précision et du rappel.
F-m = P*R / P+R :

Vrai positif (VP) : réponse positive et X est effectivement présent.
Faux négatif (FN) : réponse négative alors que X est en fait présent.
Faux positif (FP) : réponse positive alors que X est en fait absent.
Vrai négatif (VP) : réponse négative et X est effectivement absent.

- *Résultats de la méthode d'évaluation*

Ces ajouts de règles à la librairie *Spacy* améliorent un peu la reconnaissance des entités, notamment pour les noms des villages et les noms de régions particulières comme *Bobo, Ouaga, Boucle du Mouhoun, Hauts-Bassins, Faso.*

Afin de pouvoir réaliser une évaluation de la reconnaissance des entités spatiales par *Spacy* et des ajouts de règles (lexiques/post-traitements), nous avons annoté manuellement les entités spatiales sur 25 articles du site *l'Observateur paalga*. Nous avons ensuite comparé les résultats de l'annotation de *Spacy* avec les résultats de l'annotation manuelle, et nous avons vu une amélioration de la reconnaissance des entités spatiales grâce à l'ajout du dictionnaire et des post-traitements mentionnés ci-dessus.

	Précision	Rappel	F-mesure
Reconnaissance des entités spatiales par Spacy	0,389	0,663	0,475
Reconnaissance des entités spatiales par Spacy + ajout dictionnaire + post-traitements	0,461	0,674	0,519

Figure 16 : Résultats de l'évaluation de la reconnaissance des entités spatiales

2. *Entités temporelles*

2.1. Analyse

Heideltime effectue une assez bonne reconnaissance des entités temporelles. Il peut y avoir cependant quelques erreurs, comme :

il a été... : *été* étant annoté en entité temporelle *value="SU"* (été ; *summer SU*) alors que c'est une forme du verbe être et non un nom commun.

300 kits d'irrigation goutte à goutte de 1000 m² : *1000* annoté comme *type="DATE"* alors que c'est une quantité.

2.2. Amélioration

Nous devons spécifier au système la date de référence de l'article, afin d'annoter les entités temporelles en fonction de la date de parution de l'article. En effet, le script tague les entités temporelles en prenant comme date de référence celle de l'ordinateur, au lieu de la date de parution de l'article. Par exemple, si un article a paru le 09 janvier 2019, et que dans le texte il y a la le motif 'ce matin', et que nous aurions annoté automatiquement les entités temporelles avec *HeildelTime* le 21 mars 2019, alors cette expression temporelle aurait la valeur "2019-03-21TMO" au lieu de "2019-01-09TMO" (cf. Figure 17). Il faut donc interpréter les dates relatives en fonction d'une date de référence.

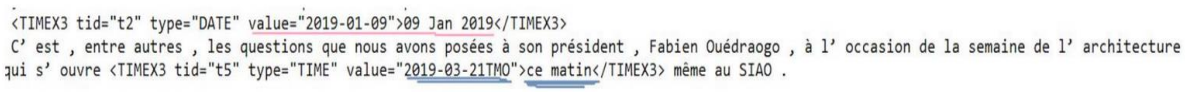


Figure 17 : Extrait d'article annotée par Heideltime avec fausse date de référence

Nous avons dû modifier le script du wrapper Python, en *parsant* d'abord la première ligne du fichier avec *Heideltime* pour extraire la date de l'article, puis *parser* l'ensemble de l'article en passant la date de l'article comme date de référence, en nettoyant les dates avec des expressions régulières, car le script accepte les dates sous une forme spécifique ('yyyy-MM-dd'). De cette manière, *HeidelTime* interprète les dates relatives présentes dans le texte en fonction de la date de référence mise en paramètre.

Également, ce système de reconnaissance peut parfois être assez sensible : par exemple pour une date de parution d'article du site l'Observateur paalga de tel écriture : *12 Aoû 2018*, *Heideltime* n'arrive pas à faire correspondre cette date au moins d'août, car il manque la dernière lettre *"t"* du mois (mais par contre le système reconnaît bien *28 avr 2018*, *25 Déc 2018, 8 Fév 2011,* etc). Nous devons donc faire un prétraitement avec des expressions régulières en remplaçant *Aoû* par *Août* dans les articles avant de les envoyer au système.

2.3. Nouvelle approche

Nous avons rencontré quelques problèmes avec l'annotateur *HeidelTime*, en effet, certains articles n'étaient pas annotés par ce système. Nous avions des articles avec l'annotation des entités spatio-temporelles avec les outils *Spacy* et d'*Heideltime*, et des articles avec seulement l'annotation de *Spacy*. Nous n'avions donc pour certains articles le manque d'annotations temporelles. N'ayant pas trouvé de solutions, et voulant avoir un résultat générique, nous avons choisi de ne plus appliquer ce système, et de seulement chercher la date de parution de l'article avec une expression régulière. Sachant que pour tous les articles, la date de parution de l'article apparaît à la deuxième ligne, cela nous permet de faciliter la recherche et d'avoir un script plus précis et générique pour tous les articles.

```
Liste_text=text.split(".")
Date=liste_text[1]
```

3. *Format des données*

Les articles que nous avons extraits respectent une structure, nous avons la première phrase correspondant au titre de l'article, la deuxième phrase est la date de parution de l'article, le reste de l'article est le texte de l'article, et le dernier motif apparaissant dans le texte de l'article distingué par la barre verticale de telle sorte : |*[09.]*+, est la probabilité d'appartenance au thème.

Campagne agricole : Entre espoir d'une bonne pluviométrie et lutte contre la menace chenille.
Publié le 20 juillet 2018 at 14 h 11 min.
La campagne agricole s'est effectivement installée dans toutes les régions du ['Burkina', 'LOC'] . C' est la principale information donnée ce vendredi 20 juillet 2018 par le ministre de l'agriculture et des aménagements hydrauliques Jacob Ouédraogo au cours d'une conférence de presse à ['Ouagadougou', 'LOC'] . Selon les relevés, le cumul de pluviométrie au 10 juillet 2018 comparativement à la même période de la campagne précédente indique une situation déficitaire dans la majeure partie du territoire, exception faite aux régions du ['Sahel', 'LOC'] , du ['Centre-Ouest', 'LOC'] et des ['Cascades', 'LOC'] qui, elles, ont enregistré un cumul pluviométrique excédentaire. Cependant, rien d'inquiétant selon le ministre Jacob Ouédraogo qui espère même une campagne agricole fructueuse. D'après l'Agence Nationale de la Météorologie (ANAM), une pluviométrie excédentaire à normale est annoncée sur la quasi-totalité du territoire national pour cette campagne agricole. Pour autant, une pluviométrie excédentaire ne signifie pas forcément une sécurité alimentaire. Et le ministre Jacob dit en être conscient. C'est d'ailleurs pour cela qu'il encourage « les producteurs à l'utilisation des semences améliorées, des engrais de qualité et des pesticides homologuées ». En plus, dans le but d'une meilleure sécurisation des exploitations agricoles face à d'éventuelles poches de sécheresse, le ministre en charge de l'agriculture invite les producteurs à la pratique de l'irrigation de complément à partir des bassins de collecte d'eau de ruissèlement et de toute retenue d'eau. La chenille légionnaire d'automne réapparait, le gouvernement riposte… Mais il convient de noter que la campagne agricole n'est pas seulement menacée par la pluviométrie capricieuse. Désormais et cela depuis 2017, la chenille légionnaire d'automne fait partie des principales menaces de l'agriculture. Pour la présente campagne, cette chenille est encore apparue dans certaines régions du ['Burkina', 'LOC'] . Face à ces prédateurs de culture, le département de l'agriculture ne compte pas en rester les bras croisés. La saison agricole a du reste été placée sous le signe de « la lutte contre les prédateurs des cultures ». Selon son premier responsable, des équipes de traitement phytosanitaire sont actuellement déployées dans les régions pour apporter une riposte précoce à ces insectes.\xa0 Pour éradiquer les chenilles légionnaires d'automne hors du ['Burkina', 'LOC'] , le département de Jacob Ouédraogo bénéficie d'un soutien de l'organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO) pour la mise en œuvre d'un projet d'élaboration d'une stratégie de gestion durable de la chenille légionnaire d'automne. Ce qui est attendu de cette campagne… Au titre de cette campagne agricole, il est attendu une production de 5 0800 0000 tonnes de céréales et de plus de 1 500 000 tonnes de cultures de rentes et 979 900 tonnes d'autres cultures vivrières. Pour que ces attentes se concrétisent, le gouvernement a consenti, selon le ministre en charge de l'agriculture, 16 0000 tonnes d'engrais, 8 155 tonnes de semences améliorées, 1 150 000 boutures de manioc et de patate, 27 400 unités d'équipements agricoles, 10 500 animaux de trait et 20 000 litres de pesticides au profit des producteurs. Ces différentes subventions vont coûter au budget de l'Etat environ 25 milliards de FCFA. Maxime KABORE ['Burkina', 'LOC'] 24|0.529963

Figure 18 : Exemple d'article du Burkina24 traitant de sécurité alimentaire et tagué par Spacy

4. *Mise en lien*

4.1. Procédure

Après avoir constitué le corpus, identifié les articles de ce corpus les plus pertinents au thème de la sécurité alimentaire au Burkina Faso et identifié les entités spatiales et temporelles dans ces articles, nous pouvons mettre ces informations en relation avec d'autre données présentes dans les systèmes d'alerte précoce.

L'objectif de cette étape est d'avoir une base de données textuelles synthétisant les informations spatio-temporelles des articles du corpus de la sorte :

Tableau 6 : Exemple du fichier csv recensant les informations spatio-temporelles des articles pertinents

Nom fichier (classé par probabilité appartenance thème sécurité alimentaire)	Date publication	Entités des deux premières phrases	Entités reste du texte	probabilité thème sécurité alimentaire
1.txt	2012	Ouagadougou...	Burkina, Cascades...	0.75
2.txt	...	….	….	

Nous cherchons donc le nom du fichier, la date de publication de l'article, les entités taguées du début de l'article, les entités taguées du reste de l'article, et la probabilité d'appartenance au thème.

Nous utilisons des expressions régulières pour rechercher les informations présentes dans le texte, que nous voulons mettre dans notre tableau.

Pour rechercher la probabilité dans un article :

probabilite=re.findall(r"\|[0-9.]+", text)

Le(s) lieu(s) référents à l'article sont généralement mentionnés au début de l'article. Nous nous concentrons plus sur le début de l'article mais nous ne négligeons pas les autres informations dans le reste de l'article.

Pour ce qui est d'extraire les entités spatiales reconnus, nous cherchons les motifs "['mot', 'LOC']" dans les cinq premières phrases

premier_nom_lieu=re.findall(r" \['[^']', 'LOC'\]",les5p)*

De même pour les entités qui ont été annotées après les cinq premières phrases, seront trouvées avec la même expression régulière, mais dans la partie du texte qui n'a pas encore été analysé :

reste_nom_lieu=re.findall(r" \['[^']', 'LOC'\]",le_reste_texte)*

Puis avec des expressions régulières, nous nettoyons les résultats de sorte à ne plus avoir les balises d'entités spatiales, et juste à avoir le nom de l'entité (soit supprimer les crochets, les guillemets simples et le tag LOC).

Une fois les informations extraites pour chaque article, nous les écrivons dans un fichier csv qui résumera les informations spatio-temporelles des articles des journaux, qui constituera notre base de données textuelles.

Nom fichier (classé par probabilité appartenance thème sécurité alimentaire)	Date publication	Entités des deux premières phrases	Entités reste du texte	probabilité thème sécurité alimentaire
1.txt	2012	Ouagadougou..	Burkina, Cascades...	
2.txt	...	....	....	

Figure 19 : Exemple du fichier csv recensant les informations spatio-temporelles des articles pertinents associées aux cartes

4.2. Résultats

Pour mettre en lien le corpus obtenu aux images satellitaires présentes dans les systèmes d'alerte précoce de sécurité alimentaire au Burkina Faso, nous avons analysé les entités spatiales apparaissant le plus dans les premiers articles (rappel : les premiers articles sont ceux ayant les probabilités les plus élevées au thème spécifique), pour voir si le/les lieu(x) correspondent avec les cartes lors de crise alimentaire.

Par manque de temps et de ressources, nous n'avons analysé que les 20 entités spatiales les plus fréquentes dans les 50 premiers articles de l'Observateur paalga, de Burkina24 et de Lefaso.net. Les fréquences des lieux sont présentées dans le tableau ci-dessous :

Tableau 7 : Résultats des fréquences des entités spatiales dans les 50 premiers articles des journaux

ObservateurPaalga			**Burkina24**			**Lefaso.net**	
Ouagadougou	88		Ouagadougou	119		Ouagadougou	91
Ouaga	30		Centre	33		Centre	68
Centre	18		Ouaga	19		Sahel	26
Pouytenga	13		Bobo	17		Bobo-Dioulasso	22
Bobo	12		Nouna	15		Diallo	11
Sahel	12		Bobo-Dioulasso	10		Gaoua	11
Bobo-Dioulasso	10		Sahel	10		Baskuy	10
Banfora	10		Tanghuin	9		Bobo	9
Logobou	9		Centre-Est	9		Koubri	9
Cascades	9		Nord	9		Nord	9
Tanzéongo	8		Bassins	8		Zorgho	7
Koudougou	8		Dédougou	8		Central	7
Diébougou	6		Bagré	7		Toma	6
Nord	6		Bombissiri	6		Dolo	6
Gaoua	4		Centre-Ouest	5		Pô	6
Gourcy	4		Djibo	5		Koudougou	5
Mané	3		Arbinda	5		Ouahigouya	5
Sanmatenga	3		Larlé	5		Yatenga	5
Centre-Nord	3		Bagassi	5		Plateau	5
Diébou	3		Sara	4		Centre-Nord	5

Pour avoir plus de détails des fréquences, voir l'annexe 3 à la page.71 pour l'Observateur paalga ; l'annexe 4 à la page 72 pour Burkina24 et l'annexe 5 à la page 73 pour Lefaso.net.

Conclusion

Ce mémoire avait pour ambition de mettre en place un système d'acquisition semi-automatique d'un corpus d'articles, tirés de sites d'information du Burkina Faso. Le corpus aurait ainsi été exploité afin d'identifier les articles les plus pertinents à la problématique de la sécurité alimentaire au Burkina Faso. Puis d'extraire les informations spatio-temporelles à partir de ces articles, pour les associer à des images satellitaires de ce pays. Cette mise en lien aurait permis d'apporter une information thématique complémentaire aux données satellitaires présentes dans les systèmes d'alerte précoce pour les pénuries alimentaires au Burkina Faso.

Le premier chapitre détaillait l'acquisition d'un corpus d'articles à partir de sites web en ligne. Le deuxième chapitre présentait l'identification des articles pertinents à la gestion des risques liés à la sécurité alimentaire au Burkina Faso dans le corpus d'articles. Enfin, le troisième chapitre exposait l'application de la reconnaissance d'entités spatio-temporelles dans les articles, afin de les extraire et de les lier aux images satellites présentes dans les systèmes d'alerte précoce.

Nous avons créé le corpus d'articles pertinents à la problématique de sécurité alimentaire au Burkina Faso, et extrait les informations spatio-temporelles présents dans les textes. Avec l'évaluation des thématiques des articles des expertes, nous avons constaté que le corpus avait des articles traitant de la problématique de la sécurité alimentaire, c'est-à-dire traitant de la famine, de la sécheresse, de la production agricole, des prix des aliments, etc. Néanmoins, les articles dans les journaux sont souvent de la publicité pour des journées/ateliers liés à la sécurité alimentaire. Cela biaise un peu la situation réelle de l'état actuel de la sécurité alimentaire. De plus, peu d'articles traitent réellement de la situation de la crise alimentaire au Burkina Faso.

Nous avons constaté que la fouille de texte est un domaine vaste, qui est toujours en évolution avec les méthodes d'apprentissage automatique. Le traitement du langage écrit peut s'avérer difficile, car les mots et segments de mots peuvent avoir plusieurs interprétations et sont donc ambigus pour les systèmes informatiques. De plus, il faut bien analyser le corpus, car le texte est parfois difficile à traiter ; il faut souvent nettoyer les données textuelles pour supprimer du bruit pouvant gêner les systèmes informatiques et fausser les résultats.

Ce projet présente plusieurs perspectives. Extraire d'autres articles de sites d'information du Burkina Faso aurait élargi le corpus et nous aurions pu avoir plus de données et peut-être plus de résultats significatifs. Nous pourrions analyser d'autres sources de données textuelles comme des données sur les réseaux sociaux (Twitter), ou notamment des vidéos journalistiques, car les documents sonores et audiovisuels représentent une nouvelle forme de source d'information de plus en plus répandue aux cours des années. Pour finir, nous pourrions nous attarder sur le traitement de l'identification des articles pertinents à la thématique recherchée *(apprentissage non-supervisé avec le modèle LDA et Word2Vec)*, en employant des lexiques différents ; par exemple influer le nombre d'entrées du lexique (plus de mots, moins de mots), ou encore peser avoir des lexiques avec des mots plus spécifiques, techniques au thème, etc. pour voir si les résultats d'identification des articles sont plus représentatifs ou non à la problématique de la sécurité alimentaire au Burkina Faso.

Bibliographie

Allahyari, M., Pouriyeh, S.A., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., & Kochut, K.J. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *ArXiv, abs/1707.02919.*

[9] Basu, M., Ghosh, K., Das S., R, Dey, Bandyopadhyay, S., & Ghosh, S. (2017). Identifying Post-Disaster Resource Needs and Availabilities from Microblogs. In Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM), pages 427–430.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3(Jan), p.993–1022.*

Chang, A & Manning, C. (2012). SUTIME: A Library for Recognizing and Normalizing Time Expressions.

Dang, S., & Ahmad, P.H. (2014). Mining : Techniques and its Application. In *IJETI International Journal of Engineering & Technology Innovations, Vol. 1 Issue 4,November 2014*

Hatmi, M. (2014). Reconnaissance des entités nommées dans des documents multimodaux, thèse de doctorat, Université de Nantes.

Imran, M. & Castillo, C. & Diaz, F. & Vieweg, S. (2015). Processing Social Media Messages in Mass Emergency: A Survey. *In ACM Computing Surveys. 10.1145/2771588.*

[7] Interdonato I., Doucet A., Guillaume J.L (2018): Unsupervised Crisis Information Extraction from Twitter Data. ASONAM 2018: 579-580

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2017). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211.

[6] Nazer T. H., Morstatter F., Dani H., and Liu H. (2016). Finding requests in social media for disaster relief. In Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM), pages 1410–1413.

[5] Nazer T. H., Xue G., Ji Y., Liu H. (2017) : Intelligent Disaster Response via Social Media Analysis A Survey. SIGKDD Explorations 19(1): 46-59

Nouvel, D. (2012). Reconnaissance des entités nommées par exploration de règles d'annotation – Interpréter les marqueurs d'annotation comme instructions de structuration locale.

Roche, M. (2016). Knowledge Discovery from Texts on Agriculture Domain. MISC'2016, Constantine, Algeria

[1] Shilpa, D. & Peerzada H.A (2015). A review of Text Mining techniques with Various Application Areas. *In International Journal of Science and Research (IJSR) 4(2) :2461-2466.*

Tawofaing, A.F. (2018). Recherche d'entités nommées complexes sur le Web - propositions pour l'extraction et pour le calcul de similarité

Sitographie

[3] Beginners Guide to Topic Modeling in Python [en ligne] analyticsvidhya.com : https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/, consultée le 3 juin 2019

[2] Introduction à la fouille de textes [en ligne]. Lattice Cnrs : http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf, consultée le 24 juin 2019

[4] Pense-bête d'apprentissage supervisé [en ligne]. Stanfortd.edu : https://stanford.edu/~shervine/l/fr/teaching/cs-229/pense-bete-apprentissage-supervise, consultée le 20 mai 2019

[8] Vector Representations of Words [en ligne]. Tensorflow.org : https://www.tensorflow.org/tutorials/representation/word2vec , consultée le 10 juillet 2019

TimeML Annotation Guidelines [en ligne]. Timeml.org : http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf, consultée le 19 avril 2019

Table des figures

Table des tableaux

Table des annexes

Annexe 1
Classification Topic Modeling LDA

Top 10 words for topic #0:['ordre', 'kadré', 'pays', 'houet', 'présidentielle', 'sanou', 'opposition', 'militants', 'président', 'parti']

Top 10 words for topic #1:['maire', 'ministre', 'municipal', 'année', 'jeunes', 'élèves', 'ville', 'cours', 'scolaires', 'parti']

Top 10 words for topic #2:['finale', 'face', 'football', 'groupe', 'coupe', 'joueurs', 'traoré', 'équipe', 'etalons', 'match']

Top 10 words for topic #3:['ordre', 'magistrat', 'comptes', 'homme', 'pays', 'faso', 'parti', 'burkina', 'etat', 'président']

Top 10 words for topic #4:['terroriste', 'assaillants', 'défense', 'gendarmerie', 'état', 'police', 'forces', 'sécurité', 'terroristes', 'attaque']

Top 10 words for topic #5:['drian', 'monde', 'président', 'nègre', 'burkinabè', 'armée', 'burkina', 'négritude', 'humains', 'droits']

Top 10 words for topic #6:['cas', 'kadhafi', 'fois', 'ouagadougou', 'etat', 'chose', 'pays', 'jours', 'ministre', 'gouvernement']

Top 10 words for topic #7:['ouagadougou', 'ministre', 'galian', 'bobo', 'produits', 'presse', 'ouest', 'faso', 'burkina', 'prix']

Top 10 words for topic #8:['cancer', 'types', 'publique', 'chrétien', 'africain', 'céréales', 'santé', 'plantes', 'alimentaires', 'aliments']

Top 10 words for topic #9:['transports', 'namibie', 'président', 'transporteurs', 'ucrb', 'windhoek', 'ville', 'otraf', 'routiers', 'chauffeurs']

Top 10 words for topic #10:['société', 'politiques', 'politique', 'burkina', 'etat', 'réformes', 'président', 'gorba', 'dialogue', 'faso']

Top 10 words for topic #11:['allergie', 'monde', 'alliance', 'manéga', 'majorité', 'politique', 'président', 'partis', 'pays', 'apmp']

Top 10 words for topic #12:['appel', 'électricité', 'pays', 'président', 'évaluation', 'commission', 'burkina', 'obésité', 'chirurgie', 'offres']

Top 10 words for topic #13:['esclavage', 'religion', 'foi', 'jésus', 'christianisme', 'noir', 'noirs', 'christ', 'esclaves', 'dieu']

Top 10 words for topic #14:['situation', 'charge', 'cancer', 'santé', 'burkina', 'parents', 'famille', 'cas', 'enfant', 'enfants']

Top 10 words for topic #15:['mosquée', 'jour', 'monde', 'prière', 'nuit', 'femmes', 'musulmans', 'islam', 'mois', 'dieu']

Top 10 words for topic #16:['consulaire', 'président', 'côte', 'ivoire', 'vote', 'burkina', 'faso', 'carte', 'pays', 'burkinabè']

Top 10 words for topic #17:['cas', 'situation', 'erdogan', 'énergie', 'personnes', 'communiste', 'turquie', 'afrique', 'chaleur', 'pays']

Top 10 words for topic #18:['nationale', 'dents', 'radio', 'jour', 'faso', 'put', 'président', 'burkina', 'justice', 'zongo']

Top 10 words for topic #19:['pays', 'décision', 'affaire', 'magistrats', 'faso', 'burkinabè', 'etat', 'président', 'justice', 'conseil']

Top 10 words for topic #20:['gestion', 'municipal', 'rencontre', 'ouagadougou', 'comité', 'gouvernement', 'etat', 'commune', 'conseil', 'maire']

Top 10 words for topic #21:['village', 'ensemble', 'communauté', 'communautés', 'violences', 'morts', 'peuls', 'victimes', 'drame', 'yirgou']

Top 10 words for topic #22:['newton', 'déclaration', 'partis', 'enrôlement', 'électoral', 'président', 'opposition', 'plateforme', 'commissaires', 'ceni']

Top 10 words for topic #23:['défense', 'culture', 'religions', 'pénitentiaire', 'dialogue', 'snc', 'ministre', 'foi', 'gsp', 'sécurité']

Top 10 words for topic #24:['taïwan', 'chinois', 'relations', 'développement', 'monde', 'faso', 'burkinabè', 'chine', 'burkina', 'pays']

Top 10 words for topic #25:['ministre', 'sud', 'etat', 'opposition', 'gouvernement', 'burkina', 'politique', 'président', 'parti', 'pays']

Top 10 words for topic #26:['etat', 'jours', 'marché', 'ramaphosa', 'président', 'parti', 'sud', 'jacob', 'réfugiés', 'anc']

Top 10 words for topic #27:['milieu', 'vérité', 'mise', 'esclavage', 'révèle', 'titre', 'ouagadougou', 'santé', 'quartiers', 'lotis']

Top 10 words for topic #28:['dossier', 'procédure', 'cour', 'audience', 'président', 'militaire', 'chambre', 'avocats', 'tribunal', 'procès']

Top 10 words for topic #29:['chef', 'secte', 'buhari', 'gouvernement', 'abubakar', 'fois', 'président', 'nigeria', 'haram', 'boko']

Top 10 words for topic #30:['film', 'festival', 'continent', 'français', 'pays', 'africains', 'france', 'africain', 'cinéma', 'afrique']

Top 10 words for topic #31:['etat', 'avocat', 'militaire', 'président', 'coup', 'témoin', 'major', 'colonel', 'diendéré', 'général']

Top 10 words for topic #32:['kinshasa', 'électorale', 'etat', 'présidentielle', 'tshisekedi', 'pays', 'congolais', 'président', 'congo', 'kabila']

Top 10 words for topic #33:['villages', 'ménages', 'centre', 'nord', 'populations', 'kaya', 'projet', 'femmes', 'personnes', 'déplacés']

Top 10 words for topic #34:['démocratie', 'candidat', 'pu', 'ancien', 'eddie', 'chef', 'congrès', 'politique', 'président', 'parti']

Top 10 words for topic #35:['parti', 'politique', 'politiques', 'conseil', 'burkinabè', 'fédération', 'pays', 'monde', 'burkina', 'président']

Top 10 words for topic #36:['œuvre', 'conseil', 'ministère', 'pays', 'burkinabè', 'projet', 'développement', 'ministre', 'faso', 'burkina']

Top 10 words for topic #37:['services', 'service', 'travail', 'syntsha', 'syndicat', 'agents', 'gouvernement', 'travailleurs', 'grève', 'santé']

Top 10 words for topic #38:['cfop', 'parlementaire', 'députés', 'chef', 'sakandé', 'bala', 'nationale', 'assemblée', 'président', 'opposition']

Top 10 words for topic #39:['maire', 'gestion', 'communication', 'institution', 'commune', 'conseillers', 'conseil', 'médias', 'presse', 'président']

Top 10 words for topic #40:['guerre', 'pierre', 'peine', 'grand', 'mai', 'monde', 'cour', 'appel', 'bemba', 'pays']

Top 10 words for topic #41:['personnes', 'coucher', 'hommes', 'journée', 'ancêtres', 'nuit', 'tolérance', 'dormir', 'général', 'sommeil']

Top 10 words for topic #42:['mesures', 'œuvre', 'fête', 'catastrophes', 'valentin', 'allah', 'vue', 'saint', 'porc', 'santé']

Top 10 words for topic #43:['présidentielle', 'mandat', 'barry', 'candidature', 'boussouma', 'vice', 'ministre', 'pays', 'parti', 'président']

Top 10 words for topic #44:['police', 'loi', 'koudougou', 'affaire', 'élève', 'enregistrement', 'débats', 'justice', 'zongo', 'procès']

Top 10 words for topic #45:['terroristes', 'parle', 'saleh', 'terroriste', 'juges', 'justice', 'magistrats', 'faso', 'pays', 'président']

Top 10 words for topic #46:['difficultés', 'élèves', 'personnes', 'pays', 'élève', 'enfants', 'classe', 'cours', 'école', 'parents']

Top 10 words for topic #47:['manifestants', 'rue', 'gaïd', 'peuple', 'salah', 'président', 'algérie', 'armée', 'général', 'bouteflika']

Top 10 words for topic #48:['statue', 'article', 'projet', 'conseil', 'mot', 'conducteurs', 'mémorial', 'heures', 'président', 'tricycles']

Top 10 words for topic #49:['aqmi', 'armée', 'niger', 'biya', 'burkina', 'travail', 'cameroun', 'ouest', 'forces', 'pays']

Top 10 words for topic #50:['ministre', 'président', 'burkina', 'comité', 'chef', 'sawadogo', 'pèlerins', 'ouédraogo', 'vol', 'ouagadougou']

Top 10 words for topic #51:['enfant', 'règles', 'exemple', 'dr', 'cas', 'traitement', 'santé', 'personnes', 'maladies', 'maladie']

Top 10 words for topic #52:['adjudant', 'soldat', 'major', 'parquet', 'accusé', 'barre', 'président', 'sergent', 'militaire', 'chef']

Top 10 words for topic #53:['burkina', 'cas', 'burkinabè', 'corruption', 'santé', 'salifou', 'etat', 'juge', 'président', 'bassolet']

Top 10 words for topic #54:['ministre', 'étape', 'édition', 'tour', 'burkina', 'faso', 'pays', 'routière', 'burkinabè', 'sécurité']

Top 10 words for topic #55:['ouagadougou', 'burkinabè', 'gounghin', 'municipal', 'tall', 'roger', 'vie', 'parents', 'hommage', 'cimetière']

Top 10 words for topic #56:['médecins', 'parti', 'ordre', 'pdp', 'peuple', 'pays', 'burkinabè', 'burkina', 'chinois', 'chine']

Top 10 words for topic #57:['section', 'bougouriba', 'nationale', 'sg', 'prix', 'service', 'forêt', 'yacouba', 'sawadogo', 'police']

Top 10 words for topic #58:['situation', 'faso', 'nationale', 'parti', 'peuple', 'burkina', 'président', 'burkinabè', 'politique', 'pays']

Top 10 words for topic #59:['pays', 'ancien', 'togo', 'politique', 'bolloré', 'lomé', 'opposition', 'gnassingbé', 'président', 'faure']

Top 10 words for topic #60:['place', 'forêt', 'affaire', 'in', 'environnement', 'durable', 'affaires', 'burkina', 'développement', 'pays']

Top 10 words for topic #61:['repas', 'nutrition', 'service', 'hôpital', 'alimentation', 'place', 'malades', 'ceni', 'jour', 'yalgado']

Top 10 words for topic #62:['chargé', 'achille', 'artisans', 'militants', 'eddie', 'bureau', 'président', 'secrétaire', 'congrès', 'parti']

Top 10 words for topic #63:['vie', 'formation', 'travail', 'évaluation', 'etat', 'techniques', 'difficultés', 'bourses', 'besoin', 'pays']

Top 10 words for topic #64:['forces', 'force', 'terroristes', 'sécurité', 'burkina', 'terrorisme', 'g5', 'pays', 'mali', 'sahel']

Top 10 words for topic #65:['vote', 'politique', 'élection', 'résultats', 'président', 'tour', 'candidat', 'scrutin', 'opposition', 'présidentielle']

Top 10 words for topic #66:['personnel', 'test', 'générale', 'environnement', 'presse', 'travail', 'nationale', 'cgt', 'recrutement', 'cnss']

Top 10 words for topic #67:['lycée', 'examen', 'centre', 'écoles', 'éducation', 'année', 'enseignants', 'candidats', 'école', 'élèves']

Top 10 words for topic #68:['guinée', '2010', 'politique', 'monnaie', 'côte', 'ivoire', 'franc', 'président', 'pays', 'gbagbo']

Top 10 words for topic #69:['blessés', 'cause', 'maladie', 'camion', 'symptômes', 'insuffisance', 'cardiaque', 'douleur', 'travail', 'facteurs']

Top 10 words for topic #70:['mai', 'monde', 'temps', 'fois', 'faso', 'avocat', 'défense', 'ouédraogo', 'président', 'office']

Top 10 words for topic #71:['ministre', 'salaires', 'commun', 'economie', 'etat', 'travailleurs', 'ministère', 'finances', 'fonds', 'agents']

Top 10 words for topic #72:['président', 'textile', 'gouvernement', 'koudougou', 'marc', 'ravalomanana', 'crise', 'faso', 'fani', 'usine']

Top 10 words for topic #73:['administrateur', 'faso', 'décret', 'conseiller', 'titre', 'inspecteur', 'conseil', 'directeur', 'classe', 'échelon']

Top 10 words for topic #74:['etat', 'sidi', 'accusé', 'parquet', 'écoutes', 'président', 'éléments', 'diendéré', 'bassolet', 'général']

Top 10 words for topic #75:['queue', 'partie', 'rat', 'bouche', 'tête', 'tukguili', 'enfant', 'terrible', 'mauvaise', 'haleine']

Top 10 words for topic #76:['culture', 'public', 'ouédraogo', 'œuvres', 'bureau', 'édition', 'musique', 'artiste', 'kundé', 'artistes']

Top 10 words for topic #77:['identifiant', 'entreprises', 'logiciel', 'direction', 'dgi', 'contribuables', 'électronique', 'documents', 'impôts', 'unique']

Top 10 words for topic #78:['information', 'fin', 'presse', 'mine', 'or', 'heures', 'mines', 'essakane', 'travailleurs', 'ministre']

Top 10 words for topic #79:['santé', 'temps', 'œuvre', 'équipe', 'chef', 'sécurité', 'nouveau', 'etat', 'gouvernement', 'ministre']

Top 10 words for topic #80:['adjoint', 'ministre', 'femmes', 'administrateur', 'civil', 'femme', 'province', 'secrétaire', 'administratif', 'département']

Top 10 words for topic #81:['personnes', 'moment', 'ouagadougou', 'circulation', 'route', 'véhicules', 'usagers', 'véhicule', 'ouédraogo', 'ville']

Top 10 words for topic #82:['gynécologue', 'dr', 'pays', 'femmes', 'fièvre', 'cas', 'monde', 'femme', 'utérus', 'fibromes']

Top 10 words for topic #83:['lutte', 'opération', 'mossé', 'gouvernement', 'paren', 'mogho', 'lac', 'ren', 'peulhs', 'corruption']

Top 10 words for topic #84:['mai', 'autorités', 'agents', 'chef', 'etat', 'barsalogho', 'sécurité', 'gouvernement', 'ministère', 'ministre']

Top 10 words for topic #85:['poa', 'carrières', 'fois', 'kyon', 'eaux', 'saison', 'village', 'route', 'pont', 'eau']

Top 10 words for topic #86:['ado', 'président', 'politique', 'eddie', 'ancien', 'suppléant', 'titulaire', 'komboïgo', 'membre', 'parti']

Top 10 words for topic #87:['vie', 'intuition', 'cour', 'pu', 'temps', 'jour', 'ouagadougou', 'femme', 'intrigante', 'cousin']

Top 10 words for topic #88:['africain', 'urbaine', 'anc', 'guide', 'kadhafi', 'libye', 'président', 'mandela', 'sud', 'afrique']

Top 10 words for topic #89:['pays', 'adama', 'président', 'dakar', 'sénégal', 'sénégalais', 'idrissa', 'ouédraogo', 'macky', 'sall']

Top 10 words for topic #90:['lutte', 'nationale', 'logement', 'coalition', 'hydrocarbures', 'vie', 'populations', 'ccvc', 'gouvernement', 'prix']

Top 10 words for topic #91:['lieu', 'burkina', 'victime', 'football', 'burkinabè', 'président', 'cameroun', 'marché', 'édition', 'etalons']

Top 10 words for topic #92:['bado', 'élucubreur', 'vote', 'puis', 'pays', 'président', 'parti', 'élucubrations', 'toégui', 'ministre']

Top 10 words for topic #93:['etat', 'jour', 'fin', 'sexualité', 'président', 'khartoum', 'béchir', 'guerre', 'soudan', 'pays']

Top 10 words for topic #94:['public', 'avocat', 'tribunal', 'capitaine', 'militaire', 'client', 'faits', 'lieutenant', 'accusé', 'parquet']

Top 10 words for topic #95:['ministre', 'mois', 'situation', 'pays', 'travail', 'mai', 'président', 'politique', 'parti', 'gouvernement']

Top 10 words for topic #96:['maître', 'mentales', 'apprentissage', 'cartes', 'président', 'education', 'kéré', 'formation', 'élections', 'ceni']

Top 10 words for topic #97:['etat', 'yaméogo', 'suspension', 'arrêté', 'générale', 'situation', 'assemblée', 'avocats', 'mai', 'audiences']

Top 10 words for topic #98:['pays', 'lopez', 'homme', 'vie', 'livre', 'fois', 'nouveau', 'situation', 'eau', 'jour']

Top 10 words for topic #99:['in', 'question', 'électoral', 'jour', 'grève', 'etat', 'ministère', 'ministre', 'burkinabè', 'gouvernement']

Annexe 2
Erreurs annotations spatiales

Extrait article Burkina24 : demo ner spacy

Sécurité alimentaire ORG : La culture de saison sèche lancée au Burkina MISC . ORG Publié le 24 novembre 2018 at 17 h 36 min. LOC Ce samedi 24 novembre 2018 a marqué le lancement officiel de la campagne agricole de saison sèche 2018-2019, à Nabadogo PER , dans la commune rurale de Sabou LOC , région du Centre-Ouest LOC sous le thème « contribution de la production de saison sèche à l'atteinte d'une sécurité alimentaire et nutritionnelle durable ». Cette cérémonie vise à partager les orientations du département de l'agriculture en matière de production de saison sèche avec l'ensemble des acteurs du monde rural. Une campagne agricole s'achève, une autre commence. LOC

Extrait article observateur paalga demo ner spacy

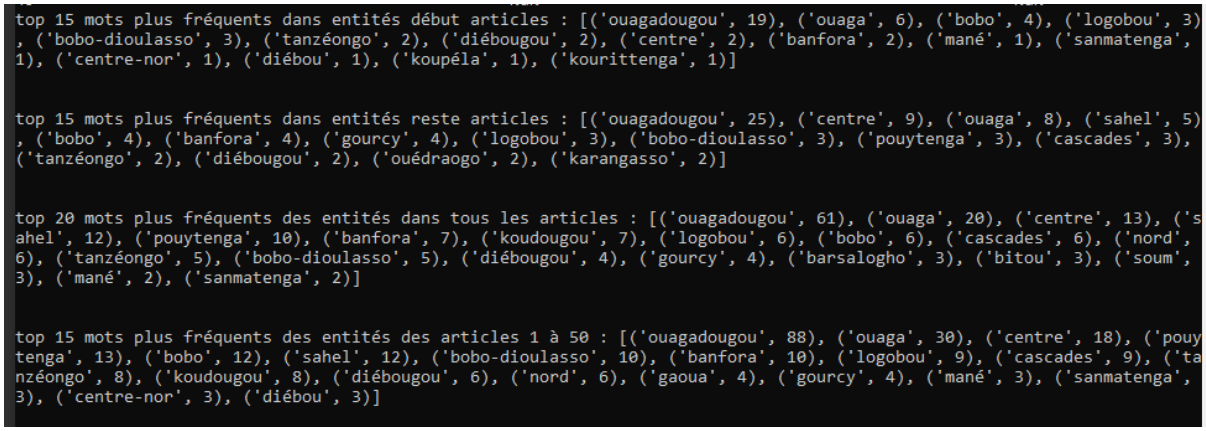
Froid PER , vent et hygiène corporelle : Xérose PER cutanée , dermatite atopique ... ces maladies qui vous collent à la peau

29 Jan 2019 LOC Ah ce temps ! Malgré la chaleur ressentie dans la journée , le froid est toujours présent et sévit au petit matin . A cela s' ajoute ce vent sec de l' harmattan qui souffle et agresse les narines , lèvres et autres parties du corps , provoquant certaines maladies . En pareilles circonstances , prendre soin de son corps , aussi bien sur les plans esthétique qu' hygiénique , devient un véritable casse-tête pour de nombreuses personnes . Quel genre de savon utiliser ? Après le bain , faut - il se passer une crème ou une pommade sur le corps ? Comment prendre soin des lèvres , des mains et des pieds pour qu' ils ne subissent pas l' effet du vent ? Quelles LOC sont les maladies de la peau liées au froid ? Nous avons consulté pour vous une spécialiste de la peau , le Dr Nomtondo Amina Ouédraogo MISC , épouse Zoungrana PER , dermatologue - vénéréologue - allergologue en service à l' hôpital Yalgado LOC . LOC

Extrait article observateur paalga demo ner spacy

RN1 LOC , à Koubri LOC , à Saaba LOC , à Pabré LOC , mais il existe hélas des abattoirs clandestins dans la commune de Ouagadougou LOC . Quelle est la politique d' hygiène et de suivi des animaux sur les sites ? Les abattages effectués à la SO.GE.A.O LOC sont suivis par des vétérinaires assermentés du ministère des Ressources LOC animales et Halieutiques PER . Ils assurent l'inspection des carcasses. Les animaux qui sont introduits à l' abattoir sont identifiés par le numéro des bouchers et ensuite par un numéro de série au moment de l' abattag . Les carcasses portent aussi le numéro du boucher à la livraison des viandes. L' A.F.O. PER abat essentiellement des bovins , des petits ruminants (ovins, caprins) et des porcs. Les bouchers sont-ils formés ? Ils ont bénéficié d' une formation du Programme LOC d'appui aux filières agro- Sylvo PER -pastorale qui a pris fin en 2017. LOC

Annexe 3
Analyse fréquence entités nommées journal Observateur paalga



top 15 mots plus fréquents dans entités situés au début articles, articles 1 à 50 : [('ouagadougou', 19), ('ouaga', 6), ('bobo', 4), ('logobou', 3), ('bobo-dioulasso', 3), ('tanzéongo', 2), ('diébougou', 2), ('centre', 2), ('banfora', 2), ('mané', 1), ('sanmatenga', 1), ('centre-nor', 1), ('diébou', 1), ('koupéla', 1), ('kourittenga', 1)]

top 20 mots plus fréquents des entités des articles en entiers, de 1 à 50 : [('ouagadougou', 88), ('ouaga', 30), ('centre', 18), ('pouytenga', 13), ('bobo', 12), ('sahel', 12), ('bobo-dioulasso', 10), ('banfora', 10), ('logobou', 9), ('cascades', 9), ('tanzéongo', 8), ('koudougou', 8), ('diébougou', 6), ('nord', 6), ('gaoua', 4), ('gourcy', 4), ('mané', 3), ('sanmatenga', 3), ('centre-nor', 3), ('diébou', 3)]

top 15 mots plus fréquents dans entités reste articles : [('ouagadougou', 25), ('centre', 9), ('ouaga', 8), ('sahel', 5), ('bobo', 4), ('banfora', 4), ('gourcy', 4), ('logobou', 3), ('bobo-dioulasso', 3), ('pouytenga', 3), ('cascades', 3), ('tanzéongo', 2), ('diébougou', 2), ('ouédraogo', 2), ('karangasso', 2)]

top 20 mots plus fréquents des entités dans tous les articles : [('ouagadougou', 61), ('ouaga', 20), ('centre', 13), ('sahel', 12), ('pouytenga', 10), ('banfora', 7), ('koudougou', 7), ('logobou', 6), ('bobo', 6), ('cascades', 6), ('nord', 6), ('tanzéongo', 5), ('bobo-dioulasso', 5), ('diébougou', 4), ('gourcy', 4), ('barsalogho', 3), ('bitou', 3), ('soum', 3), ('mané', 2), ('sanmatenga', 2)]

Annexe 4
Analyse fréquence entités nommées journal Burkina24

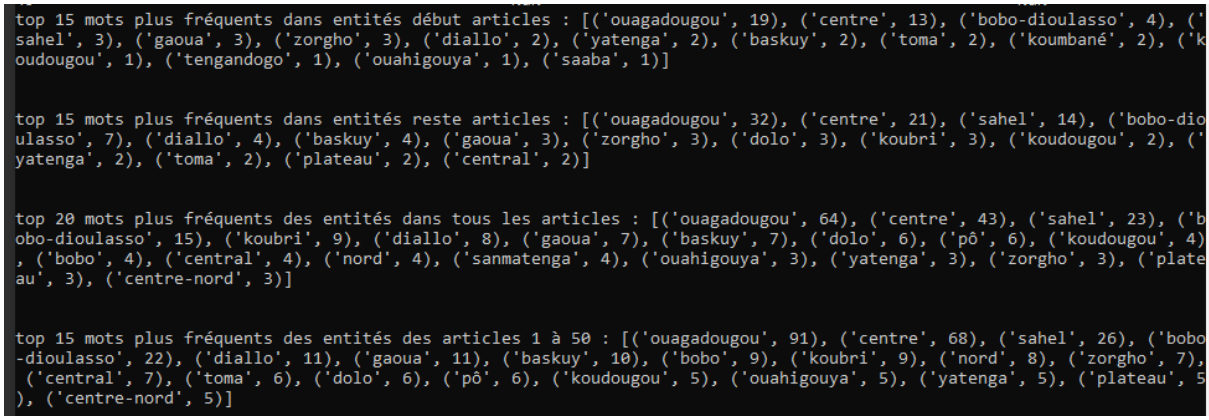
top 15 mots plus fréquents dans entités situés au début articles, articles 1 à 50 : [('ouagadougou', 28), ('nouna', 5), ('bagré', 3), ('bobo', 3), ('bassins', 3), ('dédougou', 3), ('ouaga', 2), ('kombissiri', 2), ('sara', 2), ('bagassi', 2), ('centre-est', 2), ('nabadogo', 1), ('sabou', 1), ('centre-ouest', 1), ('tanghin', 1)]

<u>top 20 entités plus fréquentes des articles (en entiers) 1 à 50</u> : [('ouagadougou', 119), ('centre', 33), ('ouaga', 19), ('bobo', 17), ('nouna', 15), ('bobo-dioulasso', 10), ('sahel', 10), ('tanghin', 9), ('centre-est', 9), ('nord', 9), ('bassins', 8), ('dédougou', 8), ('bagré', 7), ('kombissiri', 6), ('centre-ouest', 5), ('djibo', 5), ('arbinda', 5), ('larlé', 5), ('bagassi', 5), ('sara', 4)]

<u>top 15 entités plus fréquentes reste articles</u> : [('ouagadougou', 41), ('centre', 12), ('ouaga', 7), ('bobo', 5), ('nouna', 5), ('nord', 4), ('komki', 4), ('bagré', 3), ('bassins', 3), ('dédougou', 3), ('centre-est', 3), ('centre-ouest', 2), ('sahel', 2), ('kombissiri', 2), ('sara', 2)]

<u>top entités plus fréquentes dans tous les articles</u> : [('ouagadougou', 86), ('centre', 26), ('ouaga', 17), ('bobo', 10), ('nouna', 10), ('nord', 9), ('bobo-dioulasso', 8), ('sahel', 8), ('centre-est', 6), ('tanghin', 5), ('dédougou', 5), ('centre-ouest', 4), ('bagré', 4), ('bassins', 4), ('kombissiri', 4), ('komki', 4), ('boudtenga', 4), ('djibo', 3), ('arbinda', 3), ('larlé', 3)]

Annexe 5
Analyse fréquence entités nommées journal Lefaso.net



top 15 mots plus fréquents dans entités situés au début articles, articles 1 à 50 : [('ouagadougou', 19), ('centre', 13), ('bobo-dioulasso', 4), ('sahel', 3), ('gaoua', 3), ('zorgho', 3), ('diallo', 2), ('yatenga', 2), ('baskuy', 2), ('toma', 2), ('koumbané', 2), ('koudougou', 1), ('tengandogo', 1), ('ouahigouya', 1), ('saaba', 1)]

<u>top 20 entités plus fréquentes des articles (en entiers) 1 à 50</u> : [('ouagadougou', 91), ('centre', 68), ('sahel', 26), ('bobo-dioulasso', 22), ('diallo', 11), ('gaoua', 11), ('baskuy', 10), ('bobo', 9), ('koubri', 9), ('nord', 8), ('zorgho', 7), ('central', 7), ('toma', 6), ('dolo', 6), ('pô', 6), ('koudougou', 5), ('ouahigouya', 5), ('yatenga', 5), ('plateau', 5), ('centre-nord', 5)]

<u>top 15 entités plus fréquentes reste articles</u> : [('ouagadougou', 32), ('centre', 21), ('sahel', 14), ('bobo-dioulasso', 7), ('diallo', 4), ('baskuy', 4), ('gaoua', 3), ('zorgho', 3), ('dolo', 3), ('koubri', 3), ('koudougou', 2), ('yatenga', 2), ('toma', 2), ('plateau', 2), ('central', 2)]

<u>top 20 entités plus fréquentes dans tous les articles</u> : [('ouagadougou', 64), ('centre', 43), ('sahel', 23), ('bobo-dioulasso', 15), ('koubri', 9), ('diallo', 8), ('gaoua', 7), ('baskuy', 7), ('dolo', 6), ('pô', 6), ('koudougou', 4), ('bobo', 4), ('central', 4), ('nord', 4), ('sanmatenga', 4), ('ouahigouya', 3), ('yatenga', 3), ('zorgho', 3), ('plateau', 3), ('centre-nord', 3)]

Table des matières

MOTS-CLÉS : Fouille de texte, Traitement Automatique du Langage, Apprentissage automatique, Corpus, Sécurité alimentaire

RÉSUMÉ

Des systèmes d'alerte précoce sur la gestion des risques liés à la sécurité alimentaire sont mis en place pour informer les acteurs, afin d'adapter et améliorer le développement de l'agriculture. Ces systèmes analysent principalement des images satellitaires (par exemple des images de champs agricoles) et des données quantitatives (par exemple le prix du marché alimentaire). Cependant, peu de données textuelles sont intégrées dans leur traitement, or elles sont de plus en plus abondantes et accessibles sur Internet. Elles ne décrivent pas parfaitement la situation géographique, mais elles pourraient apporter une information ou une observation complémentaire aux images satellitaires. Ce mémoire propose donc d'acquérir un corpus d'articles de sites d'actualités en lien avec le thème de la sécurité alimentaire au Burkina Faso, d'extraire les informations spatio-temporelles présentes dans les articles, puis de les mettre en relation avec les données présentes dans les systèmes d'alerte précoce de sécurité alimentaire.

KEYWORDS : Text mining, Natural Language Processing, Machine learning, Corpus, Food safety

ABSTRACT

Food security early warning systems are set up to inform stakeholders in order to adapt and improve agricultural development. These systems mainly analyze satellite images (e. g. images of agricultural fields) and quantitative data (e. g. food market prices). However, few textual data are integrated into their processing, yet they are increasingly abundant and accessible on the Internet. They do not perfectly describe the geographical situation, but they could provide additional information or observations to satellite images. This thesis therefore proposes to acquire a corpus of articles from newspapers sites related to the theme of food security in Burkina Faso, to extract the spatio-temporal information contained in the corpus, and then to link them to the data contained in food security early warning systems.