TheoremKB : une base de connaissance des résultats mathématiques

Yacine Brihmouche

Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16
dauphine.psl.eu

RAPPORT DE STAGE

Année universitaire 2021/2022

Nom et Prénom de l'étudiant :

Brihmouche Yacine

Année d'études :

☐ L3 INFO ☐ L3 INFO APPR. ☐ L3 MATHS

☐ M1 MIAGE ☒ M1 I2D ☐ M1 INFO APPR. ☐ M1 MATHS

Organisme d'accueil : ECOLE NORMALE SUPERIEURE

Titre du rapport :

TheoremKB : une base de connaissance des Résultats mathématiques.

Tuteur de stage : Pr. Pierre Senellart

Dates du stage : du 30/05/2022 au 04/09/2022

1. L'INTRODUCTION :

Dans le cadre de la validation de ma première année de Master informatique, décision, données, j'ai rejoint l'équipe VALDA du département d'informatique de l'école normale supérieure d'Ulm pour effectuer un stage de trois mois du 30/05/2022 au 04/09/2022.

L'équipe VALDA est une équipe commune au département d'informatique de l'ENS, à Inria Paris et au CNRS. Une équipe qui se spécialise dans les systèmes et la théorie de la gestion des bases de données. J'ai choisi cette équipe et ce stage parce que je voulais découvrir le monde de la recherche académique pour éventuellement préparer une thèse de doctorat après mon diplôme de master.

Le sujet de ce stage m'a aussi intéressé. L'extraction de l'information sémantique des articles de recherche scientifiques implique en même temps des techniques du traitement de langage naturel et de manipulation de données semi-structurés.

Ma mission durant ce stage était le développement d'un nouveau module logiciel pour le projet TheoremKB. Un projet de recherche d'envergure qui a pour but de construire, à partir des articles de recherches scientifiques, une base de connaissances des résultats mathématiques facilement navigable. Le module à implémenter aura pour mission de créer automatiquement un lien entre deux résultats mathématiques (théorèmes, preuves, lemme ...) de deux articles différents si un résultat utilise un autre dans son énoncé ou sa preuve.

Au sein de cette équipe, j'ai été encadré par M. Pierre Senellart, Professeur d'informatique à l'École normale supérieure**.** Et j'ai été accompagné par Shrey Mishra, doctorant à l'ENS.

Le présent rapport s'articule en deux grandes parties. Dans la première partie, je présente l'établissement d'accueil, son histoire, l'équipe VALDA et différents aspects organisationnelles et communicationnelles en m'appuyant sur quelques chiffres clés.

Dans la seconde partie, j'évoque les missions réalisées durant ce stage au sein de l'équipe. Je présente également les différents outils et technologies utilisés dans le cadre de la réalisation de ses missions avec à chaque fois les résultats obtenus.

2. PRESENTATION DE L'ORGANISME D'ACCEUIL :

2.1. Organisation et structure :

Juridiquement, L'École normale supérieure (ENS) appelée aussi « ENS de la rue d'Ulm », «ENS Ulm », « Normale Sup' », parfois « ENS-PSL » ou « ENS », est un établissement public à caractère scientifique, culturel et professionnel. Le directeur de l'ENS est nommé pour cinq ans par décret du président de la République. Il est depuis 1996 assisté de deux directeurs adjoints. Il les nomme pour trois ans. La direction générale des services est responsable des finances et de l'équipe administrative de l'école. Les départements d'enseignement et de recherche de l'ENS sont traditionnellement divisés en deux sections appelées « Lettres » et « Sciences », accueillant des enseignants et étudiants en proportions relativement égales. A chaque département sont rattachés plusieurs laboratoires, ou unités mixtes de recherche (UMR), associés au Centre national de la recherche scientifique (CNRS) ou d'autres institutions de recherche, ainsi que des écoles doctorales cogérées avec d'autres universités d'Île-de-France. On compte sept départements dans chacune des deux sections :

Les sept départements de lettres sont :

- Philosophie ;
- Littérature et langages (LILA) ;
- Histoire ;
- Sciences de l'Antiquité, ou Centre d'études anciennes (CEA) ;
- Sciences sociales et économiques ;
- Géographie ;
- Histoire et théorie des arts, ou « passerelle des arts ».

Les sept départements de sciences sont :

- Physique ;
- Mathématiques et applications (DMA) ;
- Informatique ;
- Biologie ;
- Chimie ;
- Géosciences, ou « Terre atmosphère océans » (TAO) ;
- Études cognitives (DEC).

La bibliothèque de lettres et sciences humaines est considérée comme un département à part entière et ne fait pas partie de l'administration. (Archive, 2020)

2.2. Histoire de l'école normale supérieure :

On présente ci-dessous les dates marquantes de l'histoire de l'école :

- **1794 :** les origines de l'établissement remontent à l'éphémère « École normale de l'an III » créée en 1794 par Dominique Joseph Garat, de Joseph Lakanal sous le régime de la Convention nationale.
- **1808 :** Napoléon crée par décret un « pensionnat normal » au sein de l'Université de France pour « former à l'art d'enseigner les lettres et les sciences »
- **1818 :** un concours d'entrée est instauré pour la première fois mais, considéré comme un foyer de l'esprit libéral, le pensionnat est supprimé par Frayssinous en 1822.
- **1826 :** Une ordonnance du 9 mars 1826 crée une « École préparatoire », dans les locaux du collège Louis-le-Grand.
- **1830 :** À la faveur de la révolution de Juillet 1830, l'École préparatoire prend, par arrêté de Louis-Philippe, le nom d'« École normale » par référence à l'École normale de l'an II .
- **1847 :** L'institution s'installe dans de nouveaux locaux, rue d'Ulm, dans le Ve arrondissement de Paris.
- **1881 :** L'École normale supérieure de jeunes filles (ENSJF), pendant féminin de l'ENS, est créée le 26 juillet 1881 à Sèvres.
- **1879 :** La loi « Paul Bert » impose aux départements de disposer chacun d'une école normale de garçons et, ce qui est nouveau, d'une école normale de filles.
- **1985 :** l'École normale supérieure de jeunes filles de Sèvres et l'ENS (rue d'Ulm) fusionnent13 : il en résulte l'actuelle École normale supérieure, établissement mixte, dont les bâtiments principaux sont toujours à Paris, rue d'Ulm, et qui dispose également des anciens locaux de l'ENSJF, boulevard Jourdan, et à Montrouge.

Les écoles normales supérieures, parmi elles l'ENS d'Ulm, ont connu à travers leurs histoires multiples reconfigurations et évolutions qui reflètent l'évolution de la société française et du régime politique en place. Issus d'origines révolutionnaires et républicaines, l'ENS cherche toujours à faire vivre les valeurs de la république au sein de l'établissement. De nombreuses personnalités marquantes de l'histoire de France ont enseigné ou dirigé dans l'ENS. On peut citer : Marie Curie, Louis Pasteur, Simone de Beauvoir et Jean-Paul Sartre, Georges Pompidou … (Wikipedia, 2022)

2.3. Caractéristiques globales de l'ENS :

2.3.1. Effectifs :

L'École normale supérieure accueille chaque année près de 2330 étudiants : (l'ENS, 2022)

- 900 élèves normaliens (concours CPGE)
- 420 étudiants normaliens (concours universitaire)
- 60 étudiants de la Sélection internationale
- 230 mastériens non normaliens
- 490 doctorants
- 230 étudiants en échange international
- 43 Masters
- 15 écoles doctorales rattachées à l'ENS
- 720 stages longs

L'ENS compte 1350 enseignants et chercheurs permanents qui animent la vie scientifique et pédagogique de l'École au sein des 15 départements et 29 laboratoires de recherche (17 Sciences et 12 en Lettres) soit :

- 450 enseignants-chercheurs
- 890 enseignants-chercheurs (autres établissements)
- 490 agents administratifs et techniques
- 500 agents administratifs et techniques (autres établissements)
- 115 post-doctorants
- 490 doctorants
- 15 écoles doctorales co-accréditées PSL
- 124 contrats de recherche signés

2.3.2. Classements internationaux :

L'ENS est un établissement composant de l'université Paris sciences et lettres (PSL). À ce titre, elle bénéficie du classement de PSL dans les principaux classements internationaux, au côté de tous les établissements qui composent cet établissement. En effet, vu la taille humaine de l'établissement, il était difficile de la comparer avec d'autres universités qui comptent parmi leurs effectifs 10 fois plus d'étudiants et de chercheurs.

Sous le drapeau de PSL, l'ENS a significativement progressé dans les différents classements internationaux.

Pour Alain Fuchs, président de l'Université PSL :

« Ce résultat est la récompense des efforts considérables menés depuis dix ans. Le classement QS est l'un des trois plus regardés au monde, avec celui de Shanghai et celui du Times Higher Education. Figurer dans le top 30 constitue un précieux passeport pour nos étudiantes et étudiants. En témoigne encore récemment l'éligibilité des diplômés de PSL au visa high-potential individuals mis en place par le gouvernement britannique ». (Fuchs, 2022)

Classement QS 2023 : PSL est dans la 26ème position, en progression constante.

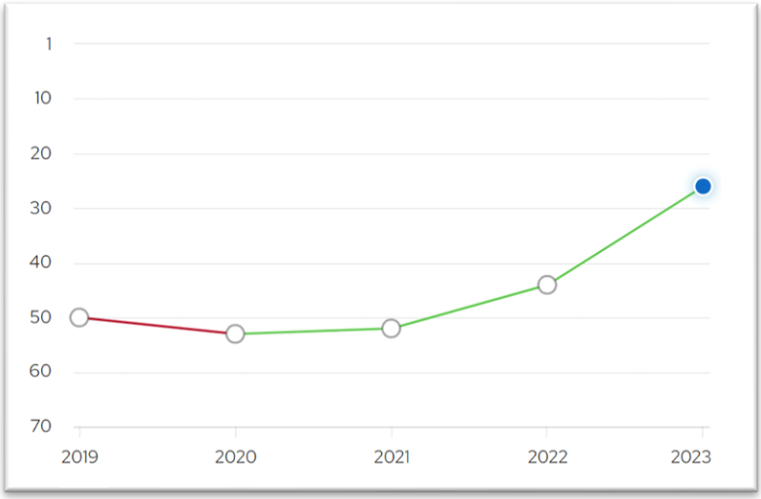


Figure 1: évolution du classement QS de PSL université au fil du temps
source : https://www.topuniversities.com/universities/universite-psl

Classement "Times Higher Education" 2022 **:** 40ème place.

Classement de Shanghai (ARWU) 2021 **:** 38ème place.

Classement Young University Rankings 2022 **:**

Ce classement liste les meilleures universités au monde qui ont moins de 50 ans d'existence. PSL prend la 1ème place du classement. (PSL, 2022)

2.4. Présentation de l'équipe VALDA du département d'informatique :

Durant mes trois mois de stage, j'ai intégré l'équipe VALDA (acronyme pour VALue from DAta), une équipe mixte de ENS/CNRS/Inria au sein du Département d'Informatique de l'École normale supérieure**.** L'équipe 'Valda' se concentre sur les aspects théoriques et système de la gestion de données complexes, en particulier les données produites par l'activité humaine.

L'équipe compte parmi ces effectifs :

- 6 membres seniors permanents entre professeurs, chercheurs ..
- 2 professeurs invités.
- 5 doctorants
- 2 chercheurs post-doctoraux
- 1 ingénieur en informatique
- 2 stagiaires

2.4.1. Un peu d'histoire :

- Créée le 1er septembre 2016 en tant qu'équipe de la DI ENS
- Considérée comme équipe Inria à partir du 1er Décembre 2016, puis équipe-projet d'Inria à partir du 1er Janvier 2018.

2.4.2. Les axes de recherches :

1. **Les fondements de la gestion des données :**
    - Conception des langages de requête
    - Spécification des contraintes sur des instances de données
    - Complexité, algorithmes, expressivité
    - Représentation de la connaissance

2. **L'incertitude et la provenance des données**
    - Gestion des bases de données probabilistes et incomplètes.
    - Traitement des données incohérentes
    - L'annotation par provenance des données.
    - L'accent est mis sur l'efficacité de la gestion des données, et la généralité de la représentation de l'incertitude
3. **Applications :**
    - Données du web, réseaux sociaux
    - Données textuels et semi-structurés (données médicales, littérature scientifique, etc.)

2.4.3. Projets de recherche en cours :

Jetons un coup d'œil rapide sur les travaux de recherches de Valda à travers le prisme de ces projets :

1) **Projet CQFD :**

   CQFD correspond à Complex ontological Queries over Federated and heterogenous Data. Le projet a pour but d'explorer les problèmes théoriques liés à la réponse aux requêtes médiées par l'ontologie.

2) **Projet QUID :**
   Le projet traite les fondements théoriques de l'interrogation efficace de données en présence de données incomplètes ou incohérentes.

3) **Projet EQUUS :**
   Le nom du Projet correspond à Efficient Query answering Under UpdateS. La problématique à résoudre est comment maintenir des index permettant une interrogation efficace des données en présence de mise à jour de la base de données.

En plus de ces projets, deux projets plutôt applicatifs :

4) **ProvSQL :**
   Une extension de la base de données PostgreSQL permettant de calculer la provenance et les probabilités des réponses aux requêtes, de manière générique et efficace. On peut consulter le projet via ce lien :
   https://github.com/PierreSenellart/provsql

5) **Collaboration avec Neo4j** (système de gestion de base de données au code source libre basé sur les graphes).

6) **TheoremKB :**
   C'est le projet sur lequel j'ai travaillé durant mon stage. Il vise à transformer la littérature mathématique en base de connaissances sous forme de théorèmes et de preuves interconnectés. Dans la suite, on présentera cela en plus de détails (VALDA, 2020) .

2.5. Analyse de certaines dimensions culturelles et communicationnelle :

L'ENS se veut être une école engagée, engagée pour la science et sa diffusion vers la société en ouvrant ses portes au public lors de divers évènements, engagée pour l'égalité des chances en menant une analyse critique des raisons des inégalités de réussite scolaire entre les différentes classes sociales et en menant des actions pour les réduire, engagée aussi pour l'égalité homme-femme en mettant en place plusieurs dispositifs pour questionner l'existant et la liste est longue (ENS s. , www.ens.psl.eu, s.d.).

Un engagement marquant de l'école est l'engagement pour une grande liberté intellectuelle, un engagement fort qui a permis la naissance de plusieurs figures notoires pour leurs idées atypiques et controversées à leur époque comme Simone Weil et Jean-Paul Sartre. Cette liberté intellectuelle avec la pluridisciplinarité en lettres et en sciences a créé un environnement propice aux débats intellectuels et scientifiques. Ce principe est mis en avant sur le site de l'école et sur les différentes chartes publiées.

De mon expérience avec l'équipe VALDA, je pense que ce principe est bien présent au quotidien. Bien qu'on ne puisse pas parler d'idées controversées dans le domaine de l'informatique, le doctorant avec qui j'ai travaillé m'a expliqué comment il avait la liberté de choisir l'approche et les outils qui lui conviennent lors de son travail. De ce fait, on peut retrouver des chercheurs dans la même équipe qui travaillent avec toutes sortes de technologies.

Aussi, l'esprit d'initiative est encouragé et chacun peut explorer les options qui lui semblent pertinentes. Je trouve que tout le monde dans le laboratoire bénéficie d'une autonomie presque totale. Chacun s'arrange pour honorer ces responsabilités et mener à bien ces travaux de recherches, formant une sorte d'auto-régulation. Par exemple, Il n'y pas forcément des réunions régulières, ni de délais bien précis pour finir son projet.

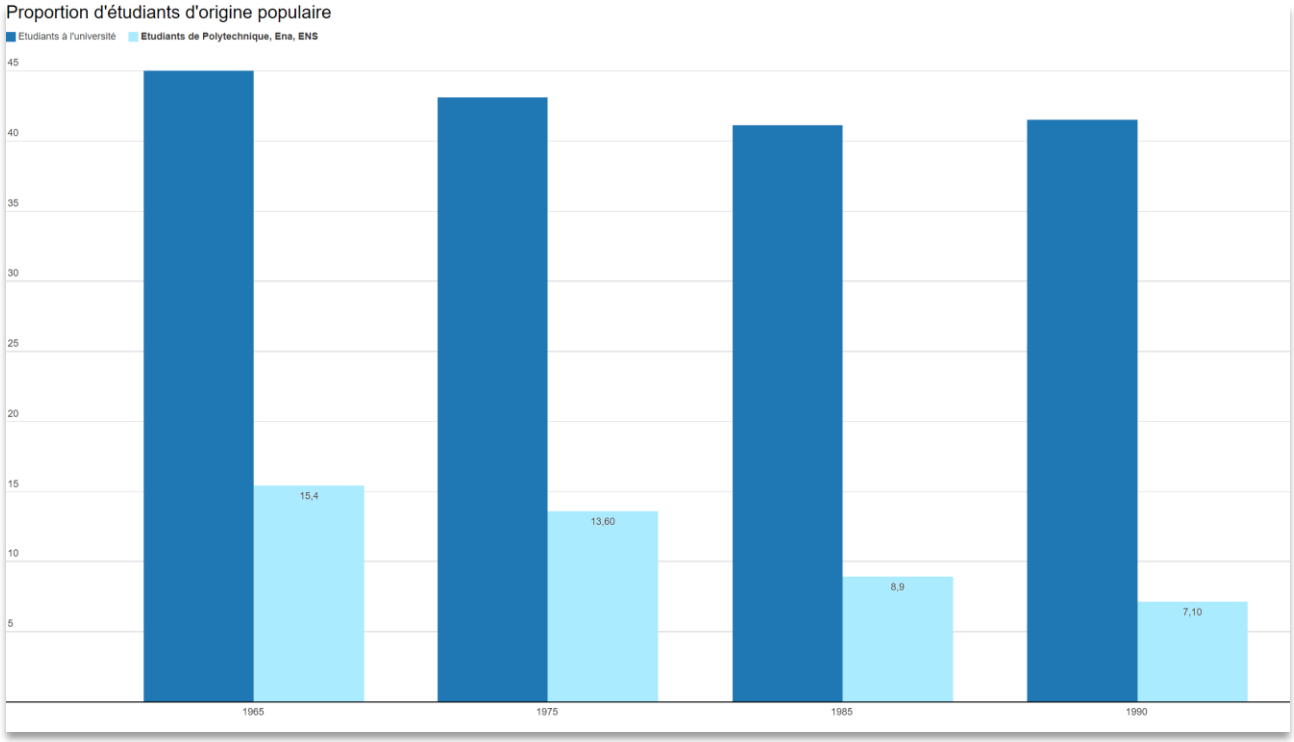
Du côté opposé, c'est vrai que le système d'évaluation des chercheurs basé sur le nombre d'articles et de citations peut significativement peser sur cette liberté individuelle des chercheurs. Il peut les contraindre à abandonner leurs travaux de recherche fondamentale et courir après ces statistiques. Mais, ce problème n'est pas propre à l'ENS.

Une autre contrainte qui n'est pas forcement liée au mode de fonctionnement du laboratoire est les délais de soumissions des travaux aux différentes revues scientifiques qui sont très strictes. Le chercheur doit faire tout pour respecter ces délais.

Un autre point important à mon avis, qui favorise cette liberté de pensée et cet échange libre avec les autres chercheurs de tout part est l'emploi de la langue anglaise dans les différents moyens de communication et même dans les évènements organisés sur le campus.

Une autre dimension culturelle de l'école importante est le principe d'égalité des chances. L'école se trouve fréquemment critiquée pour sa politique de sélection très stricte et qui, selon certains, servait avant tout à 'la reproduction des élites'.

Les données sont relativement peu abondantes. Mais, comme le montre le schéma ci-dessous tiré d'un article de LeMonde (Laudren, 2017), il y a effectivement une baisse des élèves d'origine populaire plus forte au fil du temps dans les grandes écoles qu'à l'université.



Source 1: Revue française de sociologie

Une autre étude menée par l'Institut des politiques publiques (IPP) sur la période 2006-2016 pointe dans la même direction, depuis les années 1990, pas grand-chose n'a changé. (inégalités, 2021)

Dans cette étude, les étudiants ont été rassemblés en quatre grands groupes sociaux : défavorisés (enfants d'ouvriers et de personnes sans activité), moyens (employés, agriculteurs, commerçants et artisans), favorisés (professions intermédiaires) et très favorisés (cadres supérieurs, chefs d'entreprise, professions libérales).

Origine sociale des élèves des classes préparatoires et des grandes écoles
Unité : %

	Agriculteurs, artisans, commerçants et chefs d'entreprise	Cadres supérieurs	Professions intermédiaires	Employés	Ouvriers	Autres*	Ensemble
Classes préparatoires aux grandes écoles (2019)	10,8	51,9	12,6	11,0	7,1	6,6	100
Écoles d'ingénieurs (2019)	12,1	54,4	11,9	9,1	5,4	7,2	100
Écoles normales supérieures (2019)	7,1	64,2	9,9	6,8	2,3	9,7	100
ENA (2015)	9,4	68,8	8,7	4,5	4,4	4,3	100
École polytechnique (2018)	8,7	69,8	8,8	4,1	0,4	8,2	100
Ensemble des étudiants	**10,9**	**34,4**	**14,0**	**16,8**	**11,5**	**12,5**	**100**

* Notamment : inactifs et retraités.
Lecture : 7,1 % des élèves des classes préparatoires aux grandes écoles sont enfants d'ouvriers.
Source : ministère de l'Éducation nationale, ENA, École polytechnique – © Observatoire des inégalités

D'après les auteurs de l'étude, « *Ces institutions d'élite sont restées largement fermées aux élèves des milieux sociaux défavorisés* », ils rajoutent : « *Les dispositifs d'"ouverture" qui ont été mis en place par les grandes écoles depuis le milieu des années 2000 pour diversifier leur recrutement n'ont pas atteint leurs objectifs* ».

L'ENS a mis en place plusieurs dispositifs pour participer à la réduction des inégalités de réussite scolaire, dans l'enseignement secondaire et/ou supérieur. La cellule en charge de ces questions est appelée PESU, Programmes pour l'Egalité Scolaire et Universitaire. Ce pôle gère plusieurs programmes destinés aux lycéens et aux normaliens. (ENS s. , 2017)

Mon court séjour à l'école m'a permis de confirmer ce constat. Je ne pense pas que le système de sélection est la partie défaillante du système. De ce que j'ai lu, les élèves sont recrutés sur des critères purement académiques. Et l'anonymat total est présent lors de la correction de toutes les épreuves. A l'école aussi, je pense que les programmes d'études sont offerts équitablement à tous les élèves. De ma courte expérience, je ne peux que témoigner que l'école est un milieu d'apprentissage inclusif et que tout le monde est bien accueilli. Mais, on ne peut pas tirer des conclusions d'un seul échantillon.

Une autre dimension que je voulais aborder est la communication au sein de l'équipe. Les membres de l'équipe VALDA se trouve au même étage de l'établissement, la communication verbale est généralement le moyen privilégié. Les autres collaborateurs du projet TheoremKB étant souvent disponible, il est aisé d'obtenir des réponses rapides à des questions ponctuelles ainsi que d'obtenir des entretiens voire même des réunions.

L'échange d'information se fait également par mail ou même sur Discord. La confidentialité des travaux de recherches et des systèmes d'information est primordial. Chaque nouvel arrivant aux laboratoires de l'école doit signer une charte de bon usage des systèmes d'information.

Le département d'informatique de l'ENS possède ses propres services : mail, cloud, visio-conférences … Et l'utilisation de ces ressources informatiques est soumise à une autorisation préalable. Vu les tentatives d'espionnage par les puissances étrangères, ces mesures sont totalement compréhensibles et justifiées. (ENS D. , 2016)

2.6. La question du changement au sein de l'école :

En 2019, ENS s'associe à huit autres établissements parisiens d'enseignement supérieur pour former une jeune institution : l'université PSL. Ce toit commun n'est que l'aboutissement d'une longue histoire scientifique qu'ils partagent et une volonté d'avancer ensemble en renforçant les liens scientifiques et culturels. (Wikipédia, Université Paris Sciences et Lettres, 2022)

En rétrospective, ce rassemblement semble être une évidence. Ces établissements d'excellence manquaient, tous, de la visibilité internationale. Les classements d'universités les plus connus ne prennent pas en compte dans leurs notations la taille de l'établissement. Tout simplement, ils comparent l'incomparable dans certains cas. Comme présenté dans les parties précédentes, ENS en tant que composante de PSL université occupe désormais des places de premier rang dans les différents classements. Cela implique un accroissement de l'attractivité de l'école surtout au niveau international.

Le rassemblement a aussi permis de faciliter le financement et la mise en œuvre des projets de recherches communs. Il signifie ainsi moins de barrières administratives qui empêchaient jusque-là une collaboration plus efficace et fructueuse entre les chercheurs et les enseignants des différents établissements.

Malgré ces multiples avantages, cet événement et cette nouvelle configuration représente un changement d'envergure qui touche à divers aspects de l'organisation de l'école et perturbe beaucoup de processus. Il demande une réorganisation de plusieurs services, l'adoption de nouveaux outils de travail, de nouvelles stratégies de communication et la formation d'équipes conjointes avec les autres instituts pour piloter cette évolution. C'est un travail qui a démarré dès 2015 avec une nouvelle entité nommée « Paris Sciences et Lettres - Quartier latin » qui est le résultat du regroupement de cinq grandes écoles et institutions académiques parisiennes, et qui continue jusqu'à maintenant.

On va se concentrer dans notre analyse sur un aspect bien précis de ce changement qui est le domaine technologique. En effet, la création de l'université de PSL nécessite la création et, des fois, la migration vers de nouvelles plateformes technologiques et de nouveaux outils de travail qui représenteront mieux la nouvelle identité de l'école.

Dès le départ, on peut prévoir des perturbations qui suivront l'introduction de ces changements. C'est pour cela que ce changement est lent et incrémental. La bonne gestion de la transition est primordiale pour faire en sorte que tous les collaborateurs se sentent engagés dans cette aventure.

Sur le terrain, une nouvelle interface pour le site web de l'école s'est avéré nécessaire. On peut déjà remarquer l'évolution de charte graphique des différentes pages qui référencent l'école. Les designers et les infographes sont les premiers concernés par cette évolution. Cela signifiera pour eux un grand travail de rebranding et d'adaptation des chartes graphiques existantes à la charte de la nouvelle organisation qu'est l'université de PSL.

Sous le nouveau toit, tous les services de l'ENS doivent migrer vers un nouveau nom de domaine (psl.eu). Une mission qui n'est pas du tout évidente. Les services concernés sont le service mail, le (nouveau) site web, les différents serveurs qui permettent l'accès aux ressources informatiques de l'école ... C'est une opération critique, à exécuter sans faute vu l'importance de ces services. Cette migration vers un nouveau domaine signifie, des fois, la perte d'accès aux services concernées. Les équipes en charge de cette migration veillent à ce que ces temps de perte d'accès soit la nuit.

Le personnel de l'école ainsi que les étudiants sont bien sensibilisés de l'importance de cette transition. Ils voient comment ce changement s'inscrit dans une stratégie plus globale de re-modernisation et d'accroissement d'attractivité.

En règle générale, tout le monde arrive à adopter ses nouveaux outils sans gros soucis même si cela demande au début un temps d'adaptation. Les étudiants sont invités à créer de nouveaux comptes sur les nouvelles plateformes parfois et à mémoriser leurs nouveaux identifiants.

Comme ces travaux de migration sont menés par des prestataires de services généralement. Les ingénieurs et techniciens du service informatique de l'école sont formés pour pouvoir gérer la transition et intervenir en cas de problème. Cela demande de vrais efforts et du temps. De nouvelles technologies sont à adoptées, de nouveaux processus du quotidien viennent remplacer les anciens. Et c'est là qu'une résistance au changement se manifeste chez tous les individus concernés.

Les causes de cette résistance sont multiples : pour certains individus, le changement est anxiogène dans la mesure où il est synonyme de rupture, de perte de points de repères et de futur incertain. Pour d'autres, il est synonyme de dégradation de conditions de travail et de mal fonctionnement organisationnel.

Sans une bonne compréhension de l'impact de cette évolution, ces employés vont vites se sentir désengagés. Ce n'est pas une situation idéale, surtout si vous voulez éviter de perdre des personnes au cours de ce processus. C'est pourquoi, la direction et le service RH qui gère cette transition essaye d'échanger, de dialoguer sur les préoccupations de ces collaborateurs.

Pour ce faire, l'école mène des enquêtes de satisfaction dans l'objectif d'atténuer les risques associés à cette transition et faire en sorte que leurs collaborateurs les accompagnent dans cette aventure. Les questions portent sur la compréhension des enjeux du changement, les difficultés rencontrées, les prochaines étapes du processus…

Les données collectées de ces sondages sont analysées pour parvenir à une meilleure compréhension de l'état des lieux. Ce processus est très important vu que toute décision prise dépend fortement de la connaissance qu'on a du système qu'on veut gérer.

2.7. Conclusion de la partie organisation et communication :

Dans cette première partie dédiée à l'aspect organisationnelle et communicationnelle de l'établissement que j'ai intégré durant mon stage, j'ai appris personnellement une multitude de choses. Le travail de recherche que j'ai effectué était révélateur de plusieurs aspects de l'organisation qui m'était inconnu avant. L'histoire de l'établissement est particulièrement notable avec ses origines révolutionnaires. Ensuite, de cette histoire, on peut noter le très grand nombre de reconfigurations et de restructurations qu'a connu l'école et ses différents laboratoires et facultés, difficile de se retrouver dans les dédales de ce système. Des changements qui reflètent malgré tout l'évolution de l'enseignement supérieur en France avec ses avantages et ses inconvénients.

3. PARTIE INFORMATIQUE :

3.1. Présentations des missions réalisées :

3.1.1. Contexte du projet :

L'accès à la littérature scientifique dans les domaines mathématiques (informatique théorique, mathématiques, etc.) repose surtout actuellement sur des moteurs de recherches et bases de données académiques tels que Google Scholar, Microsoft Academic, MathSciNet, DBLP, Scopus, ou Web of Science. Ces plates-formes permettent de rechercher des articles scientifiques par mots-clefs, auteurs, lieux de publication, dates et (pour certaines) de naviguer le graphe des citations d'un article à l'autre. Dans tous les cas, l'élément de base d'information est l'article scientifique lui-même (habituellement un document PDF), avec ses métadonnées, contenus textuels, et références bibliographiques. Mais l'élément de base utilisé par les scientifiques n'est pas l'article en lui-même, mais les résultats mathématiques (théorèmes, lemmes, etc.) qu'il contient : leurs énoncés, preuves et éventuellement d'autres métadonnées comme un nom ou un identifiant.

Le projet TheoremKB https://github.com/PierreSenellart/theoremkb est un projet d'envergure visant à construire automatiquement une base de connaissance des résultats mathématiques contenus dans les articles, dans lequel les résultats mathématiques sont l'objet d'intérêt, qui peuvent être explorées de nouvelles manières.

Une telle base de connaissances pourrait être très utile pour comprendre comment différents résultats dépendent les uns des autres, ou pour explorer des classes de résultats (par exemple, obtenir tous les problèmes NP-complets connus). Elle peut également aider les auteurs à comprendre les dépendances entre les résultats contenus dans leurs propres articles.

3.1.2. Contributeurs du projet :

- Mon tuteur de stage M. Pierre SENELLART, créateur du projet et premier contributeur.
- Deux anciens stagiaires.
- Et actuellement le doctorant du Département d'Informatique de l'ENS Shrey Mishra dans le cadre de sa thèse.

3.1.3. Contribution personnelle :

Mon apport personnel durant ce stage serait un module du projet TheoremKB :

Etant donné une collection d'articles scientifiques avec pour chaque article, l'ensemble des résultats (théorèmes, lemmes, corollaire, etc.) extraits, identifier s'il y a des résultats qui référencent d'autres résultats dans un autre article. Ensuite, désambiguïser quel est le résultat exact dont il est question à l'aide de techniques de traitement du langage naturel (Transormers, TF-IDF …)

Donnons quelques exemples pour clarifier cette tâche :

Dans le théorème 5.1 de l'article avec l'ID « 1710.00234», on trouve deux tags le tag [1] et [8]. Chaque tag pointe vers un article précis. Le tag [8] dans ce cas, pointe vers l'article avec ID suivant : 1003.3879. On peut récupérer cette information en se référant à la bibliographie du premier article « 1710.00234 ». Le but de notre travail est de récupérer tous ces liens et chercher à chaque fois quel résultat mathématique (théorème, lemme, etc.) est utilisé par le théorème source (dans ce cas le théorème 'source' est le 5.1 de l'article « 1710.00234 » et on cherche le résultat cité 'cible' dans l'ensemble des résultats du deuxième article « 1003.3879 »).

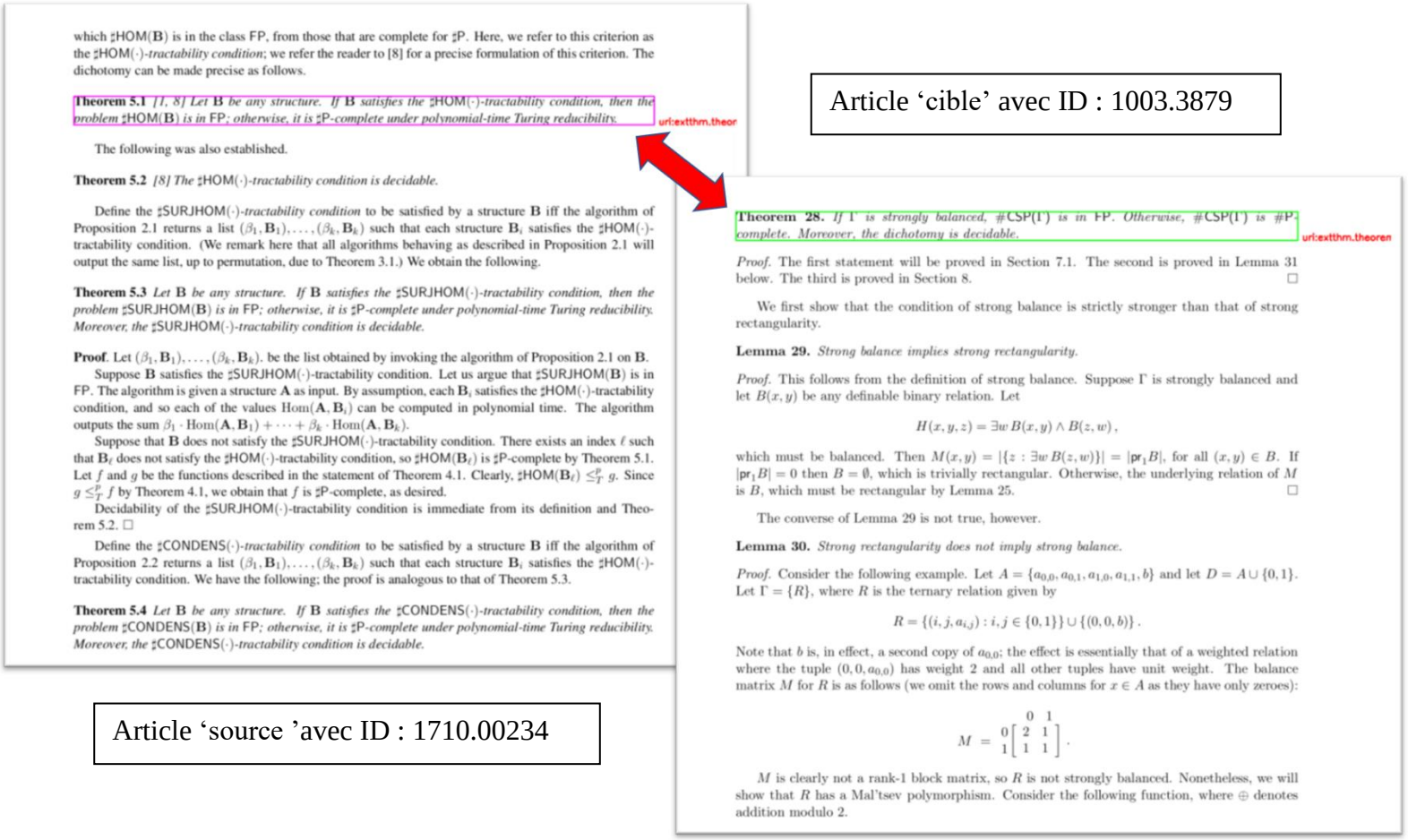


Figure 2 : Un lien entre deux résultats de deux articles 'source' et 'cible' est établi

Le résultat attendu dans cet exemple est un lien entre le théorème 5.1 de l'article « 1710.00234 » et le théorème 28 de l'article « 1003.3879 ».

Un autre exemple avec un lien facile à repérer au moins pour un être humain :

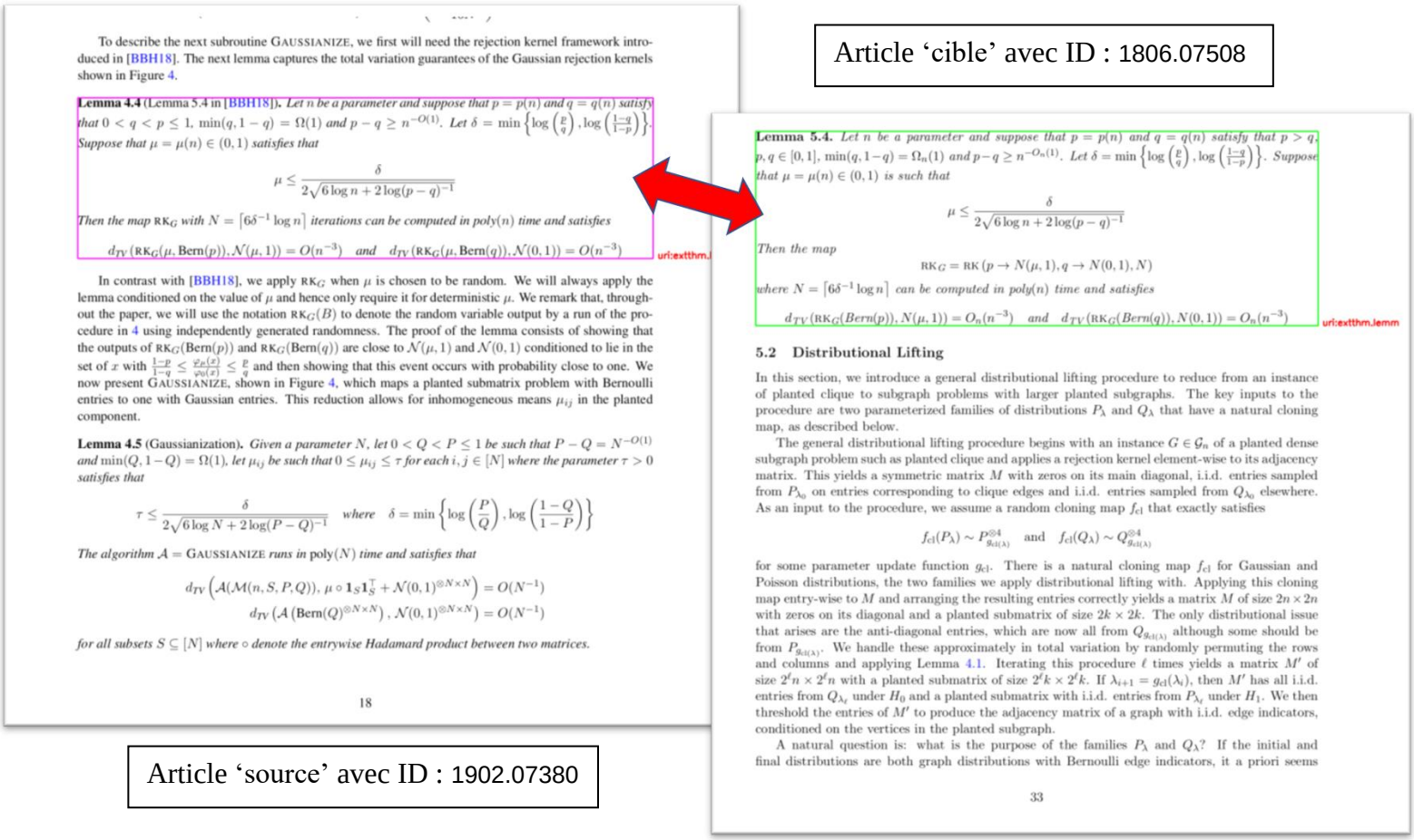


Figure 3: Un lien entre deux résultats de deux articles ‘source’ et ‘cible’ est établi

Le résultat attendu dans ce cas est un lien entre le lemme 4.4 de l’article « 1902.0738» et le théorème 5.4 de l’article « 1806.07508 ». Ce lien établi signifie que l’article « 1902.0738» se base sur le lemme 5.4 de l’article « 1806.07508 » dans ces résultats.

Le but du projet est de détecter tous ces liens et de construire ce graphe de connaissances que les chercheurs peuvent naviguer sémantiquement.

Taches nécessaires à la mise en œuvre de ce module :

- Analyser l’existant et déterminer les outils à ré-utiliser et essayer d’anticiper l’intégration du nouveau module avec le reste du système.
- Définir l’ensemble d’articles scientifiques sur lequel on effectue nos taches d’extraction et d’apprentissage.
- A partir des sorties de la partie extraction de théorèmes, filtrer et appliquer plusieurs transformations sur les données fournies pour pouvoir les exploiter par la suite.
- Extraire toutes les références qu’on peut trouver dans les articles.

- Déterminer le théorème source (le théorème qui cite un autre article).
- Déterminer le théorème cible (le théorème cité par un autre théorème d'un autre article)

Ce schéma illustre les différentes sous-tâches du module à réaliser. La partie désambiguïsation des références correspond à la détection des théorèmes sources et cibles.

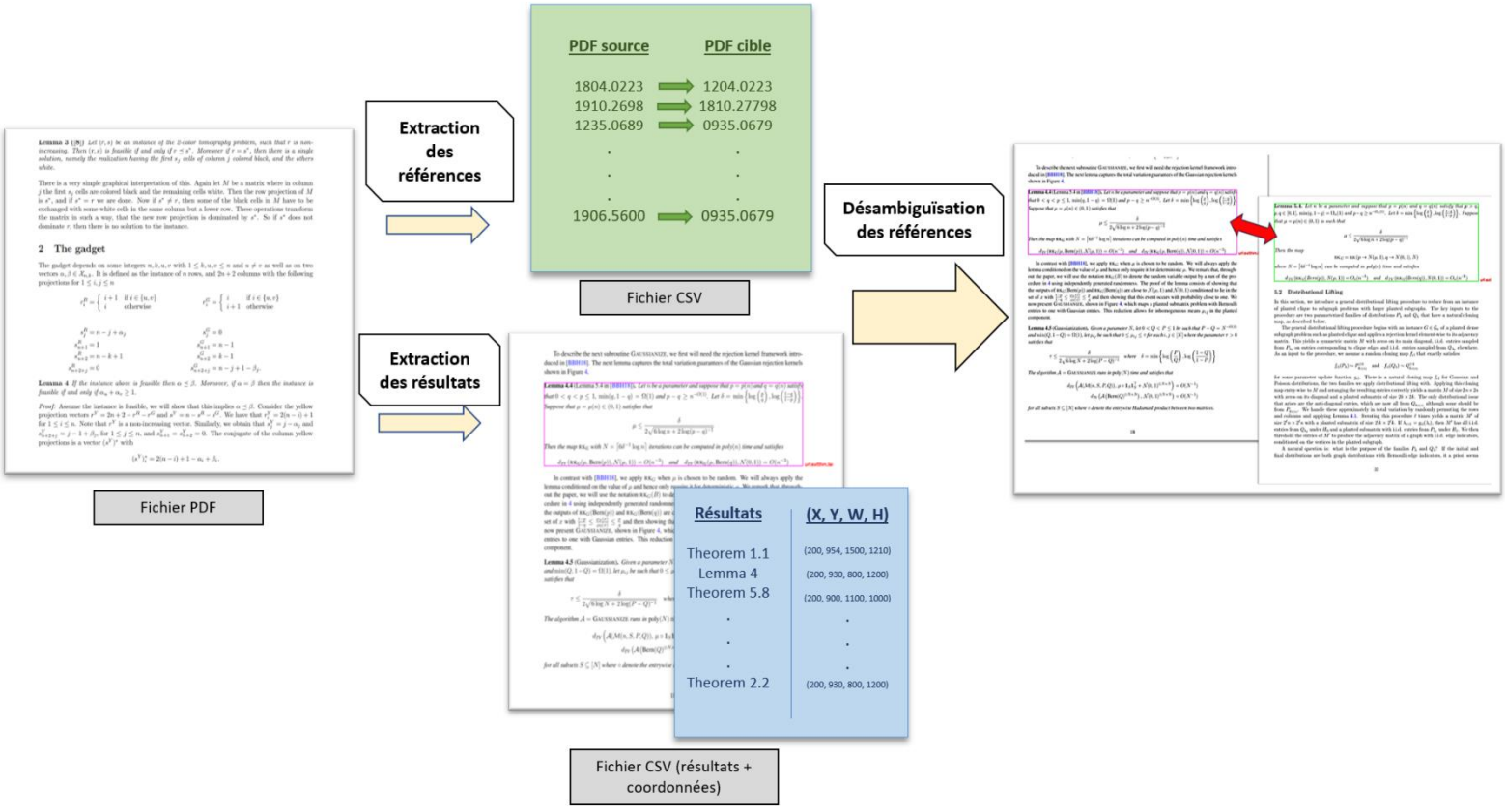


Figure 4: Schéma globale du système montrant les grandes étapes du processus d'extraction et de désambiguïsation des références

3.2. Description des missions :

Sur les trois mois de stage, j'ai dû passer par plusieurs étapes toutes autant importantes que ma tâche finale de désambiguïsation des références. Le premier mois était plutôt une phase d'exploration et de familiarisation avec le travail déjà réalisé. Ensuite, il fallait préparer un pipeline pour permettre l'apprentissage nécessaire pour la tâche.

Difficultés à surmonter :

- La littérature scientifique est très peu structurée, le format pdf des articles scientifiques est laborieusement exploitable par les machines.
- La labélisation des données pour les taches d'apprentissage est, des fois, fastidieuse et difficilement automatisable.
- Interdépendance des différents modules logiciels du projet.
- Présence de bugs dans les librairies logicielles utilisées (GROBID par ex.) dont la correction demande l'intervention des développeurs de ces librairies.

Je vais maintenant présenter dans ce qui suit le processus en détails :

3.2.1. Analyse de l'existant :

Le projet TheoremKB avait déjà démarré avant mon arrivée. Sous la supervision de mon tuteur de stage, deux stagiaires ont déjà travaillé sur ce projet lors de leurs stages de fin d'études. Chacun s'est occupé d'une partie différente :

- Théo a travaillé sur la partie extraction des références. Il a essayé à partir d'une grande base de données d'articles (presque tous les articles de l'ArXiv) de construire un graphe où chaque article représente un sommet dans ce graphe et chaque lien entre deux articles (il y a un lien quand un article cite un autre) est représenté par un arc. Il a ensuite effectué une analyse quantitative et qualitative des types de liens obtenus. Cette partie est très utile pour mon module. Le graphe des références entre articles est la première étape vers un graphe de références entre résultats mathématiques.
- Lucas, le deuxième stagiaire, s'est reposé sur les informations de style (police du texte, taille, couleur, etc.) pour extraire les résultats mathématiques des articles scientifiques. Les champs aléatoires conditionnels (conditional random fields ou CRFs) étaient la classe des modèles qu'il a décidé d'employer pour labéliser ces séquences de données (les résultats mathématiques).

Actuellement, Shrey est le doctorant qui se penche sur cette partie d'extraction des résultats. Il emploie différentes méthodes que celles qu'a utilisées Lucas pour sa partie. Il expérimente avec des techniques de l'état de l'art de l'apprentissage automatique : des techniques de Vision Ordinateur, des techniques de traitement du langage naturel et même une combinaison des deux formes d'information (image et texte) avec de l'apprentissage multimodal.

La performance de la partie d'extraction des résultats mathématiques varie d'un modèle à l'autre. Les modèles utilisés vont évoluer au fil du temps avec les nouvelles améliorations et les nouvelles

approches en test. Comme les résultats de cette partie d'extraction sont primordiaux pour la suite du processus, on considérera que la performance obtenue jusque-là est suffisante et on fera une abstraction de ces méthodes en considérant ce module comme une **boite noire** qui étant donné un PDF produira un fichier csv avec les coordonnées des résultats mathématiques (théorèmes, définition, etc.) qui se trouve dedans et plusieurs autres informations utiles : énoncé du résultat, le label du résultat (théorème, lemme, définition, remarque, preuve, etc.), la police d'écriture utilisée …

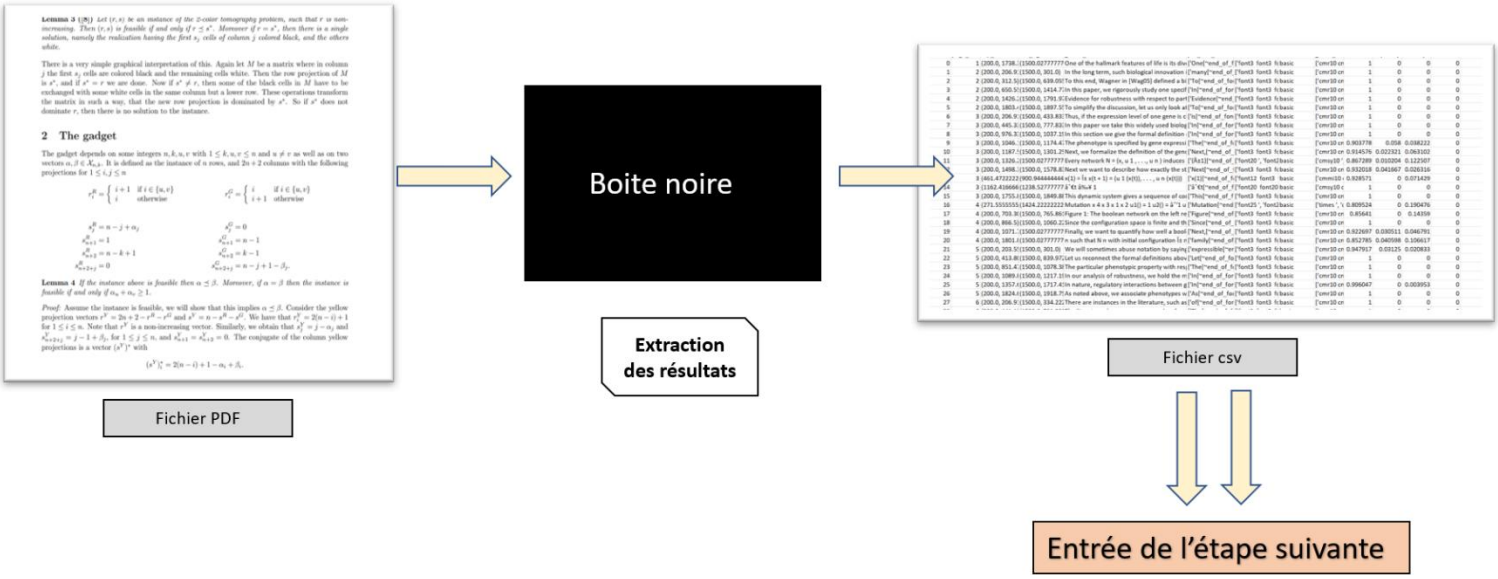


Figure 5: abstraction de la partie extraction des résultats

La librairie GROBID :

GROBID est une bibliothèque de Machine Learning pour l'extraction, l'analyse et la restructuration de documents bruts tels que les PDF en documents structurés encodés en XML/TEI, avec un accent particulier sur les publications techniques et scientifiques. Les premiers développements ont commencé en 2008 comme un hobby. En 2011, l'outil a été mis à disposition en open source. (GROBID, 2008)

On utilisera cette bibliothèque pour extraire les références qui se trouve dans la bibliographie de chaque article et déterminer la position de chaque **TAG** qu'on peut retrouver dans l'article.

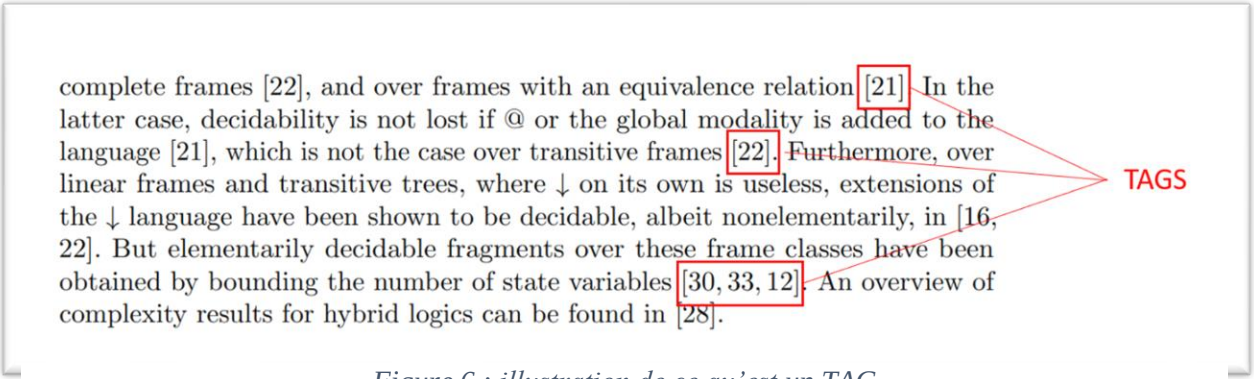


Figure 6 : illustration de ce qu'est un TAG

On peut consulter le projet sur le lien GitHub suivant : https://github.com/kermitt2/grobid.

3.2.2. Choix du dataset d'articles scientifiques :

Le dataset CS-CC :

CS-CC correspond à Computer Science - Computational Complexity , un sous-domaine de l'informatique théorique qu'on trouve dans l'archive ArXiv. Les résultats qu'on cherche (théorèmes, preuves ...) sont très communs dans ces articles. Et les chercheurs dans le domaine s'appuient fréquemment sur les résultats d'autres papiers de recherches dans leurs preuves. Ainsi, il y'a plus de chances de retrouver les liens qu'on cherche dans ce dataset. On utilisera ce sous-ensemble pour les taches d'apprentissage, de validation et de test des modèles qu'on développera.

Ces articles sont téléchargeables de l'ArXiv. Je reprends dans mon analyse le même sous-ensemble d'articles que Théo a utilisé dans son analyse : tous les PDF des articles entre 2010 et 2020 de la catégorie CS-CC. Cela revient à ~6000 articles.

Le dataset arXiv :

Le but de notre projet était de tester nos programmes sur le plus grand dataset possible d'articles et de construire le graphe des résultats le plus grand possible. ArXiv nous fournit exactement cela. ArXiv est une archive ouverte de prépublications électroniques dans différents domaines. Chaque article de cette archive a un ID unique (ArXiv ID) avec un format spécial. L'ensemble des articles de cet archive était déjà téléchargé. Il contient ~1.7 millions d'articles. Pour éviter de faire tourner nos programmes sur ce très grand nombre d'articles, l'ensemble était filtré avec une expression régulière en ignorant tous les articles qui ne contiennent pas ces mots-clés : Theorem, Lemma, Proposition. Après cette étape de filtrage, il nous restait ~500000 articles.

3.2.3. Extraction des références (liens) :

Objectif :

Etant donné un fichier PDF, on doit extraire toutes les références (citations) qui se trouvent dans la bibliographie de ce fichier. Une référence serait : un **TAG**, le titre de l'article cité (cible), et ArXiv ID de l'article cible, l'arXiv ID de l'article source (article qui cite un autre). A noter que l'arXiv ID n'est généralement pas accessible directement de la bibliographie.

La réalisation :

Cette partie était le sujet principal du stage de Théo. Et il a déjà effectué un très bon travail. Après avoir consulté son rapport de stage et après avoir visité son code source, j'ai décidé de refaire cette partie en s'inspirant de ses idées et des choix d'outils qu'il a pris. Cela était pour plusieurs raisons :

- Avoir plus de contrôle sur les sorties des algorithmes mis en place pour cette tâche.
- C'était plus simple de refaire le tout que d'explorer tout seul la base de code qui était déjà écrite.
- Manque de documentation pour certains programmes implémentés avant, ce qui a rendu leur exploitation plus difficile.

- Cette partie utilise GROBID principalement, et dans la partie que je devrais implémenter GROBID aussi est nécessaire. Je devrais donc dans tous les cas apprendre à utiliser cette librairie.
- De nouveaux algorithmes d'extraction de résultats ont été introduit au projet entre temps. Des algorithmes qui ne sont pas forcément compatibles avec les anciens programmes d'extraction de références.

Dans la nouvelle implémentation de cette partie, on utilise toujours GROBID pour le parsing de la bibliographie des références. En entrée, le service web de GROBID reçoit un fichier PDF et en sortie, il nous donne un fichier XML avec un format bien spécifique, avec des balises et des attributs bien documentés qu'on peut exploiter pour récupérer le contenu de la bibliographie : le TAG + le titre de l'article cité. Il nous manque maintenant la correspondance arXiv ID + titre de l'article.

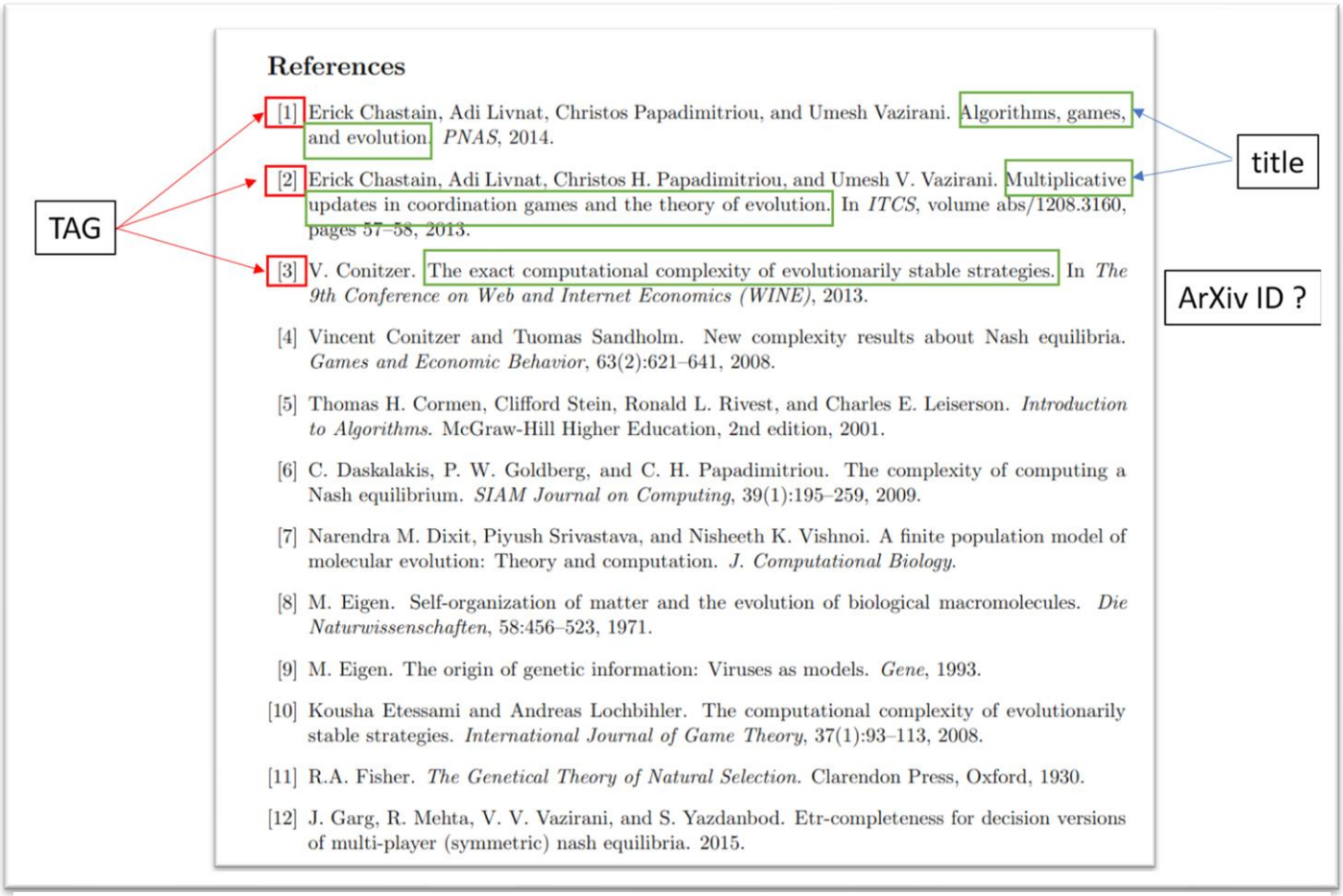


Figure 7 : illustration des informations TAG et titre qu'on peut trouver directement dans la bibliographie

Pour ce faire, on introduit un nouvel outil et un nouveau dataset :

L'outil s'appelle Semantic Scholar. C'est un moteur de recherche de publications universitaires basé sur l'intelligence artificielle. Il utilise les progrès dans le traitement du langage naturel pour fournir des résumés d'articles scientifiques et beaucoup d'autre information sémantique extraite de ces articles.

A l'aide de cet outil, les développeurs de ce projet ont construit un très grand dataset appelé : **S2ORC** - The Semantic Scholar Open Research Corpus. Un corpus destiné à être utilisé pour des taches de traitement de langage naturel (NLP) et de text mining sur les articles scientifiques. La structure des articles est préservée (les sections, l'abstraction, les références ..). Ce corpus couvre plus de 136 millions de nœuds d'articles avec plus de 12,7 millions d'articles en texte, reliés par plus de 467 millions d'arêtes de citation, en unifiant des données provenant de nombreuses sources différentes couvrant de nombreuses disciplines universitaires. (Lo, 2020)

Notre objet d'intérêt est toujours la correspondance (titre de l'article + arXiv ID) qu'on ne peut pas trouver directement dans la bibliographie avec GORBID. C'est là qu'intervient S2ORC. Pour chaque article dans ce corpus, on peut récupérer toutes les références qui se trouvent dans la bibliographie, mais cette fois, avec les ArXiv ID correspondants. Notre dataset initial de la catégorie CS-CC de l'arXiv n'étant qu'un sous-ensemble de ce corpus. La tâche est simple.

La question qui se pose : pourquoi ne pas prendre simplement ces références extraites par l'équipe de Semantic Scholar et qui sont dans le corpus de S2ORC ?

On a remarqué que parfois pour le même article, la version e l'article de S2ORC et celle de l'ArXiv n'est pas la même. Ainsi, des fois, les tags ne sont pas les mêmes, des fois, il y'a une ou deux références en plus ou en moins … On a essayé de pallier ce problème en faisant une jointure des références extraites avec GROBID et avec S2ORC et en procédant par du cas par cas, on décide quelle référence on doit garder.

<u>Résultats :</u>

A partir du dataset de ~4700 articles compilés avec réussite des 6000 articles initiaux de la catégorie CS-CC, on a :

- 4705 liens entre les articles de ce même dataset pour lesquelles on a les ArxivID.
- 3711 liens parmi ces 4705 avec un TAG non NULL (78,81%) => liens exploitables pour les prochaines étapes.

3.2.4. Détection des théorèmes 'source' :

Dans cette étape, on parcourt l'ensemble des liens entre articles trouvés dans l'étape précédente et à chaque fois, pour chaque référence qu'on peut trouver dans l'article 'source' (l'article qui cite un autre), on vérifie s'il y'a un tag qui pointe vers cette référence et qui se situe dans le bounding box d'un résultat mathématique par exemple dans un théorème, qu'on appellera dans ce cas : théorème source.

Donnons un exemple d'un théorème 'source' :

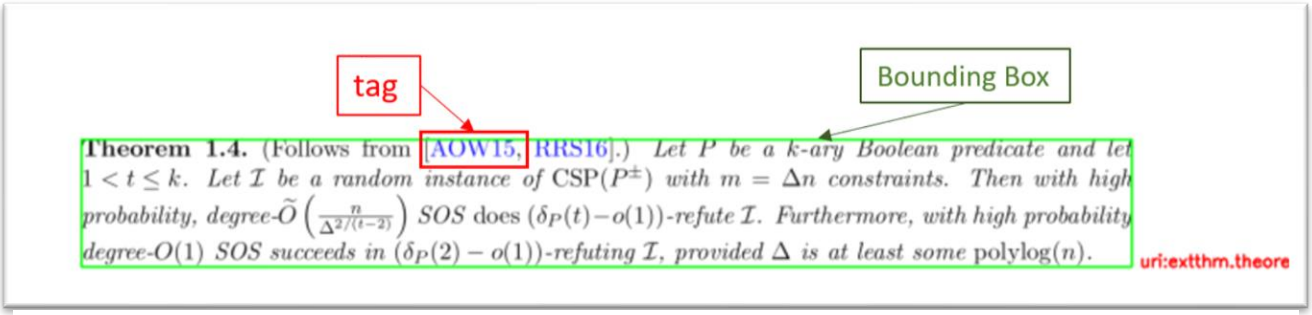


Figure 8: un exemple des théorèmes sources détectés

Commentaire **:** Ce théorème est parmi ceux détectés par le programme. Dans ce cas, on savait qu'il y'a un lien entre l'article « 1701.04521 » contenant ce théorème est l'article « 1505.044 » auquel pointe le tag **[AOW15]** (parce que AOW15 figure dans la bibliographie du premier).

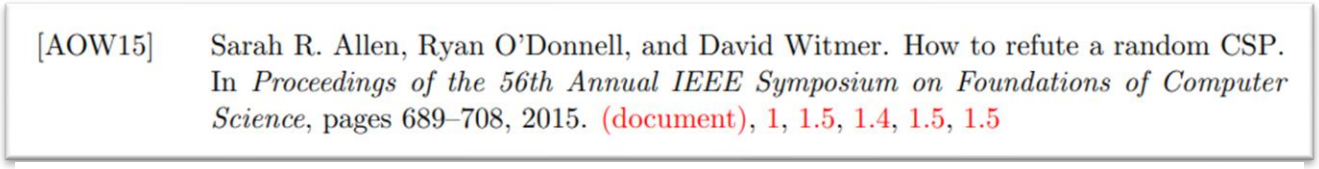

[AOW15] Sarah R. Allen, Ryan O'Donnell, and David Witmer. How to refute a random CSP. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, pages 689–708, 2015. (document), 1, 1.5, 1.4, 1.5, 1.5

Figure 9: un aperçu d'une partie de la bibliographie de "1701.04521"

A l'aide de GROBID, on peut parcourir tous les TAGS avec leur position (les coordonnées du bounding box) et tester si ce TAG tombe dans un bounding box de l'un des résultats comme dans cet exemple. Une fois le théorème source trouvé, on rajoute cette information à la base de données des liens entre articles qu'on avait.

Résultats :

Parmi les 3711 liens entre articles retrouvés précédemment, en appliquant les méthodes décrites juste avant, on a pu trouver le théorème 'source' pour 607 liens.

3.2.5. Détection des théorèmes ‘cible’ :

Objectif :

Dans cette étape, notre but est de décider exactement quel résultat mathématique dans l’article ‘cible’ est cité par le théorème ‘source’ de l’article ‘source’. Après cette étape, on aura des liens entre différents résultats mathématiques et on pourra naviguer ce graphe et découvrir les dépendances qui existent entre les articles de notre dataset.

Réalisation **:**

C’est cette partie du projet qui nécessitera des techniques de Machine Learning et d’apprentissage. En effet, on doit choisir parmi tous les résultats de l’article ‘cible’ lequel est le plus proche du résultat ‘source’, lequel il est le plus probable d’être celui référencé par le résultat ‘source’, illustrons cela en reprenant cet exemple :

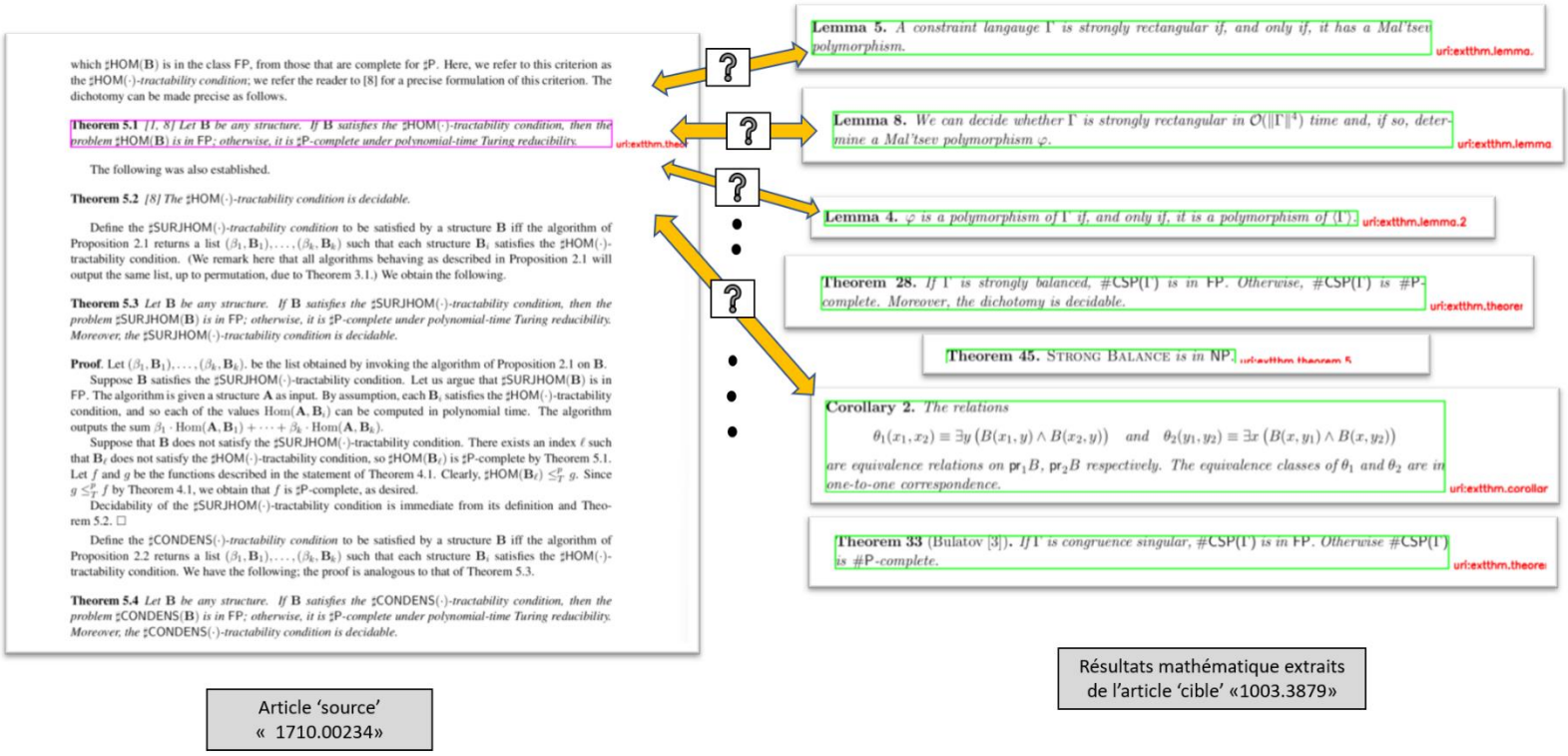


Figure 10: schéma qui illustre la partie détection des théorèmes cibles

Plusieurs techniques se présentaient devant nous pour cette tâche, explorons les plus prometteuses :

Vectorisation TF-IDF + similarité cosine :

Pour pouvoir comparer les résultats entre eux, il faut d’abord une représentation vectorielle de ces résultats. Une solution basique mais puissante est d’utiliser un *Vectorizer* **TF-IDF.**

TF-IDF correspond à *Term Frequency Inverse Document Frequency.* Etant donné un ensemble de documents (dans ce cas de résultats), on construit un vocabulaire représentatif de ces documents et pour chaque terme *t* dans le vocabulaire et chaque document *d*, on calcule un poids tfidf (t, d) (Wikipédia, TF-IDF, s.d.)

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

Ayant calculée la représentation vectorielle de tous les résultats mathématiques : d'une part du résultat 'source' et d'autre part de tous les résultats de l'article 'cible', on peut les comparer à l'aide de la distance Cosine. Le résultat le plus proche sera notre prédiction pour le théorème 'cible'.

Vectorisation DistillBERT + similarité cosine :

Une autre méthode pour calculer la représentation vectorielle des résultats est d'utiliser un modèle DistillBERT pré-entrainé sur un très large corpus de texte et pour lequel les poids et le vocabulaire (le tokenizer) ont été ajustés sur un autre corpus de texte tiré des articles scientifiques. C'est ce qu'a fait mon collègue Shrey. Ce modèle était utile pour sa tâche d'extraction de théorèmes. Et il s'avère utile pour ma tâche aussi.

On prend la sortie de l'avant dernière couche de ce modèle comme représentation vectorielle et on compare les différents vecteurs comme pour TF-IDF avec la similarité Cosine.

Évaluation de la performance :

Pour pouvoir évaluer la performance de ces différentes méthodes, il nous faut un dataset de liens pour lesquels on connait les théorèmes 'source' et 'cible'. Pour avoir ce dataset de validation, il faut labéliser les 607 liens de la dernière étape. On a mis en place un outil avec pour aider à la labélisation manuelle de ces liens vu que la labélisation automatique n'est pas une option dans notre cas. L'outil est bâti avec la librairie Python de traitement d'images : OpenCV.

L'outil te permet avec les touches de clavier de :

- Naviguer dans l'ensemble des liens avec un théorème 'source'
- Naviguer dans l'ensemble des résultats mathématiques de l'article 'source' et 'cible'.
- Créer un lien entre deux théorèmes.
- Signaler diverses erreurs qu'on peut constater
- Sauvegarder les labels créer dans un fichier CSV.

Voici un aperçu de l'outil :

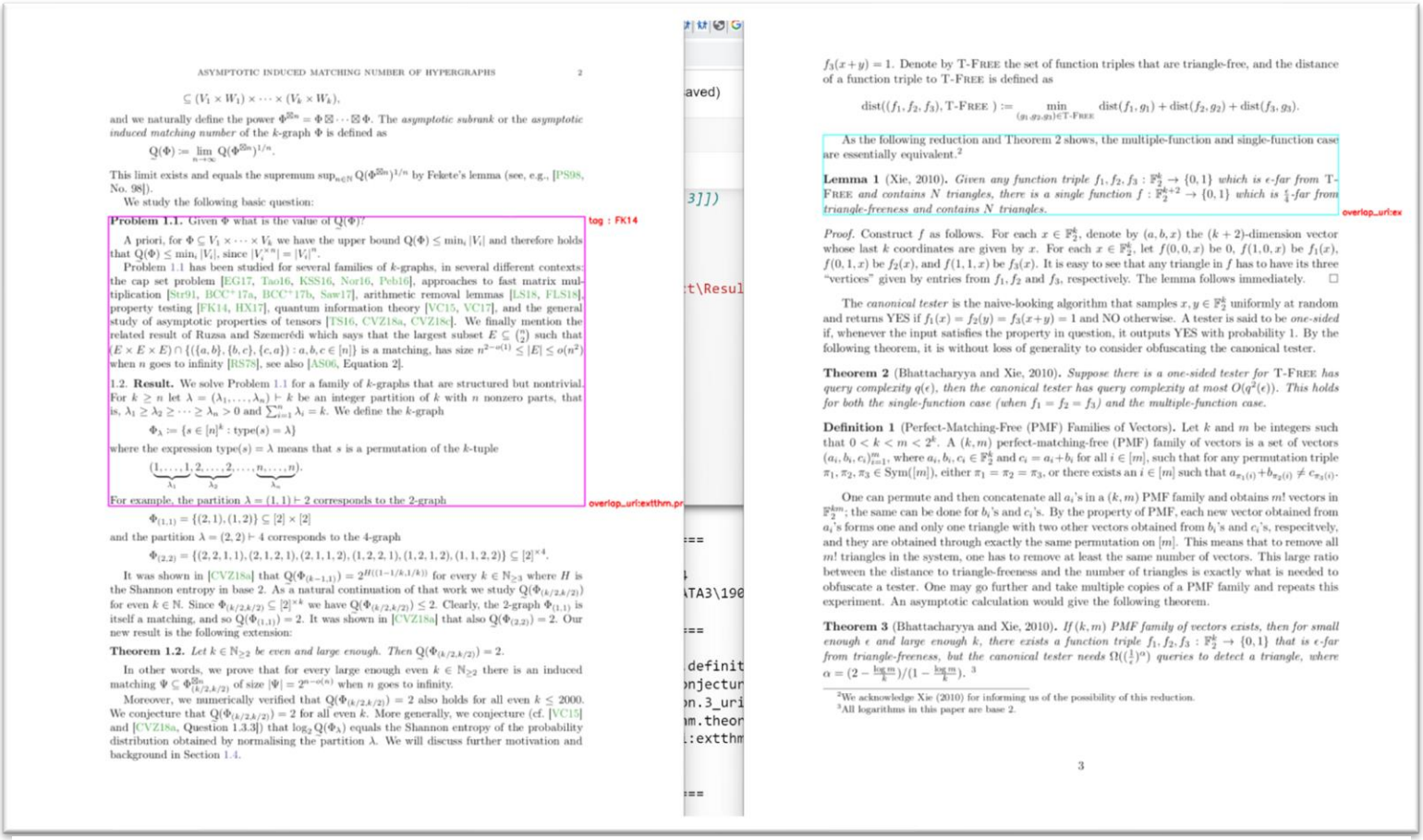


Figure 11 : outil de labélisation implémenté

Beaucoup d'erreurs apparaissent quand on essaye de labéliser les exemples manuellement : des erreurs de la partie extraction des résultats, des erreurs de GROBID et dans d'autres cas des erreurs, même si rares, des auteurs. Et parfois, même pour un être humain, la labélisation est difficile. Il faut que tu sois expert dans le domaine et que tu lises tout le papier pour pouvoir dire avec confiance quel résultat 'cible' est cité par un théorème 'source'.

Résultats :

Au moment d'écriture de ce rapport, on a les résultats suivants :

Pour la labélisation :

- 150 exemples labélisée sur 607 liens initiaux.
- 45 exemples pour lesquels on a trouvé le théorème 'cible' avec succès.
- 40 erreurs de GROBID détectées
- 27 erreurs d'extraction de résultats détectées.
- 36 fois où le théorème cible n'est pas trouvé (référence difficile à disambiguiser)
- 2 autres erreurs différentes

Pour la tâche de prédiction :

Pour TF-IDF **:**

- 17/45 résultats trouvés avec succès (37.7%)

Pour DistillBERT :

- 7/45 résultats trouvés avec succès (15%)

3.3. Conclusion et perspectives :

Durant ce stage, j'ai travaillé sur le module de désambiguïsation des références. Plusieurs sous-tâches étaient nécessaires à la mise en œuvre de ce module : d'abord, une analyse de l'existant, ensuite le choix du dataset de validation, après cela, l'extraction des liens et la détection des théorèmes cibles et sources. Il reste beaucoup de choses à faire, beaucoup d'améliorations à prévoir et d'autres expérimentations à mener. Pour donner quelques exemples, la partie labélisation prend beaucoup de temps et doit être optimisé ou externalisée (sur une plateforme d'annotation de données par Crowdsourcing) Dès qu'on a plus de données labélisées, on pourra essayer d'autres techniques d'apprentissage (pourquoi pas profond) pour entrainer des classifieurs qui décideront pour chaque paire (resultat source, resultat cible), si le lien existe bien dans la paire ou pas (classifieur binaire).

Il nous faut aussi beaucoup plus d'articles scientifiques dès la première étape d'extraction. Les 6000 articles de notre dataset ne seront pas suffisants pour accomplir ce qui était prévu. En ce moment meme, avec Shrey (le doctorant), on essaye de faire tourner les algorithmes d'extraction et détection sur le Cluster Inria pour plus de 300 000 articles. Cela prend du temps et ne se déroule pas de façon très fluide à chaque fois.

Un autre point à améliorer est la performance de GROBID dans certain cas. Pour cela, Shrey s'est déjà entretenu avec un des membres de l'équipe de développeurs, espérant que les bugs signalés seront résolus prochainement.

4. LA CONCLUSION GENERALE :

Durant mon stage de validation de M1, j'ai eu la chance d'intégrer une équipe de recherche de l'ENS. Cette aventure m'a permis de découvrir le monde de la recherche académique de plus près et de me familiariser avec le fonctionnement organisationnel d'un laboratoire de recherche. Le travail que j'ai fait pendant la rédaction de ce rapport m'a permis aussi de mieux comprendre l'histoire et les origines de cet établissement qu'est l'ENS et son positionnement à l'échelle nationale et internationale.

D'un point de vue technique, cette expérience était enrichissante. Le développement d'un module pour le projet TheoremKB m'a permis de mettre en pratique mes compétences techniques en programmation et en machine learning et d'acquérir de nouvelles compétences en matière de communication, gestion de projet et de développement. Le projet a nécessité dans un premier temps un travail de recherche et de documentation pour choisir les outils et les techniques adaptés à la tache en main. Ensuite, il fallait se servir de tous ces outils et techniques pour accomplir le travail demandé. Cette phase de développement comportait une composante de machine learning qui intervient dans la désambiguïsation des liens. Malheureusement, on n'avait pas assez d'exemples labélisés pour pouvoir entrainer des modèles de langages. On s'est contenté donc d'utiliser des modèles prés entrainés. C'est pourquoi, pour cette partie de labélisation, on a dû implémenter un outil qui permet d'accélérer ce processus.

Notre perspective pour ce projet et ce module en particulier est d'acquérir plus de données et d'utiliser des modèles scalables pour pouvoir monter en performance. La labélisation demandera l'effort de tout un groupe de personne. On pense notamment à du Crowdsourcing pour cette tâche. Il faut aussi qu'on mobilise les ressources informatiques à notre disposition notamment notre quota dans le supercalculateur Jean Zay.

Gagnant et très satisfait de l'expérience acquise durant ce stage, j'ai une idée beaucoup plus claire du monde de la recherche en France. J'ai eu la chance durant ce stage de partir au séminaire du département de l'informatique de l'ENS où il y avait les dix équipes de recherche du département, chacun dans sa spécialité. Leurs présentations des différents projets de recherche en cours m'ont donné un avant-gout de ce que serait la poursuite d'une thèse doctorale dans le domaine. Cette occasion m'a permis aussi de faire la connaissance de beaucoup personnes formidables qui ont eu la gentillesse de partager avec moi leurs expériences et leurs conseils.

J'ai hâte de découvrir le monde de l'entreprise l'année prochaine dans le cadre du master 2 Intelligence artificielle, Systèmes et Données en apprentissage. Cela me permettra de gagner une autre expertise et de prendre une décision informée quant au choix de la carrière que je veux entreprendre par la suite.

Pour conclure, ce travail aurait été incomplet voire impossible sans la collaboration précieuse de l'équipe VALDA et surtout de M. Pierre Senellart qui m'a gratifié de son hospitalité et de sa disponibilité.

Bibliographie

Archive, F. (2020, 02 10). Récupéré sur https://francearchives.fr/fr/authorityrecord/FRAN_NP_003779

ENS, D. (s.d.). Récupéré sur https://www.di.ens.fr/laboratory

ENS, D. (2016, 6 13). Récupéré sur https://www.ssi.ens.fr/charte/

ENS, s. (s.d.). Récupéré sur www.ens.psl.eu: https://www.ens.psl.eu/l-ecole-normale-superieure-psl/une-ecole-engagee

ENS, s. (2017, 04 26). Récupéré sur https://www.ens.psl.eu/l-ecole-normale-superieure/une-ecole-engagee/l-egalite-des-chances

Fuchs, A. (2022). Récupéré sur https://psl.eu/actualites/classement-qs-2023-psl-dans-le-top-30-mondial

GROBID. (2008). Récupéré sur https://github.com/kermitt2/grobid

inégalités, o. d. (2021, 04 09). Récupéré sur https://www.inegalites.fr/spip.php?page=article&id_article=1601

Laudren, J. (2017, 02 17). MARINE LE PEN : « Dans les années 1960, dans les grandes écoles, il y avait 25 % de fils d'ouvriers et d'employés. Aujourd'hui, c'est 5 % ». *Lemonde*. Récupéré sur https://www.lemonde.fr/blog/factoscope/2017/02/17/marine-le-pen-dans-les-annees-1960-dans-les-grandes-ecoles-il-y-avait-25-de-fils-douvriers-et-demployes-aujourdhui-cest-5/

l'ENS, S. d. (2022, 6 7). Récupéré sur https://www.ens.psl.eu/l-ecole-normale-superieure-psl/chiffres-cles-et-classements-internationaux/chiffres-cles#:~:text=L'ENS%20compte%201350%20enseignants,enseignants%2Dchercheurs%20(autres%20%C3%A9tablissements)

Lo, K. a. (2020, 07). S2ORC: The Semantic Scholar Open Research Corpus. doi:10.18653/v1/2020.acl-main.447

PSL, S. d. (2022). Récupéré sur https://psl.eu/actualites/classement-qs-2023-psl-dans-le-top-30-mondial

VALDA, é. (2020). *2020 activity report.* Paris: Inria Paris.

Wikipedia. (2022 , 08 03). Récupéré sur https://fr.wikipedia.org/wiki/%C3%89cole_normale_sup%C3%A9rieure_(Paris)

Wikipédia. (2022, 07). *Université Paris Sciences et Lettres*. Récupéré sur https://fr.wikipedia.org/wiki/Universit%C3%A9_Paris_Sciences_et_Lettres#Histoire

Wikipédia. (s.d.). *TF-IDF*. Récupéré sur https://fr.wikipedia.org/wiki/TF-IDF