Fact-checking dans les médias

Sanda Hachana

Rapport de stage

Effectué au sein de l'entreprise **BUSTER.AI**

Sanda
HACHANA

Tuteur universitaire : **Claude Ponton**

Tuteur professionnel : **Aurélien Cluzeau**

UFR LLASIC

Mémoire de master 2 mention Sciences du langage- 20 crédits

Parcours : Industries de la langue, orientation professionnelle

Année universitaire 2020 - 2021

Rapport de stage

Effectué au sein de l'entreprise **BUSTER.AI**

Sanda
HACHANA

Tuteur universitaire : **Claude Ponton**

Tuteur professionnel : **Aurélien Cluzeau**

UFR LLASIC

Mémoire de master 2 mention Sciences du langage- 20 crédits

Parcours : Industries de la langue, orientation professionnelle

Année universitaire 2020 - 2021

Remerciements

Je tiens, avant tout, à remercier mon tuteur universitaire Claude Ponton de m'avoir encadrée, conseillée et aidée tout au long de mon stage et d'avoir surtout été présent lorsque j'en avais besoin pendant les moments les plus difficiles du stage et ce, jusqu'à la dernière minute.

Je remercie l'entreprise Buster.Ai, et plus précisément mon tuteur Aurélien Cluzeau de m'avoir accueillie pendant ces cinq derniers mois et de m'avoir donné l'opportunité de découvrir le monde professionnel. Je remercie aussi Mostafa, Léa ainsi que tous les autres membres de Buster de m'avoir intégrée au sein de leur équipe.

Je remercie particulièrement mon enseignante Agnès Tutin qui m'avait envoyé l'offre de stage pour ce poste et m'y avait encouragée à y postuler. Je remercie également tous mes autres enseignants du Master, sans qui, je n'en serai pas là aujourd'hui.

DÉCLARATION ANTI-PLAGIAT

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.

PRENOM : Sanda

NOM : HACHANA

DATE : 2021

Sommaire

Introduction

« Il est interdit de chasser la baleine au Japon » ;

« La tarte au fruit est le dessert préféré des français.» ;

« Marie Curie est la première femme à recevoir un prix Nobel.»

Ces citations sont-elles vraies ou fausses ? S'agit-il de fausses informations (autrement dit de « fake news ») ? Comment peut-on différencier les fausses informations des vraies informations ? Il n'est pas si simple de répondre à ces questions, en effet, lorsque nous lisons un document, nous ne savons plus s'il s'agit d'une vraie ou d'une fausse information, nous remettons souvent la fiabilité de cette actualité en question. La vérification des faits (ou le « fact-checking ») est une technique qui examine des données afin d'en déterminer leur véracité et leur exactitude. Cette technique est utilisée en France depuis 1995 par des journalistes, mais certains chercheurs en intelligence artificielle s'y intéressent également. C'est le cas des chercheurs de l'entreprise Buster.Ai.

Dans le cadre de mon Master Sciences du langage, parcours Industries de la langue à l'Université Grenoble Alpes, j'ai effectué mon stage de fin d'études dans cette entreprise qui traite cette problématique avec l'intelligence artificielle, me permettant ainsi d'appliquer mes connaissances acquises lors de ma formation universitaire durant ce stage. Du 1er Mars 2021 au 30 Juillet 2021 j'ai donc rejoint l'entreprise Buster.AI qui se situe à Paris. Au cours de ce stage, j'ai pu apprendre des nouvelles techniques de l'intelligence artificielle et découvrir les méthodes du travail en équipe.

Afin de mieux analyser ces 5 mois de stage passés au sein de Buster.AI, il sera présenté, dans un premier temps, le cadre du stage, soit une description de l'entreprise Buster.AI et les missions qui m'ont été confiées. Ensuite, une explication plus précise du travail effectué tout au long de cette période sera faite, pour enfin terminer ce rapport avec un bilan personnel sur cette expérience professionnelle.

Partie 1

-

Contexte du stage

Chapitre 1. L'entreprise

1. Buster.AI

Buster.AI est une jeune entreprise qui a pour but de devenir un antivirus de l'information en vérifiant l'intégrité et la véracité des contenus textuels (principalement des articles), des contenus vidéos (les vidéos de deepfake par exemple) et des contenus visuels (tels que des photos que l'on peut retrouver sur twitter) grâce à l'intelligence artificielle.

En 2019 Julien Madras et Aurélien Cluzeau fondent Buster.Ai et créent une équipe constituée d'experts en intelligence artificielle afin de trouver une solution au problème que l'on rencontre de nos jours majoritairement dans les médias. En effet, des informations importantes susceptibles de déclencher des conflits internationaux comme économiques sont diffusées sans aucune vérification. Buster veut donc limiter ce phénomène et créer un produit qui sera capable de vérifier et traiter chaque information qui sort chaque jour, en différenciant les vraies informations des fausses et donc créer une sorte d'antivirus de l'information.

Buster.AI, située à Paris, participe à plusieurs challenges chaque année afin d'évoluer et se faire une place dans ce domaine. Elle est pour l'instant une des rares entreprises à travailler sur cette problématique.

L'équipe de Buster a pour l'instant créé un produit en ligne nécessitant une identification afin d'y accéder. Faisant partie de l'équipe durant mon stage, des identifiants m'ont été assigné dès le début du stage dans l'intention de me permettre de travailler sur le produit et le tester directement. Voici un aperçu de la page d'accueil du produit[1] :

[1] https://coverity.test.buster.ai/

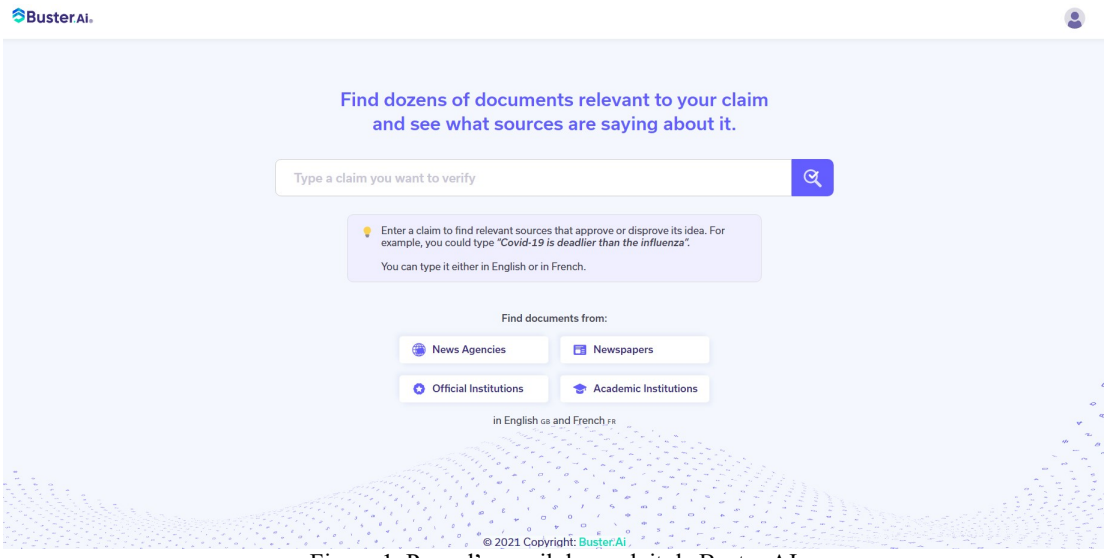

Figure 1. Page d'accueil du produit de Buster.AI

Pour vérifier la véracité d'une information, il suffit d'entrer une requête contenant cette information dans la barre de recherche pour obtenir les résultats.

Figure 2. Exemple de recherche sur le produit

On peut voir sur la capture d'écran ci-dessus, un exemple de résultats lorsqu'on exécute une recherche sur le produit de Buster, on ne peut pas voir si l'information est fausse ou vraie ici, la mise en page du site, n'étant pas encore terminée, n'est pas encore parfaite et ne contient pas toutes les informations. Les ingénieurs de Buster, répartis en différents groupes, travaillent toujours sur le produit pour proposer des versions sans cesse améliorées. Chaque groupe s'occupe d'une tâche.

2. L'équipe

Buster est divisée en trois équipes. La première équipe, l'équipe des gestionnaires du produit et de l'entreprise, est principalement composée de Cedra et Léa mais aussi de Julien. Ils cherchent toujours de nouvelles améliorations pour le produit, mais gèrent aussi les clients et l'organisation de l'entreprise. Ensuite, la deuxième équipe regroupe les programmeurs de l'entreprise, on y retrouve Aurélien, Alexandre, Georges et Vireya. Ce sont des ingénieurs de backend[2] et full-stack[3] qui travaillent directement sur le code du produit. Enfin, Amine, Hugo et Mostafa forment la dernière équipe, celle de l'équipe de recherche. Ils travaillent principalement avec le traitement automatique des langues et ont déjà remporté la première place d'un concours de NLP appliqué au Fact-Checking.

L'entreprise est composée à 75% d'hommes et seulement 25% de femmes. L'âge moyen des employés est de 27 ans. Les trois équipes de Buster évoluent et intègrent constamment de nouveaux membres.

Les équipes de Buster ont toutes le même objectif, celui de créer un produit qui détecte les fausses informations grâce au Fact-Checking.

[2] https://fr.wikipedia.org/wiki/Backend
[3] https://fr.wikipedia.org/wiki/D%C3%A9veloppeur_full_stack

Chapitre 2. Fact Checking

1. Définition

Aujourd'hui, avec le développement important d'Internet, on trouve énormément d'informations qui circulent sur le web et les réseaux sociaux, souvent des fausses informations (des « fake news »). Ces informations fallacieuses ont pour but de tromper le lecteur, elles sont de plus en plus nombreuses puisque tout le monde peut publier des documents sur Internet, il n'y a pas besoin d'être journaliste. On n'est pas toujours capable de savoir si ces informations sont vérifiées ou non, et si oui, comment et par qui est-ce qu'elles le sont. Le Fact-Checking ou « vérification des faits » est une technique qui permet, comme son nom l'indique, de vérifier des faits politiques principalement. Cette technique est utilisée par des journalistes qui, depuis 2013, ont recourt à des robots pour automatiser cette vérification. Les journalistes ne sont plus les seuls praticiens du Fact-Checking, des chercheurs en intelligence artificielle commencent petit à petit à s'y intéresser pour aider les journalistes et s'occuper de la partie automatisation des robots.

2. Présentation de l'existant

Buster n'est pas la première ni la seule entreprise à travailler avec le Fact-Checking. De nombreuses presses françaises (comme Le Monde[4] ou Libération[5]) utilisent le Fact-Checking pour certaines informations, mais elles n'offrent pas la possibilité de vérifier la véracité de chaque information. Ce que propose Buster, c'est cette partie manquante chez les agences de presse, celle de traiter toutes les informations. De plus, la plateforme de Buster ne nous propose pas seulement des articles de presse mais des informations de tout genre comme les tendances que l'on retrouve sur les réseaux sociaux par exemple ou encore des sources provenant d'encyclopédies. Ce qui diffère aussi entre Buster et les agences de presse, c'est que, sur la plateforme de Buster, l'utilisateur cherche lui-même l'information qu'il veut vérifier. Cette entreprise n'est pas une agence de presse, elle ne rédige pas d'articles mais renvoient à des articles existants. On pourrait donc presque comparer Buster à un moteur de recherche, mais les chercheurs de Buster veulent éviter cette comparaison car ils proposent un produit qui vérifie avec des sources officielles les

[4] https://www.lemonde.fr/
[5] https://www.liberation.fr/

informations contrairement aux moteurs de recherches où l'on peut retrouver toute sorte d'information.

Les chercheurs de Buster travaillent avec l'outil Elasticsearch[6] pour construire leurs algorithmes de Fact-Checking. Elasticsearch s'exécute comme un moteur de recherche, cet outil permet de faire des recherches sur toute la base de données et il est bien évidemment possible de mettre à jour cette base de données au fur et à mesure de l'avancement du projet.

La prise en main de ces outils ne m'a pas été si simple. On m'a demandé dans un premier temps de travailler sous système Windows pour Linux avec Ubuntu mais n'ayant pas l'habitude de travailler sous Linux, les installations des outils et des environnements virtuels m'ont pris beaucoup de temps. Le télétravail n'a pas favorisé cette prise en main car je n'ai pas toujours eu l'aide dont j'avais besoin des autres membres de l'équipe pour avancer plus vite. Cependant, cela m'a permis de découvrir de moi-même les différentes étapes et problèmes que l'on peut rencontrer lors des installations. J'ai ensuite pu m'intégrer au cœur du projet pour découvrir comment les chercheurs de Buster utilisent la méthode de Fact-Checking.

3. Buster AI

Afin de vérifier si une information est vraie ou fausse, la technique utilisée par Buster se déroule en plusieurs étapes. Dans un premier temps, l'utilisateur doit rédiger une requête contenant l'information à vérifier. De là, cette requête est récupérée par un algorithme qui va chercher une réponse dans sa base de données qui est composée de données textuelles que l'on retrouve sur le web. Ces données sont majoritairement des articles de presse, mais peuvent aussi provenir de sources officielles (comme le site du gouvernement intérieur), des encyclopédies (Wikipédia[7] par exemple), des agences de presse, des documents de recherche ou de bases de données. L'algorithme lance une recherche qui est faite à travers des vecteurs de similarité qui sont calculés en comparant les mots que l'on retrouve dans la requête et dans un document. Lorsqu'un document est choisi par l'algorithme, il est découpé en plusieurs passages, car en effet pour relever une réponse plus précise, il est plus intéressant de se concentrer seulement sur un passage du

[6] https://www.elastic.co/fr/elasticsearch/
[7] https://fr.wikipedia.org/

texte plutôt que sur le texte entier. Après avoir trouvé une réponse à la requête dans un paragraphe, la prochaine et dernière étape consiste à voir si ce paragraphe vérifie ou réfute l'information de la requête. On parle donc ici d'étiquette ou d' « entailment » (qui est le mot employé par les chercheurs de Buster). Pour cela, Buster propose la liste d'étiquettes suivante :

- SUPPORTS : si l'article vérifie l'information ;

- REFUTES si l'information est fausse ;

- NOT_ENOUGH_INFO : s'il n'y a pas suffisamment d'information dans l'article pour répondre à la requête ;

- NA : si l'article ne répond pas à la requête.

La réponse à la requête correspond donc à cette étiquette, qui a été attribuée à un paragraphe, lui-même extrait d'un document.

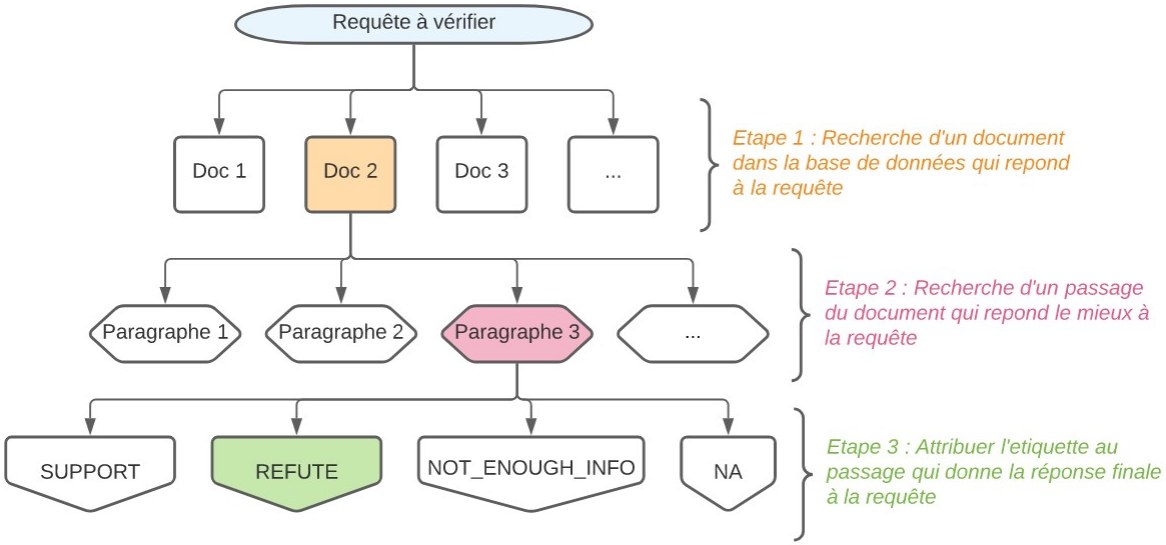

Figure 3. Schéma explicatif des différentes étapes traitées par le produit

Chapitre 3. Ma mission

1. Mon poste

J'ai intégré Buster.AI en tant que stagiaire linguiste. J'ai rejoint l'équipe de recherche qui travaille sur le traitement automatique du langage avec l'intelligence artificielle. Les trois chercheurs, Amine, Hugo et Mostafa, ont pour but d'améliorer le produit. A partir d'une phrase donnée en entrée, ils vont voir si cette phrase est vérifiée ou non. Ils travaillent avec Elasticsearch qui est une technologie qui permet l'indexation et la recherche de données. J'ai rejoint cette équipe (avec Célia Martin) pour les aider à mieux analyser les différents résultats provenant d'une même requête.

Tout au long du stage, j'ai donc principalement travaillé en équipe avec Célia, Clément (un autre stagiaire qui nous a rejoint en mai) les trois chercheurs mais surtout Mostafa (qui nous donnait les missions sur lesquelles nous devions travailler) puis Nadine (une jeune bénévole qui avait rejoint Buster pour quelques mois).

En raison de la situation sanitaire, il m'a été possible de travailler à distance. J'ai donc dû me déplacer à Paris qu'une seule fois durant mon stage, ce qui m'a surtout permis de rencontrer les membres de Buster et découvrir leur environnement de travail.

2. Mon travail

Les chercheurs ont construit des algorithmes retournant de nombreux résultats en français et en anglais (ils n'ont pas encore assez de compétences linguistiques pour travailler avec d'autres langues). Cependant, ils ont observé que ces résultats ne sont pas toujours cohérents car pour une même information il peut y avoir différents résultats. Par exemple, en recherchant la requête « le PIB français a-t-il chuté en 2020 ? » on retrouve ces deux résultats dans la liste des résultats :

0. « France: le PIB a rebondi de 18,2% au troisième trimestre (Insee) » (REFUTE)

1. « -13,8% : chute historique du PIB de la France au deuxième trimestre » (NA)

Mon but pendant ce stage était de les aider à résoudre ce problème et donc chercher à comprendre pourquoi, pour une même information, les résultats peuvent-ils varier (certains résultats vérifient l'information et d'autres la réfutent), qu'est-ce qui cause cette

irrégularité ? Nos premières hypothèses étaient basées sur la formulation de la requête, en effet, si une requête est formulée de différentes manières peut-être que l'algorithme renverra des résultats différents pour chaque formulation. Par exemple :

- « C'est un été pluvieux qui est prévu pour 2021. » = SUPPORTS
- « Il ne va pas pleuvoir cet été. » = REFUTES

Tout au long de mon stage j'ai donc aidé les chercheurs de Buster.AI à étudier ce problème.

Partie 2

-

Travail réalisé

Chapitre 4. Découverte du sujet

J'ai travaillé sur plusieurs tâches pour étudier la problématique posée par les chercheurs. Toutefois, certaines étant plus importantes que d'autres, elles m'ont pris plus de temps. Les premières taches que l'on m'a confiées m'ont surtout aidé à mieux comprendre sur quoi travaille Buster avant de me plonger dans le cœur du problème.

1. Vérification d'informations

Buster.AI a pour but principal de vérifier des informations automatiquement et d'indiquer s'il s'agit d'une information vraie ou fausse. Ma première mission consistait à trouver des sources, manuellement, en faisant des recherches simples sur Google, qui justifient que des documents contiennent des fausses informations. J'ai travaillé sur un corpus composé de quatre documents : une capture d'écran d'un tweet, un rapport, un document officiel et une photo. J'ai travaillé avec Cedra, Célia et Léa sur un fichier Excel où nous avons partagé nos sources trouvées.

2. Analyse statistique textuelle

La première analyse que j'ai faite m'a permis de découvrir sur quels types de données et surtout sur quels types de corpus j'allais travailler durant mon stage. Elle consistait tout simplement à faire une analyse statistique de données textuelles sur un corpus composé de plus de 6 000 articles. J'ai travaillé sur cette analyse de données avec Célia, nous nous sommes répartis les tâches, afin d'avancer plus vite. Nous avons créé des scripts Python et utilisé la librairie Spacy[8] pour faire plusieurs analyses. Notre analyse statistique est composée d'un comptage sur les mots et les phrases du corpus, d'une étude sur les catégories morpho-syntaxiques que l'on retrouve dans le corpus, d'un relevé des lemmes et des formes les plus fréquents, d'une analyse sur la richesse lexicale du corpus et enfin, de proportions sur des différents temps de conjugaison des verbes que l'on retrouve dans le corpus. Dans un premier temps, nous avons fait cette analyse sur le contenu des articles puis dans un deuxième temps, sur les titres des articles.

[8] https://spacy.io/models/fr

Nous avons déposé nos scripts sur GitLab[9] et avons fait une page d'analyse sur Confluence[10] ou nous avons mis les graphiques que renvoient chacun de nos scripts. Par exemple :

- le graphique correspondant au comptage des phrases est :

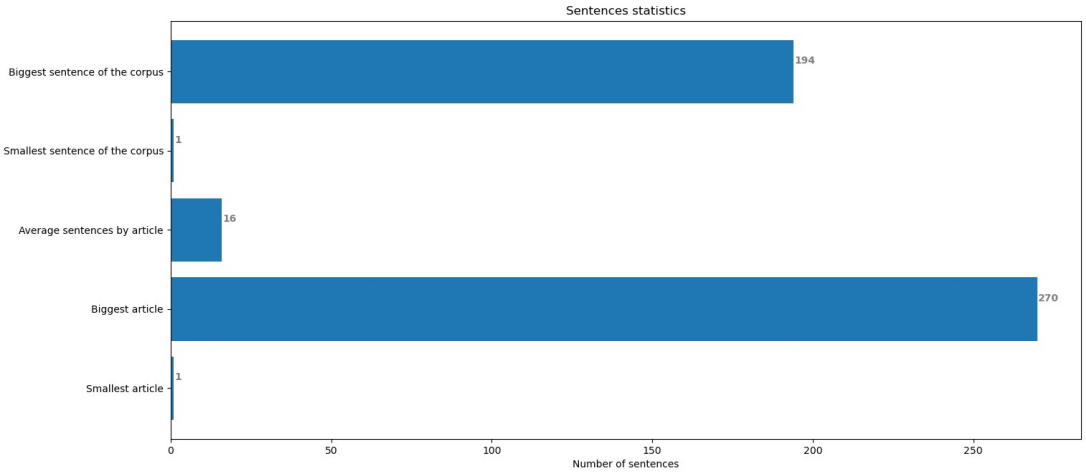

Figure 4. Graphique généré lors de l'analyse statistique portant sur le comptage des phrases

- le graphique correspondant à notre dernière analyse sur les temps de conjugaison est :

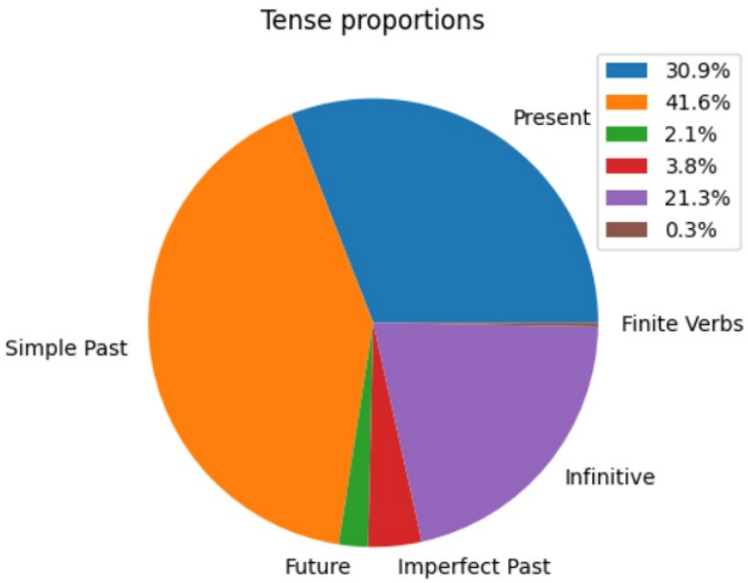

Figure 5. Graphique généré lors de l'analyse statistique portant sur les différents temps de conjugaison

[9] https://gitlab.com/gitlab-org/gitlab
[10] https://www.atlassian.com/fr/software/confluence

Chapitre 5. Data set

Une fois l'analyse statistique terminée et après avoir étudié le type de corpus sur lequel les chercheurs de Buster travaillent, je me suis concentrée sur la vraie problématique de mon stage. J'ai principalement travaillé sur les parties du "passage retriever" et du "entailment label" de l'entreprise qui correspondent respectivement à retrouver le bon passage qui répond à la requête et lui donner une étiquette pour savoir si ce passage vérifie ou non la requête. Pour cela, j'ai travaillé avec plusieurs jeux de données qui ont servi de données d'entraînement puis de test pour le modèle.

1. Data set général

Premièrement, Célia et moi avons créé un data set de plus de 200 requêtes qui a été utilisé pour entraîner le modèle afin de répondre au problème majeur du produit : pourquoi obtient-on différents résultats pour une même information. Le but de ce jeu de données était donc de le construire avec différentes formulations de requêtes pour étudier et comparer les résultats de ces différentes formulations. Nous avons créé un fichier Excel pour la construction de ce jeu de données. Avant de remplir l'Excel, la première étape consistait à rechercher les différentes manières de poser une question en français et à les lister. Les étapes pour construire ce jeu de données sont assez longues. Premièrement, il fallait trouver un article contenant une information médiatique, pour ensuite lire cet article et rédiger une requête dans le fichier Excel ayant sa réponse dans l'article. Après, il fallait donner un code identifiant associé à la requête (Q7F par exemple), donner un code associé au document (D16 par exemple) (si nous l'avions déjà utilisé nous reprenions le code déjà écrit dans l'Excel), écrire le contenu du passage répondant à la requête, donner un code associé au passage, décider du "label d'entailment", écrire l'url de l'article, le titre de l'article et enfin sa source (nom de la source). Pour chaque type de phrase, nous avons ajouté sa négation et l'avons traitée comme tous les autres types (id, article, passage etc.). Une fois ces informations rédigées dans l'Excel, nous avons repris ces requêtes pour les traduire une à une en anglais avec Deepl[11]. Pour chaque traduction nous avons donné également sa grammaticalité (si elle est correcte ou non), trouvé un article qui répondrait à la requête, et suivi les mêmes étapes que pour le français (donner un id au document, trouver le passage qui répond à la requête etc.). Nous avons finalement repris chaque ligne

[11] https://www.deepl.com/fr/translator

de l'Excel pour noter les codes IPTC[12] (International Press & Communications Council) et lister toutes les entités nommées présentes dans l'article en leur associant un tag basé sur ceux de la librairie Spacy (nous avons choisi les tags que Spacy propose pour l'anglais car cette liste est bien plus complète que la liste des tags pour le français). Cette tâche a donc consisté à :

- lire l'article intégralement en notant chaque entité nommée

- donner un tag à chacune des entités relevées en se basant sur les tags de la librairie Spacy et en l'adaptant si besoin était (par exemple il n'y a pas de tags pour les noms d'animaux, nous l'avons donc ajouté). Nous avons imaginé des tags plus fins pour la catégorie des « PERSON » pour les entités nommées : par exemple, avec des catégories comme PERSON_POLITICIAN ou PERSON_SCIENTIST.

- une fois l'article lu, trouver les tags IPTC pertinents en se limitant à une profondeur de niveau 3 et en trouvant le code correspondant au tag dans la liste donnée fournie sur le site iptc.org

Pour avoir suffisamment de données, à savoir environ 200 requêtes, nous avons réalisé ces étapes 3 fois, c'est-à-dire que nous avons au final 3 requêtes par type de phrase.

General Query ID	Specific Query ID	Query	Grammaticality	Query Type	Doc ID	Passage ID	Passage
Q1	Q1F	est-ce qu'il y aura des examens écrits en juin pour le bac 2021 ?	correct	question totale directe (polaire) : "est-ce que"	D1	P1	"Il y aura des examens écrits au mois de juin avec l'épreuve de philosophie et le Grand Oral qui est l'une des innovations de ce baccalauréat et qui va permettre aux élèves de montrer leurs compétences orales", a indiqué le ministre de l'Éducation nationale au micro de Thomas Sotto dans RTL Soir.
Q1	Q1E	will there be written exams in June for the 2021 bac?	correct		D105	P2	Terminale students will start in March 2021 with the specialty exams, followed in June by the written exams and the grand oral.
Q2	Q2F	est-ce que la dette n'a pas augmenté en 2020 ?	correct	question totale directe (polaire) negative : "est-ce que"	D2	P1	Sans surprise, après avoir déjà fortement augmenté au premier trimestre 2020, la dette publique française a explosé au deuxième trimestre, atteignant un nouveau sommet historique de 2.638,3 milliards d'euros : 114,1% du PIB français, soit 12,7 points de plus qu'au premier trimestre 2020 et 49,6 points de plus qu'en 2007

Figure 6. Exemple du data set avec les huit premières colonnes

[12] Cf. Annexe 1 page 50

Entailment	URL	Title	Source	IPTC	Named Entity	Comments
Support	https://www.rtl.fr/actu/debats-societe/bac-2021-il-y-aura-des-examens-ecrits-en-juin-promet-blanquer-sur-rtl-7900006850	Bac 2021 : "Il y aura des examens écrits en juin", promet Blanquer sur RTL	RTL	["20000404","20000480"]	Jean Michel Blanquer:PERSON_POLITICIAN;mois de juin:DATE;2021:DATE;ministre de l'Education nationale:PERSON_POLITICIAN;Thomas Sotto:PERSON_JOURNALIST;le ministre:PERSON_POLITICIAN	
Support	https://www.faxinfo.fr/en/nouvelle-formule-du-bac-2021-parcoursup-les-grandes-dates-a-retenir/	New Bac 2021 formula, Parcoursup: key dates to remember!	faxinfo	["20000404"]	the Ministry of National Education:PERSON_POLITICIAN;last Wednesday:DATE;between March 15th and 17th:DATE;the 17 June:DATE;From Monday June 21 to Friday July 2:DATE;the 6 July:DATE;from July 7 to 9:DATE;the ministry:PERSON_POLITICIAN;final year:DATE;Jean-Michel Blanquer:PERSON_POLITICIAN;Parcoursup:ORG;January 20:DATE;Between May 27 and July 16, 2021:DATE;from June 16, 2021 until September 16, 2021:DATE	
Refute	https://www.capital.fr/economie-politique/la-dette-publique-de-la-france-explose-qui-va-payer-1381487#:~:text=Sans%20surprise%2C%20apr%C3%A8s%20avoir%20d%C3%A9j%C3%A0,p	"La dette publique de la France explose : qui va payer ?"	Capital	["20000346"]	d euros:MONEY;114,1%:PERCENT;Marc Touati:PERSON_ORG;ACDEFI:ORG;trimestre 2020:DATE;deuxième trimestre:DATE;12,7:CARDINAL;premier trimestre 2020:DATE;49,6:CARDINAL;2007:DATE;fin 2020:DATE;125%:PERCENT;Etat français:NORP;dix ans:DATE;Japon:GPE;240%:PERCENT;milliards:MONEY;trillards:MONEY;premier	

Figure 7. Exemple du data set avec les sept dernières colonnes

Après avoir recueilli les données d'entrainement qui forment un jeu de données assez complet, avec de nombreuses formulations différentes pour écrire une requête, les chercheurs ont entrainé le modèle avec ces données. J'ai étudié les résultats pour voir quelles formulations étaient les plus intéressantes et pertinentes.

2. Test sur le modèle

La prochaine étape de l'apprentissage machine consiste à tester le modèle avec des données de tests. C'est la tâche que l'on m'a confiée après avoir fini l'étude sur le premier jeu de données. On m'avait demandé d'en construire un nouveau qui a, par la suite, servi de test pour le modèle. Le but ici était de tester un jeu de données directement sur le produit de Buster pour analyser les résultats. J'ai utilisé le notebook Jupyter[13] pour faire ce travail. Le jeu de données était composé d'une requête que nous avons formulée soit Célia soit moi, en nous inspirant des formulations les plus intéressantes du précédent data set, et d'un article, plus précisément, tous les passages d'un article entrés un par un dans le data set afin de les étudier séparément. Pour cela, nous avons dû réfléchir à comment découper les passages d'un article. Nous avons utilisé trois méthodes de découpage :

- En fonction des sauts de lignes (un passage = une balise <p> dans le code html) ;

- Créer des passages de 2 phrases maximum ;

- Regrouper plusieurs passages (découpage entre les balises <h2> par exemple).

[13] https://jupyter.org/

Les articles sélectionnés pour ce data set, devait faire partie de la base de données du produit de Buster car nous avons travaillé sur le produit avec une partie de son code.

Les tests consistaient à donner en entrée une requête et tous les passages d'un article contenant la réponse à la requête, pour avoir en sortie un classement de ces passages. En première position, on retrouve le passage qui répond au mieux à la requête et en dernière position celui qui y répond le moins. Ces passages ont donc chacun un score (le passage en première position a le meilleur score). Le total des scores de tous les passages est égal à 1.

Afin de réussir au mieux le test, nous avons pour chaque article, choisi nous-mêmes un passage, le « passage attendu » que l'on a considéré comme étant le meilleur passage de l'article pour répondre à la requête. Nous avons ensuite comparé ce passage avec le passage que le produit retourne en première position. Les résultats de ce test sont plutôt bons. Le passage attendu, pour plus de 50% des résultats, était retourné en première position. Lorsque la réponse est évidente, c'est-à-dire lorsqu'un article est composé de plusieurs passages contenant des informations indépendantes les unes des autres, le score du passage attendu est largement supérieur à ceux des autres passages de l'article. On le retrouvait souvent pour les autres cas dans les trois premiers passages avec un bon score. On retrouve ces cas de figure pour des articles avec des passages contenant des informations similaires ou, qui se complètent, le modèle a donc du mal à les différencier. Parfois, un mot dans un passage peut faire toute la différence. En effet, si deux passages sont éligibles à répondre à une requête, si un même mot est utilisé dans la requête et dans un de ces deux passages, c'est ce passage-là qui aura un meilleur score. Si un synonyme est employé dans un passage, le modèle aura plus de difficulté à le relever.

Voici ci-dessous, deux exemples, le premier illustre un bon résultat et le second, un résultat médiocre. Ces exemples ont été inventés pour cause de confidentialité.

Exemple 1 :

Requête : «*Le mois de juillet a été un mois pluvieux en France cet été*.»

Titre de l'article : «*Juillet 2021 a été le 3e mois de juillet le plus chaud jamais mesuré et le 2e en Europe»*

Source de l'article : *https://fr.news.yahoo.com/juillet-2021-%C3%A9t%C3%A9-3e-mois-133000511.html*

Passage attendu : « *Le mois de juillet en France a été marqué par un temps assez maussade. « Les températures sont restées inférieures aux normales de saison une grande partie du mois, notamment du 12 au 16 avec un pic de fraîcheur marqué », observe Météo France. Le déficit d'ensoleillement a dépassé les 20 % dans certaines régions et le mois a été particulièrement pluvieux.* »

Résultats :

Score	Passage	Position du passage
0.9	*Le mois de juillet en France a été marqué par un temps assez maussade. «* Les températures sont restées inférieures aux normales de saison une grande partie du mois, notamment du 12 au 16 avec un pic de fraîcheur marqué *», observe Météo France. Le déficit d'ensoleillement a dépassé les 20 % dans certaines régions et le mois a été particulièrement pluvieux.*	**1e**
0.07	*Une météo pourrie qui offre un contraste saisissant avec la situation mondiale et européenne. Selon le réseau de surveillance Copernicus, le mois de juillet 2021 a été le troisième mois de juillet le plus chaud jamais mesuré, derrière ceux de 2019 et 2016. Au niveau européen, c'est même le deuxième mois de juillet le plus chaud après celui de 2010. La température mensuelle moyenne a été 0,33 °C plus élevée que la moyenne normale constatée sur la période 1991-2020.*	2e
0.03	*Un peu partout dans le monde, des vagues de chaleur exceptionnelles ont fait grimper le thermomètre à des niveaux stratosphériques. Lundi 2 août, le Premier ministre grec Kyriakos Mitsotakis a prévenu que son pays était frappé par « la pire canicule depuis celle de 1987 », avec des températures maximales de 44 à 45 °C dans le Péloponnèse et en Thessalie. La Turquie a battu son record de température absolu avec 49,1 °C enregistrés le 20 juillet à Cizre au sud-est du pays. Des records ont aussi été franchis au Maroc, au Canada, au Japon, aux États-Unis ou en Sibérie, qui a connu plusieurs journées à plus de 39 °C. Le record historique de température maximale quotidienne a été battu en Irlande du Nord, et les températures ont été bien supérieures à la moyenne dans l'est de l'Islande et dans certaines parties de l'est du Groenland, note également Copernicus.*	3e

Exemple 2 :

Requête : « La vaccination des adolescents a débuté en France. »

Titre de l'article : «*Covid-19 : la vaccination des moins de 12 ans n'est "pas d'actualité", assure Jean-Michel Blanquer»*

Source de l'article : *https://www.francetvinfo.fr/sante/maladie/coronavirus/vaccin/la-vaccination-des-moins-de-12-ans-n-est-pas-d-actualite-selon-jean-michel-blanquer_4742195.html*

Passage attendu : «"Aujourd'hui, l'objectif c'est d'avoir les plus de 12 ans qui soient vaccinés en France", a poursuivi le ministre. Plus de la moitié des adolescents français de

12 à 17 ans ont reçu au moins une dose de vaccin cet été, avance le ministre, ce qui devrait permettre de maintenir les établissements scolaires ouverts, selon lui. "On dépasse aussi les 30% de ceux qui ont eu deux [doses de] vaccin. C'est évidemment favorable à une année [scolaire] la plus normale possible", a-t-il détaillé.»

Résultats :

Score	Passage	Position du passage
0.41	*La vaccination des moins de 12 ans contre le Covid-19 n'est "pas d'actualité", a assuré jeudi 19 août le ministre de l'Education nationale, Jean-Michel Blanquer, à l'occasion d'un déplacement dans les Hauts-de-Seine, deux semaines jour pour jour avant la rentrée scolaire.*	1e
0.39	*Le ministre annoncera dans les prochains jours le protocole sanitaire retenu pour les écoles, collèges et lycées, parmi quatre scénarios possibles, en fonction du degré de circulation du virus. "Nous sommes organisés aussi pour des campagnes de vaccination dès le mois de septembre pour les élèves qui ne seraient pas vaccinés, sur le mode de l'incitation", a souligné Jean-Michel Blanquer, tout en rappelant qu'il n'y avait "évidemment pas de pass sanitaire pour aller à l'école".*	2e
0.2	*"Aujourd'hui, l'objectif c'est d'avoir les plus de 12 ans qui soient vaccinés en France", a poursuivi le ministre. Plus de la moitié des adolescents français de 12 à 17 ans ont reçu au moins une dose de vaccin cet été, avance le ministre, ce qui devrait permettre de maintenir les établissements scolaires ouverts, selon lui. "On dépasse aussi les 30% de ceux qui ont eu deux [doses de] vaccin. C'est évidemment favorable à une année [scolaire] la plus normale possible", a-t-il détaillé.*	**3e**

Pour cette tâche, je n'ai travaillé que sur le bon passage à retrouver (la partie sur le « passage retriever ») et non pas sur l'étiquette de vérification (l'« entailment »).

3. Data sets spécifiques

Les premiers jeux de données que j'avais construit jusqu'à présent, étaient plutôt généraux et ne ciblaient pas un point spécifique. Les chercheurs de Buster m'ont donc demandé par la suite de travailler sur de nouveaux jeux de données mais en se focalisant sur une partie bien précise.

J'ai commencé le premier jeu en ne me concentrant uniquement sur l'étiquette de vérification puisque cette partie a été exclue lors des tests sur le bon passage à retrouver. Pour cela, j'ai donc construit un jeu de données avec différentes requêtes, en français et en anglais comme à chaque fois, mais en faisant varier l'étiquette de vérification à chaque type de requête pour avoir un maximum de modèles. Pour rappel, la liste des étiquettes de vérification que Buster utilise est la suivante : SUPPORT, REFUTE,

NOT_ENOUGH_INFO et NA. La construction de ce jeu de données m'a parue plus compliqué que les autres puisque toutes les étiquettes de vérification devaient être traitées, et en plusieurs fois. Ce qui signifie qu'il fallait écrire des requêtes pour lesquelles l'information n'était pas connue (notamment pour utiliser l'étiquette NEI (NOT_ENOUGH_INFO)). Pour cette catégorie, les requêtes portaient souvent des informations sur l'astrologie ou des problèmes concernant le futur.

Par la suite, ce sont des requêtes portant la négation que j'ai étudiées. Cependant, je me suis vite rendu compte qu'il n'y a avait pas une différence majeure entre les requêtes formulées avec des phrases affirmatives ou négatives. Ce data set ne comporte pas beaucoup de données et a été très rapidement terminé.

Enfin, avec les chercheurs de Buster, on se demandait si rajouter des tokens précis dans la requête ne renverrait pas d'autres résultats. Pour étudier ce point, on a décidé de créer un data set composé de requêtes écrites avec des tokens de localité et de temporalité. Dans le data set, plusieurs lieux et plusieurs dates étaient précisés. Par exemple, si aujourd'hui j'écris la requête suivante « Barack Obama a été élu en 2008. », on retrouvait ces informations concernant la requête dans le data set :

Lieu où a été rédigée la requête	**Date de la rédaction de la requête**	**Lieu renvoyé par la requête**	**Information sur le lieu dans la requête**	**Date renvoyée par la requête**	**Information sur la date dans la requête**
France	2021	Etats-Unis	Implicite	2008	Explicite

Tableau 1. Exemple du data set de temporalité et localité

Et si, j'avais écrit, l'année dernière, la requête suivante « Le premier confinement a débuté en mars 2020 en France. » voici les informations que l'on obtiendrait dans le data set :

Lieu où a été rédigée la requête	**Date de la rédaction de la requête**	**Lieu renvoyé par la requête**	**Information sur le lieu dans la requête**	**Date renvoyée par la requête**	**Information sur la date dans la requête**
France	2020	France	Explicite	Mars 2020	Explicite

Tableau 2. Deuxième exemple du data set de temporalité et localité

Pour les informations de temporalité et localité retrouvées dans les articles, elles ont également toutes été relevées et pour chacune d'entre elles, il a été précisé s'il s'agit d'une information explicite, implicite ou relative.
J'ai apprécié travailler sur la construction de ce data set puisqu'il est riche en information et distinct des autres. Ce data set est moins riche en nombre de requêtes, il est plus petit que les précédents, mais il est plus riche en terme de données car on retrouve des informations spécifiques pour chacune des données. Il n'était composé que d'une centaine de requête.

Dès lors que j'ai terminé d'étudier les différents points pour résoudre le problème majeur à savoir celui de retrouver le bon passage d'un article pour une requête posée par l'utilisateur via la construction de plusieurs data set, j'ai travaillé sur d'autres types de taches que j'ai trouvé plus complexes mais plus intéressantes. Etant donné que les data set étaient tous terminés et que nous disposions de nos propres résultats pour le produit de Buster, un travail de comparaison entre nos résultats et d'autres résultats provenant d'autres produits serait intéressant pour étudier l'efficacité du produit. C'est en utilisant la technique de Web Scraping que j'ai effectué cette tâche.

Chapitre 6. Web Scraping

Le Web Scraping[14] est une méthode qui consiste à, comme son nom l'indique, extraire des données provenant de sites web qui pourront être réutilisées pour diverses raisons.

En ce qui me concerne, on m'a demandé de travailler avec la technique du Web Scraping pour créer un nouveau projet dont le but était de récupérer les résultats de la première page retournée par un moteur de recherche suite à une recherche à partir d'une requête. Pour cela, j'ai utilisé l'outil Selenium[15] dont je connaissais les fonctionnalités puisque je l'avais déjà découvert lors d'un examen du Master. Cet outil automatise l'extraction de données en suivant le schéma HTML de la page.

Les chercheurs de Buster nous ont confié à Célia et moi ce projet de Web Scraping qui avait pour but d'être un projet de comparaison entre les résultats de différents moteurs de recherche connus et les résultats du produit de Buster. Avec ce travail, ils ne cherchaient pas à comparer leur produit aux moteurs de recherche, car le produit de Buster n'est pas un moteur de recherche, mais tout simplement à étudier les résultats afin d'identifier le niveau de performance du produit par rapport à de grands moteurs de recherche.

Pour cela, nous avons travaillé avec les moteurs de recherche Google[16] et Bing[17]. Nous avons écrit un code Python[18] que nous avons divisé en deux parties. La première partie, que l'on appelé « Fetcher » et qui a été gérée par Célia, consistait à écrire un code qui prend une entrée une URL et une requête (ou une liste de requête) pour donner en sortie une autre URL (ou liste d'URL dans laquelle chaque URL correspondait au résultat de chaque requête). Cette partie du code permettait de récupérer la liste des résultats pour chaque requête. La deuxième partie du projet est la partie « Parser » sur laquelle j'ai travaillé. Elle est la suite de la partie Fetcher puisqu'en entrée elle récupère l'URL ou la liste d'URL qui a été générée en sortie par le code Fetcher qui contient les résultats de la requête, de la première page du moteur de recherche uniquement. Cette partie Parser du

[14] https://fr.wikipedia.org/wiki/Web_scraping
[15] https://selenium-python.readthedocs.io/
[16] https://www.google.com/
[17] https://www.bing.com/
[18] https://www.python.org/

projet va donc étudier les résultats de la requête. En effet, elle cherche, pour chaque résultat, son rang dans la page de recherche, son titre, son URL, l'extrait qui est affiché, les informations complémentaires (telles que la date et l'auteur lorsqu'ils sont indiqués) et enfin les mots clés de la requête.

Finalement pour ce projet, mon rôle a été d'écrire des codes Python afin de créer des listes contenant tous les résultats nécessaires pour le travail de comparaison. Les chercheurs ont préféré poursuivre le projet pour étudier la comparaison des résultats eux-mêmes. Ils ont récupéré ces listes qu'ils ont ensuite comparé aux résultats de leur produit.

Cette tâche a été différente des autres mais pas moins intéressante car elle m'a permis d'améliorer mes compétences en programmation et de découvrir un tout autre type de travail et d'analyse. Elle a d'ailleurs été ma dernière tâche avec l'équipe de recherche concernant le produit. J'ai travaillé durant la fin de mon stage sur un tout autre sujet.

Chapitre 7. FEVEROUS

Pour clôturer mon stage, j'ai rejoint l'équipe de Buster sur un challenge assez important, le challenge FEVEROUS (Fact Extraction and VERification Over Unstructured and Structured information) appartenant à FEVER[19] (Fact Extraction and VERification) qui a pour but d'évaluer la capacité d'un système à vérifier des informations à l'aide de preuves structurées et non structurées provenant de Wikipédia[20]. Buster participe à ce challenge, en concurrence avec d'autres équipes travaillant sur la même problématique, en espérant le remporter mais surtout pour tester la performance de son produit. Ce challenge consiste à vérifier les informations que l'on retrouve sur les pages WEB. La base de données de FEVER est composée des pages Wikipédia dont ils ont structuré les résultats. On peut comment est présenté un résultat dans la base de données de FEVER sur la capture d'écran ci-dessous. Pour une requête exécutée sur le web (« text ») et son code d'identification (« id ») on obtient la réponse à cette requête avec « entailment_label » qui définit si la requête est vérifiée ou réfutée sur une ou plusieurs pages Wikipédia (« label ») avec « label_ids » qui correspond aux endroits précis de l'information dans la page Wikipédia (avec « sentence » qui correspond à une phrase, « cell » à une cellule d'un tableau, « item » à une liste et « table_caption » au titre d'un tableau).


```
"14901": {
    "_id": 14901,
    "text": "Gymnopilus nashii is a part of the plantae kingdom and belongs to the basidiomycota family.",
    "entailement_label": "REFUTES",
    "label": [
        "Gymnopilus nashii",
        "Gymnopilus nashii",
        "Gymnopilus nashii",
        "Gymnopilus nashii",
        "Gymnopilus nashii"
    ],
    "label_ids": [
        "cell_0_6_0",
        "cell_0_6_1",
        "cell_0_2_0",
        "cell_0_2_1",
        "sentence_0"
    ]
},
```


Figure 8. Extrait d'un résultat de la base de données de FEVER

[19] https://fever.ai/
[20] https://www.wikipedia.org/

Mon rôle durant ce challenge devait être d'entraîner les modèles de Buster à retrouver les bons passages sur la bonne page Wikipédia à partir d'une requête écrite par l'utilisateur (comme habituellement). Pour cela, afin de mieux comprendre le challenge, j'ai brièvement étudié les méthodes d'entraînement de Quin+, BM25 et Facebook MDR. Mais je n'ai finalement pas pu travailler sur ces trois méthodes car il m'a été impossible d'installer tous les outils nécessaires sur mon ordinateur qui n'est pas assez puissant.

Le problème principal était d'entrainer les modèles afin qu'ils détectent si une requête a besoin d'un tableau ou d'une phrase ou des deux comme réponse. Je n'ai cependant pas pu participer à ces taches puisque, comme pour la tâche d'initiation au challenge, mon ordinateur n'est pas assez performant pour ce type de travail. J'ai donc simplement travaillé sur des analyses de données en m'aidant de quelques articles pour mieux m'instruire sur ce challenge.

La première analyse consistait à relever des données statistiques basiques telles que le nombre moyen de phrases, de tableaux ou encore la longueur des phrases pour chaque requête.

La deuxième analyse sur laquelle j'ai travaillé durant le challenge de FEVER était d'étudier la formulation de la requête, en fonction du type de l'information qui la vérifie ou la réfute, c'est-à-dire, si on observe des formulations différentes pour les tableaux et pour les phrases. Et en effet, les formulations des requêtes ayant leur réponse dans des phrases sont grammaticalement correctes et contiennent plus d'informations, alors que les formulations des requêtes pour lesquelles la réponse se trouve dans des tableaux, sont plus courtes et contiennent moins d'information (uniquement l'information recherchée est précisée dans la requête).

Par exemple, on voit que la requête « *Tammy Garcia was born in California but currently lives in Taos, she comes from a long line of Santa Clara Pueblo artists and her great-great-great grandmother Sara Fina Tafoya was a potter.* » est riche en information, grammaticalement correcte et composée de phrase complexe. Elle contient des verbes et des compléments de verbe. En effet, sa réponse se trouve dans des phrases (« sentence »).

```
"13061": {
    "_id": 13061,
    "text": "Tammy Garcia was born in California but currently lives in Taos,
            she comes from a long line of Santa Clara Pueblo artists and her
            great-great-great grandmother Sara Fina Tafoya was a potter.",
    "entailement_label": "SUPPORTS",
    "label": [
        "Tammy Garcia",
        "Tammy Garcia",
        "Tammy Garcia",
        "Tammy Garcia"
    ],
    "label_ids": [
        "sentence_0",
        "sentence_3",
        "sentence_4",
        "sentence_5"
    ]
},
```


Figure 9. Capture d'écran extraite de de la base de données illustrant l'exemple cité ci-dessus

Dans l'exemple ci-dessous, pour la requête « *Participating teams in the 2012\2013 Macedonian First Football League included club Bregalnica, managed by Dobrinko Ilievski, and club Shkendija, managed by Artim Shakiri, a retired football midfielder from North Macedonia.* », on peut voir que la phrase est aussi longue que celle de l'exemple précèdent, pourtant, c'est une phrase simple et qui contient moins de verbes. La réponse à cette requête se trouve principalement dans un tableau.


```
"20773": {
    "_id": 20773,
    "text": "Participating teams in the 2012\u201313 Macedonian First Football League included club
            Bregalnica, managed by Dobrinko Ilievski, and club Shk\u00ebndija, managed by Artim Shakiri,
            a retired football midfielder from North Macedonia.",
    "entailement_label": "SUPPORTS",
    "label": [
        "2012\u201313 Macedonian First Football League",
        "2012\u201313 Macedonian First Football League",
        "2012\u201313 Macedonian First Football League",
        "2012\u201313 Macedonian First Football League",
        "Artim \u0160akiri"
    ],
    "label_ids": [
        "cell_1_1_0",
        "cell_1_1_1",
        "cell_1_8_0",
        "cell_1_8_1",
        "sentence_0"
    ]
},
```


Figure 10. Capture d'écran extraite de de la base de données illustrant le deuxième exemple cité

Ensuite, les chercheurs de Buster m'ont demandé de travailler sur une autre analyse statistique me concentrant cette fois-ci uniquement sur les résultats sous forme de titres de tableaux (les « table_caption ») et sous forme de liste (les « item ») afin d'étudier si les informations que l'on retrouve dans les titres des tableaux ou dans les listes suffisent pour répondre à la requête ou si elles doivent être complétées par d'autres informations (sous forme de phrase par exemple).

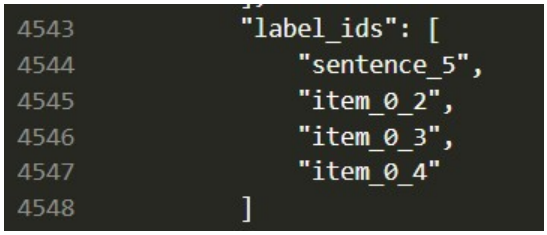

Figure 11. Exemple d' « item» dans la base de données

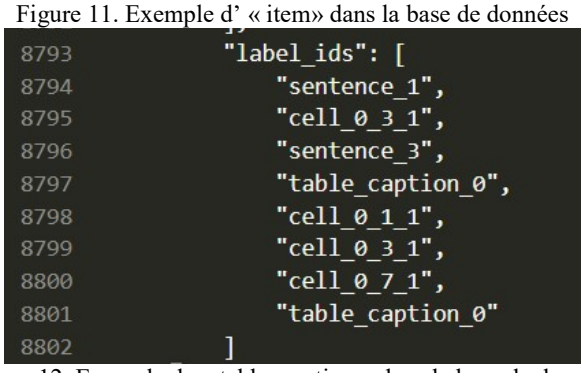

Figure 12. Exemple de « table_caption » dans la base de données

J'ai réalisé cette analyse en deux étapes : une première où j'ai cherché à retrouver la réponse à la requête et une deuxième pour relever la bonne étiquette de vérification. Les titres des tableaux ne sont pas vraiment utiles lorsqu'ils sont seuls puisque l'information que l'on retrouve dans la « table_caption » est incomplète : elle ne peut répondre à une information que si elle est complétée par les cellules de ce même tableau ou par des phrases. L'information du titre seul ne peut pas aider à répondre à la demande. Lorsqu'une information est utile dans le titre du tableau, c'est souvent qu'il s'agit d'une information numérique (comme une année, une quantité, un âge, etc.). En ce qui concerne l'étiquette de vérification, le titre d'un tableau ne réfute pas vraiment une requête puisqu'il s'agit simplement du thème de l'information.

En revanche les informations que l'on retrouve dans les listes sont meilleures que celles des titres des tableaux. Pour répondre aux requêtes, les éléments peuvent être complétés soit avec des phrases, soit avec des cellules d'un tableau, soit des deux, mais ce n'est pas nécessaire. L'information relevée dans une liste peut avoir la même importance que celle relevée dans une phrase ou dans une cellule de tableau. L'étiquette de vérification est souvent correcte pour une information provenant d'une liste. L'information d'une liste est particulièrement utile lorsqu'elle réfute la requête.

Enfin, en travaillant sur ce data set, les chercheurs de Buster ont remarqué que pour certaines requêtes, la réponse pouvait être dans différentes pages Wikipédia et non pas qu'une seule. La dernière analyse sur laquelle j'ai travaillé, en compagnie de Célia

pour ce challenge, était donc de voir si, pour ces requêtes-là, celles qui ont besoin de plus d'une page Wikipédia en « label », ces autres pages (secondaires) sont accessibles à partir d'un lien que l'on retrouverait dans la page Wikipédia principale. Par exemple, si la requête C nécessite les pages Page1, Page2 et Page3 pour y répondre, il faut chercher si dans la page Wikipédia Page1, il y a un lien cliquable vers la Page2 et vers la Page3, puis potentiellement les autres pages.

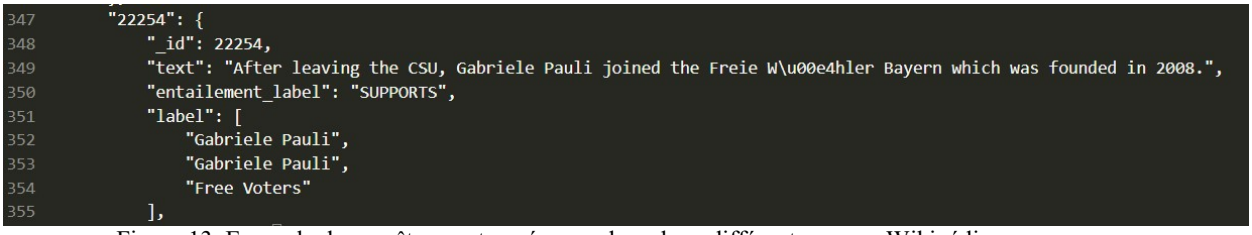

Figure 13. Exemple de requête ayant sa réponse dans deux différentes pages Wikipédia

Sur cet exemple, on voit qu'il y a deux titres de page différents dans la catégorie « label ». Les informations pour répondre à cette requête, se trouve donc dans deux pages Wikipédia. On compte au total trois « label » ce qui signifie en fait qu'il faut consulter trois endroits différents (phrases, tableaux, etc.) pour répondre complètement à la requête, deux informations se trouvent sur la même page (« Gabriele Pauli), c'est pour cela que le nom de cette page est écrit deux fois, puis une information sur la page « Free Voters ». Il faut ensuite parcourir les pages Wikipédia en s'aidant des « label_ids » pour retrouver ces informations.

J'ai fait quelques observations sur le jeu de données pour essayer de trouver un modèle qui pourrait trouver automatiquement les requêtes qui nécessitent plusieurs pages afin d'y trouver une réponse. Mais avant cela, j'ai écrit un code python[21] qui permet de relever la liste des « label » pour chaque requête, ce qui correspond à la liste des pages Wikipédia ayant une réponse pour la requête. Les observations ont surtout été faites dans les étiquettes « label » et « label_ids » du jeu de données, mais je n'ai pas relevé de points intéressants. Avec Célia, nous avons vérifié si les « label_ids » sont listés par ordre croissant, ce qui pourrait signifier que les « labels_ids » des secondes pages et des pages qui suivent seraient listés après les « label_ids » de la première page, mais ce n'est pas toujours le cas. Nous avons également vérifié si les noms de la deuxième page et des autres

[21] Cf. Annexe 2 page 51

pages étaient listés après la première page mais non, les étiquettes sont listées dans l'ordre des informations que nous trouvons dans la requête.

Cette tâche a été très intéressante. C'était un vrai travail de recherche qui nécessitait des réflexions assez complexes pour se concentrer sur un maximum de détails et faire en sorte de n'en éviter aucun, afin que le travail soit le plus cohérent.

Chapitre 8. Missions secondaires

Lors de mon stage, j'ai également pu participer à d'autres missions qui ne concernaient pas (ou pas uniquement) l'équipe de recherche et qui ont aidé à améliorer le produit.

Le produit de Buster a été programmé pour exécuter des recherches en français et en anglais principalement, mais les ingénieurs de Buster cherchent à améliorer leur produit en ajoutant d'autres langues. Pour cela, on m'a proposé de rejoindre Cedra, Célia et Léa sur une mission où nous avons, sur un fichier Excel, fait la liste de sources internationales provenant d'articles ou d'agences de presse. Ce travail sera ensuite repris par les développeurs afin d'améliorer le produit de Buster pour qu'il possède un maximum de sources et donc d'informations pour toutes les langues et pour tous les pays. Buster souhaite que sur son produit on puisse retrouver toutes sortes d'informations concernant n'importe quel pays, et non pas seulement des informations françaises ou européennes. Sur ce fichier, nous avons indiqué pour chaque article ou agence, son nom, son pays, les langues dans lesquelles l'article peut être lu, son lien, éventuellement d'autres sources reliées s'il en existe et le type d'information que produit cet article ou cette agence (s'il s'agit par exemple d'information politique, économique, sportive ou générale). Pour chaque source, nous avons également rajouté les liens du flux RSS ou des Sitemap, afin de faciliter le travail pour les développeurs ensuite lors du Web Scraping pour récupérer les informations. Je me suis principalement occupée des pays africains, européens et sud-américains. J'ai donc majoritairement travaillé avec des articles rédigés en français ou en anglais, puis en espagnol et arabe.

Un travail que j'ai réalisé sur la quasi-totalité de mon stage était de tester le produit en écrivant chaque semaine dix requêtes sur le produit de Buster et d'identifier si un article répondant à cette requête était retrouvé ou non, et si c'était le cas, si cet article confirme ou réfute l'information. Je devais donc réaliser les tests sur le produit[22] directement et ensuite faire une analyse des résultats dans des fichiers JSON[23]. Cette mission consiste à analyser les résultats que renvoie le produit et donc d'étudier sa performance.

[22] https://coverity.test.buster.ai/analyse
[23] Cf. Annexe 3 page 52

Une dernière mission qui ressemble à la précédente, était de compléter un autre fichier Excel (tous les membres de Buster ont participé à cette mission) qui consiste à donner pour chaque requête la meilleure source qui vérifie son information. Il fallait également indiquer sur le fichier si le produit de Buster trouve cette source dans ses résultats ou non lorsque l'on teste cette requête.

Partie 3

-

Prise de recul

Chapitre 9. Prise de recul sur l'entreprise

1. L'Entreprise

Grâce à ce stage j'ai pu découvrir ce qu'est vraiment une entreprise et plus précisément, ce qu'est une start-up. Buster.AI est encore une jeune entreprise, avec une quinzaine de salariés seulement. Elle est divisée en trois équipes, mais certaines personnes travaillent dans plusieurs équipes en même temps (elles ne sont pas fixées sur un seul objectif uniquement). Cependant, l'organisation du travail et l'avancement de projet chez Buster se déroule assez bien. Les ingénieurs de Buster utilisent la méthode Sprint, qui consiste à se fixer un objectif de travail sous deux semaines et faire un point, avec l'équipe entière, après ces deux semaines. Ce point permet de valider ce travail effectué et de fixer un nouvel objectif. J'ai trouvé cette méthode très intéressante mais je n'ai malheureusement pas pu la tester, puisqu'en tant que stagiaire, je n'avais pas de temps limité pour effectuer mes tâches.

Ce stage m'a appris que travailler dans une start-up n'est pas toujours évident. Surtout lorsqu'il s'agit d'une entreprise comme Buster qui se fixe énormément d'objectifs en si peu de temps. J'ai rapidement relevé le fait que les ingénieurs de Buster étaient souvent débordés par leur charge de travail. En effet, ils dépassent le temps de travail moyen d'une journée pour avancer au plus vite. Mais ceci ne les dérange pas vraiment car ils sont tous passionnés par leur travail, ils aiment donc anticiper sur les projets de Buster en dehors de leurs horaires de bureau. J'ai essayé de suivre ce rythme pendant les premiers mois de mon stage et j'ai trouvé que c'était assez fatigant et gênant car je n'avais plus de temps pour ma vie personnelle.

Par contre, j'ai aussi remarqué que Buster est encore en train de se former. En effet, des nouvelles recrues arrivaient régulièrement dans l'équipe durant mon temps de stage, ce qui est un point positif pour l'entreprise car cela montre que Buster s'agrandit et fera surement bientôt partie de la liste des grandes entreprises.

2. Intégration et relation avec l'équipe de Buster

Buster est une petite entreprise et les membres de cette entreprise sont très aimables, mon intégration au sein de Buster s'est donc très bien déroulée, on m'a très bien accueillie au début de mon stage. J'étais également très satisfaite de l'équipe des

chercheurs que j'avais intégrée. Lors de ma première semaine de stage, les trois chercheurs avaient pris la majorité de leur temps pour m'accueillir et me présenter leur travail ainsi que ce que j'allais faire durant mon stage. Malgré le télétravail, mon intégration au sein de l'entreprise au début du stage était assez positive et motivante. Aussi, j'avais parfois l'impression que dans l'entreprise ils ne nous considèrent pas, Célia et moi, comme des stagiaires mais plutôt comme n'importe quel autre membre de l'équipe. C'est-à-dire qu'ils nous donnaient de vraies tâches avec de réelles responsabilités, on a aussi eu la chance d'assister à de nombreuses réunions.

Concernant ma relation avec les chercheurs (et même avec les autres membres de Buster lorsque j'avais l'occasion de travailler avec eux), elle était plutôt bonne. Ils étaient toujours présents pour répondre à chacune de mes interrogations. Ils étaient très compréhensifs et gentils, tout au long de mon stage, et, en me laissant travailler à mon rythme, ils ne m'ont jamais mis aucune pression.

Avant mon stage, j'appréhendais un peu le fait d'être en télétravail alors qu'eux, ils étaient quasiment tous en présentiel. J'avais peur d'être parfois oubliée ou mise de côté, ce qui n'est presque jamais arrivé avec les chercheurs, ni avec Léa (avec qui j'ai souvent travaillé durant mon stage) qui a été très présente pour moi tout en long du stage et a toujours fait en sorte de me faire assister aux réunions et me transmettre toutes les informations. En revanche, certaines personnes n'utilisaient pas la même méthode. Il est vrai que je travaillais la majorité du temps avec Célia, qui elle, était en présentiel. On m'a donc parfois transmis les informations via Célia, ce qui n'a pas toujours été évident ni pour elle ni pour moi. Il m'est arrivé de travailler plusieurs jours sur une tâche qui n'avait pas lieu d'être car elle avait mal compris et/ou transmis ces informations.

Enfin, la communication avec les supérieurs de Buster n'était pas aussi bonne que celle avec les chercheurs et ingénieurs. Les conditions de travail n'ont pas toujours été correctes. Mon ordinateur n'était malheureusement pas assez performant pour travailler sur les tâches que me confiaient les chercheurs. J'en avais pourtant parlé avec les membres de Buster et essayé de réclamer un nouvel ordinateur (le mien étant tombé en panne en essayant de travailler sur un projet beaucoup plus puissant que mon ordinateur). Ma demande n'ayant pas reçu de réponse positive, j'ai dû m'attribuer mes propres taches pour ne travailler uniquement que sur des projets simples, ne nécessitant pas un ordinateur de grande performance. J'ai donc été déçue de l'entreprise sur la fin de mon stage puisque je devais me débrouiller seule pour régler un problème qui n'aurait pas dû exister.

Chapitre 10. Prise de recul personnelle

Ce stage en télétravail m'a permis de découvrir de nombreux points sur le monde professionnel mais aussi sur moi-même. Je ressors de cette expérience avec de nouveaux objectifs et une nouvelle vision sur l'environnement professionnel, mais aussi de nouveaux traits de caractères.

1. Points positifs du stage

Effectuer un stage en fin d'études m'a permis de faire un point sur ce que je suis capable de faire et de juger mes compétences après avoir acquis de nombreuses connaissances tout au long de ma formation universitaire.

Grâce à cette expérience professionnelle, j'ai appris à gérer mon organisation et à apprendre à travailler en équipe. Il est vrai qu'avant j'appréciais surtout le travail en autonomie et j'avais du mal à travailler avec d'autres personnes sur un même projet. Pendant mon stage, j'ai réalisé la majorité de mes missions avec Célia avec qui je n'avais jamais travaillé auparavant et qui partage avec moi un trait de caractère assez fort. Pour que le stage se déroule dans de bonnes conditions, j'ai appris à partager mes idées mais surtout à accepter d'autres idées qui ne sont pas forcément en accord avec les miennes et aussi à prendre sur moi lors des désaccords.

Buster avait normalement besoin de moi pour les aider sur le côté linguistique de leur produit, mais je n'ai finalement que très peu apporté mes connaissances en linguistique durant mon stage. Au début de mon stage, j'étais un peu déçue à l'idée de ne pas faire plus de linguistique que ce que j'avais pensé faire. J'ai rapidement changé d'avis puisque mon stage m'a permis de me pencher sur les autres côtés du TAL. J'ai surtout développé mes connaissances en programmation et en intelligence artificielle. En effet, j'ai écrit de nombreux programmes en langage Python, des programmes que je n'avais jamais écrits avant. Ces missions ne m'ont pourtant pas toujours été évidentes puisque les chercheurs ont tous un niveau de programmation très élevé par rapport au mien. J'avais parfois un peu de mal à suivre le rythme. Buster m'a donc permis d'améliorer mes compétences en programmation. J'ai aussi appris à construire des jeux de données et brièvement vu comment entrainer des modèles pour l'apprentissage automatique. Cette partie a été la meilleure de mon stage, j'ai découvert l'intelligence artificielle dans un vrai

contexte et pour une véritable utilisation avec de réels résultats. L'intelligence artificielle est un domaine assez vaste et complexe, mais grâce à mon stage qui m'a permis de travailler dans le cœur de ce domaine, je suis fière d'avoir rejoint cette entreprise qui m'a aidée à développer de nouvelles compétences.

2. Points négatifs du stage

Comme dans beaucoup d'expériences, on trouve toujours quelques points qui ne se sont pas déroulés comme on l'aurait souhaité. Pendant ces cinq derniers mois, je n'ai presque jamais travaillé seule. Il m'a donc été particulièrement difficile, de mettre en avant mes compétences et mon propre travail. Vers la fin de mon stage, j'ai surtout travaillé avec Célia et Clément, qui ont plus de compétences que moi en programmation, j'avais donc parfois un peu de mal à suivre leur rythme.

Ce qui m'a surtout posé problème durant mon stage était les pannes que j'ai eues avec mon ordinateur, qui m'ont empêché de travailler sur les taches qui m'étaient normalement assignées. Le fait que l'entreprise me demande de trouver une solution moi-même à ce problème, sachant que tous les autres membres de Buster (y compris les stagiaires) avaient reçu le matériel nécessaire pour travailler sur les projets professionnels, m'a déçue et démotivée sur la fin de mon stage.

Conclusion

Je tire un bilan positif de cette expérience professionnelle qui a été très enrichissante.

Pendant cinq mois, j'ai fait partie de l'entreprise Buster.AI qui m'a appris à utiliser mes connaissances acquises lors de mon Master pour les aider à traiter le problème principal qu'ils rencontraient, celui de retrouver le passage exact dans un article qui répond au mieux à la requête entrée par l'utilisateur ainsi que de relever la bonne étiquette de vérification. Grâce aux nombreux jeux de données que j'ai construit ainsi qu'aux différents tests sur les modèles que j'ai effectués tout au long de mon stage, cette problématique a bien avancé et on a pu observer de meilleurs résultats pour le produit de Buster.

Pendant ce stage, j'ai aussi acquis de nouvelles compétences personnelles, comme celle de travailler en équipe et de me fixer des objectifs sur une courte durée. Cette expérience m'a également confirmé que le traitement automatique des langues est un domaine qui m'intéresse énormément.

Aujourd'hui, après avoir réalisé ces cinq mois de ce stage avec Buster.Ai, je sais que je souhaite m'orienter vers une entreprise qui traite le domaine de l'intelligence artificielle, ce qui me permettra de développer mes compétences acquises lors de ma formation universitaire et professionnelle.

Bibliographie

Benveniste, E. (1966). *Problèmes de linguistique générale*. Paris : Gallimard.

Beyssade, C. (2007). *La structure de l'information dans les questions : quelques remarques sur la diversité des formes interrogatives en français.* Linx, Presses Universitaires de Paris Nanterre. https://jeannicod.ccsd.cnrs.fr/ijn_00356329

Bigot, L. (2017). Communication & langages. In *Le fact-checking ou la réinvention d'une pratique de vérification.* (p. 131-156). https://doi.org/10.4074/S0336150017012091

Coveney, A. (1997). La variation en syntaxe. In *L'approche variationniste et la description de la grammaire du français : Le cas des interrogatives.* (p. 88-100). https://www.persee.fr/docAsPDF/lfr_0023-8368_1997_num_115_1_6224.pdf

Dagnac, A. (2013). *La variation des interrogatives en français.* https://hal-univ-tlse2.archives-ouvertes.fr/hal-00988751v2/document

Lambrecht, K., & Michaelis, L. A. (1998). Sentence Accent in Information Questions : Default and Projection. *Springer*, *21*(5), 477-544.

Tellier, C., & Valois, D. (2006). 9. L'inversion du sujet. In *Constructions méconnues du français* (Presses de l'Université de Montréal).

Zumwald, G. (2010). *Traiter la diversité des formes interrogatives en français : Apports et limites de l'approche variationniste* [Université de Neuchâtel Faculté des Lettres et Sciences humaines]. https://doc.rero.ch/record/20611/files/G_Zumwald_m_moire.pdf.

Sitographie

BUSTER.AI : http://www.dei.unipd.it/~ferro/CLEF-WN-Drafts/CLEF2020/paper_134.pdf (dernière consultation juillet 2021)
Produit de Buster : https://coverity.test.buster.ai/ (dernière consultation juillet 2021)
FEVEROUS : https://arxiv.org/pdf/2106.05707.pdf (dernière consultation juillet 2021)
FEVEROUS : https://arxiv.org/pdf/2004.07347.pdf (dernière consultation juillet 2021)
FEVEROUS : https://arxiv.org/pdf/2103.12011.pdf (dernière consultation juillet 2021)
FEVEROUS : https://fever.ai/ (dernière consultation juillet 2021)
FEVEROUS : https://eval.ai/web/challenges/challenge-page/1091/overview (dernière consultation juillet 2021)
Web Scraping : https://www.actualitesdudroit.fr/browse/tech-droit/start-up/9404/le-web-scraping-une-technique-d-extraction-legale (dernière consultation juillet 2021)
Web Scraping : https://fr.wikipedia.org/wiki/Web_scraping (dernière consultation aout 2021)
Fact-Checking : https://fr.wikipedia.org/wiki/V%C3%A9rification_des_faits (dernière consultation aout 2021)
Fact-Checking : https://www.lemonde.fr/big-browser/article/2018/08/23/a-l-agence-france-presse-plongee-dans-le-service-fact-checking_5345538_4832693.html (dernière consultation aout 2021)
Formulation des requêtes : https://jeannicod.ccsd.cnrs.fr/ijn_00356329/document (dernière consultation mai 2021)
Formulation des requêtes : http://www.semantique-gdr.net/dico/index.php/Structure_informationnelle (dernière consultation mars 2021)
Formulation des requêtes : https://doi.org/10.4074/S033615001701209 (dernière consultation mars 2021)
Formulation des requêtes : https://doc.rero.ch/record/20611/files/G_Zumwald_m_moire.pdf (dernière consultation mars 2021)
Formulation des requêtes : https://hal-univ-tlse2.archives-ouvertes.fr/hal-00988751v2/document (dernière consultation mars 2021)
Formulation des requêtes : https://www.persee.fr/docAsPDF/lfr_0023-8368_1997_num_115_1_6224.pdf (dernière consultation mars 2021)
Ingénieurs Backend : https://fr.wikipedia.org/wiki/Backend (dernière consultation juillet 2021)
Ingénieurs full stack : https://fr.wikipedia.org/wiki/D%C3%A9veloppeur_full_stack (dernière consultation juillet 2021)
Le Monde : https://www.lemonde.fr/ (dernière consultation juillet 2021)
Libération : https://www.liberation.fr/ (dernière consultation juillet 2021)
Wikipédia : https://fr.wikipedia.org/ (dernière consultation juillet 2021)
Google : https://www.google.com/ (dernière consultation juillet 2021)
Bing : https://www.bing.com/ (dernière consultation juillet 2021)
Elastic Search : https://www.elastic.co/fr/elasticsearch/ (dernière consultation juillet 2021)

Spacy FR : https://spacy.io/models/fr (dernière consultation juillet 2021)
Spacy EN : https://spacy.io/models/en (dernière consultation juillet 2021)
GitLab : https://gitlab.com/gitlab-org/gitlab (dernière consultation juillet 2021)
Confluence : https://www.atlassian.com/fr/software/confluence (dernière consultation juillet 2021)
DeepL : https://www.deepl.com/fr/translator (dernière consultation aout 2021)
Selenium : https://selenium-python.readthedocs.io/ (dernière consultation mai 2021)
Python : https://www.python.org/ (dernière consultation juillet 2021)

Sigles et abréviations utilisés

NEI : Not Enough Info

NA : No Answer

IPTC : International Press & Communications Council

URL : Uniform Resource Locator

FEVER : Fact Extraction and VERification

FEVEROUS : Fact Extraction and VERification Over Unstructured and Structured information

JSON : JavaScript Object Notation

TAL : Traitement Automatique des Langues

Table des illustrations

Table des annexes

Annexe 1
Extrait du fichier Excel comprenant tous les codes IPTC

A1156 | http://cv.iptc.org/newscodes/mediatopic/20001311

	NewsCode-URI	NewsCode-QCode (flat	Level1NewsCode	Level2NewsCode	Level3NewsCode	Level4NewsCode	Level5NewsCode	Level6NewsCode	RetiredDate	Name (ar)	Definition (ar)	Name (de)	Definition (de)
1133	http://cv.iptc.org/newscodes/mediatopic/20000982 (retired)	medtop:20000982				medtop:20000982			2015-02-24T12:00:00+00:00	رياضي		K4 (Kajak)	Ein geschlossenes
1134	http://cv.iptc.org/newscodes/mediatopic/20001306	medtop:20001306			medtop:20001306								
1135	http://cv.iptc.org/newscodes/mediatopic/20000986	medtop:20000986			medtop:20000986					لاكروس		Lacrosse	Zwei Teams mit ge
1136	http://cv.iptc.org/newscodes/mediatopic/20000987	medtop:20000987			medtop:20000987					الزحافات الثلجية		Rodeln	Wird als Einsitzer
1137	http://cv.iptc.org/newscodes/mediatopic/20000988	medtop:20000988			medtop:20000988					ماراثون		Marathon	Straßenrennen, be
1138	http://cv.iptc.org/newscodes/mediatopic/20001157	medtop:20001157			medtop:20001157							Kampfsport	Kampfkünste sind
1139	http://cv.iptc.org/newscodes/mediatopic/20001084	medtop:20001084				medtop:20001084				تايكواندو		Taekwon-Do	Kampfsportart, urs
1140	http://cv.iptc.org/newscodes/mediatopic/20000969	medtop:20000969				medtop:20000969				جودو		Judo	Eine verteidigende
1141	http://cv.iptc.org/newscodes/mediatopic/20000970 (retired)	medtop:20000970					medtop:20000970		2015-02-24T12:00:00+00:00	الوزن الخفيف (جودو)		Extra Leichtgewicht (Judo)	Bis zu 60 kg bei M
1142	http://cv.iptc.org/newscodes/mediatopic/20000971 (retired)	medtop:20000971					medtop:20000971		2015-02-24T12:00:00+00:00	الوزن نصف الثقيل		Halb-Schwergewicht (Judo)	Bis zu 100 kg bei M
1143	http://cv.iptc.org/newscodes/mediatopic/20000972 (retired)	medtop:20000972					medtop:20000972		2015-02-24T12:00:00+00:00	الوزن نصف الخفيف		Halb-Leichtgewicht (Judo)	Bis zu 73 kg bei M
1144	http://cv.iptc.org/newscodes/mediatopic/20000973 (retired)	medtop:20000973					medtop:20000973		2015-02-24T12:00:00+00:00	الوزن نصف المتوسط		Halb-Mittelgewicht (Judo)	Bis zu 81 kg bei M
1145	http://cv.iptc.org/newscodes/mediatopic/20000974 (retired)	medtop:20000974					medtop:20000974		2015-02-24T12:00:00+00:00	الوزن الثقيل (جودو)		Schwergewicht (Judo)	Gewöhnlich über 1
1146	http://cv.iptc.org/newscodes/mediatopic/20000975 (retired)	medtop:20000975					medtop:20000975		2015-02-24T12:00:00+00:00	الوزن الخفيف (جودو)		Leichtgewicht (Judo)	Bis zu 66kg bei M
1147	http://cv.iptc.org/newscodes/mediatopic/20000976 (retired)	medtop:20000976					medtop:20000976		2015-02-24T12:00:00+00:00	الوزن المتوسط (جودو)		Mittelgewicht (Judo)	Bis zu 90kg bei M
1148	http://cv.iptc.org/newscodes/mediatopic/20000977	medtop:20000977				medtop:20000977				جوكندو		Jukendo	Japanische traditio
1149	http://cv.iptc.org/newscodes/mediatopic/20000983	medtop:20000983				medtop:20000983				كاراتيه		Karate	Eine Kampfsportar
1150	http://cv.iptc.org/newscodes/mediatopic/20000984	medtop:20000984				medtop:20000984				كندو		Kendo	Traditionelle japan
1151	http://cv.iptc.org/newscodes/mediatopic/20001310	medtop:20001310				medtop:20001310							
1152	http://cv.iptc.org/newscodes/mediatopic/20001308	medtop:20001308				medtop:20001308							
1153	http://cv.iptc.org/newscodes/mediatopic/20000985	medtop:20000985				medtop:20000985				كيودو		Kyodo	Japanische traditio
1154	http://cv.iptc.org/newscodes/mediatopic/20001231	medtop:20001231				medtop:20001231						gemischte Kampfkünste	Ein Vollkontakt-Ka
1155	http://cv.iptc.org/newscodes/mediatopic/20001009	medtop:20001009				medtop:20001009				ناعيتا		Naginata	Japanische traditio
1156	http://cv.iptc.org/newscodes/mediatopic/20001311	medtop:20001311				medtop:20001311							
1157	http://cv.iptc.org/newscodes/mediatopic/20001102	medtop:20001102				medtop:20001102				ووشو		Wushu	Grundsätzliche chi
1158	http://cv.iptc.org/newscodes/mediatopic/20000989	medtop:20000989			medtop:20000989					بينتاثلون حديث		Moderner Fünfkampf	Der moderne Fünfk
1159	http://cv.iptc.org/newscodes/mediatopic/20000991	medtop:20000991			medtop:20000991					سباق الورارق السريعة		Motorsport	Rennen mit Autos
1160	http://cv.iptc.org/newscodes/mediatopic/20000993 (retired)	medtop:20000993				medtop:20000993			2017-07-04T12:00:00+00:00	فورمولا3000		F3000	Mit weniger starker
1161	http://cv.iptc.org/newscodes/mediatopic/20000994 (retired)	medtop:20000994				medtop:20000994			2017-07-04T12:00:00+00:00	فورمولا		Formel 1	Einzelfahrer Autor
1162	http://cv.iptc.org/newscodes/mediatopic/20000995 (retired)	medtop:20000995				medtop:20000995			2017-07-04T12:00:00+00:00	انديكار		Indy	Nordamerikanische
1163	http://cv.iptc.org/newscodes/mediatopic/20000992 (retired)	medtop:20000992				medtop:20000992			2017-07-04T12:00:00+00:00	سباق التحمل		Ausdauer	Rennen mit einem
1164	http://cv.iptc.org/newscodes/mediatopic/20000996 (retired)	medtop:20000996				medtop:20000996			2017-07-04T12:00:00+00:00	سيارات مسافات طويلة		Crossrallye	Rennen im Stadion
1165	http://cv.iptc.org/newscodes/mediatopic/20000997 (retired)	medtop:20000997				medtop:20000997			2017-07-04T12:00:00+00:00	سباق سيارات		Rallye	Ein Saison von Au
1166	http://cv.iptc.org/newscodes/mediatopic/20000990	medtop:20000990			medtop:20000990					زورق سريع		Motorboot	Rennen zwischen
1167	http://cv.iptc.org/newscodes/mediatopic/20000998	medtop:20000998			medtop:20000998					سباق دراجات نارية		Motorradrennen	Rennen mit 2,3 ode
1168	http://cv.iptc.org/newscodes/mediatopic/20001000 (retired)	medtop:20001000				medtop:20001000			2017-07-04T12:00:00+00:00	سباق الطرق الوعرة		Enduro	Geschwindigkeitsr
1169	http://cv.iptc.org/newscodes/mediatopic/20001001 (retired)	medtop:20001001				medtop:20001001			2017-07-04T12:00:00+00:00	مضمار عشبي		Grasbahn	Rennen, was auf G
1170	http://cv.iptc.org/newscodes/mediatopic/20001002 (retired)	medtop:20001002				medtop:20001002			2017-07-04T12:00:00+00:00	كرة الدراجات النارية		Motoball	Mannschaftssport
1171	http://cv.iptc.org/newscodes/mediatopic/20001003 (retired)	medtop:20001003				medtop:20001003			2017-07-04T12:00:00+00:00	موتوكروس		Motocross	Rennen über schm
1172	http://cv.iptc.org/newscodes/mediatopic/20001004 (retired)	medtop:20001004				medtop:20001004			2017-07-04T12:00:00+00:00	الجائزة الكبرى لسباق الدراجات النارية		MotoGP	Hubraumklasse 10
1173	http://cv.iptc.org/newscodes/mediatopic/20000999 (retired)	medtop:20000999				medtop:20000999			2017-07-04T12:00:00+00:00	سباق تحمل للدراجات النارية		Ausdauer	Straßenrennen mit
1174	http://cv.iptc.org/newscodes/mediatopic/20001005 (retired)	medtop:20001005				medtop:20001005			2017-07-04T12:00:00+00:00	رالي		Rallye	Rennen über regul
1175	http://cv.iptc.org/newscodes/mediatopic/20001008 (retired)	medtop:20001008				medtop:20001008			2017-07-04T12:00:00+00:00	دراجات نارية لقفز الحواجز		Trial	Rennen mit Belast
1176	http://cv.iptc.org/newscodes/mediatopic/20001006 (retired)	medtop:20001006				medtop:20001006			2017-07-04T12:00:00+00:00	دراجة بمقعد جانبي		Beiwagen	Mit verschiedenen
1177	http://cv.iptc.org/newscodes/mediatopic/20001007 (retired)	medtop:20001007				medtop:20001007			2017-07-04T12:00:00+00:00	مضمار		Speedway	Rennen mit bis zu
1178	http://cv.iptc.org/newscodes/mediatopic/20000884	medtop:20000884			medtop:20000884					تسلق		Bergsteigen	Sich am Berg hoch
1179	http://cv.iptc.org/newscodes/mediatopic/20000885	medtop:20000885				medtop:20000885				تسلق الجليد		Eisklettern	Klettern auf Eisflä
1180	http://cv.iptc.org/newscodes/mediatopic/20000886	medtop:20000886				medtop:20000886				تسلق الجبال		Alpinismus	Bergsteigen im Ho

Annexe 2
Code python générant la liste des pages Wikipédia par requête

```
import json
import numpy as np

with open('train_questions.json') as t:
    train = json.load(t)

with open('validation_questions.json') as v:
    validation = json.load(v)

all_labels = {}

dataset = {'train': train,'validation' : validation}

for data_name in dataset :
    current_data = dataset[data_name]
    for key in current_data:
        labels = current_data[key]["label"]
        labels_unique = np.unique(labels)
        for label in labels_unique:
            if label in all_labels:
                if data_name in all_labels[label] :
                    all_labels[label][data_name].append(key)
                else :
                    all_labels[label][data_name] = [key]

            else:
                all_labels[label] = {data_name : [key]}

common_labels = {}
for key in all_labels:
    value = all_labels[key]
    if 'train' in value and 'validation' in value:
        common_labels[key] = all_labels[key]

#to json file
with open ('labels.json', 'w') as jf:
    json.dump(common_labels, jf, indent=4)
```

Annexe 3
Exemple d'un fichier JSON envoyé chaque semaine


```
{
    "scenario": "ANALYZE_TEXT__NA__100_NIGHTS_PROTEST_PORTLAND",
    "request": [
        {
            "event": "ANALYZE_TEXT__NA__100_NIGHTS_PROTEST_PORTLAND",
            "content": {
                "text": "100 nights of protest in Portland"
            }
        }
    ],
    "response": [
      {
        "event": "ANALYZE_TEXT_RESPONSE",
        "checks": [
          {
            "key": "content.factCheck.result",
            "value": "NA"
          },
          {
            "key": "content.factCheck.factCheckScore",
            "min": -1
          },
          {
            "key":
"content.factCheck.explanations.otherEvidences[0].info.meta.title",
            "value": "Gaz lacrymogène et cocktails molotov pour la 100e nuit
de manifestation à Portland"
          },
          {
            "key":
"content.factCheck.explanations.otherEvidences[0].info.meta.id",
            "value": "http://doc.afp.com/1X097D"
          },
          {
            "key":
"content.factCheck.explanations.otherEvidences[0].info.meta.kind",
            "value": "NEWS_AGENCY"
          },
          {
            "key":
"content.factCheck.explanations.otherEvidences[0].info.meta.url[0]",
            "value": "https://www.afp.com/"
          },
          {
            "key":
"content.factCheck.explanations.otherEvidences[0].matchingScore",
            "value": 0.9470999999999776
          },
          {
            "key":
"content.factCheck.explanations.otherEvidences[0].entailmentScore",
            "value": -1
```

Table des matières

MOTS-CLÉS : intelligence artificielle, vérification des faits, jeu de données, traitement automatique des langues

RÉSUMÉ

Durant cinq mois j'ai travaillé sur une thématique portant sur l'intelligence artificielle qui consistait à vérifier des faits dans des articles de presse majoritairement. C'est en intégrant l'entreprise Buster.AI, dont le but premier étant de réaliser mon stage de fin d'études, que j'ai pu traiter ce sujet. Tout au long de mon stage, j'ai travaillé avec des bases de données importantes et ai construit des jeux de données qui ont permis de tester, d'entrainer et d'améliorer les modèles de l'entreprise. Ce travail a développé de nouvelles performances au produit de Buster grâce aux résultats obtenus. Ce stage a également approfondie mes compétences et m'en a appris des nouvelles dans le domaine du traitement automatique des langues. J'ai aussi, grâce à cette expérience professionnelle, pu découvrir le monde professionnel et l'organisation d'une entreprise.

KEYWORDS : artificial intelligence, fact-checking, data set, natural language processing

ABSTRACT

During five months I worked on a topic related to artificial intelligence which was about the fact-checking in newspaper mainly. It is by joining the company Buster.AI, whose first goal was to carry out my end of studies internship, that I was able to study this subject. Throughout my internship, I worked with large databases and built many datasets that allowed me to test, train and improve the company's models. This work developed new performances to the Buster product thanks to the new results obtained. This internship has also improved my skills and taught me new ones in the domain of natural language processing. Thanks to this work experience, I was also able to discover the professional world and the organization of a company.