



RAPPORT PROJET PET

Nancyclotep CHRU de Nancy Brabois



02 SEPTEMBRE 2020

INTRODUCTION	3
PRÉSENTATION DES ENTREPRISES	3
Nancyclotep	3
CHRU Nancy-Brabois.....	3
LE BESOIN CLIENT	3
L'ÉTAT DE L'ART	4
Général.....	4
Non supervisé.....	5
Supervisé.....	6
LES ÉLÉMENTS DE CONCEPTION TECHNIQUE.....	8
Les examens TEP	8
Contenu des comptes-rendus.....	8
Le jeu de données.....	9
L'anonymisation	10
Le découpage en sections	11
Le découpage en phrases.....	11
L'annotation.....	11
LES CHOIX TECHNIQUES LIÉS AU PROJET	13
Frontend	13
Backend	13
Middleware	14
Autres	15
Conclusion.....	16
LES RÉALISATIONS	16
Les modèles.....	16
Le Département de l'information Médicale	19
Notre Application	20
ORGANISATION TECHNIQUE ET ENVIRONNEMENT DE DÉVELOPPEMENT	29
Gestion de projet.....	33
Notre organisation.....	29
Les sprints	30
BILAN ET AMÉLIORATION.....	31
Problématiques rencontrées concernant l'organisation	31
Retours d'expérience sur les outils, techniques et compétences à l'œuvre.....	31
Problématiques rencontrées pendant le développement.	32
Amélioration possible ou futur	33
CONCLUSION.....	33

REMERCIEMENTS 33
SOURCES..... 34
ANNEXES 36

I. INTRODUCTION

Dans le cadre de notre formation de développeur en Intelligence Artificielle (IA), j'ai eu l'opportunité de faire mon contrat de professionnalisation au sein du Centre Hospitalier Régional Universitaire (CHRU) de Nancy Brabois et au sein de NancyClotep, groupement d'intérêt en relation directe avec le service de médecine nucléaire du CHRU. Nous sommes cinq à avoir été recrutés et deux d'entre nous ont travaillé activement sur le projet.

Le Pr Karcher a comme objectif la création d'une base de données d'images de Tomographie par Émission de Positons (TEP). L'objectif principal de cette base de données est de structurer et d'associer des informations clés aux images (pathologie, anatomie, ...). Cela permettra par la suite de composer et d'utiliser différents jeux d'images dans la recherche en vision par ordinateur. Pour ce faire, le Pr Karcher nous a demandé d'essayer d'extraire un maximum d'informations pertinentes dans les rapports médicaux associés à chaque examen et donc à chaque image.

Dans les parties qui suivent nous allons vous détailler notre parcours au sein du CHRU de Nancy Brabois et de Nancyclotep pour la réalisation de ce projet.

II. PRÉSENTATION DES ENTREPRISES

A. NANCYCLOTEP

Fondé en 2007, Nancyclotep^[1] développe et propose des solutions allant de la R&D jusqu'aux études cliniques et à la production, pour répondre aux besoins de transfert en clinique des radiopharmaceutiques.

Nancyclotep dispose d'infrastructures de pointe avec un laboratoire de radiochimie et radiopharmacie, une plateforme pour les études précliniques, un laboratoire de production industrielle, 3 PET-scans, un laboratoire d'e-learning proposant des solutions innovantes et un environnement hospitalier permettant un accès immédiat aux compétences médicales et aux patients pour les applications technologiques d'utilité clinique.

Nancyclotep est un groupement d'intérêt économique public-privé, installé sur le site du CHRU de Nancy (France) et emploie 18 personnes.

B. CHRU NANCY-BRABOIS

Construit entre 1968 et 1973, Le CHRU de Brabois^[2] est inauguré en septembre 1973. Il assure au quotidien des missions de soins de recours et de proximité, grâce à une expertise médicale pluridisciplinaire de qualité au service des patients à tous âges de la vie.

Il comporte de nombreux autres sites dispersés dans la ville de Nancy comme la maternité de Nancy, l'hôpital central, l'hôpital saint julien et le centre Émile Gallé. Mais aussi des sites à l'extérieur de Nancy comme l'hôpital de Lunéville ou celui de Pont-à-Mousson par exemple.

III. LE BESOIN CLIENT

L'objectif final du Pr. Karcher est une base de données orientée recherche pour les images TEP et le service de médecine Nucléaire. C'est un projet complexe qui s'inscrit sur le long terme et dont nous avons posé les premiers jalons.

Dans un premier temps, il a été convenu d'analyser les comptes-rendus. L'objectif était de déterminer s'il y avait des choses possibles et utiles à extraire pour les associer aux images. Le résultat de cette analyse est détaillé plus bas.

Nous avons ensuite effectué un état de l'art dans le but de trouver les méthodes appropriées pour extraire les informations pertinentes des comptes-rendus. Compte tenu de la spécificité de ceux-ci, deux approches se sont distinguées :

- La première, dite "non supervisée" consiste en l'utilisation de dictionnaire ontologique et de différentes méthodes de similarité pour standardiser les rapports et les indexer.
- La seconde, dite "supervisée" consiste en l'annotation des rapports et l'utilisation de modèles de deep learning du type Named Entity Recognition (NER) pour extraire les informations souhaitées.

Notre client a décidé que nous utiliserions l'approche supervisée. La raison de ce choix est simple : L'approche non supervisée aurait permis d'indexer les textes pour pouvoir par exemple développer un moteur de recherche. Celle supervisée permet l'extraction d'informations dans les textes ce qui correspond plus au besoin énoncé plus haut.

Nous avons aussi défini avec notre client qu'il serait intéressant d'avoir un outil de visualisation des comptes-rendus. Celui-ci affichera les informations extraites par les différents modèles et permettra aux médecins de pouvoir les corriger. Le Pr. Karcher nous a demandé que chaque compte-rendu puisse être corrigeable en moins de 5 minutes.

Enfin, nous nous sommes aperçus qu'en plus du service de Médecine nucléaire, d'autres projets similaires émergeaient au CHRU. Ils ont tous pour point commun de vouloir extraire de l'information dans du texte en langage naturel. Il y aurait donc besoin à terme d'un logiciel de gestion et d'annotation de documents correspondant à des besoins spécifiques : facilité de prise en main, déployable en interne, connecté avec la base de données du CHRU. C'est pour cela que nous avons décidé de penser notre application pour qu'à terme, elle puisse être enrichie de plusieurs fonctionnalités (composition de corpus, annotation de corpus, design de modèle...).

IV. L'ÉTAT DE L'ART

A. GENERAL

1. NATURAL LANGUAGE PROCESSING (NLP)

Le Natural Language Processing (NLP) sert à étudier la compréhension, la manipulation et la génération du langage naturel par les machines. Le langage naturel est le langage utilisé par les humains dans notre communication de tous les jours par opposition au langage formel comme les langages de programmation.

2. LES REGULAR EXPRESSION (REGEX)

Constitués d'une séquence de caractère spécifique appelée paterne de recherche. Les regex sont un bon moyen d'extraire des informations standardisées à l'intérieur d'un texte.

Avantages :

- Très rapide.
- Suffisant en présence de schémas de mots répétitifs.

Inconvénients :

- Sensible à l'orthographe.
- Pas de flexibilité sémantique (Synonyme, etc...).
- Nécessite des compétences pour rédiger les patronnes de recherches.

3. LA TOKENISATION

C'est la segmentation du texte en token. La tokenisation cherche à transformer un texte en une série d'objets individuels tels que les mots, la ponctuation, etc. C'est une étape essentielle pour tout algorithme de NLP.

Exemple :

La phrase : J'aime les gaufres au sucre

Donne : j', aime, les, gaufres , au, sucre

4. WORD EMBEDDING

Le Word Embedding^[3] est une méthode d'apprentissage d'une représentation de mots utilisée en NLP. Cette technique permet de représenter chaque mot d'un dictionnaire par un vecteur de nombres réels. Cette représentation a pour particularité que les mots apparaissant dans des contextes similaires possèdent des vecteurs correspondants qui sont relativement proches. Ce procédé est utilisé pour rendre compréhensible un texte par la plupart des algorithmes de deep learning.

D'autres techniques existent comme le très connu TF-IDF mais elles ne sont pas développées ici puisque nous ne les avons pas utilisées.

B. NON SUPERVISE

1. SIMILARITE

Mesurer la similarité^[4] entre deux phrases consiste à évaluer jusqu'à quel point le sens de ces phrases est proche en termes de proximité de surface "similarité lexicale" et de signification "similarité sémantique". Plusieurs méthodes existent :

- Similarité cosinus^[5] qui donne la similarité entre deux vecteurs de taille n. Souvent utilisée avec ceux produits par le word embedding pour calculer une valeur de similarité sémantique.
- Distance de Levenshtein^[6] qui est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.
- Distance de Jaro-Winkler^[7], semblable à celle de Levenshtein dans son objectif

Avantages :

- Pas d'entraînement
- Simple à mettre en place
- Permet de construire un système de recommandation^[8]

Inconvénients :

- Difficile à évaluer
- Pas de certitude d'obtenir des résultats probants.
- La similarité peut-être biaisée par les différents auteurs.
- Pas tout à fait en accord avec le besoin.

2. ONTOLOGIE

L'objectif premier d'une ontologie^[9] est de modéliser un ensemble de connaissances dans un domaine donné, qui peut être réel ou imaginaire. Les ontologies sont employées dans l'intelligence artificielle, le web sémantique, ... comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde. Les ontologies décrivent généralement :

- Concepts : les objets de base
- Classes : ensembles, collections de concepts
- Attributs : propriétés, fonctionnalités, caractéristiques ou paramètres que les concepts peuvent posséder et partager
- Relations : les liens que les concepts peuvent avoir entre eux
- Événements : changements subis par des attributs ou des relations

Il existe plusieurs dictionnaires ontologiques dans le domaine médical (MeSH, SNOMED, SNMIFRE ...)^[10]. MeSH est par exemple utilisé par le site PubMed^[11] (moteur de recherche d'articles scientifiques orientés médecine et biologie) pour indexer les articles.

Avantages :

- Pas de labellisation
- Possibilité de faire du concept matching, utilisable pour indexer les textes^[12,13,14]
- D'autres CHRU utilisent déjà ces techniques^[15].

Inconvénients :

- Pas de certitude d'obtenir des résultats probants.
- Difficile à évaluer
- Pas tout à fait en accord avec le besoin

C. SUPERVISE

1. TEXT CLASSIFICATION

Possibilité de labelliser au préalable les rapports selon une liste de catégories données puis d'entraîner un algorithme (Deep Learning ou Machine Learning classique) pour les classer^[16, 17, 18].

Avantage :

- Contrôle sur les variables extraites.

Inconvénients :

- Nécessité d'un grand jeu de données préalablement labellisé (1000 - 10 000).
- Pas d'application possible hors classification.

2. NAMED ENTITY EXTRACTION (NER)

La famille de modèle NER^[19,20,21] comprend plusieurs types d'architecture (Experte, Machine Learning, Deep Learning ...). Dans notre cas, nous nous intéressons aux modèles de Deep Learning.

A partir d'un corpus de phrases annotées (surligner les informations importantes), on peut entraîner ce type de modèle à reconnaître dans des textes "inconnus" les mêmes parties importantes. Pour ce faire, il s'entraîne à reconnaître les similarités sémantiques entre différents groupes de mots.



Apple **ORG** is looking at buying U.K. **GPE** startup for \$1 billion **MONEY**

Ici on voit le résultat d'un NER entraîné à extraire les organisations (Apple), les pays (U.K) et les références à la monnaie (\$1 billion)

Il existe globalement deux types d'architectures NER, les modèles classiques et les transformers^[22]. Les transformers utilisent le mécanisme d'attention^[23] et sont globalement plus performants que les modèles classiques.

Avantages :

- Flexibilité sémantique (Synonyme, etc...).
- Peu sensible à l'orthographe.

Inconvénients :

- Marge d'erreur induite dans tout type de modèle de Deep Learning (les meilleurs NER anglais a .90 de f1 score sur Spacy)
- Nécessite un corpus de phrases préalablement annotées pour pouvoir s'entraîner.

3. ENTITY LINKING

L'entity linking^[24,25,26] consiste à attribuer une identité unique aux entités (telles que des personnes, des lieux ou des entreprises célèbres) mentionnées dans le texte.

Par exemple :

- Dans la phrase « Paris est la capitale de la France »
- L'idée est de déterminer que « Paris » fait référence à la ville et non pas au prénom « Paris »

L'entity linking est différent de NER : NER identifie l'occurrence d'une entité nommée dans le texte. Entity Linking permet de lever leurs possibles ambiguïtés.

D. THE CANCER IMAGING ARCHIVE (TCIA)

TCIA^[27] est un service qui anonymise et héberge un grand nombre d'images médicales de cancer. Les images sont toutes accessibles en téléchargement gratuit. Elles sont organisées en collections. Par maladie (Cancer de la prostate, Cancer de la langue), par type d'image (MRI, CT, ...), par anatomie (Chest, Breast, Legs ...). Toutes les images sont au format DICOM.

Ce site nous a beaucoup inspiré, il correspond plutôt bien au besoin de Nancyclotep (base de données d'images). C'est notamment en le parcourant que nous avons décidé de cibler l'extraction des localisations anatomiques comme détaillé plus bas.

V. LES ÉLÉMENTS DE CONCEPTION TECHNIQUE

A. LES EXAMENS TEP

Le TEP est un examen combinant 2 images (Scintigraphie et scanner), il consiste à injecter, par voie intraveineuse, un médicament radioactif qui va se fixer sur les tissus cancéreux et inflammatoires.

Le TEP est une méthode complémentaire des techniques radiologiques (IRM, échographie ...) mais ne les remplace pas. Il a diverses indications :

- Diagnostic
- Bilan d'extension
- Evaluation de la réponse au traitement
- Délimitation des lésions pour la radiothérapie.

Pour un examen TEP, le manipulateur vous injecte une petite dose de médicament radioactif qui va se fixer sur les cellules ayant un métabolisme anormal, en particulier les cellules cancéreuses.

B. CONTENU DES COMPTES-RENDUS

La première étape du projet était d'analyser les informations contenues dans les rapports. C'était avant tout nécessaire pour savoir où concentrer nos efforts et quel type de méthode d'extraction convient.

Généralement, un rapport se découpe en six paragraphes :

- **Une description technique** de la machine qui a permis l'examen.
- **Une introduction** qui communique des informations sur le patient et l'examen.
 - Nom et Prénom du patient.
 - Date de naissance et âge du patient.

- Civilité du patient (et donc Genre si celui-ci est adulte)
- Date de l'examen
- Type d'examen
- **Un rappel du contexte de l'examen.** Il contient des informations sur les antécédents du patient, ainsi que les raisons pour lesquelles il passe l'examen. L'ensemble de ces informations est décrit en langage naturel, ce qui rend difficile leur extraction.
 - Antécédent de pathologie, de traitement et d'examen.
 - Raison de l'examen (initial, contrôle, suspicion, symptômes)
- **Le détail de la technique employée** pour l'examen.
 - Intervalle de temps entre l'injection et l'acquisition.
 - Mesure de l'activité radioactive du produit injecté (MBq).
 - Nature du produit injecté.
 - Mesure de la glycémie du patient lors de l'injection (mg/dl).
 - Mesure de la Computed Tomography Dose Index (CTDI, mGy)
 - Mesure du Produit dose longueur (DLP, mGy.cm)
- **Une liste détaillée des observations.** C'est la valeur ajoutée du rapport. La plupart des autres informations peuvent être retrouvées (parfois difficilement) dans d'autres bases de données (fichier patient, VNA ...). Le travail d'analyse de l'image par le médecin est seulement consigné dans ce rapport. On y retrouve une énumération des problèmes et/ou absence de problème identifiés à l'image. Pour la plupart, ils suivent ce schéma : Nature, Localisation et Quantification du problème. Les observations sont décrites en langage naturel, ce qui rend difficile leur extraction.
 - Ex : Hypermétabolisme focal du col utérin (SUVmax à 7,4 pour une référence hépatique à 3,2) en rapport avec l'atteinte tumorale connue à ce niveau.
- **Une conclusion.** Elle résume les observations importantes. Peut proposer un diagnostic et recommander des examens supplémentaires ou même des traitements. Si une question est clairement posée dans le contexte, elle y répond. L'ensemble des informations est décrit en langage naturel, ce qui rend difficile leur extraction.

Après échange avec l'équipe de Nancyclotep, nous avons conclu que la valeur ajoutée du rapport est située principalement dans la partie "Observations". La plupart des autres informations étant trouvable ailleurs (Xplore, Dossier patient ...), notre objectif principal est donc devenu la réalisation d'un inventaire standardisé des problèmes cités dans l'observation.

Dans l'idéal, nous devons concevoir un modèle qui analyse l'observation rédigée en langage naturel et en tirer des informations sous la forme : [Nature] [Localisation] [Quantification].

Exemple :

Hypermétabolisme focal du col utérin (SUVmax à 7,4 pour une référence hépatique à 3,2) en rapport avec l'atteinte tumorale connue à ce niveau.

[Nature] Hypermétabolisme focal [Localisation] Col utérin [Quantification] SuvMax : 7,4/3,2

Nous avons simplifié cet objectif très ambitieux en nous concentrant sur l'extraction des localisations grâce à la conception d'un modèle NER.

C. LE JEU DE DONNEES

Nous avons reçu de la part de la DSI un jeu de données sous format CSV contenant 43908 comptes-rendus d'examens TEP. Cela représente l'ensemble des examens effectués depuis janvier 2016 jusqu'à décembre 2020. En plus du texte du compte-rendu, nous avons à notre disposition le numéro d'examen, son libellé, sa date et son heure.

Le graphique en annexe 1 nous donne une idée de l'évolution du nombre d'examens sur les dernières années. La diminution relative entre 2019 et 2020 s'explique pour deux raisons (annexe 2) : Notre dernier rapport date du 14 décembre 2020, il nous manque donc la moitié du mois de décembre. La crise Covid a ralenti le service, on le voit très bien sur les mois de mars, avril et mai (grosse reprise visible en juin).

D. L'ANONYMISATION

1. OBJECTIF

L'une des problématiques rencontrées durant le projet concerne l'anonymisation des comptes-rendus. Bien que celle-ci ne soit pas nécessaire pour que nous puissions travailler (nous avons signé une charte), il nous a été demandé de développer une solution d'anonymisation. L'objectif est de pouvoir sortir les comptes-rendus du réseau CHRU afin notamment de présenter le projet à l'extérieur.

Nous nous sommes basés sur une partie des recommandations de la CNIL^[29] pour construire notre solution.

2. METHODE

Nous avons ciblé trois caractères discriminants à l'intérieur du texte. Le nom, le prénom ainsi que la date de naissance. Le texte donne d'autres informations qui permettent d'identifier la personne : certaines pathologies parfois très rares. Mais nous ne pouvions pas chercher à les éliminer sans perdre de l'information utile à notre projet.

Pour pouvoir censurer le nom, le prénom et la date de naissance, nous avons conçu une regex qui ciblait la partie introduction du compte-rendu afin de remplacer nom, prénom et date de naissance par quatre x. (annexe 4)

La deuxième étape a été de rechercher dans le reste du rapport une autre mention de ces trois valeurs. Si nous trouvions quelque chose, nous supprimons l'intégralité du rapport.

Cette méthode a pour défaut de créer des faux positifs (ex : Mr Deauville qui peut aussi correspondre à Score de Deauville) mais c'est la seule solution que nous avons trouvée et elle nous permet de conserver un grand nombre de comptes-rendus.

3. CONCLUSION

Le nombre de suppressions (annexe 5), peut se résumer en un chiffre : nous supprimons 3.29% des rapports totaux à cause d'un problème d'anonymisation. Cette valeur respecte totalement la marge de suppression autorisée par l'équipe de NancyClotep qui est de 5 à 10% maximum. Le graphique (annexe 6) représente l'évolution de cette suppression par rapport aux années. On peut voir que le nombre de comptes-rendus supprimés diminue largement après 2016. Cela nous permet de supposer que :

- Les comptes-rendus poseront moins de problèmes pour l'anonymisation.
- L'introduction a tendance à se standardiser avec le temps.

Enfin il faut préciser que nous stockons les numéros d'examen associés à chaque rapport et qu'ainsi il est possible de retrouver le patient avec ce numéro. C'est donc plus de la pseudo anonymisation que de l'anonymisation.

E. LE DECOUPAGE EN SECTIONS

Pour pouvoir simplifier l'annotation, nous avons décidé de découper les comptes-rendus selon les sections énoncées dans la partie "Contenu des comptes-rendus". Cela nous a permis par la suite de cibler notre échantillonnage sur certaines sections précises. En plus, il peut être intéressant de comparer les résultats de différents modèles sur différentes sections.

Nous avons utilisé des regex pour extraire la position de quatre phrases dont le début et la fin nous ont permis de décomposer les 5 sections d'un compte-rendu (annexe 4). Les médecins n'adoptent pas tous exactement le même standard. Les regex ne sont pas parfaites.

Un graphique (annexe 5) montre l'efficacité de nos regex. Nous pouvons voir que pour les années 2016, 2017 et 2018 le taux d'erreur est relativement faible ($\leq 1\%$) et qu'il augmente un peu en 2019 ($\leq 3.5\%$). Cela s'explique très certainement par l'agrandissement du service de médecine nucléaire et l'arrivée de nouveaux médecins. Les regex sont bien évidemment améliorables mais elles ne pourront pas couvrir l'ensemble des patrons présents dans les rapports.

Pour conclure, le taux de compte-rendu problématique (anonymisation + découpage en section) s'élève à 5.05 % du total reçu. Cela correspond à la marge fixée par NancyClotep qui est de 5 à 10 %. De plus, nous ne supprimons pas les comptes-rendus non découpés, nous les excluons simplement de notre échantillonnage. Les modèles développés s'appliqueront quand même dessus.

F. LE DECOUPAGE EN PHRASES

Le dernier point important avant d'aborder l'annotation est celui du découpage en phrases. En effet, comme l'explique la littérature et comme le montrent les projets de NER que nous avons trouvés sur internet : Il est important de découper le texte en phrases avant d'y appliquer un modèle NER (Annexe X).

En effectuant une recherche et des tests sur les différentes solutions techniques disponibles, nous en avons retenu deux :

- Python Sentence Boundary Disambiguation (PySBD) pour les modèles non transformers. (Annexe 7)
- Le modèle lui-même quand il s'agit d'un transformer (Annexe 8).

Nous n'avons pas évalué nous-même les deux solutions. Elles ont été évaluées dans d'autres contextes et l'annotation d'un total de 6000 phrases par nos soins nous a permis d'avoir une vision d'ensemble des erreurs possibles et d'en être satisfait.

G. L'ANNOTATION

Notre client ayant opté pour une solution supervisée, nous avons dû apprendre à annoter un corpus de façon cohérente et homogène. Plusieurs problématiques nous sont apparues assez rapidement.

1. QUEL LOGICIEL D'ANNOTATION UTILISER ?

Il existe plusieurs logiciels d'annotation de texte. Les deux qui reviennent le plus sont :

- Doccano^[29], une solution open source, dockeriser et qui permet d'annoter un texte de trois manières : NER, TextCat et SentenceToSentence
- Prodigy^[30], une solution payante qui permet d'annoter un texte de quatre manières : Ner, TextCat, SentenceToSentence et EntityLinking.

Nous avons décidé d'opter pour Doccano, parce qu'il est gratuit, open source et déployable facilement en interne. De plus, il permet de labelliser les textes en mode NER ce qui est notre besoin principal.

2. QU'EST-CE QU'ON ANNOTE ?

Cette partie a été supervisée par le Pr. Olivier qui nous a apporté toute son expérience pour que nous puissions labelliser les données qui seront utiles pour le personnel de santé. Il nous a aussi offert une meilleure compréhension du vocabulaire médical que nous rencontrons. Il nous a fallu plusieurs réunions pour avoir une idée de ce qu'il fallait annoter.

Le résultat de ces échanges a déjà été expliqué dans la partie besoin client. Nous avons convenu d'essayer d'extraire la localisation des problèmes listés dans la partie observation du compte-rendu. Avant cela et pour mieux appréhender la technologie, nous avons annoté un corpus avec comme entité les "Traitements".

3. COMBIEN DE PHRASES ?

En regardant dans la littérature et à travers les différents exemples que nous avons pu trouver sur internet : Nous avons conclu qu'en utilisant un modèle NER pré-entraîné (transfert learning/fine tuning), il était possible d'obtenir de bons résultats à partir de 1500 phrases contenant au moins une annotation.

- 1735 phrases pour le NER_fashion_brand de spacy^[50].
- 1977 phrases pour le NER_drug de spacy^[50].
- 949 phrases pour le NER_ingredients de spacy^[50].

Entraîner un modèle de zéro demande beaucoup plus de phrases annotées. Le nombre est difficile à estimer, il dépend de la difficulté de l'entité à extraire. On peut citer comme exemple WikiNer qui s'entraîne sur 7200 articles Wikipédia découpés en phrase.

4. COMMENT ETRE HOMOGENE ?

Quand nous avons commencé à annoter, nous avons déjà bien regardé comment étaient annotés d'autres corpus et nous avons lu plusieurs fois les guidelines associées ^[51,52,53,54,55]. Malgré toutes ces précautions, nous nous sommes vite aperçus que nous ne procédions pas tout à fait de la même manière.

Du fait de la complexité des comptes-rendus et de l'ambiguïté du langage naturel, il est important de rédiger une guideline stricte et de s'y tenir au maximum. Celle-ci doit se

construire au fur et à mesure de l'annotation et être le résultat d'échanges continus entre les annotateurs. Il est aussi important de regarder les résultats du modèle pour l'affiner.

A ce jour nous n'avons pas trouvé le temps de rédiger formellement la nôtre. Des exemples des problèmes fréquemment rencontrés sont disponibles en annexe 9.

5. CONCLUSION

Nous retenons trois points principaux de cette expérience :

- Annoter un corpus pour entraîner un NER prend du temps (100 à 200 phrases par heure soit 5 à 10 heures pour 1000 phrases, plusieurs relectures sont nécessaires).
- L'annotation d'un texte médical ne doit pas forcément être réalisée par un médecin. Nous avons annoté nous même les textes même si le Pr. Olivier nous a beaucoup aidé au démarrage.
- Vouloir annoter plusieurs entités différentes dans un même corpus complexifie grandement le travail. Il est préférable dans un premier temps d'aborder chaque entité séparément puis dans un second temps de chercher à les regrouper.

VI. LES CHOIX TECHNIQUES LIÉS AU PROJET

A. FRONTEND

1. REACT.JS

React.js^[31] est un framework javascript open source, sous licence MIT. Ces avantages et inconvénients sont :

- Evolutif : React.js permet d'obtenir un code facile à maintenir et à faire évoluer, du fait de son orientation composant. Ce qui a pour avantage une maintenance plus simple du code. Les composants React.js sont combinables, testables et réutilisables.
- Performant : L'un des avantages est qu'avec React le rechargement complet de la page ne s'opère que lorsqu'il y a des changements importants. React calcule les changements et rafraîchit seulement les parties impactées. L'avantage est que les pages ayant un grand contenu dynamique ne devront pas entièrement se recharger à chaque changement.
- Apprentissage difficile : La prise en main de React est complexe, elle nécessite des connaissances en HTML, JS, pour pouvoir l'utiliser correctement. Certains concepts (State, Hook, Rafraîchissement ...) sont difficiles à assimiler.

2. MATERIAL-UI

Matériel-UI^[32] est une bibliothèque de composants basée sur React. Ils peuvent s'utiliser indépendamment les uns des autres, et ils ne requièrent aucune librairie supplémentaire pour fonctionner.

Matériel-UI a pour avantage un développement web plus rapide et plus simple. Il permet d'avoir un site en design material, sans avoir à s'occuper du style.

B. BACKEND

1. DJANGO REST (PYTHON)

Django Rest Framework^[33] est une extension de Django faite pour développer rapidement des API REST. Elle reprend la philosophie Django, c'est-à-dire la prise en main rapide et l'efficacité. Elle nous a servi à développer toute la partie API de notre projet. C'est-à-dire toutes les connexions entre le frontend et la base de données.

Le framework rend disponible plusieurs méthodes d'authentification dont celle JWT que nous avons utilisée. Il met aussi à disposition un panel admin plutôt fourni ce qui permet de concentrer les efforts frontend sur les besoins client et pas sur le management du site. Exemple : Nous utilisons le panel admin par défaut pour gérer les utilisateurs et les rôles.

Nous avons choisi cette solution parce que nous avons déjà développé plusieurs projets avec Django et donc que la montée en compétence s'avérait plus simple.

2. POSTGRESQL

PostgreSQL^[34] est un système de gestion de base de données relationnelle. Il est libre et sous licence BSD.

Nous avons choisi ce système car il permet facilement de mixer relationnel et non relationnel. Cette fonctionnalité était nécessaire pour remplir toutes les cases de la certification. Cela s'est fait en utilisant le type champs JSON comme "sous système non relationnel" (plus de détails dans la partie architecture des données).

Il est prévu de passer la base de données en mode full relationnelle par la suite (via MongoDB ou CosmosDB). Cela ne devrait pas poser de problème puisque les interactions base de données sont entièrement encapsulées par Django.

C. MIDDLEWARE

1. FLASK (PYTHON)

Flask^[35] est un micro-framework basé sur Python. Il s'efforce donc d'être le plus simple et le plus petit possible. À l'opposé de Django, il est composé d'une poignée de modules, et n'offre pas de template par défaut, pour démarrer le développement d'une application, mais simplement une page vide.

Nous avons choisi d'utiliser flask comme serveur middleware. En effet cette partie ne devait comporter qu'un seul endpoint, il semblait donc plus approprié d'utiliser flask que django-rest.

Nous aurions pu utiliser node.js, mais Python est beaucoup plus approprié pour faire du deep learning et surtout il offre la possibilité d'utiliser le framework Spacy qui est une référence dans le domaine de la NLP.

2. SPACY (PYTHON)

Spacy^[36] est un framework python spécialisé dans le NLP. Open source et publié sous licence MIT, il est sûrement l'outil le plus abouti pour faire de la NLP. La liste des fonctionnalités qu'il propose est vaste :

- Tokenization, lemmatization, Word Embedding

- Découpage en phrases
- Analyse Syntaxique (Tagger, Dependency Parsing, Attribute Ruler ...)
- Analyse Sémantique (NER, Entity Linking, Similarity)
- Solution Experte (Pattern Matching, Rule based Matching ...)

Le tout est intégré dans un design pattern pipeline qui permet de composer facilement des modèles répondant à des besoins spécifiques. Cela nous a d'ailleurs beaucoup aidé pour structurer notre middleware.

Spacy propose aussi un système d'entraînement de modèle rapide à mettre en place. On peut entraîner des modèles depuis zéro ou utiliser le transfert learning.

D. AUTRES

1. GIT

Git^[37] est un logiciel de gestion de versioning de code. Git crée différentes versions de notre projet au fur et à mesure que nous travaillons sur les différents fichiers de notre code. Git nous a permis de pouvoir travailler en parallèle sur nos projets.

2. DOCKER

Docker^[38] est un logiciel de conteneurisation, c'est à dire qu'il a la capacité d'empaqueter une application et ses dépendances dans un conteneur isolé. Il est libre et sous licence apache.

Étant donné la variété technologique de notre projet, il était essentiel d'utiliser docker et docker-compose pour pouvoir déployer en développement et en production notre application.

3. JENKINS

Jenkins^[39] est un logiciel open source d'intégration continue développé en Java. L'intégration continue est une méthode de développement avec laquelle nous pouvons intégrer régulièrement la modification de notre code à un référentiel centralisé.

L'un des avantages est que des opérations de création et de test sont automatiquement menées. Jenkins nous permet aussi via son interface de visualiser facilement si notre site est fonctionnel ou non.

The screenshot shows the Jenkins dashboard with a search bar and user information. The main content is a table of build jobs for the 'DevIA-PET' project. The table has columns for status (S), matrix (M), project name, last success, last failure, last duration, and favorite status. There are four rows of build jobs, all with a success status. Below the table, there are options for 'Légende', 'Atom feed pour tout', 'Atom feed de tous les échecs', and 'Atom feed juste pour les dernières compilations'.

S	M	Nom du projet	Dernier succès	Dernier échec	Dernière durée	Fav
✓	⚙️	DEV-DevIA-PET-Middleware	4 j 19 h - #63	1 mo. 7 j - #19	2.1 s	🌐 ☆
✓	⚙️	DEV-DevIA-PET-Middleware-Deployment	4 j 19 h - #74	8 j 16 h - #49	1.4 s	🌐 ☆
✓	⚙️	DEV-DevIA-PET-Platform	4 j 19 h - #114	8 j 16 h - #101	3.3 s	🌐 ☆
✓	⚙️	DEV-DevIA-PET-Platform-Deployment	4 j 19 h - #91	8 j 18 h - #74	1 mn 4 s	🌐 ☆

E. CONCLUSION

Nous sommes contents des choix techniques que nous avons faits. Ils ont dans l'ensemble répondu aux attentes du projet. De plus, leur architecture solide et adaptable nous a permis de nous affranchir du choix et de la mise en place de design pattern. Nous nous sommes contentés de suivre la logique et les recommandations de développement des différents frameworks.

VII. LES RÉALISATIONS

A. LES MODELES

Avant de présenter nos réalisations, il est important de préciser que nous n'avons pas trouvé de modèle déjà entraîné qui correspondait à nos attentes. Il en existe plusieurs dans le domaine biomédical mais ils sont pour la plupart en anglais et aucun ne satisfait le besoin.

1. LES METRIQUES

Pour évaluer un NER, on utilise les mêmes métriques qu'une classification.

- Precision (P) : Capacité d'un modèle à éviter les faux positifs
- Recall (R) : Capacité d'un modèle à détecter tous les positifs
- F-score (F1) = Moyenne Harmonique de la P et du R

Ces valeurs sont calculées à partir d'autres métriques :

- Message Understanding Conference (MUC) du nom de la conférence où elles ont été mises au point.
- SemEval correspondant aussi au nom de la conférence qui a vu sa mise au point.

Plus de détails sur les calculs dans ce billet de blog^[40].

2. LE MODELE DE BASE

Avant de détailler les deux modèles que nous avons entraînés, il nous a fallu choisir un modèle de base dans notre pipeline Spacy.

Nous avons testé le "fr_dep_news_trf"^[41] qui est basé sur camenBERT mais n'ayant pas été déjà entraîné sur un NER, le transfert learning est indisponible. Les résultats pour nos petits corpus sont autour de .5 de f1-score.

Nous nous sommes donc rabattus sur le "fr_core_news_lg"^[42] qui est déjà entraîné à reconnaître les entités classiques (LOC, MISC, ORG, PER) avec comme métriques :

- Precision : .84
- Recall : .84
- F-Score : .84

A ce jour, nous n'avons pas encore testé les modèles anglais et multi-langage.

3. L'ENTRAINEMENT

Spacy propose un système facilitant l'entraînement de nouveau modèle. Ce système est basé sur la rédaction d'une config^[43] qui détermine la manière dont le framework va gérer l'entraînement. Quelques paramètres sont importants à lister :

- Optimiser : Adam
- Batch : Géré par l'optimiser
- Learning Rate : Géré par l'optimiser
- Loss function : Hard coder par spacy, optimiser pour le texte
- Poids des métriques : 100% f-score, 0% precision, 0% recall

4. TRAITEMENT : INTRODUCTION.

Dans un premier temps et à titre d'exemple pour nous et Nancyclotep, nous avons cherché à concevoir un modèle qui extrait les traitements. Nous les avons choisis parce qu'ils étaient plus simples à appréhender dans le texte, ce qui nous a apporté une relative autonomie dans l'annotation du corpus.

5. TRAITEMENT : CORPUS

Le corpus a été composé avec des phrases prises aléatoirement dans la section Introduction des comptes-rendus. Nous avons annoté 3500 phrases, nous en avons conservé 3494 et seulement 985 phrases comprennent au moins un traitement. Au total il y a 1462 traitements annotés. Il y a donc :

- 28% des phrases contiennent au moins un traitement.
- 0.43 traitement par phrase.
- 1.49 traitement par phrase contenant au moins un traitement.

Il nous faudrait annoter beaucoup plus de phrases avec traitement, au moins 1500 pour avoir un modèle fiable. Nous avons décidé de garder l'ensemble des phrases sans traitement parce que pour cette entité il est important de lui donner un large contexte afin d'éviter les faux positifs.

Nous avons découpé aléatoirement le corpus en trois parties.

- Train (2 236 phrases, 64%) qui sert de jeu d'entraînement au modèle

- Dev (629 phrases, 18%) qui permet au modèle de s'évaluer entre les epochs
- Test (629 phrases, 18%) qui permet d'évaluer le modèle à la fin de l'entraînement.

6. TRAITEMENT : RESULTATS

Nous avons obtenu les résultats suivants :

- Precision : .893
- Recall : .859
- F-score : .875

Ce sont des résultats corrects malgré les défaillances de notre corpus. Bien sûr ils sont à mettre en relation avec le faible montant d'entités, notamment la précision qui devrait baisser à mesure que l'on en rajoute. Si Nancyclotep avait pour intérêt de développer un NER traitement. Il pourrait se baser sur notre méthode d'annotation pour y parvenir.

7. TRAITEMENT : CONCLUSION

Bien que ce modèle ne corresponde pas à un besoin client, il nous a permis d'approfondir nos connaissances sur la manière d'annoter et d'entraîner un NER.

Dû à une grande variété de mots pour désigner un même groupe de traitement, le modèle doit être couplé à une standardisation (ontologie/similarité), ou à un Entity Linker, ou à une classification pour être plus exploitable.

En effet, on trouve une grande variété de mots différents qui désigne un traitement mais on peut aisément les regrouper en catégories. Voilà le début d'un travail de catégorisation des traitements :

- Trois catégories majeures :
- Thérapie : Radio, Chimio, Immuno, Curie, bio, Cortico
- Chirurgie : Exérèse, X-ctomie ...
- Produits (sans doute liés aux différentes thérapies)
- Autres :
- Prothèse, souvent liée au cœur (chirurgie ?)
- Greffe (chirurgie ?)
- Antibiotiques
- Traitement spécifique : antituberculeux

Les quinze traitements les plus fréquents sont disponibles en annexe 11

8. MEDLOC : INTRODUCTION

Le deuxième modèle que nous avons conçu est une demande de Nancyclotep. Après avoir vu ce que nous avons pu faire pour le modèle traitement, ils nous ont demandé d'essayer d'extraire les localisations anatomiques associées aux différents problèmes listés dans la partie observation du compte-rendu.

9. MEDLOC : CORPUS

Le corpus a été composé avec des phrases prises aléatoirement dans la section "Observation" des comptes-rendus. Les phrases sélectionnées devaient aussi avoir la mention

du terme "SUV". Ce choix s'est fait en anticipation du développement d'un autre NER pour extraire les différentes valeurs "SUV". Ces perspectives sont détaillées plus bas.

Nous avons annoté 2001 phrases, nous en avons conservé 1996 et 1921 phrases comprennent au moins une localisation. Au total il y a 3928 localisations annotées.

Il y a donc :

- 96% des phrases contiennent au moins une localisation.
- 1.96 localisation par phrase.
- 2.04 localisations par phrase contenant au moins une localisation.

Ce corpus est plus solide que celui utilisé pour les traitements. Il présente l'avantage d'une grande densité d'entités par phrase. L'objectif à terme est d'y inclure 1000 nouvelles phrases.

Nous avons découpé aléatoirement le corpus en trois parties.

- Train (1298 phrases, 64%) qui sert de jeu d'entraînement au modèle
- Dev (349 phrases, 18%) qui permet au modèle de s'évaluer entre les epochs.
- Test (349 phrases, 18%) qui permet d'évaluer le modèle à la fin de l'entraînement.

10. MEDLOC : RESULTATS

Nous avons obtenu les résultats suivants :

- Precision : .834
- Recall : .770
- F-score : .800

Ce sont des résultats bruts, le corpus n'a pas été encore entièrement relu et doit être soumis à quelques modifications. Il devrait être possible d'atteindre une précision, un recall et un f-score autour de .85.

Un autre point intéressant est que l'usage de ce modèle sera toujours validé par un médecin. L'objectif étant simplement de lui faciliter le listing. On pourrait imaginer accorder un poids plus important au recall durant l'entraînement. Cela permettrait de s'assurer d'extraire la quasi-totalité des entités quitte à augmenter le nombre de faux positifs. C'est assez facile à réaliser avec spacy puisqu'il suffit de modifier une ligne dans la config de l'entraînement du modèle.

11. MEDLOC : CONCLUSION

Il reste encore un peu de travail pour obtenir un modèle correspondant au besoin de Nancytotep. Comme pour les traitements, il y a une grande variété d'entités différentes. Il nous faudra les regrouper comme expliqué plus haut pour les traitements. Nous pourrions nous appuyer sur le site cancerimagenet et leurs répartitions anatomiques des images pour réaliser cette tâche. Les quinze localisations les plus fréquentes sont disponibles en annexe 12.

B. LE DEPARTEMENT DE L'INFORMATION MEDICALE

L'intérêt de notre application réside en partie dans le fait qu'il puisse être adapté aux projets d'autres services du CHRU. En fonction des besoins, leur propre modèle pourrait être implémenté dans notre application afin de leur permettre d'en visualiser le résultat.

Notre collègue Adeline, qui travaillait sur un projet pour le Département de l'information Médicale (DIM), nous a offert l'opportunité de tester cette approche. En effet, nous nous sommes aperçus que nous travaillions sur des projets similaires : le projet d'Adeline consiste à faire ressortir des mots provenant de compte-rendu médical par rapport à un dictionnaire de mots clés. Il s'avère que le DIM aussi pouvait être intéressé par un outil de visualisation des résultats obtenus par un modèle NLP. Nous avons donc décidé de travailler de concert pour ajouter une page de création de différents projets et de sélections des modèles à utiliser.

Cette opportunité nous a permis de tester l'envie initiale d'implémenter des modèles venant d'autres services et de confirmer que nous pouvions le faire relativement facilement.

C. NOTRE APPLICATION

1. LES FEATURES

Dans notre application nous appelons toutes les valeurs extraites par nos modèles de NLP. Exemple de feature extraite :

- Age (par regex)
- Genre (par regex)
- Les différentes sections (par regex)
- Les localisations anatomiques (par NER)
- Les traitements médicaux (par NER)

2. LES FONCTIONNALITES DE L'APPLICATION

A) MODULE AUTHENTIFICATION

Basé sur JSON Web Token^[44] (JWT), il est en grande partie géré par django-rest. Il offre la possibilité de se connecter via un nom d'utilisateur et un mot de passe et de rester connecté en dehors des cinq minutes d'activation par défaut du token si l'option "rester connecté" est cochée. La gestion des utilisateurs se fait directement via le panel admin de django-rest.

B) MODULE IMPORT

Ce module permet d'importer un fichier CSV depuis un export de la DSI et de stocker les documents qu'il contient dans la base de données de l'application. Il doit pouvoir prendre en compte plusieurs contenus différents (actuellement, PET et DIM). Les CSV doivent tous au moins comporter un titre et un corps de texte. Le reste des valeurs va directement dans la feature Meta. Au moment de l'import, l'utilisateur choisit le projet associé au document.

C) MODULE PROJET

L'objectif de ce module est de permettre le choix des modèles de NLP à appliquer sur les différents documents. Lorsque l'on crée ou édite un projet, on peut choisir de lui associer ou non chaque modèle présent sur le middleware. Ainsi on **filtre les modèles** qui seront appliqués à différents types de documents. Cela évite un temps de calcul trop long. On peut aussi trier les documents selon le projet. Cela permet de naviguer plus simplement et d'affiner la page "statistiques".

D) MODULE GESTION DOCUMENT

Module basique qui permet de trier et de supprimer des documents. Le tri peut s'effectuer selon plusieurs critères :

- Projet associé
- Date (ajout et modification)
- Utilisateur (ajout et modification)
- Mot dans le titre ou le corps de texte
- Valeurs des features (si on a le temps)

Le tri se répercute sur la liste des documents, la fonctionnalité document aléatoire, le module statistique. Chaque filtre appliqué est conservé et visible en haut du document. Il suffit de cliquer sur la petite croix pour le supprimer. Cela permet à l'instant T de trier les comptes-rendus avec plusieurs critères tout en pouvant revenir en arrière pour certains.

E) MODULE VISUALISATION DOCUMENT

Ce module concerne la visualisation d'un document unique. Il se divise en trois parties :

- La visualisation du corps du texte, avec surlignage d'une feature lorsque celle-ci est sélectionnée.
- La visualisation de l'ensemble des features extraites, la possibilité de les sélectionner.
- La visualisation des mots les plus récurrents (filtrer par stopwords), sous la forme d'un wordcloud.

C'est le cœur de l'application. Il permet pour le moment d'avoir une vision d'ensemble des valeurs extraites par les différents modèles. C'est aussi la partie sur laquelle nous avons le plus d'idées d'amélioration pour la suite (détails plus bas).

F) MODULE GESTION DES FEATURES

Il permet l'ajout et la suppression de feature de type utilisateur. C'est pour le moment comme cela que les médecins pourront compléter les résultats du modèle. L'utilisateur peut ajouter deux formes de features. La première en cliquant sur le bouton "+" du tableau de feature, il peut alors choisir le type et le label de la feature qu'il veut ajouter. La seconde en sélectionnant une partie du texte, il choisit alors la aussi le type et le label de la feature. Les délimitations de la partie du texte choisie seront automatiquement ajoutées.

Il est important de noter qu'un utilisateur ne peut pas et ne doit pas supprimer une feature de type modèle. Dans cet objectif, nous avons implémenté un système de validation des features modèles par l'utilisateur. Ce choix est fait pour deux raisons. La première est que le modèle peut se ré-appliquer à chaque visionnage du document. Autoriser la suppression des features modèles aurait complexifié grandement l'algorithme d'ajout des features. La deuxième est qu'un autre utilisateur pourrait vouloir consulter des features non valides. Dans cet objectif, nous avons implémenté un système de validation des features modèles par l'utilisateur.

G) MODULE STATISTIQUES

A chaque application des modèles, nous récupérons un ensemble de statistiques lié au document. Une page entière utilise ces statistiques afin d'avoir une vision d'ensemble de ce que contiennent les documents. Quatre graphiques sont disponibles :

- Trois graphiques sont à titre d'exemple de ce qui peut être fait. Ils prennent en compte seulement les documents où les valeurs libellées, genre et âge, sont extraites. (Actuellement les comptes-rendus TEP) :
 - Répartition des libellés.
 - Répartition du genre des patients
 - Répartition de l'âge en fonction du genre des patients
- Un qui répond au besoin :
 - Wordcloud de l'ensemble des documents filtrés.

H) MODULE RANDOM

Un bouton qui envoie vers un document aléatoire (soumis au filtrage). Une fonctionnalité qui ne correspond pas au besoin mais qui nous a permis de rapidement tester certains aspects de l'application.

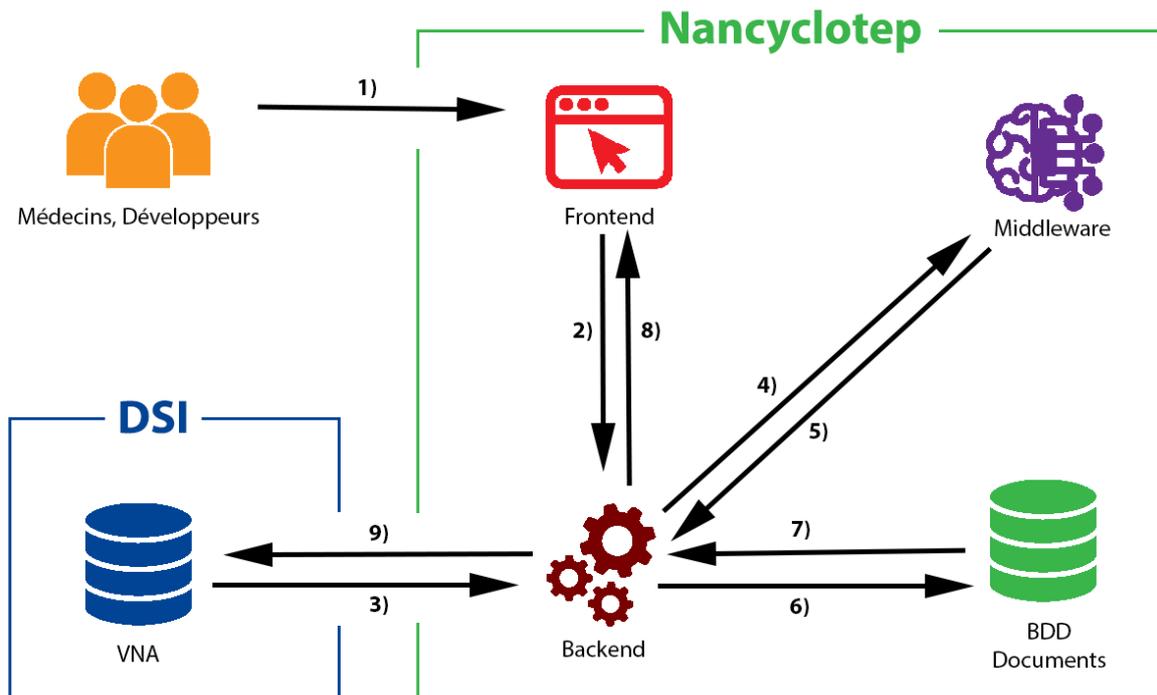
3. LE BACKLOG

En relation avec le product owner (PO) nous avons construit le backlog de notre projet. Le backlog est essentiel pour la planification de notre projet. Celui-ci nous permettra de définir nos phases de sprint et l'organisation de notre travail d'équipe. Le PO définissant nos priorités pour le futur avancement du projet.

Pour le backlog complet cf. annexe 10 et source^[45]

4. ARCHITECTURE TECHNIQUE

L'architecture de l'application a été conçue avec le soutien de notre product owner qui est aussi administrateur Système et réseaux. Elle est détaillée dans le schéma ci-dessous.



1) Les interactions entre les utilisateurs et le frontend peuvent se résumer en 3 points :

- Importer des rapports au format csv.
- Visualiser les informations extraites.
- Corriger les informations erronées.

2) Les rapports importés sont directement envoyés au serveur backend.

3) Idéalement, mise en place d'un push récurrent des rapports présents sur le VNA

4), 5) Le serveur backend envoie les rapports au serveur middleware qui applique l'ensemble des traitements et les renvoie au serveur backend.

6) Une fois les traitements effectués, le resultat est stocké dans une base de données relationnelle.

7), 8) Affichage des rapports ainsi que des valeurs extraites par le frontend.

9) Idéalement, un export (automatique ?) des informations extraites des rapports ainsi que leurs identifiants pour pouvoir les recouper avec les images TEP.

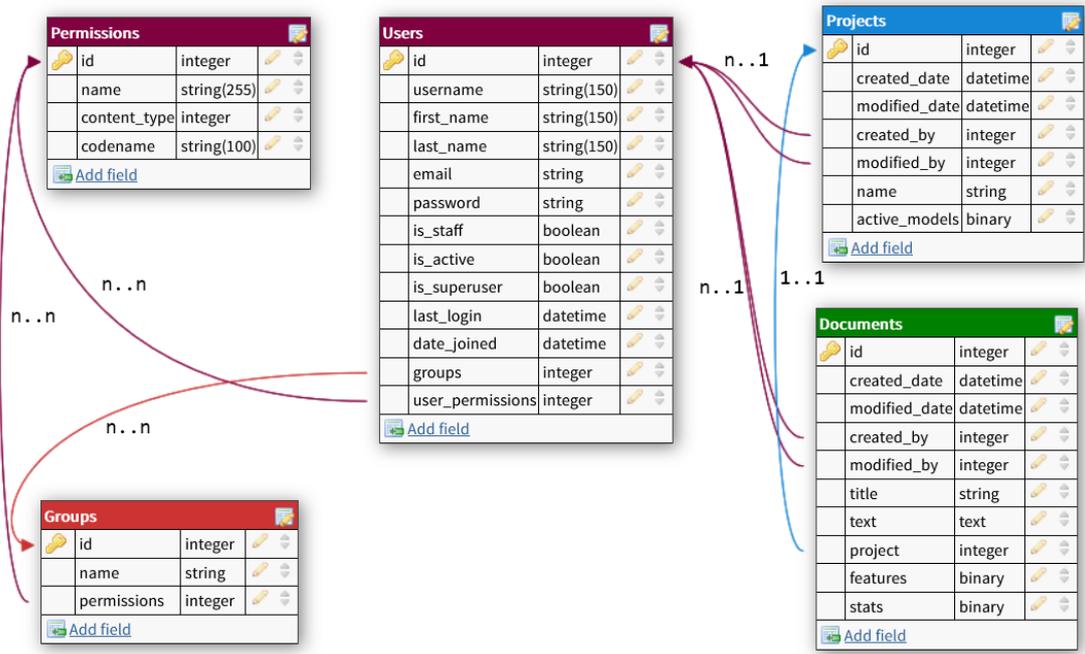
L'application est divisée en **4 container** :

- **Container BDD** : Base de donnée relationnelles. Technologie à déterminer.
- **Container Backend** : API en Python (django-rest). Sert d'intermediaire entre les trois autres container.
- **Container Middleware** : Service en python (flask). A priori, un seul endpoint qui appelle la pipeline de traitement.
- **Container Frontend** : Web App en React.

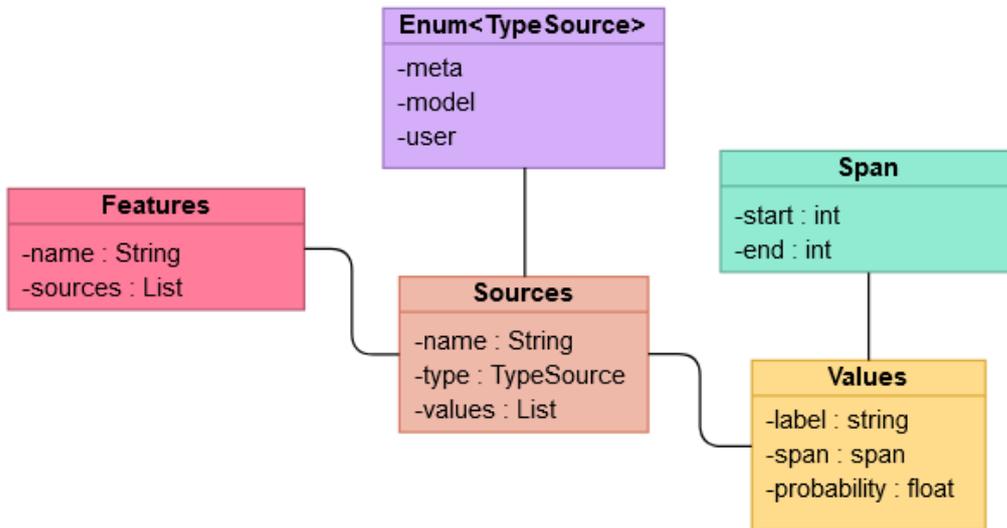
Ces 4 container seront contrôlés par un **docker-compose** pour assurer une facilité de deployment

5. ARCHITECTURE BASE DE DONNEES

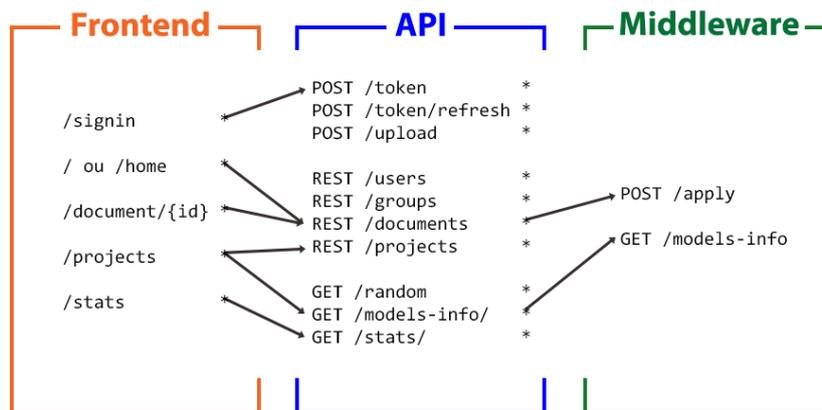
Le coeur de notre base de données suit le schéma ci-dessous :



Les features sont stockées à l'aide du champ JSON de PostgreSQL. Cela permet de les structurer comme ceci :



6. ORGANISATION DES ROUTES

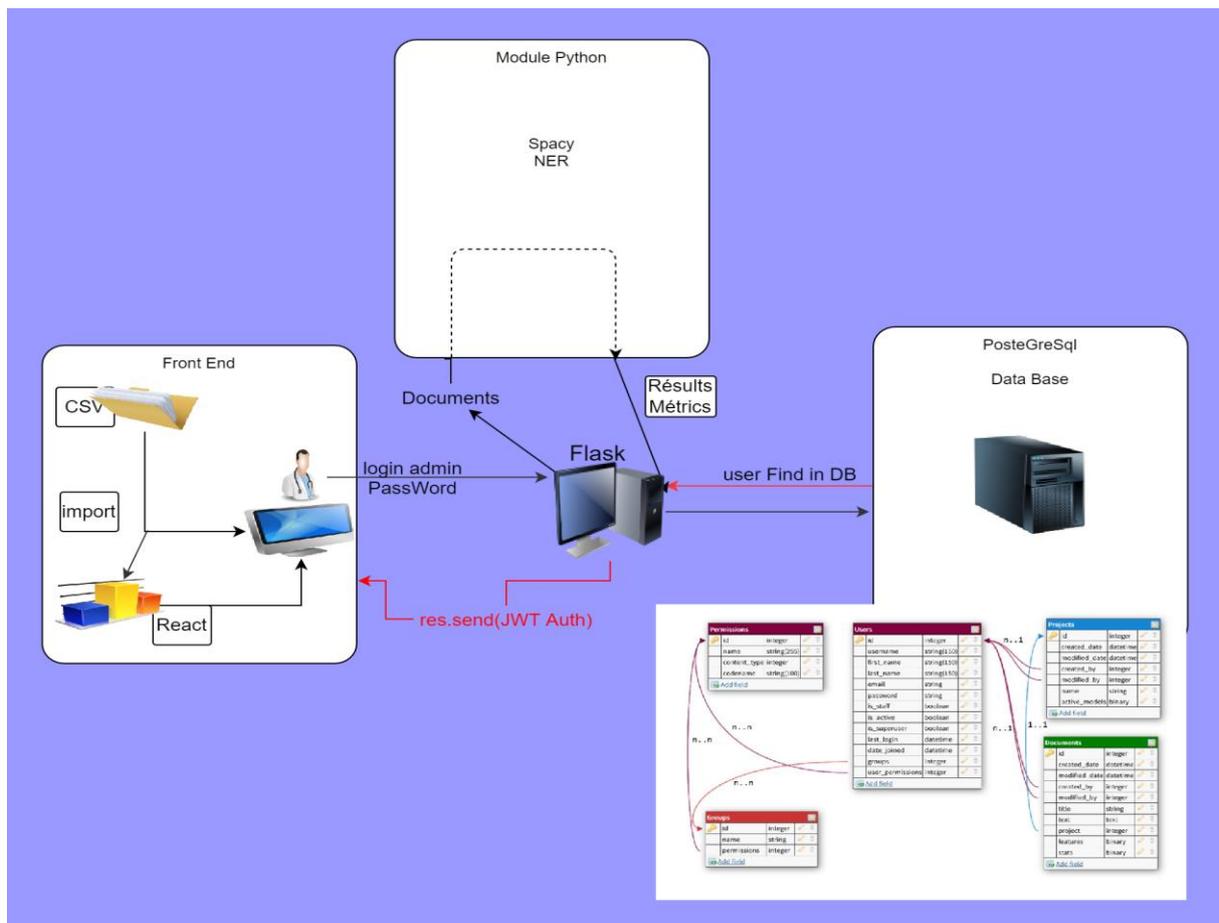


Les routes API random et upload sont accessibles depuis le header du frontend, sur toutes les pages.

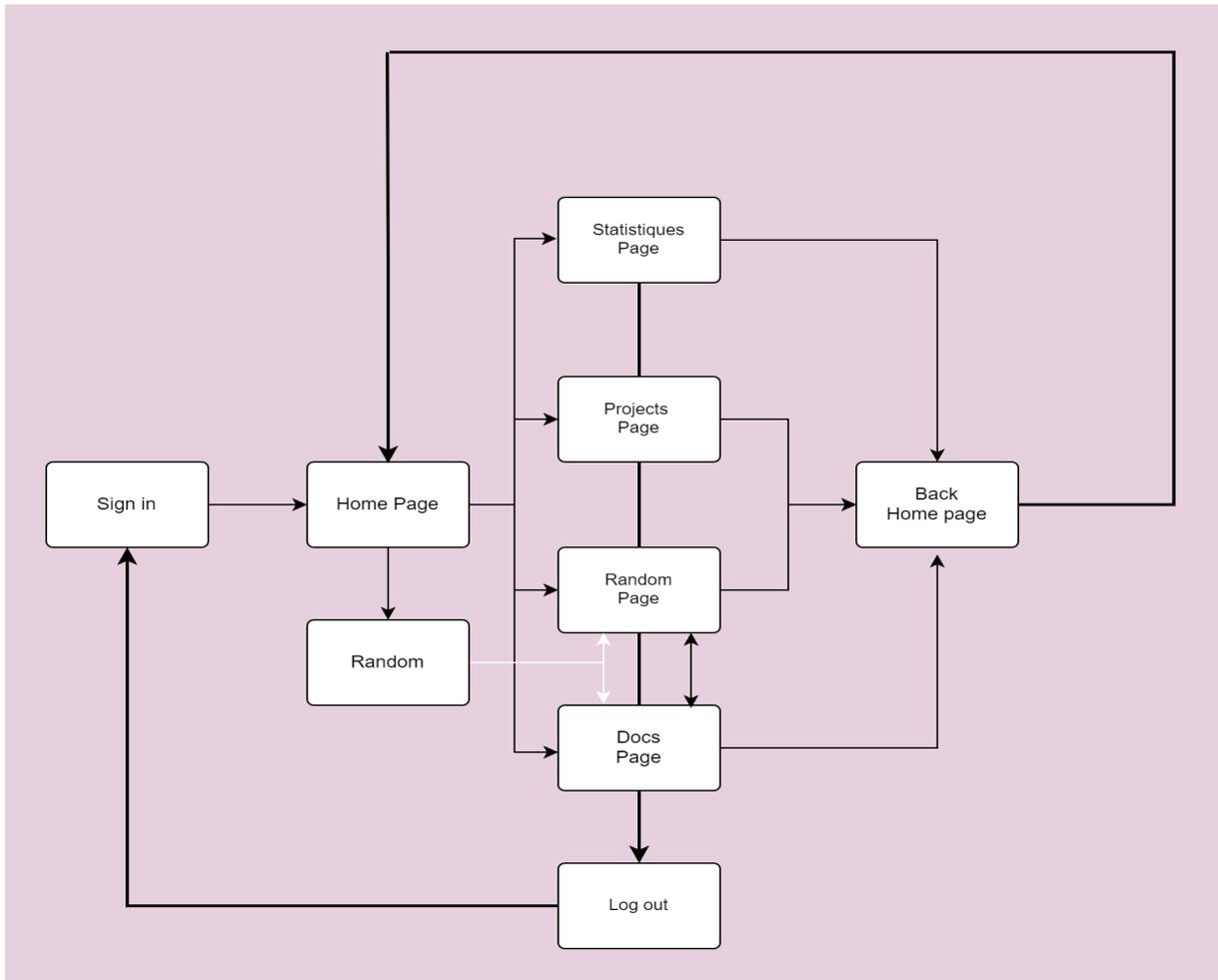
La route API token/refresh est appelée automatiquement par le Frontend quand le token d'accès n'est plus valable

Les routes API users et groups sont utilisées à travers le panel admin de django

7. DIAGRAMME FONCTIONNEL



8. DIAGRAMME DE NAVIGATION



9. LES PAGES DE L'APPLICATION

Notre application comporte actuellement 5 pages :

Une page de connexion :


 Sign in

Email Address

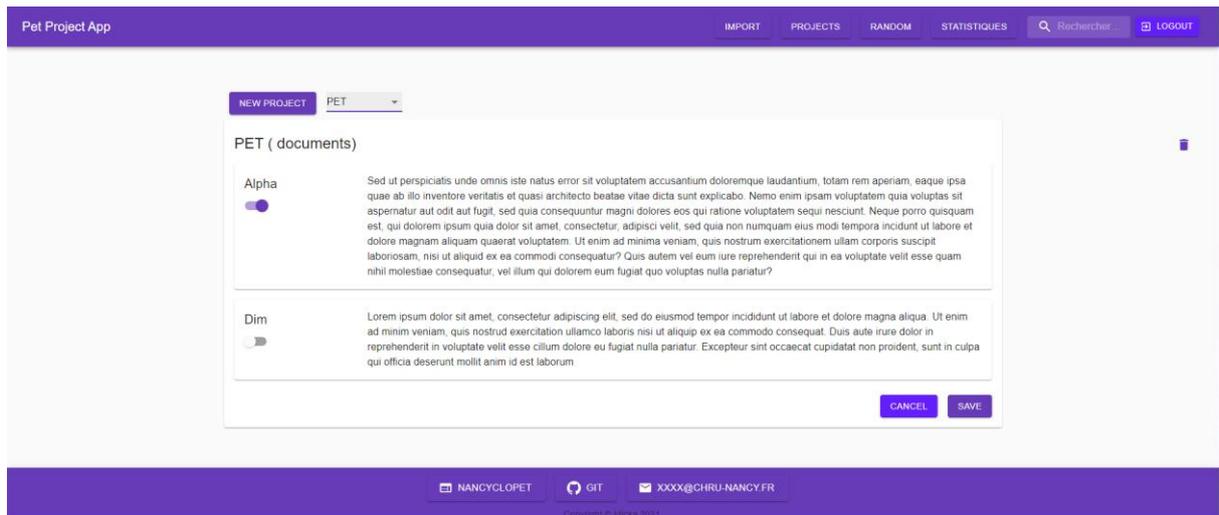
Password

Remember me

SIGN IN

Copyright © Micka 2021.

Une page de sélection des projets :



Une page liste des documents

<input type="checkbox"/>	Title	Project	Created By	Last Modification	
<input type="checkbox"/>	A11453682646	PET	admin	8/14/2021, 11:57:39 AM	
<input type="checkbox"/>	A11463275930	PET	admin	8/13/2021, 5:18:54 PM	
<input type="checkbox"/>	A11461873261	PET	admin	8/13/2021, 5:18:53 PM	
<input type="checkbox"/>	A11463278267	PET	admin	8/13/2021, 5:18:52 PM	
<input type="checkbox"/>	A11470083189	PET	admin	8/13/2021, 5:18:51 PM	
<input type="checkbox"/>	A11463279135	PET	admin	8/13/2021, 5:18:49 PM	
<input type="checkbox"/>	A11471164256	PET	admin	8/13/2021, 5:18:48 PM	
<input type="checkbox"/>	A11475105694	PET	admin	8/13/2021, 5:18:47 PM	
<input type="checkbox"/>	A11475077361	PET	admin	8/13/2021, 5:18:45 PM	

Une page d'affichage du document sélectionné

Pet Project App

IMPORT PROJECTS RANDOM STATISTIQUES Rechercher LOGOUT

Installation répertoriée sous le n° M540006 Autorisation CODEP-STR-2020-037740
VEREOS N° identification 900087 PHILIPS année 2019
SPGCodes CCAM ZZQL016

Madame et Cher Confrère,

Nous vous remercions de nous avoir adressé Mademoiselle XXXX xxxx, née le xxxx (21 ans), pour la réalisation, le 14/12/2020, d'une tomoscintigraphie par émission de positons au FDG (Morpho-TEP).

Contente dans lequel l'étude est réalisée :

Bilan de réévaluation d'une adénopathie inguinale gauche persistante chez cette patiente aux antécédents de maladie de Hodgkin, considérée en rémission complète depuis juin 2019. Précédent morphoTEP du 7 septembre 2020.

Technique de l'examen :

Les acquisitions ont été débutées 60 minutes après injection de 259 MBq de 18FDG.
La glycémie lors de l'injection était de 88 mg/dl.
CTDI : 6,0 mGy - DLP : 631,4 mGy.cm

Cet examen apporte les informations suivantes :

À l'étage cervico-thoracique :
Absence d'hypermétabolisme anormal de la filière pharyngo-laryngée, de l'aire thyroïdienne et des chaînes ganglionnaires jugulo-carotidiennes et spinales.
Absence d'hypermétabolisme suspect mammaire ou pleuro-parenchymateux.
Absence de structure ganglionnaire hypermétabolique suspecte axillaire, sus-claviculaire et médiastino-hilaire.
Régression partielle de l'hypermétabolisme thyroïdienne (SUVmax = 3,7 pour une référence hépatique max = 2,9 pour des valeurs respectives précédentes à 3,8 et 2,7).
À l'étage abdomino-pelvien :
Absence d'hypermétabolisme anormal hépatique, splénique et surrénalien.
Absence de structure ganglionnaire hypermétabolique suspecte coelo-mésentérique, lombo-aortique, iliaque et inguinale.
Absence d'hypermétabolisme anormal digestif ou intrapévien.
Stabilité morphologique du ganglion inguinale gauche, en régression métabolique complète, mesurant 12 mm versus 13 mm précédemment (SUV max = 1,2 versus 4,3).
Stabilité morphologique de ganglions mésentériques en fosse iliaque droite, non métaboliques et morphologiquement stables (SUV max = 1,5).
Régression de l'hypermétabolisme appendiculaire et cœcal.

Résultats

Search

<input type="checkbox"/>	Label	Span/Value	Probability
>	<input type="checkbox"/>	Meta	
>	<input type="checkbox"/>	Sections	
>	<input type="checkbox"/>	Sents	
>	<input type="checkbox"/>	Identity	
>	<input type="checkbox"/>	Medical	

5 rows | < > 1-5 of 5

XXXX morphologique appendiculaire complète 2020 iliaque anormal étage mm

hépatique mgy stabilité régression suv précédemment suspecte gauche

absence inguinale versus

hypermétabolisme 12 structure injection

NANCYCLOPET GIT XXXX@CHRU-NANCY.FR

Copyright © Micka 2021

Et une page de data visualisation

Pet Project App

IMPORT PROJECTS RANDOM STATISTIQUES Rechercher LOGOUT

Suv ref Hypermetabolisme Nodules Chimio Tep Multiple Adénopathie SuV Radiothérapie Localisation

Hypermutation Chirurgie

10. MONITORING DE L'APPLICATION

Un fichier log est un fichier comprenant différentes informations liées à l'utilisation d'un serveur, d'une application, d'un logiciel ou d'un système informatique. Dans la plupart des cas, il s'agit d'un fichier texte dans lequel les événements vont être classés de façon chronologique ligne par ligne.

- Backend nous nous servons des avantages que nous offre Django dans la gestion des login.
- Concernant le frontend React possède un système de login performant et efficace.
- Nous avons concentré nos efforts sur le Middleware, où sont loguées toutes les étapes de nos différents modèles en utilisant la pipeline de Spacy.

De plus, la mise en place de test unitaire sur les parties les plus sensibles de l'application sera faite pendant le dernier sprint :

- Test de la partie validation des features qui s'effectue avant leur écriture en base de données
- Test des fonctions vitales du middleware
 - Découpage en section
 - Extraction des valeurs (regex, NER)
 - Similarités

VIII. ORGANISATION TECHNIQUE ET ENVIRONNEMENT DE DÉVELOPPEMENT

A. NOTRE ORGANISATION

Après avoir discuté de notre projet avec Mr. Didier qui est notre référent en méthode agile dans la formation, nous avons décidé d'adopter la méthode ScrumBan^[46]. C'est un mixte entre la méthode Scrum et la méthode Kanban. Elle offre la possibilité d'organiser des sprints sans planifier entièrement leur contenu. C'est particulièrement efficace lorsqu'on développe une application qui nécessite une montée en compétence sur certains sujets. C'était notre cas pour React, M-UI, Docker et Jenkins.

Nous verrons un peu plus loin qu'il y a une différence entre la pratique et la théorie.

1. LA GOUVERNANCE

(Pr Pierre Olivier)

Le Pr Pierre Olivier est la gouvernance du projet, il a décidé de nommer Steeven dans le rôle du product owner. Il supervise chaque étape d'avancement du projet, pour vérifier que celui-ci corresponde toujours bien aux besoins initiaux.

2. LE PRODUCT OWNER

(Steeven Frezier) :

Il représente les clients et les utilisateurs, il est responsable de l'aspect fonctionnel du projet. Le product owner explique et priorise les fonctionnalités du projet (User Stories). Il est aussi responsable du backlog produit et de valider ce qui est réalisé.

3. LE SCRUM MASTER

(Mickael Gerard)

Il est responsable de la compréhension et de l'application de la méthode ScrumBan. Le scrum master doit jouer un rôle de facilitateur, il élimine les obstacles pouvant nuire à l'efficacité de l'équipe.

4. LE LEAD TECHNIQUE

(Pierre Epron)

Il est responsable du choix des technologies, des algorithmes et de la qualité du code produit. Il s'assure de la montée en compétence de chacun des membres de l'équipe de développement.

5. L'EQUIPE DE DEVELOPPEMENT

(Pierre Epron, Mickael Gerard, Adeline Dupres)

Elle est autonome, elle implémente les besoins exprimés par le product owner. L'équipe de développement veille aussi à la cohérence de l'architecture logicielle.

Sur les deux premiers sprints elle était composée de Mickaël Gérard et de Pierre Epron. Pour les sprints trois et quatre Adeline Dupres les a rejoints.

B. LES SPRINTS

Nous avons décidé de prévoir quatre sprints en utilisant la méthode ScrumBan. Cela nous a permis de rester vagues sur les étapes à inclure dans nos sprints. Chaque fin de sprint a donné lieu à un entretien avec le PO et nous avons pu voir le Pr. Olivier à la fin du premier, du troisième et du quatrième sprint.

Le développement des deux modèles de NLP s'est déroulé tout au long du projet, principalement en dehors des sprints. Un diagramme de gantt est disponible en annexe 13.

Nos quatre sprints :

Du 03 Au 17 Mai

- Mise en place architecture (Docker)
- Mise en place React (Frontend)
- Mise en place material UI avec les premières pages (Frontend)
- Mise en place Django Rest avec les premiers modèles base de données (API)

Du 21 mai au 4 juin

- Mise en place Flask (Middleware)
- Mise en place de la base modèle NLP (Middleware)
- Connection Api/Frontend
 - Authentification
 - Requête document

Du 16 au 30 juin

- Connection API/Middleware
 - Apply des modèles NLP
 - Collecte des statistiques
- Ajout du wordcloud document (Frontend)
- Ajout du dashboard (Frontend)
- Ajout de la fonctionnalité projet
- Implémentation du projet DIM
 - Ajout du modèle
 - Adaptation bdd

Du 16 au 21 août

- Correction des bugs relevés par les utilisateurs
- Finir fonctionnalité filtrage de documents
- Mise en place de test unitaire sur certaines parties sensibles

- Conception de script SQL pour initialiser la base de données avec des utilisateurs et des projets.
- Finir de documenter le code

IX. BILAN ET AMÉLIORATION

A. PROBLEMATIQUES RENCONTREES CONCERNANT L'ORGANISATION

1. ORGANISATION

Quand nous avons commencé notre projet nous n'avions pas clairement défini le rôle de chacun. Nous ne restions pas concentrés sur nos tâches respectives. Nous faisons des daily meetings trop longs, ils apportent souvent plus de questions que de réponses.

Après une rétrospective, nous avons donc identifié les différents problèmes et apporté des solutions :

- Échanges plus fréquents avec les clients
- Toujours planifier même dans le flou
- Mise en place d'un Trello
- Mettre les personnes aux bons postes en fonction de leur compétence.
- Daily plus courte

2. VULGARISATION

Nous nous sommes aussi aperçus que nous n'avions pas assez approfondi le besoin client. La raison principale est la difficulté à se comprendre entre le personnel médical et nous. Dès le départ nous avons eu du mal à vulgariser les principes de base de la NLP (word embedding, regex, réseaux de neurone). Et nous avons aussi rencontré des difficultés à comprendre certains aspects spécifiques du milieu médical.

L'une des pistes possibles pour améliorer la vulgarisation entre les développeurs et le personnel soignant serait d'avoir un document type "L'ia pour les nuls" contenant des définitions accessibles et structurées aidant à appréhender le monde merveilleux de l'intelligence artificielle.

B. RETOURS D'EXPERIENCE SUR LES OUTILS, TECHNIQUES ET COMPETENCES A L'ŒUVRE

1. TRELLO

Après le premier sprint, nous avons décidé de créer un trello^[47] pour l'organisation de notre équipe, cet outil nous a permis de séparer les tâches, de les attribuer aux différentes personnes en fonction de leur compétence. A mettre des deadlines sur certaines tâches à faire et d'organiser nos différents sprints pour l'ajout de nouvelles fonctionnalités venant de notre backlog.

2. EXPERIENCE UTILISATEUR

Nous avons conçu cette application de A-Z : que ce soit au niveau du code, du design ou de l'expérience utilisateurs (UX). Nous avons pu constater les limites de multiplier les casquettes. Un spécialiste dans le domaine de l'UX nous aurait grandement aidé pour rendre notre application plus intuitive et répondre aux défis que pose le besoin du client.

3. BALSAMIQ

Balsamiq^[48] est un logiciel de conception de wireframes qui permet de créer des maquettes et des prototypes interactifs. L'outil est collaboratif et permet à tous ceux qui travaillent sur le projet de suivre l'avancée des wireframes. (Cf. images annexe 14). Nous nous sommes servis de balsamiq pour avoir une idée générale de la représentation de notre application, pour différentes pages.

4. ANNOTATION ET NER

Contrairement aux modèles de machine learning classiques où beaucoup d'importance est mise dans le choix du modèle et de ses hyperparamètres, les modèles de type NER ont comme principale difficulté la manière d'annoter le corpus. Nous en avons déjà beaucoup parlé dans la partie dédiée plus haut mais il nous semblait important de revenir dessus car bien annoter son corpus est crucial. On peut résumer ce que l'on a appris sur le sujet en trois points :

- Échanger régulièrement avec des spécialistes métiers tout au long du processus.
- Prendre le temps de tester au fur et mesure de l'annotation ce que donne le modèle.
- Un NER seul ne présente que peu d'intérêt, il faudra très certainement le post process (entity linking, similarité, ontologies)

C. PROBLEMATIQUES RENCONTREES PENDANT LE DEVELOPPEMENT.

1. L'INTEGRATION CONTINUE

L'une des demandes de notre client NancyClotep était que nous travaillions en faisant de l'intégration continue. N'en ayant jamais fait, cela n'a pas été simple au début. Heureusement Steeven, notre collègue de NancyClotep, nous a aidé pour la mise en place de plusieurs paramètres de configuration pour l'intégration continue.

Après beaucoup de temps passé à assimiler les subtilités de Jenkins, de la patience et l'aide de Steeven, nous avons réussi à tout configurer pour que l'intégration continue fonctionne. Nous avons pu déployer notre site qui est hébergé à ce jour sur les serveurs de NancyClotep.

2. VERSIONS DES LIBRAIRIES

Durant le développement, deux des bibliothèques majeures que nous utilisons ont été mises à jour : Spacy et Material-UI. Nous avons fait le choix de les mettre à jour de notre côté. Cela nous a pris beaucoup de temps pour adapter le code. Cela nous a permis d'apprendre qu'il faut prendre ce type de facteur dans notre planification.

3. TEMPS DE CALCUL ET CELERY

Une des problématiques que nous n'avons pas anticipée est le temps de calcul des différents modèles lorsqu'il s'applique sur un document (5 à 15 secondes sur le serveur de NancyClotep). Pour que la fonctionnalité de recherche prenne en compte les features, il faut avoir appliqué les modèles sur l'ensemble des documents après l'import. Cela peut prendre du temps et mobiliser entièrement la couche middleware. Telle que notre architecture est conçue, il est impossible de mettre en place un système de file d'attente qui récupérerait l'ensemble des documents quand ils sont importés pour appliquer les modèles progressivement.

Une solution que nous avons découverte par la suite est l'utilisation du framework Celery^[49] pour la couche middleware. Il nous permettrait de gérer de manière asynchrone l'application des modèles sur les nouveaux documents.

D. AMELIORATION POSSIBLE OU FUTUR

- Compléter les modèles avec de l'entity linking, des calculs de similarité et/ou de la standardisation ontologique
- Possibilité de transformer l'application en un équivalent de Doccano qui permettrait au personnel de santé de pouvoir en plus annoter directement les documents
- Création d'une page qui affiche ce qui a été annoté par le personnel de santé et ce qui a été trouvé par notre modèle. Pour permettre au personnel de santé de voir directement les erreurs du modèle.
- La possibilité de désigner des modèles de NLP, en se basant sur le composant primaire de Spacy, directement depuis le site.
- La possibilité de réentraîner en un clic un modèle en prenant en compte les corrections de l'annotation effectuées sur le site.

X. CONCLUSION

Ce fût une année riche en enseignements, jalonnée d'obstacles que nous avons dû franchir un à un. La compréhension du besoin qui semblait acquise lors de la première réunion mais qui en fait ne le sera que plus tard. Grâce à un travail d'écoute, d'analyse et d'autres réunions nous avons enfin su saisir le besoin de notre client.

Notre manque de facilité à vulgariser les différents concepts d'IA, ne nous a pas facilité la tâche tout le long de notre parcours mais nous avons toujours réussi à nous faire comprendre et à comprendre nos interlocuteurs.

L'organisation pour travailler en équipe fut délicate au début, mais avec la mise en place d'outils comme Git, Trello et en appliquant une méthode ScrumBan. Nous avons réussi à trouver notre rythme et à progresser dans notre projet.

J'ai dû appréhender beaucoup de technologies différentes, REACT, Docker, DjangoREST, Flask etc. En avançant étapes par étapes, j'ai pu monter en compétence sur toutes les technologies que nous avons dû utiliser pour le développement de notre projet. Bien sûr, je n'ai pas acquis une maîtrise totale de toutes ces technologies. J'ai acquis une compréhension plus solide dans certaines React, Spacy et un peu moins solide dans d'autres PosteSql, Docker par exemple.

Mais cela m'a permis de voir, la construction d'un projet de A à Z et de mieux comprendre toutes les interactions qu'il peut y avoir entre les différentes couches. Et toute l'énergie nécessaire qu'il faut déployer pour finir ce type de projets.

Ce fût une longue année avec des périodes de doute, d'autres de joie, de frustration parfois. Mais notre persévérance et notre enthousiasme pour ce projet, nous a permis d'aller jusqu'au bout et de finir notre application. Qui je l'espère sera pour faire gagner un temps précieux aux personnels médical qui l'utiliseront.

XI. REMERCIEMENTS

Nous tenons à remercier Steeven Frezier qui nous a apporté son expertise technique et son soutien tout au long du processus de développement. Il nous a mis la musique d'ambiance qu'il fallait pendant de longue période de build.

Le professeur Olivier pour son implication dans notre processus d'annotation et sa patience à toute épreuve face à notre manque de vulgarisation dans le domaine de l'intelligence artificielle. En espérant que notre application lui serve à gagner un temps précieux dans son futur travail.

Merci au professeur Karcher de nous avoir offert l'opportunité de travailler et de monter en compétence au sein de Nancyclotep. Merci aussi de nous avoir donné la chance de participer à son projet de création d'une future base de données IA. Nous sommes fiers d'avoir posé l'une des premières pierres de ce projet.

Merci à toute l'équipe de Nancyclotep, qui nous a accueilli avec bienveillance et à Mme Fougère qui nous a suivis et aidés dans nos démarches tout au long de l'année.

Enfin merci à Mr Chatourel et au CHRU de Nancy Brabois qui nous ont accompagnés toute la durée de notre parcours et qui nous ont offert l'opportunité de voir le fonctionnement de plusieurs services au sein de leurs différents sites.

Et merci google.

XII. SOURCES

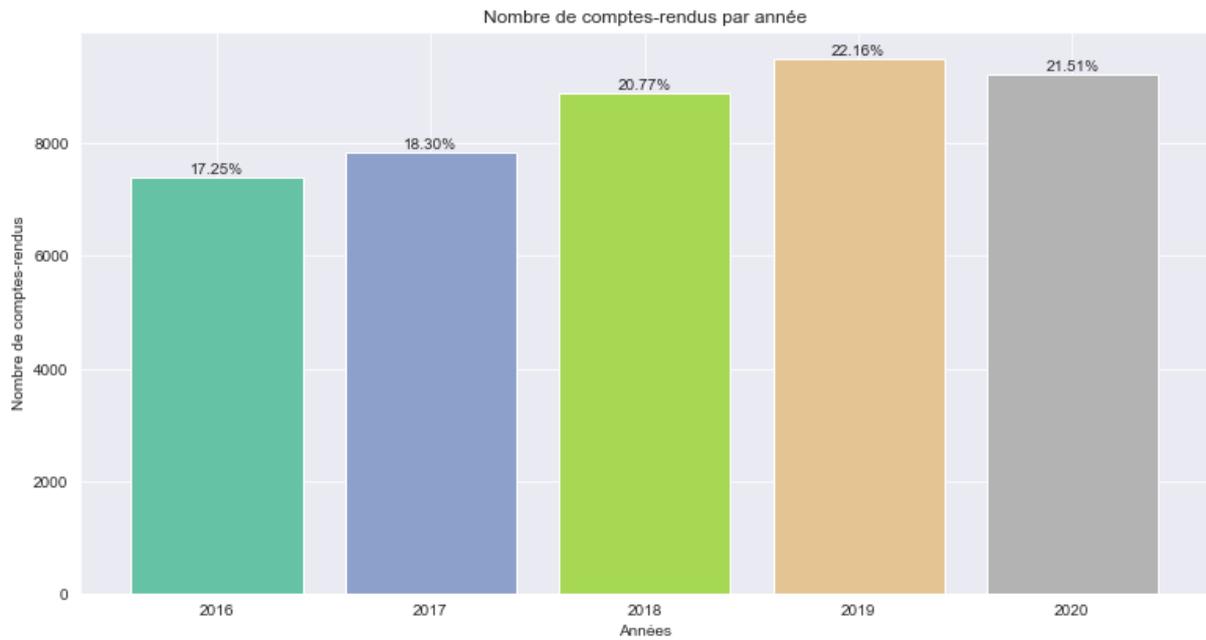
1. "Site web de Nancyclotep" <https://nancyclotep.com/>
2. "Site web du CHRU nancy" <http://www.chu-nancy.fr/>
3. "A Comprehensive Survey on Word Representation Models" <https://arxiv.org/pdf/2010.15036.pdf>

4. "Comparaison de textes: quelques approches..." <https://hal.archives-ouvertes.fr/hal-00874280/document>
5. "Cosine Similarity" https://en.wikipedia.org/wiki/Cosine_similarity
6. "Levenshtein Distance" https://en.wikipedia.org/wiki/Levenshtein_distance
7. "Jaro-Winkler Distance" https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance
8. "Introduction aux algorithmes de recommandation" <https://blog.octo.com/introduction-aux-algorithmes-de-recommandation-lexemple-des-articles-du-blog-octo/>
9. "Introduction to Ontology Concepts and Terminology" <http://dcevents.dublincore.org/IntConf/dc-2013/paper/download/140/105>
10. "Bioportal" <http://bioportal.lirmm.fr/>
11. "PubMed" <https://pubmed.ncbi.nlm.nih.gov/>
12. "Concept recognition and coding in French texts" <https://repub.eur.nl/pub/100036/>
13. "Indexation de textes médicaux par extraction de concepts, et ses utilisations" <https://tel.archives-ouvertes.fr/tel-00932922/document>
14. "Évaluation de l'indexation des comptes rendus médicaux à l'aide d'un outil états-unien adapté pour le français" https://www.researchgate.net/publication/268382861_Evaluation_de_l%27indexation_des_comptes_rendus_medicaux_a_l%27aide_d%27un_outil_etats-unien_adapte_pour_le_francais
15. "Site de la faculté de médecine de Rennes" <https://medecine.univ-rennes1.fr/>
16. "Medical Text Classification using Convolutional Neural Networks" <https://arxiv.org/ftp/arxiv/papers/1704/1704.06841.pdf>
17. "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks" <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0781-4>
18. "Unstructured Medical Text Classification using Linguistic Analysis: A Supervised Deep Learning Approach" <https://ieeexplore.ieee.org/document/9035282>
19. "Medical Text Summarization System based on Named Entity Recognition and Modality Identification" <https://aclanthology.org/W09-1324.pdf>
20. "SCIBERT: A Pretrained Language Model for Scientific Text" <https://arxiv.org/pdf/1903.10676v3.pdf>
21. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining" <https://arxiv.org/ftp/arxiv/papers/1901/1901.08746.pdf>
22. "CamenBERT" <https://camembert-model.fr/>
23. "Attention is all you need" <https://arxiv.org/abs/1706.03762>
24. "COMETA: A Corpus for Medical Entity Linking in the Social Media" <https://arxiv.org/pdf/2010.03295v2.pdf>
25. "Improving Biomedical Pretrained Language Models with Knowledge" <https://arxiv.org/pdf/2104.10344v1.pdf>
26. "Rare Disease Identification from Clinical Notes with Ontologies and Weak Supervision" <https://arxiv.org/pdf/2105.01995v3.pdf>
27. "The Cancer Imaging Archive" <https://www.cancerimagingarchive.net/>
28. "Recommandation CNIL pour l'anonymisation" <https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles>
29. "Doccano" <https://github.com/doccano/doccano>
30. "Prodigy" <https://www.prodigygame.com/main-en/>
31. "React.js" <https://fr.reactjs.org/>
32. "Material-ui" <https://material-ui.com/>
33. "Django REST Framework" <https://www.django-rest-framework.org/>
34. "PostgreSQL" <https://www.postgresql.org/>
35. "Flask" <https://flask.palletsprojects.com/en/2.0.x/>
36. "Spacy" <https://spacy.io/>
37. "Git" <https://git-scm.com/>
38. "Docker" <https://www.docker.com/>
39. "Jenkins" <https://www.jenkins.io/>

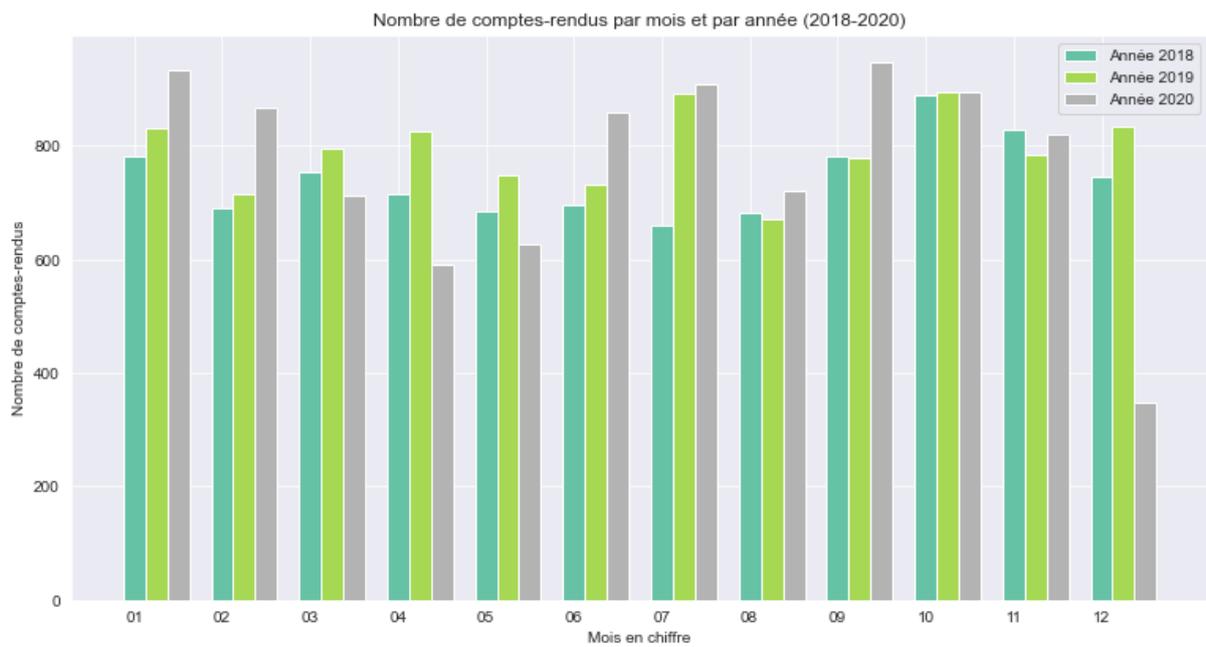
40. "Named Entity Evaluation" http://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/
41. "fr_dep_news_trf" https://spacy.io/models/fr#fr_dep_news_trf
42. "fr_core_news_lg" https://spacy.io/models/fr#fr_core_news_lg
43. "Spacy Config" <https://spacy.io/api/data-formats#config>
44. "JSON Web Token" <https://jwt.io/>
45. "Backlog" https://docs.google.com/spreadsheets/d/1v8Pa5aYq8PLlwzfWSVtPPgCIK_elwPHeYtQ_XGgDAqcs/edit?usp=sharing
46. "ScrumBan" <https://en.wikipedia.org/wiki/Scrumban>
47. "Trello" <https://trello.com/>
48. "Balsamiq" <https://balsamiq.com/>
49. "Celery" <https://docs.celeryproject.org/en/stable/>
50. "Spacy Examples" <https://github.com/explosion/projects/tree/v3/tutorials>
51. "Healthcare Text Annotation Guidelines" <https://github.com/google/healthcare-text-annotation>
52. "2010 i2b2 / VA Challenge Evaluation Concept Annotation Guidelines" <https://aclanthology.org/L16-1272.pdf>
53. "Annotating Evidence Based Clinical Guidelines" http://ceur-ws.org/Vol-952/paper_13.pdf
54. "A French clinical corpus with comprehensive semantic annotations" <https://link.springer.com/content/pdf/10.1007/s10579-017-9382-y.pdf>
55. "The Quaero French medical corpus" http://nactem.ac.uk/biotxtm2014/presentations/Neveol_pres.pdf

XIII. ANNEXES

Annexe 1 :

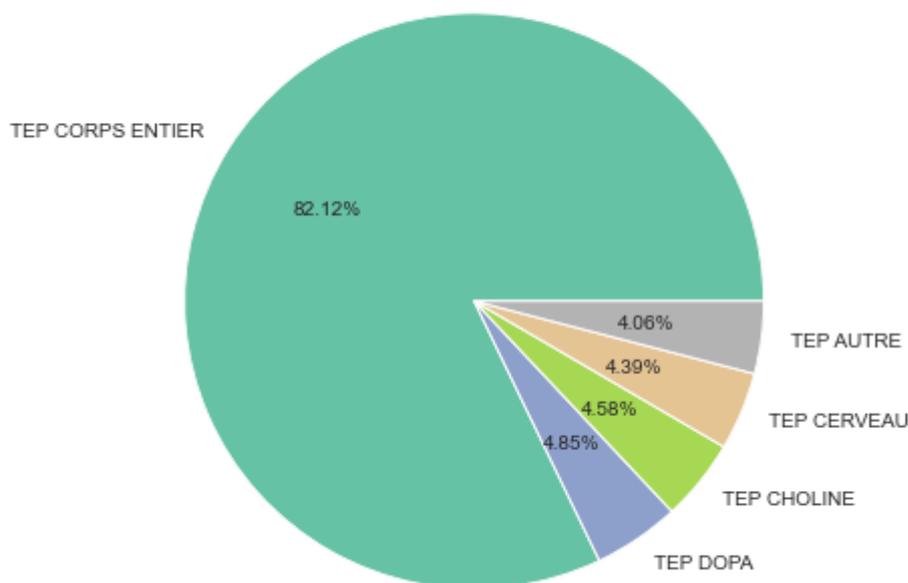


Annexe 2 :



Annexe 3 :

Pourcentage du nombre de libellés de comptes-rendus



Annexe 4 :

Installation répertoriée sous le n° M540008 Autorisation CODEP-STR-2020-037740
 VEREOS N° d'identification 900087 PHILIPS année 2019
 SPGCodes CCAM ZZQL016

Cher Confrère,

Nous vous remercions de nous avoir adressé Monsieur XXXX xxxx, né le xxxx (39 ans), pour la réalisation, le 14/12/2020, d'une tomoscintigraphie par émission de positons au FDG (Morpho-TEP).

Contexte dans lequel l'étude est réalisée :
 Caractérisation et bilan d'extension d'adénopathies médiastinales de découverte fortuite au décours du bilan d'hémostase.

Technique de l'examen :
 Les acquisitions ont été débutées 60 minutes après injection de 333 MBq de 18FDG.
 La glycémie lors de l'injection était de 118 mg/dl.
 CTDI : 6,9 mGy - DLP : 724,5 mGy.cm

Cet examen apporte les informations suivantes :

À l'étage cervico-thoracique :
 Hypermétabolisme symétrique des loges amygdaliennes, non spécifique.
 Absence d'autre hypermétabolisme anormal de la filière pharyngo-laryngée et de l'aire thyroïdienne.
 Ganglions jugulo-carotidiens modérément hypermétaboliques, prédominant à droite (SUV max = 2,7 au sein du groupe II droit pour une référence hépatique max = 2,6).
 Adénopathies modérément hypermétaboliques médiastino-hilaires bilatérales de siège latéro-trachéal droit (SUV max = 3,7), de la loge de Baretz (SUV max = 4,2), de la fenêtre aorto-pulmonaire, hilare bilatéral et sous-carinaire (SUVmax= 3,6).
 Absence de structure ganglionnaire hypermétabolique suspecte axillaire.
 Adénopathie modérément hypermétabolique sus-claviculaire droite (SUV max = 3,8).
 Absence d'hypermétabolisme anormal pleuro-parenchymateux.

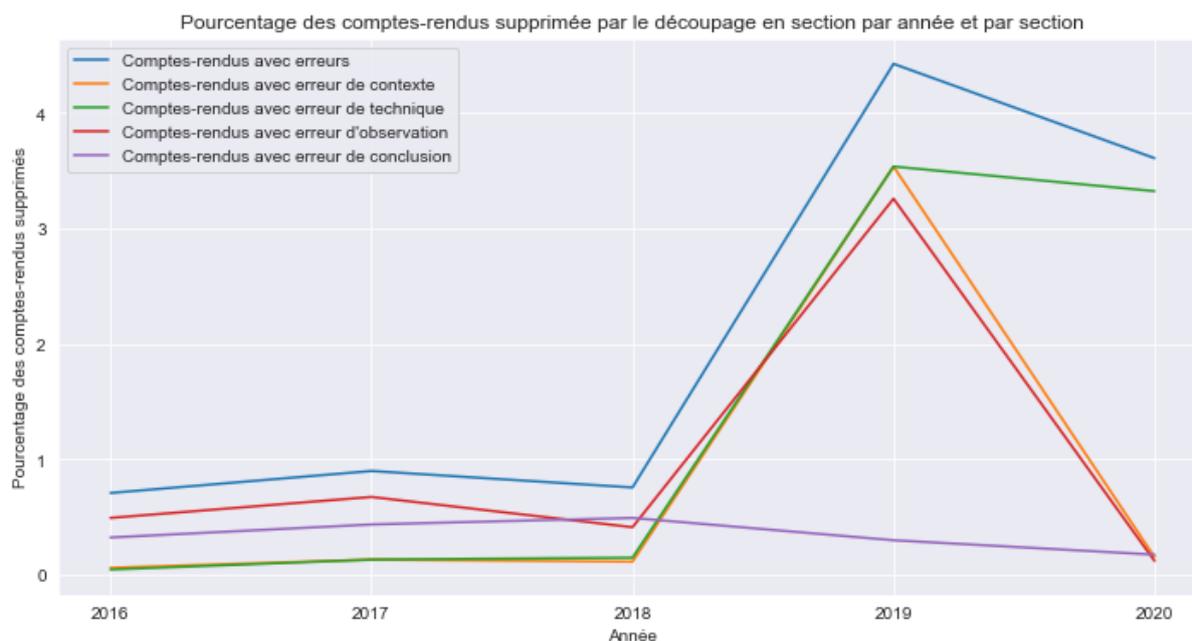
À l'étage abdomino-pelvien :
 Absence d'hypermétabolisme anormal hépatique, splénique et surrénalien.
 Absence de structure ganglionnaire hypermétabolique suspecte coelio-mésentérique, lombo-aortique, iliaque et inguinale.
 Deux foyers digestifs d'allure grélique en regard de la paroi abdomino-pelvienne antérieure, peu spécifiques.

À l'étage musculo-squelettique :
 Absence d'hypermétabolisme osseux suspect.

En conclusion cet examen identifie des adénopathies modérément hypermétaboliques jugulo-carotidiennes bilatérales, sus-claviculaire droite et médiastino-hilaires bilatérales compatibles avec une hémopathie de bas grade. Il n'est pas noté d'hypermétabolisme anormal viscéral associé.

Bien Confraternellement. Docteur Solène PARIS-GRANDPIERRE

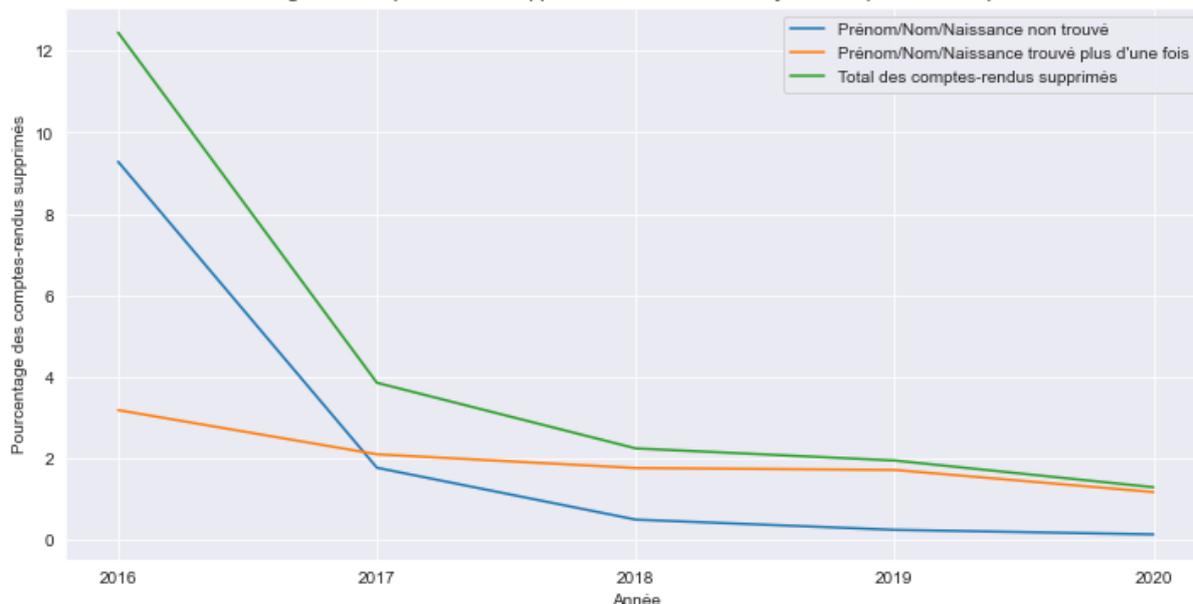
Annexe 5 :



Annexe 6 :

	Nombre comptes-rendus	Par rapport au total (%)	Delta avec l'etape précédente (%)
Total	52353	100	0
Après suppression des comptes-rendus vide	50147	95,79	4,21
Après suppression des comptes-rendus en double	43906	83,87	11,92
Après suppression des comptes-rendus avec pour libellés ['PANNE DE MACHINE', 'PAS D'EXAMEN', 'PAS DE PRODUITS', 'EXAMEN ANNULE', 'PAS VENU', 'VENU, NON REALISE']	43372	82,85	1,02
Après suppression des comptes-rendus qui ont le même numéros d'examen	42909	81,96	0,88
Après suppression des comptes-rendus où la regex anonymisation ne trouve pas nom/prenom/naissance	42008	80,24	1,72
Après suppression des comptes-rendus où nom/prenom/naissance apparaissent plus d'une fois	41184	78,67	1,57
Découpable en section	40262	76,90	1,76

Pourcentage des comptes-rendus supprimée au cours de l'anonymisation par année et par raison



Annexe 7 :



PySBD: Pragmatic Sentence Boundary Disambiguation



EPISOURCE

Nipun Sadvilkar, Mark Neumann
nipun.sadvilkar@episource.com, markn@allenai.org



Allen Institute for AI

Introduction

PySBD - a rule-based sentence boundary disambiguation Python package that works out-of-the-box for 22 languages.

Features:

- Domain agnostic rules
- Non Destructive Segmentation
- Multilingual Support
- Robust with 98% test coverage

Results & Comparison to alternatives

Tool	GRS	GENIA
blingfire	75.00	86.95
syntok	68.75	80.90
spaCy	52.08	76.80
spacy dep	54.17	39.20
stanza	72.92	63.40
NLTK	56.25	87.95
PySBD	97.92	97.00

Accuracy (%) of PySBD compared to other open source SBD packages

Tool	Speed(ms)
blingfire	85.24
syntok	1764.11
spaCy	1523.20
spacy dep	26850.69
stanza	48383.46
NLTK	780.49
PySBD	9483.96

Speed benchmark on the entire text of 'The adventures of Sherlock Holmes'

Language	Accuracy (%)
Amharic	80.95%
Arabic	70.40%
Armenian	63.75%
Bulgarian	93.35%
Burmese	48.05%
Chinese	85.35%
Danish	91.40%
Deutsch	80.95%
Dutch	91.40%
French	91.90%
Greek	91.05%
Hindi	88.50%
Italian	90.55%
Japanese	96.45%
Kazakh	63.20%
Marathi	92.60%
Persian	84.95%
Polish	55.48%
Russian	88.55%
Spanish	92.65%
Urdu	77.55%

Accuracy on OPUS 100 multilingual corpus test sets, containing 2000 sentences per language.

Conclusion

- **PySBD** has interpretable rules and are easy to modify
- Highly accurate - 97% English GRS - irrespective of domain
- Robust codebase with 98% test coverage
- Lightweight, easy to integrate with existing NLP pipelines
- Multilingual with 22 language support
- Already being used by 79 projects
- Extensible to handle more edge cases and languages in community driven way

Processing Stages

```

" My name is Jonas E. Smith. Please turn to p. 55."
  |
  | preprocess
  |
  | " My name is Jonas E# Smith. Please turn to p# 55."
  |
  | segment
  |
  | [" My name is Jonas E# Smith.", " Please turn to p# 55." ]
  |
  | postprocess
  |
  | [" My name is Jonas E. Smith.", " Please turn to p. 55." ]
                    
```

Data

- English Golden Rules Set (GRS)** - 48 hand crafted rules by considering various domains.
- GENIA Corpus:** Linguistically annotated biomedical papers Speed Benchmark
- OPUS-100 Multilingual Parallel Data**

Usage

```

import pysbd
text = " My name is Jonas E. Smith. Please turn to p. 55."
seg = pysbd.Segmenter(language="en", clean=False)
print(seg.segment(text))
# [" My name is Jonas E. Smith.", " Please turn to p. 55." ]
                    
```

Further information

GitHub Repo:
<https://github.com/nipunsadvilkar/pysbd>

Research Paper:
<https://arxiv.org/abs/2010.09657>

Annexe 8 :

SENTS_P	Sentence segmentation (precision)	0.95
SENTS_R	Sentence segmentation (recall)	0.97
SENTS_F	Sentence segmentation (F-score)	0.96

Annexe 9:

76 TRUE

Recherche d'argument pour une infection de prothèse totale de hanche gauche chez un patient ayant présenté une cellulite du membre inférieur gauche avec drainage percutané à la mi-septembre 2017 avec en parallèle une spondylodiscite L1-L2.

76 PRED

Recherche d'argument pour une infection de prothèse totale de hanche gauche chez un patient ayant présenté une cellulite du membre inférieur gauche avec drainage percutané à la mi-septembre 2017 avec en parallèle une spondylodiscite L1-L2 .

100 TRUE

TEP TDM à la FDOPA après une cure de radiothérapie interne vectorisée par LUTATHERA pour une tumeur neuroendocrine du grêle bien différenciée G1 métastatique, hépatique et ganglionnaire mésentérique.

100 PRED

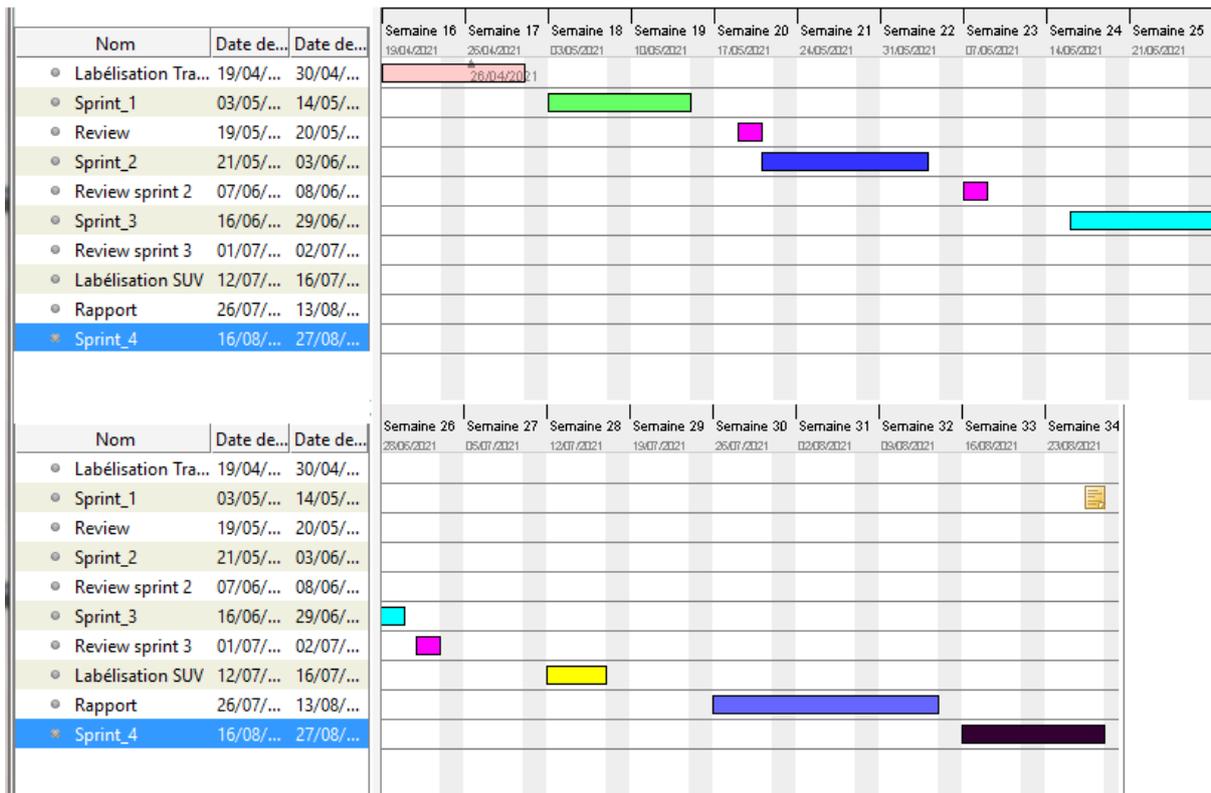
TEP TDM à la FDOPA après une cure de radiothérapie interne vectorisée par LUTATHERA pour une tumeur neuroendocrine du grêle bien différenciée G1 métastatique, hépatique et ganglionnaire mésentérique.

chimiothérapie : 200, 13.40
radiothérapie : 119, 7.97
chirurgie : 80, 5.36
radio-chimiothérapie : 41, 2.75
hormonothérapie : 39, 2.61
chimiothérapie adjuvante : 20, 1.34
prostatectomie radicale : 18, 1.21
immunothérapie : 17, 1.14
cyberknife : 14, 0.94
radiochimiothérapie : 14, 0.94
curiethérapie : 13, 0.87
thérapie ciblée : 13, 0.87
corticothérapie : 12, 0.80
r-chop : 12, 0.80
prostatectomie : 11, 0.74

Annexe 12

à gauche : 36, 0.92%
à droite : 32, 0.81%
sous-carinaire : 31, 0.79%
gauche : 28, 0.71%
droite : 26, 0.66%
hilaire droit : 25, 0.64%
hiatus de winslow : 24, 0.61%
lombo-aortiques : 21, 0.53%
axillaire gauche : 21, 0.53%
sus-claviculaire gauche : 20, 0.51%
iliaque externe gauche : 20, 0.51%
splénique : 19, 0.48%
tissulaire : 18, 0.46%
axillaires bilatéraux : 18, 0.46%
axillaires droites : 17, 0.43%

Annexe 13



Annexe 14 :

The screenshot shows a software interface with a navigation bar at the top and a table of data below. The navigation bar includes a logo, an 'Import CSV' button, an 'export metrics' button, a 'Statistiques' button, and a search bar. The table below lists exam data with columns for Date, Médecin, Ref dossier, and type d'examen.

Date	Médecin	Ref dossier	type d'examen
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep
xx/xx	Professeur xxx	10xx72xx56	tep

14/1000

1) A Nancy
Mr XXXX

Le xx/xx/xxx

Par le professeur xxxxxx

Accedebant enim eius asperitati, ubi imminuta vel laesa amplitudo imperii dicebatur, et iracundae suspicionum quantitati proximorum cruentae blanditiae exaggerantium incidentia et dolore impendio simulantium, si principis pericillitur vita, a cuius salute velut filo pendere statum orbis terrarum fictis vocibus exclamabant.

Inter haec Orfitus praefecti potestate regebat urbem aeternam ultra modum delatae dignitatis sese effertens insolenter, vir quidem prudens et forensium negotiorum oppido gnarus, sed splendore liberalium doctrinarum minus quam nobilem decuerat institutus, quo administrante seditiones sunt concitatae graves ob inopiam vini: huius avidis visibus vulgus intentum ad motus asperos excitatur et crebros.

Cumque pertinacius ut legum gnarus accusatorem flagitaret atque sollempni, doctus id Caesar libertatemque superbiam ratus tamquam obtractatorem audacem excarnificari procepit, qui ita evisceratus ut cruciatibus membra deessent, inplorans caelo iustitiam, torum renidens fundato pectore mansit immobilis nec se incusare nec quemquam alium passus et tandem nec confessus nec confutatus cum ablecto consorte poenali est morte multatus: et ducebatur intrapitius temporum iniquitati insultans, imitatus Zenonem illum veterem Stoicum qui ut mentiretur quaedam laceratus dilutus, ovilem sedibus linguam suam cum cruento sputamine in oculos interrogantis Cypril regis inpegit.

Etenim si attendere diligenter, existimare vere de omni hac causa volueritis, sic constituets, iudices, nec descensurum quemquam ad hanc accusationem fuisse, cui, utrum vellet, liceret, nec, cum descendisset, quicquam habiturum spei fuisse, nisi alicuius intolerabili libidine et nimis acerbo odio niteretur. Sed ego Atracino, humanissimo atque optimo adolescenti meo necessario, ignoaco, qui habet **excusationem** vel pietatis vel necessitatis vel aetatis. Si voluit accusare, pietati tribuo, si iussus est, necessitati, si speravit aliquid, pueritiae. Ceteris non modo nihil ignoscendum, sed etiam occiter est resistendum.

Dum haec in oriente aguntur, Arelate hiemem agens Constantius post theatralis ludos atque circenses ambizioso editos apparatus diem sextum idus Octobres, qui imperii eius annum tricensimum terminabat, insolentiae pondera gravius librans, siquid dubium deferabatur aut falsum, pro liquido accipiens et conperito, inter alia excarnificatum Gerontium Mogentianae comitem partis exulati maerore multavit.

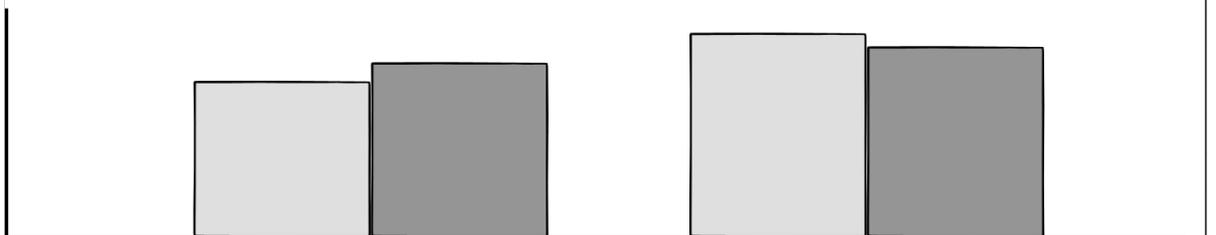
Orientis vero limes in longum protentus et rectum ab Euphratis fluminis ripis ad usque supercilio porrigitur Nilii, laeva Saracenis conterminans gentibus, dextra pelagi fragoribus patens, quom plagom Nicator Seleucus occupatam auxit mognum in modum, cum post Alexandri Macedonis obitum successorio iure teneret regna Persidis, efficaciae inpetrabilis rex, ut indicat cognomentum.

carcinome mammaire para-prothétique gauche externe en mars 2017 avec atteinte ganglionnaire inter-pectorale homolatérale traitée par **chimiothérapie**, **Herceptin**, **chirurgie** et **radiothérapie**, **para-prothétique gauche externe** en mars 2017 avec atteinte ganglionnaire inter-pectorale homolatérale traitée par chimiothérapie, Herceptin, chirurgie et radiothérapie chez une patiente aux antécédents de carcinome mammaire gauche en 2005.

Information extra	Précision (en%)	Résultat
Score de Deauville	68%	3
Traitement	38%	Chimiothérapie
		Herceptin
		Chirurgie
		Etc...

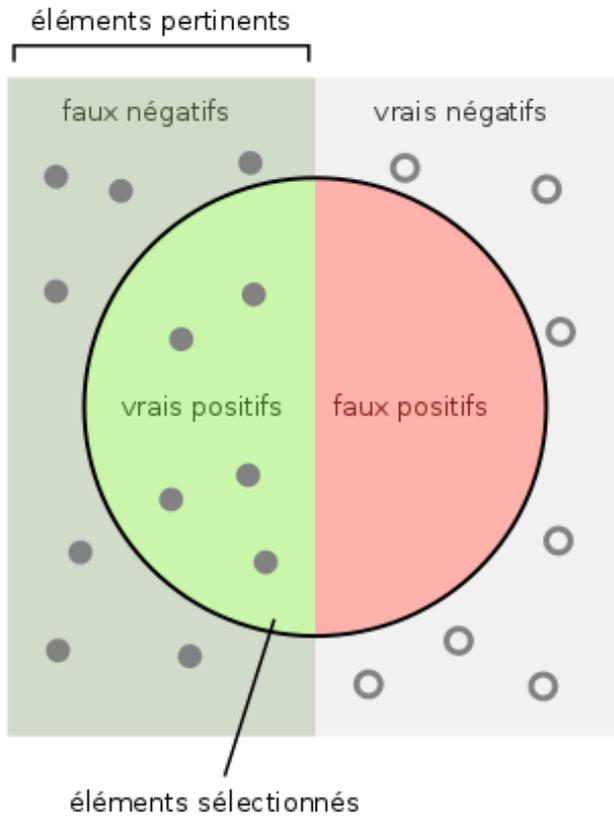
1/1000

[Retour accueil](#)



software statistics teaching **technology** tips tool **toread** travel tutorial **tutorials** tv twitter typography ubuntu usability **video** videos visualization web **web 2.0** web design **webdev** windows
wordpress work writing youtube software statistics teaching **technology** tips tool **toread** travel **tutorials** tv **twitter** typography ubuntu usability **video** videos visualization web **web 2.0** web design **webdev**
 wiki windows **wordpress** work writing youtube software statistics teaching **technology** tips tool **toread** tutorial **tutorials** tv **twitter** typography ubuntu **usability** **video** videos visualization web **web 2.0** web design **webdev**
webdev wiki windows **wordpress** work writing youtube software statistics teaching **technology** tips tool **toread** tutorial **tutorials** tv **twitter** typography **usability** **video** videos visualization web **web 2.0** web design

[Retour accueil](#)



Combien de candidats sélectionnés sont pertinents ?

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

Combien d'éléments pertinents sont sélectionnés ?

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$