Artefacts 2ème sprint :

Objectifs du programme à réaliser :

Analyser les fichiers en sortie du convertisseur pour en sortir les éléments suivants :

- ✓ Le nom du fichier d'origine -> 1 ligne
- ✓ Le titre du papier -> 1 ligne
- ✓ Le.s auteur.s de l'article
- ✓ Le résumé de l'auteur -> 1 ligne
- ✓ Le client doit pouvoir utiliser ce système en ligne de commande sous GNU/Linux en lui passant en paramètre un dossier contenant des fichiers, PDF ou non.

BackLog produit :

Voici la liste des fonctionnalités, par ordre de priorité, que ce système devra être capable d'effectuer

- **Vérifier qu'il y a un paramètre et que c'est un dossier non vide**

Récupération du nombre d'arguments qui doit être de 2, l'appel du parser et le dossier. Récupération du paramètre afin de vérifier qu'il correspond bien à un dossier, qui plus est, non vide.

La fonction est terminée quand : Elle empêche le programme global de continuer si l'utilisateur ne donne pas de nom en paramètre, si le nom donné n'est pas celui d'un dossier ou bien qu'il est vide.

- **Détecter les fichiers au format PDF pour ne traiter qu'eux**

Récupération du nom de dossier fourni et parcours de tous les fichiers de ce dossier afin de vérifier un par un leur type. Création d'une liste qui ne va contenir seulement les noms des fichiers de type PDF.

La fonction est terminée quand : La fonction sait reconnaître si le type de fichier est un PDF ou non, qu'elle entre les noms de fichiers en PDF dans une liste et qu'elle empêche un bug du programme à cause de fichiers qui ne sont pas en PDF.

- **Création d'un fichier texte et écriture à l'intérieur**

Ajout d'un fichier au format txt portant le nom du PDF traité. 1 fichier txt par fichier PDF traité, pas de création de fichiers txt pour les autres. Récupération du nom du PDF d'origine pour le réutiliser

dans le nom du fichier créé. Ouverture du fichier en écriture, afin qu'il contienne les informations trouvées.

La fonction est terminée quand : Le fichier créé porte le bon nom, est bien au format texte et comporte les éléments demandés avec une présentation correcte et lisible.

- **Récupérer les résumés**

Récupération du texte situé après la fin du titre, après le mot abstract, s'il y en a un, et avant le premier titre ou l'introduction.

La fonction est terminée quand : Elle récupère tout ce qui est contenu dans l'abstract, sans qu'il ne manque des mots ou des phrases et sans d'une autre partie du PDF y soit intégré, comme le titre ou l'introduction.

- **Récupérer les noms des fichiers d'origine**

Récupération du nom du fichier dans le Corpus, en enlevant l'extension.

La fonction est terminée quand : Elle récupère le nom de fichier sans extension ni chemin d'accès, qu'il soit relatif ou absolu.

- **Récupérer les titres**

Récupération du titre, que l'on peut trouver en utilisant les balises html : <Title> </Title>, ou encore en récupérant le.s première.s lignes du PDF.

La fonction est terminée quand : Elle récupère le titre en entier, même quand il est réparti sur plusieurs lignes et qu'il n'y a pas de mots en trop.

- **Récupérer les auteurs**

Récupération de la liste des auteurs, que l'on peut trouver dans les données de la fonction pdfinfo, ou bien en récupérant ce qui se situe entre le titre et l'abstract, en trouvant des solutions pour enlever les informations supplémentaires s'il y en a.

La fonction est terminée quand : Elle récupère la liste des auteurs, sans manquement, et sans informations supplémentaires comme l'adresse mail ou l'université…

- **Création du dossier « results » contenant les sorties du parser**

Vérification de l'existence ou non d'un dossier « results » dans le dossier fourni en paramètre, s'il existe le supprimer et dans tous les cas, le créer. Une fois les fonctions de récupération terminées, les fichiers au format texte créés sont placés à l'intérieur.

La fonction est terminée quand : Le dossier de résultat est bien un sous dossier du dossier fourni en paramètre, qu'il ne comporte que les fichiers txt précédemment créés dans le programme.

Affichage à l'utilisateur des informations ou des problèmes, mauvaise utilisation de la commande, problème avec dossier ou fichier → Tous, dans nos fonctions quand elles fonctionnent

Mise en œuvre, dans chaque fonction, de la capacité à communiquer avec l'utilisateur.

- Explication du fonctionnement du parser si son appel est mal effectué
- Indications des erreurs/défauts rencontrés, paramètre donné qui est un dossier vide, fichiers dans le dossier qui ne sont pas des PDF
- Détail de l'action en cours, affichage du PDF en cours de traitement
- Indication du dossier qui donne accès aux fichiers parsés par le programme
- Indication de l'existence d'un fichier listant les noms des fichiers qui ne sont pas des PDF

La fonction est terminée quand : Les messages affichés suffisent à la compréhension de l'utilisateur d'une erreur au lancement du programme ou de la manière de consulter le résultat du programme.

- **Création d'un fichier listant les noms des fichiers du dossier non PDF**

Ajout d'un fichier au format txt, dans le même dossier que les autres fichiers txt créés pour lister les informations principales des PDF. Parcours de tous les fichiers du dossier passé en paramètre afin de détecter ceux qui ne sont pas des PDF et de récupérer leur nom, dans le but de les lister dans le fichier.

La fonction est terminée quand : Le fichier n'est créé que si besoin, autrement dit : si tous les documents sont des PDF, ce fichier n'a pas lieu d'exister. Le fichier doit également être compréhensible et ne lister seulement les noms des fichiers qui ne sont pas des PDF.

- **Mesure de la performance du programme**

Tests de la performance en temps d'exécution du programme avec différents scénarios : exécution du programme avec le dossier fourni par le client, un dossier contenant de nombreux fichiers et un dossier contenant des fichiers volumineux. Exécutions répétées de notre programme, mesurées avec la commande time pour en écrire un tableau récapitulatif des vitesses.

Incrément : Traçage de l'évolution du programme

03/02/2021 : Choix du langage de programmation

12/02/2021 : Vérification dossier

12/02/2021 : Vérification PDF

06/02/2021 : Création fichier texte contenant les informations triées

08/02/2021 : Récupération des résumés

06/02/2021 : Recherche des noms des fichiers

06/02/2021 : Récupération des titres

20/02/2021 : Récupération des auteurs

06/02/2021 : Création du dossier « results »

20/02/2021 : Communication à l'utilisateur des erreurs/informations

14/02/2021 : Lister les fichiers non PDF

20/02/2021 : Tests des performances

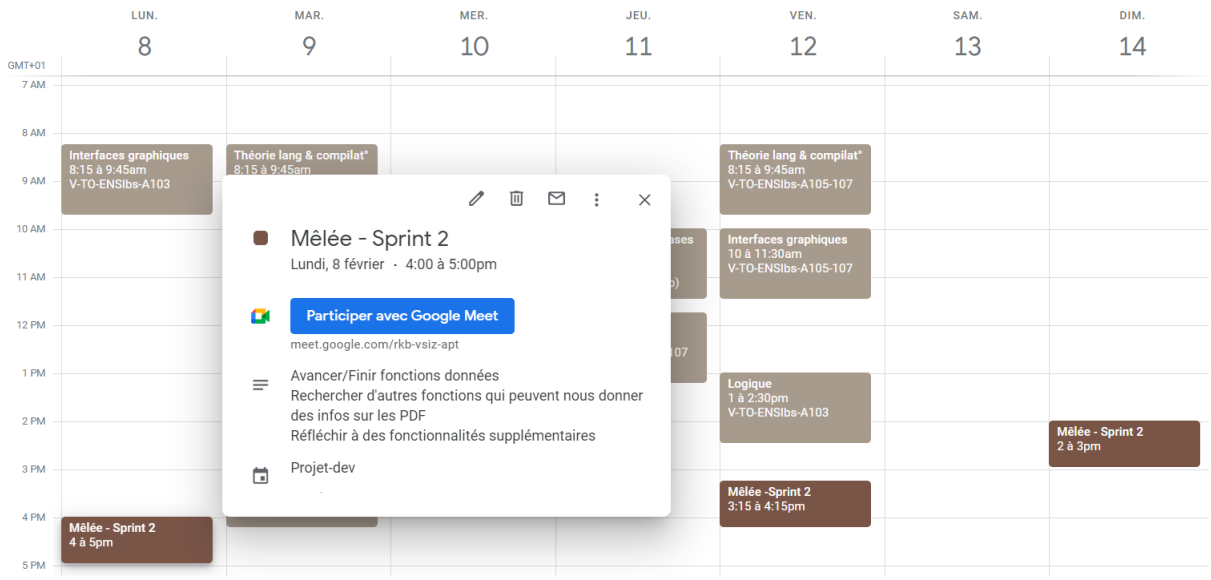
21/02/2021 : Rédaction du rapport de sprint et des tests de performance

Méthode de travail :

Maître artefact pour ce sprint

Mêlée :

Consulter sur l'agenda pour ne pas manquer une mêlée et cliquer sur l'évènement pour se rappeler des missions à réaliser d'ici là. Le lien vers la visio se trouve également en cliquant sur l'évènement. Avancer également sur les fonctions à faire et préparer ses questions ou ses problèmes pour les mettre en commun. Mise en commun de l'avancée et des éventuelles nouvelles propositions de fonctionnalités. Choix de la date de la prochaine mêlée puis déterminations des objectifs jusqu'à celle-ci et enfin rédaction de la conclusion de la mêlée.



Entre les mêlées :

Pour toute question ou information, passer par Messenger, possibilité de taguer une personne en particulier si la demande est précise. Possibilité de faire des vocaux pour se faire comprendre plus facilement. Possibilité d'envoyer des photos, pour montrer une erreur ou un défaut.

GitHub :

Pour chaque nouvelle fonction terminée, l'ajouter à la version précédente dans une nouvelle version afin de pouvoir revenir consulter le programme précédent. Prévenir les autres d'un nouvel ajout pour qu'ils testent le programme chez eux et puissent réfléchir à des améliorations.

Conclusions mêlées :

Début Sprint 2 : Lundi 1 février

Mêlée : 21/02/2021

- Finalisation des tests de performance
- Rédaction du rapport

Mêlée : 20/02/2021

- Point sur les fonctions qui marchent : **la recherche des auteurs, les affichages des informations pour communiquer avec l'utilisateur**
- Décision de la manière de faire les tests

Mêlée non prévue : 18/02/2021

- Point non prévu sur la récupération des auteurs qui présentent quelques bugs, pas tous les auteurs, ou noms non présents sur le PDF
- Réflexion sur les solutions pour réparer ces erreurs

Mêlée : 17/02/2021

- Point sur les fonctions qui marchent : **la recherche des auteurs**
- Réflexion sur les pistes abordées pour tester l'efficacité

Mêlée : 14/02/2021

- Point sur les fonctions qui marchent : **La création de la liste des fichiers non PDF**
- Point sur l'avancement de la recherche des auteurs
- Commencer à réfléchir aux moyens de tester l'efficacité

Mêlée : 12/02/2021

- Point sur les fonctions qui marchent : **la vérification PDF et dossier**
- Point sur les recherches des auteurs

Mêlée : 08/02/2021

- Point sur les fonctions qui marchent : **Récupération du résumé**
- Nouvelles idées à traiter : Faire un fichier texte pour lister les fichiers non PDF

Mêlée : 06/02/2021

- Point sur les fonctions qui marchent : **Le dossier result avec les .txt, récupération du titre et recherche des noms des fichiers**
- Tests et recherches des améliorations sur les fonctions terminées
- Nouvelles idées à traiter : Vérifier que le paramètre est un dossier et ses fichiers des PDF

Mêlée : 03/02/2021

- Mise en commun des compréhensions du sujet
- Choix du langage de programmation > **Python**
- Répartition des tâches

Mêlée : 30/01/2021

- Lecture du sujet, début de réflexion à poursuivre en solo