AI & DATA

LA CLEF DU SUCCÈS D'UN PROJET NLP

Le Natural Language Processing (NLP) est une technologie d'IA dont le but est de permettre aux machines d'analyser et comprendre le langage humain écrit ou parlé. A travers cet article, nous présenterons l'importance de l'annotation dans un projet NLP et comment mener au mieux cette partie du projet.

Dans le cadre d'un projet NLP, il existe différents projets et certains vont avoir besoin de données annotées en entrée pour la phase d'apprentissage.

Parmi ces projets, nous retrouvons :

- **La classification supervisée de documents***
- **L'étiquetage de séquences**
- **La création de relations**

Dans cet article, nous nous concentrerons sur le cas de la **classification supervisée de documents** qui est le plus courant.

La classification de documents consiste à **faire correspondre des thèmes à des documents donnés.**

Exemple :

Lors d'enquêtes de satisfaction avec un champs de texte libre pour le client, nous obtenons un verbatim qui sera classé dans différents thèmes tels que « le savoir-faire d'un conseiller », « le savoir-être d'un conseiller », « la satisfaction », « l'insatisfaction », ou des thèmes beaucoup plus spécifiques.

Cette **phase d'annotation** va permettre de **créer un modèle** qui permettra de **prédire et raccorder un verbatim à un ou plusieurs thèmes.** Sur un projet NLP qui se découpe en plusieurs phases, la partie la plus importante, qui permettra d'avoir in fine des verbatims bien classés, des thèmes bien prévisibles, est donc cette phase d'annotation. Sans une annotation et un corpus d'apprentissage propre, de qualité et conséquent, peu importe le modèle déployé, les résultats seront déceptifs. Et ce n'est pas seulement une recommandation a priori, c'est ce que nous avons constaté sur plus d'une vingtaine de projets.

L'annotation est donc une condition sine qua non à la réussite de votre projet NLP. Pour qu'une annotation soit de qualité, il faut qu'elle soit encadrée et travaillée afin de répondre au mieux aux besoins et exigences des métiers.

Nous identifions **3 étapes clés** pour la réussite d'un projet d'annotation :

1. **PRÉPARER SA CAMPAGNE**
   Rédiger le guide d'annotation ; définir le process de la campagne
2. **CONDUIRE LA CAMPAGNE**
   Former les annotateurs ; encadrer, animer et superviser la campagne
3. **S'ASSURER DE LA QUALITÉ FINALE**
   Mesures qualitatives ; finaliser le corpus

** La notion de document doit être pris au sens large comme un texte comportant un certain nombre de lignes : document au sens courant, mais aussi email, avis client sur plateforme, article en ligne, tweet, etc…*

1. Préparer la campagne

RÉDIGER LE GUIDE D'ANNOTATION

Toute campagne commence par la **rédaction du guide d'annotation,** étape qui va permettre de définir **quels sont les thèmes** à aller chercher, **quoi en faire** et **comment les définir.** Le guide d'annotation est un document qui vous accompagnera tout au long de votre projet. Ce guide, souvent un document Word, est à créer avec les métiers, les utilisateurs finaux du projet NLP.

Si le métier sait ce qu'il attend, si il connaît les données sur lesquelles l'équipe NLP va travailler ainsi que les thèmes qu'il veut retrouver, ces derniers seront définis avec lui avant vérification dans les données que ces thèmes existent en volumes suffisants pour aboutir sur des modèles qui fonctionnent.

A savoir que pour qu'un modèle soit pérenne et efficace, on estime **le volume suffisant à environ 3-4%** du thème représenté dans le corpus.

Dans le cas où le métier n'a aucune indication ou attentes spécifiques, les données seront soumises à une classification non supervisée, avec un clustering pour découvrir les thèmes et les termes qui pourraient les intéresser et ainsi rédiger le guide.

Pour rédiger le guide d'annotation, il faut :

1. Créer une suite en parallèle des sessions de travail avec le métier

2. Que le guide soit lu, relu et validé par l'équipe projet

3. Que ce soit un document Word clair et détaillé, le guide servant de référence pour les annotateurs en cas de doute dans le classement d'un verbatim afin que toute l'équipe ait les mêmes idées, les mêmes notions derrière chaque thème. Le guide doit comporter :
   - Des explications des règles d'annotation : mono ou multi-thèmes, priorisation si existante
   - Définition du thème, des bornes et points d'attention, les listes d'exemples

4. Que le guide soit testé en amont de la campagne pour vérifier qu'il sera clair pour un annotateur

5. Quand il y a beaucoup de thèmes, le guide peut être accompagné d'une fiche récapitulative ou d'une infographie qui va pouvoir aider l'annotateur à avoir un cheminement d'idées lors de son annotation. Tout élément non intégré dans le guide au démarrage d'une campagne ne sera pas annoté et ne pourra entrer dans le processus.

DÉFINITION DU PROCESS DE CAMPAGNE

Une fois le guide d'annotation rédigé, il faut **définir comment mener la campagne d'annotation.** On aborde alors la phase de démarrage avec les annotateurs jusqu'à la fin de campagne lorsqu'il est estimé qu'il y a suffisamment d'annotations pour obtenir des modèles pertinents et performants.

Il faut ainsi définir les éléments suivants :

> La durée de la campagne

Elle pourra être modifiée au fur et à mesure que le projet avance mais il est important de fixer ce dernier au démarrage pour que l'équipe sache comment avancer, pourquoi et avoir une idée globale de la charge de travail à répartir tout au long de la campagne. La durée peut se définir en semaines, plusieurs sessions ou un objectif

> Le nombre d'annotateurs

Ils participeront au projet

> Le choix du type d'annotation

Simple, double ou triple. La simple annotation peut parfois être biaisée par le jugement seul de l'annotateur. La double annotation permet d'avoir un corpus plus propre.

> La stratégie des échantillons

À annoter en aléatoire ou à définir. A savoir que dans le cas de l'aléatoire, certains thèmes ressortiront plus souvent que d'autres et afin d'avoir un corpus complet, il faudra compléter en allant chercher des verbatims spécifiques.

> La pré-annotation

Définir si on part de l'annotation brute, soit un verbatim vierge que l'annotateur devra classer lui-même, ou d'une pré-annotation, soit un verbatim préanalysé avec un nombre choix de thèmes pouvant correspondre que l'annotateur devra valider ou non. L'idéal est l'annotation brute afin d'éviter toute influence externe. Dans les faits, le plus optimal est de mélanger les deux afin de gagner du temps : commencer par une annotation brute jusqu'à obtenir un volume déjà assez conséquent par thèmes pour pouvoir basculer sur de la pré-annotation.

> L'active learning

Se baser sur de la donnée brute aléatoire en début de campagne afin d'entraîner rapidement les modèles. Ces derniers ne seront pas très performants au départ, mais le deviendront avec le temps. L'active learning permet d'identifier les modèles rapidement opérationnels, d'éviter les annotations utiles afin de pouvoir se concentrer sur les modèles plus complexes et les améliorer.

2. Conduire la campagne

La campagne est prête lorsqu'on a défini les thèmes à annoter, la fréquence, qui va les annoter, comment gérer l'annotation en simple, double ou triple.

Nous entrons dans la phase de conduction et d'ouverture de la campagne.

FORMER LES ANNOTATEURS

Lors de la phase de conduite de campagne, il est primordial de commencer par former les annotateurs à travers ces différentes étapes :

1 **Présentation du projet, des enjeux, du rôle clé de l'annotation, ...**

Lors de cette première étape, il faut expliquer aux annotateurs pourquoi ils annotent, l'intérêt de l'annotation et comment il s'inscrit dans ce projet d'IA afin de leur communiquer l'étendue projet auxquels ils prennent part et les aider à se sentir valorisé sur une tâche qui peut paraître rébarbatif. Sans les annotateurs, le projet n'existe pas.

2 **Prise en main de l'outil d'annotation**

Les annotateurs peuvent être des infolinguistes et des data scientists d'une même équipe, les métiers, les utilisateurs finaux du projet NLP

3 **Appropriation du guide d'annotation**

Si les annotateurs n'ont pas participé à la création du kit d'annotation, la lecture de l'intégralité du guide d'annotation est requise afin de se l'approprier et être efficace sur les annotations à faire.

4 **Sessions encadrées**

Lecture du guide d'annotation, questions/ réponses

Test sur un petit échantillon défini lors de la pré-campagne afin de s'assurer que la tâche a bien été comprise par les annotateurs et vérifier que leur schéma d'annotation est clair et non ambigu.

Cette phase peut amener à réexpliquer ou à corriger le plan d'annotation.

ENCADRER, ANIMER ET SUPERVISER

La campagne va pouvoir être lancée. Pour le bon déroulé de la campagne, il est important d'organiser tout le long de cette dernière :

- Une dynamique d'échange entre l'infolinguiste (ou la personne référente du projet) et les annotateurs via des outils de partage afin de pouvoir :
  - Echanger sur des questions ou difficultés rencontrées
  - Être disponible durant les sessions d'annotation
- Des sessions régulières communes de partage sur :
  - Des points d'attention concernant l'annotation
  - L'avancée du projet

3. S'assurer de la qualité finale

Pour obtenir un corpus de qualité, des mesures doivent être mises en place durant toute la campagne pour vérifier que tout se passe bien et que les résultats sont pertinents pour créer le corpus de référence, d'apprentissage.

A noter qu'**il faut approximativement *500 verbatims* par thème pour avoir un modèle robuste et pertinent.**

Afin de s'assurer de la qualité finale, il faut :

1 **Mettre en place la double** (ou bien que plus rare, triple) **annotation**

La double annotation est conseillée en début de campagne afin de :

- Juger la pertinence / qualité de la tâche
- Vérifier la qualité du guide d'annotation, qui est le socle commun porteur du projet
- Faire apparaître le plus tôt possible les divergences entre annotateurs, ce qui permet de réexpliquer ou écarter certaines annotations trop complexes. En raison de la complexité du verbatim entraînant des interprétations différentes, ce dernier doit être exclu de la phase d'apprentissage car si un être humain n'arrive pas à l'annoter, un modèle IA n'y arrivera pas non plus.
- Obtenir une annotation de référence commune

2 **Vérifier la conformité en comparant à des mini-références annotées par les experts et infolinguistes.**

3 **Comparer grâce au score « Accord inter-annotateur »** (Kappa de Cohen*)

4 **Comparer grâce au score « Accord intra-annotateur »** (plus rare)

Comparer les annotations d'un même annotateur en début, milieu et fin de campagne. Les annotations en fin de campagne sont souvent plus fiables que celles du début.

Kappa de Cohen (pour 2 annotateurs) :

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

κ	Interprétation
<0	Désaccord
0.0 - 0.20	Accord particulièrement faible
0.21 - 0.40	Accord faible
0.41 - 0.60	Accord modéré
0.61 - 0.80	Accord fort
0.81 - 1.00	Accord presque parfait

Où PR(*a*) est l'accord relatif entre annotateurs et PR(e) la probabilité d'un accord aléatoire. Si les annoteurs sont complètement en accord κ = 1. S'ils sont complètement en désaccord (ou en accord dû seulement au hasard) κ ≤0.

** Le Kappa de Cohen permet de définir la proportion de l'accord entre deux annotateurs par rapport à un accord aléatoire. Pour ce faire, le score est défini entre 0 et 1 avec une table d'interprétation qui n'est pas dans un consensus scientifique, donc bien que non fiable à 100%, il permet d'avoir une idée. En dessous de 0.4, l'accord est très faible, impliquant la révision des verbatims concernés. Au-dessus de 0.8, l'accord est presque parfait, ce qui est rare. Les scores les plus souvent obtenus varient entre 0.7 et 0.8.*

D’autres indicateurs peuvent aussi être intéressants à prendre en compte :

Informations sur le corpus annotés >

- Nombre de documents annotés
- Nombre de documents passés
- Temps moyen pour une annotation
- Temps pour chaque annotation

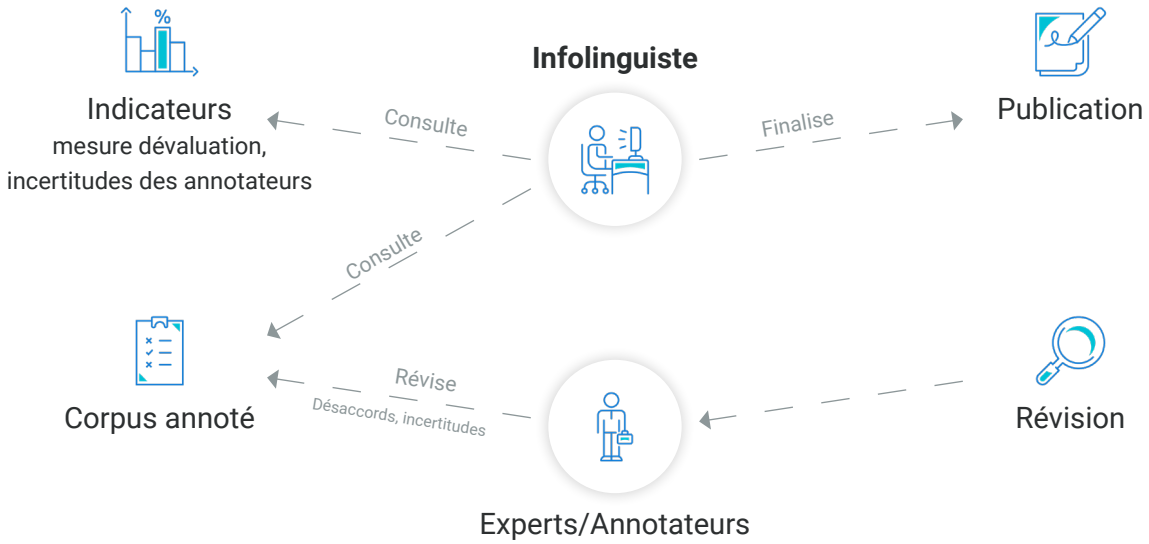
Cela permet d’avoir une vision sur la manière dont l’annotateur travaille et annote.

> Distribution des thèmes, et suivi des annotations par thème

- Calcul d’erreurs les plus fréquentes entre thèmes
- Calcul des thèmes les plus consensuels

Cela permet d’identifier les thèmes qui fonctionnent bien et se focaliser sur les thèmes qui fonctionnent moins bien dans l’annotation, nécessitant de trouver d’autres exemples pour ces derniers et pouvoir continuer à ce que l’annotation reste claire.

Finaliser le corpus annoté



L’infolinguiste a un rôle clé dans la campagne d’annotation.

Il va consulter les indicateurs (mesures d’évaluation, incertitudes des annotateurs), le corpus annoté. Il doit procéder à des ajustements continuellement pendant la campagne avec les annotateurs, mais va également pouvoir se retourner vers les experts en cas de désaccords, de certitudes, d’incertitudes sur des données précises dans le but d’obtenir le corpus le plus fiable possible. Une fois ce corpus considéré comme fiable, il va pouvoir être finalisé et publié en tant que corpus de référence et ainsi rentrer dans le processus de modélisation.

Quelques outils open source et payants

OUTILS OPEN SOURCE

OUTILS PAYANTS

Si vous **travaillez avec des fichiers texte**, ce que vous voulez faire peut-être catégoriser comme **classification de document, étiquetage de séquence ou séquence à séquence** et vous n'avez **pas besoin de relations**, mais vous voulez **commencer à étiqueter dès que possible** sans longues configurations, alors vous choisissez ***doccano***.

Si vous voulez **travailler avec des fichiers texte**, vous voulez **garder les choses aussi simples que possible** mais avez **besoin de plus de fonctionnalités** que celles fournies par doccano, alors essayez ***brat***.

Si, pour une raison quelconque, vous souhaitez **travailler avec des fichiers PDF (natifs)**, ou si vous n'avez pas peur d'**un outil d'annotation plus complexe** qui prend du temps à se familiariser avec mais vous offre **une gamme étendue de fonctionnalités**, ***INCEpTION*** est fait pour vous.

Kili Technology, très complet et de plus en plus utilisé car permet aussi de faire de l'image. Outil d'annotation dans le sens global du terme.