

UNIVERSITE DE MONTPELLIER
RAPPORT DE PROJET

Claims checking

Belkassim BOUZIDI
Chakib ELHOUITI
Massili KEZZOUL
Abdelkader NEDJARI
Ramzi ZEROUAL

Encadrant :
M^r Konstantin TODOROV



UNIVERSITÉ
DE MONTPELLIER



10 mai 2020

Remerciements

Tout d'abord nous souhaitons adresser nos remerciements au corps professoral et administratif de la faculté des sciences de Montpellier qui déploient des efforts pour assurer à leurs étudiants une formation actualisée.

En second lieu, nous tenons à remercier notre encadrant M^r Konstantin TODOROV pour ses précieux conseils et son aide durant toute la période du travail.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre projet en acceptant d'examiner notre travail.

Nous remercions M^r Yahia Zeroual pour sa relecture attentive de ce rapport.

Table des matières

1	Organisation du projet	3
1.1	Méthodes d'organisation	3
1.2	Découpage du projet	3
1.2.1	Phase de modélisation	3
1.2.2	Phase de développement	3
1.2.3	Finalisation du projet	3
1.3	Outils de collaboration	4
2	Introduction au sujet	5
2.1	Fact-checking	5
2.1.1	Présentation du principe de fact-checking	5
2.1.2	Présentation de ClaimsKG	5
2.1.3	Travail à réaliser	6
2.2	Technologies utilisées	6
3	Conception et implémentation du projet	7
3.1	Conception et Modélisation	7
3.1.1	Analyse de ClaimKG	7
3.1.2	Conception	9
3.2	Implémentation	9
4	Analyse des résultats	10
4.1	Résultats	10
4.2	Problèmes rencontrés	11
5	Bilan et conclusions	12
5.1	Perspective	12
A	Annexe	13

Chapitre 1

Organisation du projet

1.1 Méthodes d'organisation

Afin de mener à bien le développement du projet, nous avons décidé de travailler un maximum de temps ensemble et de manière très régulière. Nous nous sommes réunis trois à quatre fois par semaine, en vue de faire le point sur l'avancement du projet et de définir les objectifs restants à atteindre.

Ainsi, selon l'état de progression de la conception, nous réalisions les tâches en retard durant le week-end pour ne pas cumuler de retard et respecter l'intégralité du cahier de charges.

Toutes les semaines, nous nous sommes réunis avec notre encadrant, M^r Konstantin TODOROV. Lors de ces réunions de mises au point relatifs au projet, nous furent prodigués, cela nous a permis de bénéficier de précieux conseils.

1.2 Découpage du projet

Nous avons découpé la réalisation du projet en trois grandes phases :

1.2.1 Phase de modélisation

Durant cette étape, nous nous sommes réunis pour définir les fonctionnalités demandées par le projet. Notamment séparer les fonctionnalités importantes de celles moins importantes. Nous avons également choisi les outils de travail collaboratifs et les principales technologies utilisées, ainsi qu'une première modélisation du projet.

1.2.2 Phase de développement

Durant cette phase, nous avons commencé à implémenter les différentes fonctionnalités que nous avons modélisées lors de l'étape précédente, tout en améliorant la modélisation au fur et à mesure de l'avancement de notre projet. Nous avons notamment réalisé des tests pour les différents modules afin de s'assurer de leur bon fonctionnement.

1.2.3 Finalisation du projet

Cette étape a consisté en la réalisation des tests finaux afin de s'assurer que les différents scripts fonctionnent en toute circonstance et éventuellement corriger les bogues qui peuvent apparaître.

1.3 Outils de collaboration

Afin de s'organiser, nous avons décidé d'utiliser Git à travers le serveur GitLab hébergé par le service informatique de la faculté. En effet le logiciel libre Git a facilité grandement la collaboration entre nous. Le serveur GitLab quant à lui est fourni gratuitement par le service informatique de la faculté.

En ce qui concerne la rédaction de ce rapport, nous avons utilisé \LaTeX , système de composition de documents créé par Leslie Lamport, pour faciliter la rédaction à plusieurs.



Schéma 1.1 – Logo du GitLab



Schéma 1.2 – Logo de Latex

Chapitre 2

Introduction au sujet

2.1 Fact-checking

2.1.1 Présentation du principe de fact-checking

Le fact-checking ou la vérification des faits, est une technique consistant à vérifier la véracité des faits et l'exactitude des chiffres présentés dans les médias, les différents réseaux sociaux, les blogs, etc... Cette notion est apparue aux États-Unis dans les années 1990. Elle a été mise en pratique par des journalistes d'investigation dans le cadre de leur profession, la méthode s'est démocratisée grâce à des logiciels aidant les particuliers à vérifier les faits.

L'entrée officielle du fact-checking en France date de 1995, quand est créée l'association Acrimed, qui se présente comme « l'observatoire des médias ».

En vue de la prolifération très rapide des informations, il devient de plus en plus important de s'assurer de la véracité des informations qui se trouvent partout sur Internet et autres médias. Tant du point de vue de la société que de celui de la recherche. De nombreuses approches récentes dans diverses communautés scientifiques portent sur des problèmes tels que la vérification des faits, la détection de la pertinence ou de point de vue des documents par rapport à des revendications particulières.

2.1.2 Présentation de ClaimsKG

Le LIRMM¹ en collaboration avec 2 équipes allemandes (L3S Hannover et l'institut de sciences sociologiques GESIS à Cologne) a construit et mis à disposition la base de connaissance ClaimsKG² qui regroupe les informations et méta-données provenant d'un grand nombre de sites journalistiques internationaux de fact checking, tels que Politifact³ ou Snopes⁴. ClaimsKG est un graphe de connaissances d'assertions annotées et liées qui facilite la création de requêtes structurées sur les assertions, leurs valeurs de vérité (True, Mostly False, etc...), leurs auteurs, date de publication, etc... ClaimsKG est généré par un pipeline entièrement automatisé qui collecte des assertions et des métadonnées à partir des sites de fact-checking, il transforme les données en graphes de connaissances selon un modèle établi, et annote les assertions avec des entités DBpedia (Wikipedia). La base actuelle comprend plus de 32 000 assertions publiées depuis 1996 et est mise à jour régulièrement.

1. Le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier

2. <https://github.com/claimskg>

3. <https://www.politifact.com/>

4. <https://www.snopes.com/>

2.1.3 Travail à réaliser

Le sujet du TER consiste en l'enrichissement de cette base de connaissances avec des nouvelles données provenant des sites web suivants :

Fatabyyano ⁵ est un site jordanien en arabe. Fatabyyano (terme en arabe qui veut dire "Alors montrez-le") est la première et la seule plateforme arabe certifiée par l'IFCN ⁶ ;

Vishvas.news ⁷ est un site Internet de vérification des faits multilingue qui s'engage à combattre la désinformation et les informations erronées.

Le but du TER sera d'identifier les assertions individuelles dans chaque histoire, ainsi que leur label de véracité et par la suite identifier les relations entre elles, ainsi que les relations entre ces histoires. Les données produites par ce projet seront intégrées à la base de connaissance ClaimsKG.

La principale difficulté qu'on s'attend à rencontrer est la gestion des différentes langues proposée par ces sites web. Notamment dans l'identification des différentes relations entre les assertions. En effet afin de reconnaître les différents mots-clés du sujet traité par les articles, on utilise habituellement TAGME, un puissant outil de reconnaissance d'entités nommées dans un texte. Les langues utilisées par ces sites web ne sont pas prises en charge par cet outil. Il s'agit donc de trouver une alternative afin de reconnaître ses différentes assertions. Ce problème ainsi que sa résolution seront détaillés plus tard dans ce rapport.

2.2 Technologies utilisées

Nous avons implémenter l'application en Python, choix qui s'impose de lui même car couplé à la bibliothèque BeautifulSoup, il devient très facile de faire du web scraping ⁸.

En ce qui concerne la traduction, nous avons choisi d'utiliser Yandex Translator ⁹. En effet en vue de la taille des données à traduire, nous n'avons pas pu utiliser l'API de traduction de Google puisque cette dernière devient payante à partir d'un certain nombre de caractères.

Enfin, la reconnaissance des différents entités concernées est faite grâce à l'outil TAGME ¹⁰ a travers son API python ¹¹.



Schéma 2.1 – Logo de Python



Schéma 2.2 – Logo de BeautifulSoup



Schéma 2.3 – Logo de Yandex

5. <https://fatabyyano.net/>

6. International Fact-Checking Network : <https://www.poynter.org/ifcn/>

7. <https://www.vishvasnews.com/>

8. Le web scraping est une technique d'extraction du contenu de sites Web, via un script ou un programme

9. <https://translate.yandex.com/>

10. <https://tagme.d4science.org/tagme/>

11. Lien Github vers l'api Python de TAGME : <https://github.com/marcocor/tagme-python>

Chapitre 3

Conception et implémentation du projet

3.1 Conception et Modélisation

3.1.1 Analyse de ClaimKG

Lors de cette première phase, le plus important a été de comprendre et s’imprégner du code et de la structure déjà mis en place et mis à notre disposition (Claimskg), comprendre les outils utilisés ainsi que la structure déjà établie.

La structure

Comme le projet est d’une envergure immense, bien comprendre la structure était primordiale afin de respecter un maximum les outils utilisés. Notre encadrant nous à mis a disposition un écrit (que vous trouverez dans le dossier autre/ sous le nom de Claimskg.pdf) qui explique la structure globale du projet dont voilà ci-dessous un résumé.



Schéma 3.1 – ClaimKG

Comme vous pouvez le voir sur le schéma, Claimskg est découpé en plusieurs phases. La première consiste à extraire les données brutes à partir de site web de fact-checking et ensuite les structurer afin de produire un fichier CSV. La deuxième phase quant à elle, consiste à faire passer ce fichier CSV dans un générateur de graphes de connaissances¹. Dans notre cas, on s’occupe de la première phase, c’est-à-dire, extraire les données brutes du site web afin de produire le fichier CSV.

Le fichier CSV en question contient l’ensemble des faits traités par un site web. Chaque fait est passé à un script de scraping afin d’extraire chaque information et la structurer. La structure d’un fait est réalisée selon un modèle précis dont voici une partie :

rating value : La valeur de véracité du fait ;

1. Knowledge Graph Generator

creativeWork author name : Auteur de la claim ;
creativeWork datePublished : Date de publication de la claim ;
claimReview author : Auteur de l'article/analyse de la claim ;
extra entities claimReview : Les différentes entités présentes dans la review.

Vous trouverez ci-dessous la structuration complète.

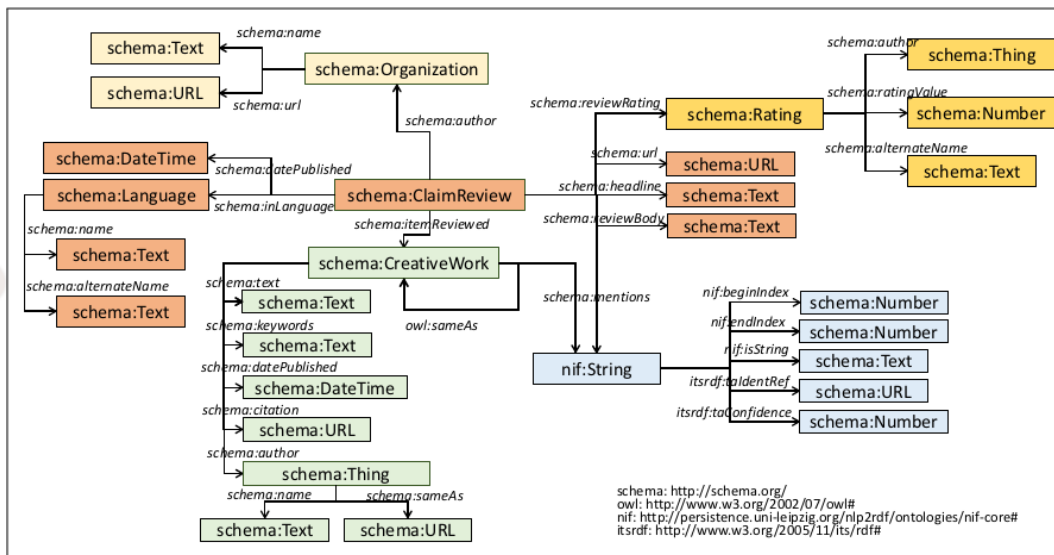


Schéma 3.2 – Structuration complète d'un fait

L'implémentation

Après la partie théorique de l'analyse nous avons commencer à regarder les implémentations déjà faite². Nous avons analysé les classes principales comme celle de « Claim.py », puis comprendre comment l'exécution de l'extraction se faisait avec « __init__.py ». nous avons ensuite analysé les différents scripts d'extraction de sites de fact-checking tel que « africacheck.py ».

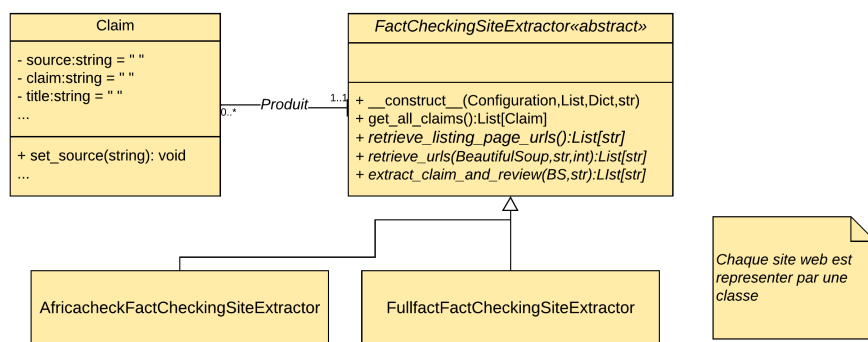


Schéma 3.3 – Structuration de claimsKG-extractor

2. Vous trouverez cette implémentation à cette adresse : <https://github.com/claimskg/claimskg-extractor>

3.1.2 Conception

Dans cette phase, nous allons présenter les étapes de conception que nous avons suivies pour bien implémenter notre application. La première étape a été de concevoir quelques diagrammes UML utiles à la bonne compréhension du projet.

Diagrammes UML

Lors de la phase de la modélisation, nous avons vu qu'il était primordial de concevoir un diagramme de classe pour connaître la manière dont les classes interagissent. Comme la figure suivante le montre, des classes ont été rajoutées par rapport aux classes déjà développées.

Dans les sous-classes de FactCheckingSiteExtractor, la méthode "extract_claim_and_review" a été implémentée pour transformer une page donnée en paramètre en une instance de claim. La méthode "get_all_claims" quant à elle, extrait les liens des articles du site web. Les données, une fois extraites seront stockées dans un fichier CSV où chaque ligne représente un article et chaque colonne les informations le concernant.

Le déroulement de l'extraction des données a été quant à elle modélisée par un diagramme de séquence. La figure ci-dessous représente le fonctionnement de l'extraction du site Fatabayyano.

Puis parler du format des données que notre prog doit rendre

La deuxième étape consistait à trouver un site web respectant les critères de fact-checking et qui nous permet d'extraire les données essentielles (la véracité, les claims écrites, un auteur, etc...) au noyau du projet (ClaimsKG). Pour apporter un maximum de diversité ainsi que notre contribution personnelle, nous avons décidé, avec l'accord de notre encadrant, d'extraire les données d'un site d'une langue qui n'est pas encore présente dans le projet. Vu que tous les étudiants de ce projet sont arabophones, la langue arabe était un choix logique. Le seul site officiel de fact-checking en arabe est : <https://fatabyyano.net/>.

Nous avons donc analysé la structure du site web puis commencé son implémentation. L'étape suivante fût de trouver comment faire une "named entity recognition" sur les mêmes principes que celles déjà présentées (TagMe qui renvoie vers des liens wikipedia).

Le second site web sur lequel nous avons travaillé est <https://www.vishvasnews.com/>, il regroupe 11 langues différentes : L'anglais, Hindi, Pendjabi, Ourdou, Bengali, Tamoul, Malayalam, Goudjerati, Télougou, Marathi et L'odia, nous avons analysé la structure du site puis extrait ses données.

3.2 Implémentation

Chapitre 5

Bilan et conclusions

5.1 Perspective

Le scrapping des sites s'effectue parfaitement, aucune amélioration n'est nécessaire au script en lui-même, la seule perspective pour notre projet est la traduction des 10 langues présentes dans le deuxième site non prises en compte par le programme "tagme", deux idées peuvent être avancées :

Continuer à utiliser Yandex malgré ces capacités limitées tel qu'expliqué plus haut en changeant uniquement le script de traduction du premier site web ou utiliser une autre API, de préférence Google Traduction qui est cependant payante, mais dont l'efficacité est sans conteste la meilleure.

Annexe A

Annexe