Quels outils et méthodes en SHS ?

Cycle *intelligence artificielle*
Atelier 3/3

jeudi 12 juin, 10h-12h Bâtiment de recherche Nord, salle 0.004

Mate-SHS Condorcet

Méthodes, Réseaux, Terrains, Enquêtes en sciences humaines et sociales (Mate-SHS)

Campus Condorcet, Paris-Aubervilliers (93)

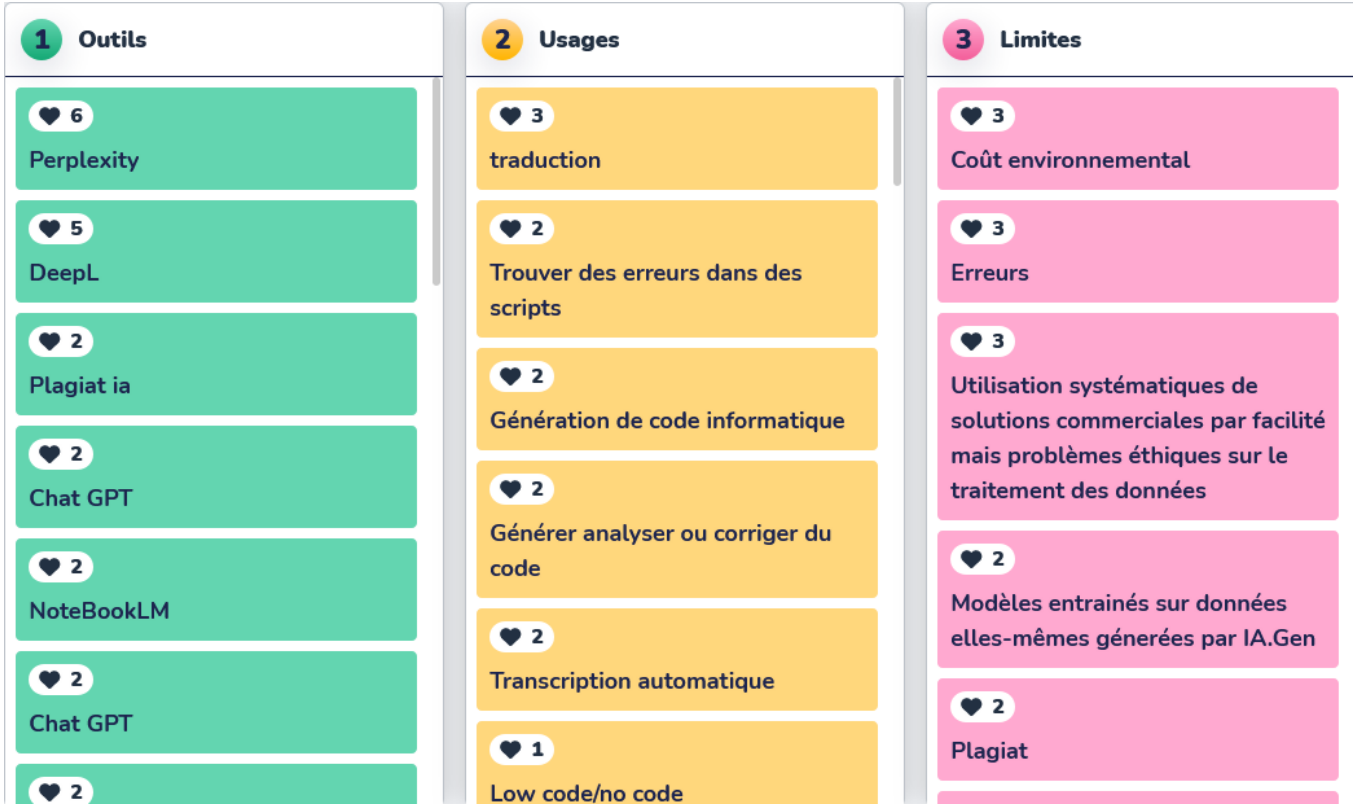
Liste de diffusion

mate-shs-condorcet.gitpages.huma-num.fr

Qu’est-ce qu’un « atelier-garage » ?

- Proposer une **discussion facilitée** sur un thème.
- **Échanger sur nos pratiques** : regarder « sous le capot » des outils utilisés.
- Partager **des cas d'usage et des ressources** dans un esprit de **mise en relation** et d’**entraide**.
- **Prendre notes** de nos échanges

Atelier 1 : Quel est l'impact de l'IA dans votre travail ? (avril)



Atelier 2 : Quels outils et méthodes en SHS ? (mai)

Svetlana Yatsyk (IRHT) *Transcription automatique de manuscrits*

Discussion : *Est-ce que l'HTR nous fait vraiment gagner du temps ?*

Lise Bernard (IRHT) : *Détection automatique d'images*

Discussion : *Qu'est-ce qu'un "bon" modèle pour un.e historien.ne ?*

Lionel Kesztenbaum (Ined) : *Reconnaissance d'écriture manuscrite*

Discussion : *Comment mesurer la qualité des données d'entrainement ?*

Bastien Sepulveda & Seïd Smatti (INALCO) : *Génération automatisée de contenus textuels : la plateforme LaCAS*

Atelier 3 : Quels outils et méthodes en SHS ? (juin)

Guillaume Rollat (Ined), *Retour d'expérience sur l'utilisation de l'IA à l'Ined – Regards des DPO*

Alexandre Wauthier (Humathèque), *Commande Regex pour une édition en Markdown + script Python pour le traitement de fichiers PDF*

François Gandolfi (Humathèque), approches simultanées de plusieurs API avec des outils no-code/low-code et ChatGPT, pour la prise en main du Java Script

Cynthia Pedroja (Humathèque), ChatGPT n'est pas bête, ce n'est qu'une machine !

Karim Noui (Humathèque), à partir de 11h, sous réserve

Commande Regex pour une édition en Markdown

Contexte :

- Copie d’un site HTML qui n’existe plus, hébergé par l’Institut d'histoire du temps présent (IHTP).
- Récupération de l’historique grâce à la *Wayback Machine* d’Internet Archive.
- Transformation des pages HTML en markdown
- Édition numérique et adaptation des pages ⟵
- Déploiement d’un nouveau site ⟵

Une Regex, c’est quoi ?

Une expression régulière (*regular expression*) est une séquence de caractères qui définit un motif (*pattern*) permettant de rechercher, d’extraire ou de modifier des chaînes de caractères spécifiques dans un texte.

Exemples d’utilisation :

- Rechercher des textes ou motifs particuliers (par exemple, tous les mots commençant par une majuscule),
- Remplacer ou supprimer certaines parties d’un texte (comme des balises HTML, des espaces, des retours à la ligne),
- Valider des formats (adresses mail, numéros de téléphone, etc.).

Prompt : « Je souhaite trouver une commande regex qui recherche les sauts de lignes et les remplace par un espace. »

- **Expression régulière :**

```text
(?<!\n)\n(?!\n)
```

- **Remplacer par :**

```text
(un espace)
```

Cette regex cible spécifiquement les sauts de ligne simples (pas ceux qui séparent des paragraphes, c'est-à-dire pas les doubles sauts de ligne), ce qui permet de conserver l'espacement entre les paragraphes tout en supprimant les retours à la ligne inutiles à l'intérieur des paragraphes.

Cette méthode reste parfaitement adaptée pour ton besoin de garder les paragraphes espacés tout en "aplatissant" les lignes à l'intérieur de chaque paragraphe.

La commande Regex : (?<!\n)\n(?!\n)

- (?<!\n)

assertion négative arrière (*negative lookbehind*).

« Il ne doit pas y avoir un saut de ligne juste avant ».

- \n

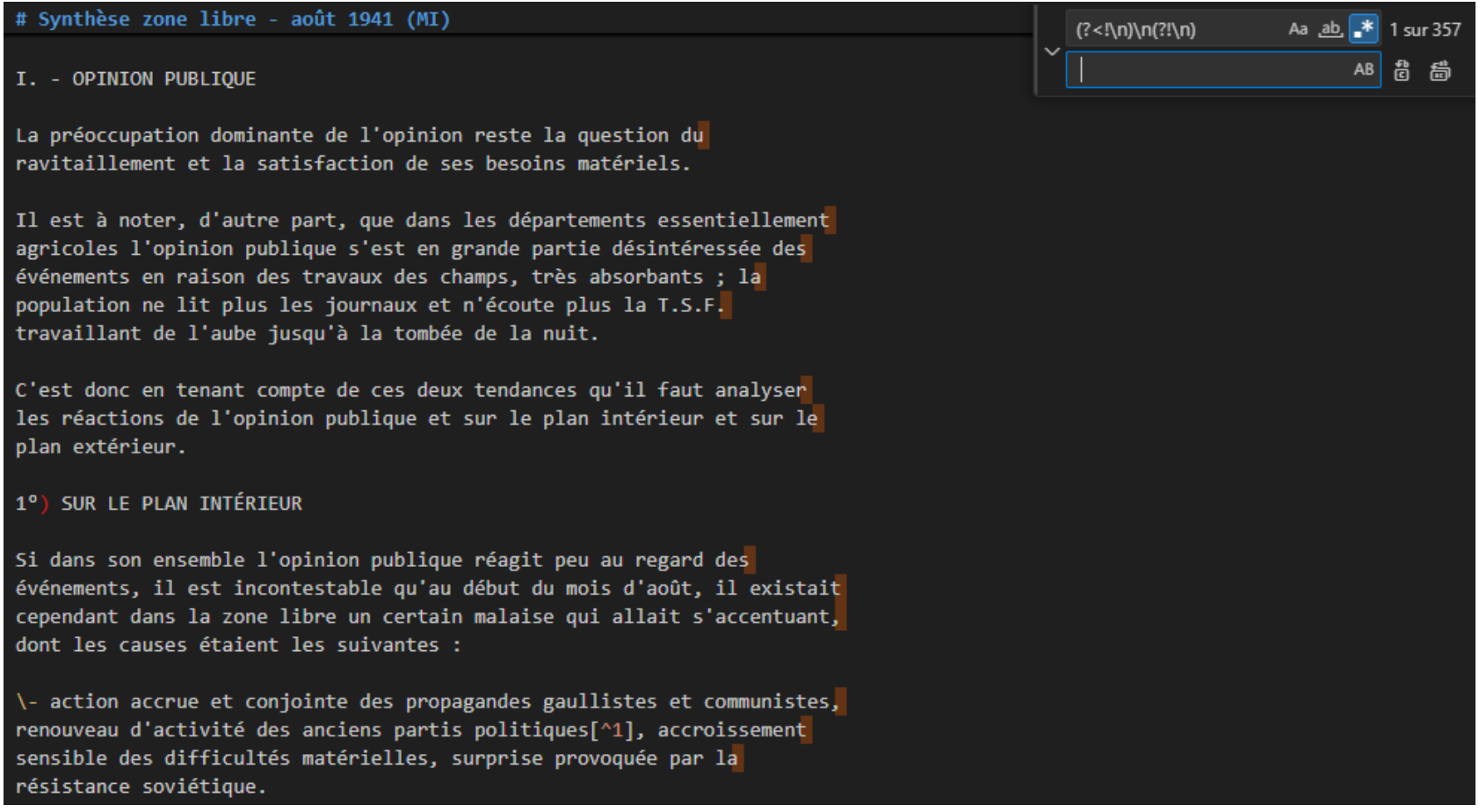
Représente un saut de ligne (caractère *newline*)

- (?!\n)

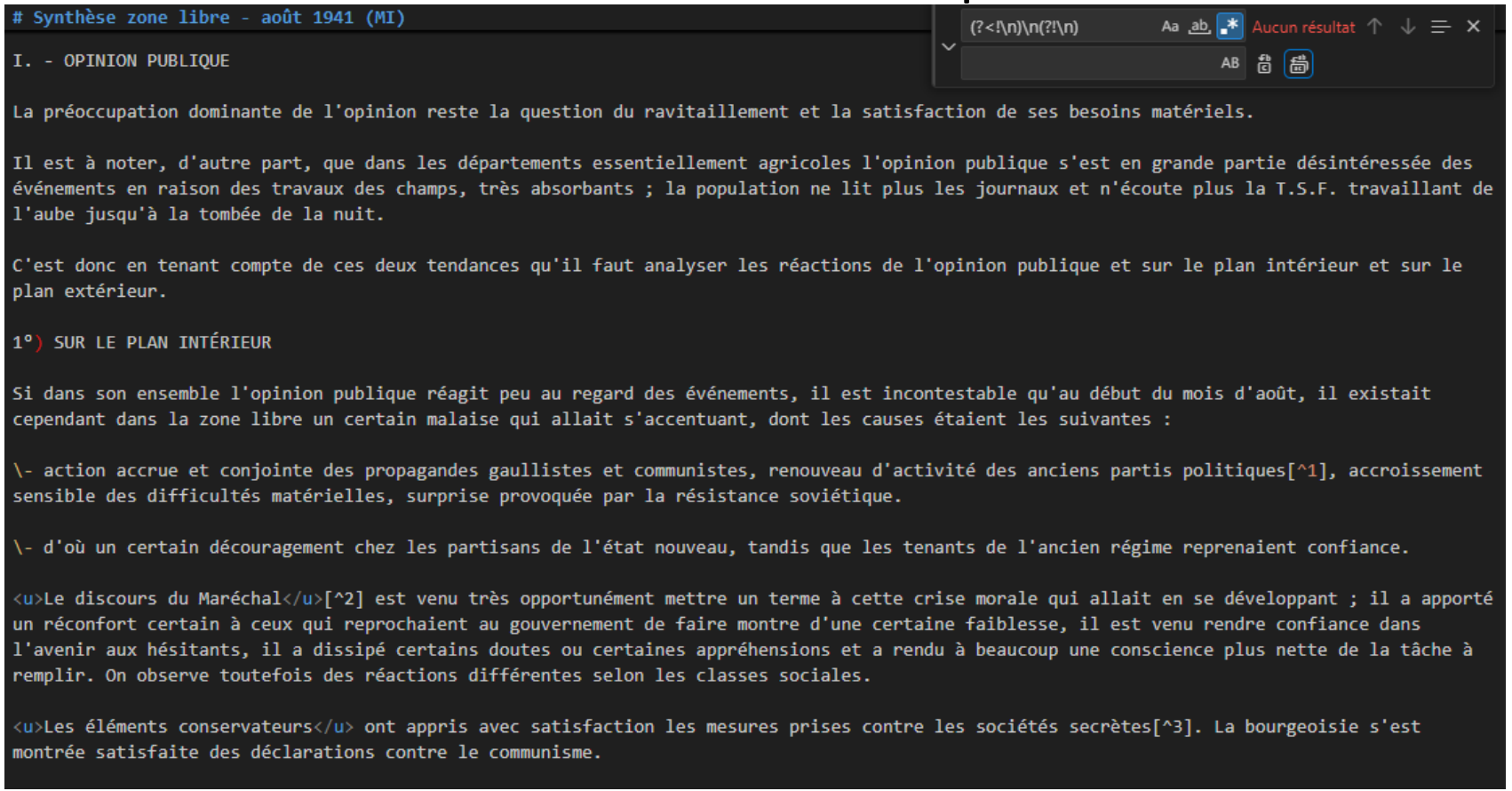
"assertion négative avant" (*negative lookahead*).

« Il ne doit pas y avoir un saut de ligne juste après ».

Démonstration sur VSCode : avant



Démonstration sur VSCode : après



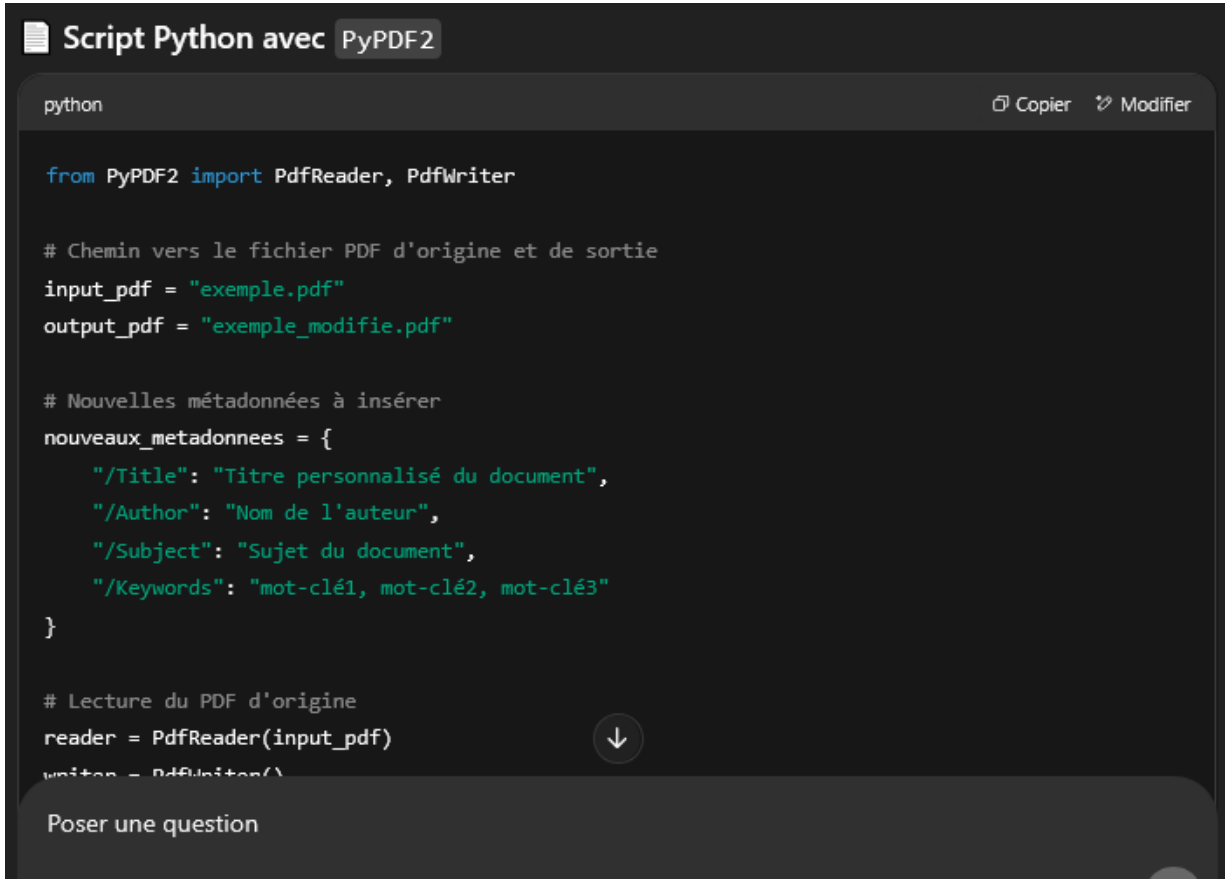

```
# Synthèse zone libre - août 1941 (MI)

I. - OPINION PUBLIQUE

La préoccupation dominante de l'opinion reste la question du ravitaillement et la satisfaction de ses besoins matériels.

Il est à noter, d'autre part, que dans les départements essentiellement agricoles l'opinion publique s'est en grande partie désintéressée des événements en raison des travaux des champs, très absorbants ; la population ne lit plus les journaux et n'écoute plus la T.S.F. travaillant de l'aube jusqu'à la tombée de la nuit.

C'est donc en tenant compte de ces deux tendances qu'il faut analyser les réactions de l'opinion publique et sur le plan intérieur et sur le plan extérieur.

1°) SUR LE PLAN INTÉRIEUR

Si dans son ensemble l'opinion publique réagit peu au regard des événements, il est incontestable qu'au début du mois d'août, il existait cependant dans la zone libre un certain malaise qui allait s'accentuant, dont les causes étaient les suivantes :

\- action accrue et conjointe des propagandes gaullistes et communistes, renouveau d'activité des anciens partis politiques[^1], accroissement sensible des difficultés matérielles, surprise provoquée par la résistance soviétique.

\- d'où un certain découragement chez les partisans de l'état nouveau, tandis que les tenants de l'ancien régime reprenaient confiance.

<u>Le discours du Maréchal</u>[^2] est venu très opportunément mettre un terme à cette crise morale qui allait en se développant ; il a apporté un réconfort certain à ceux qui reprochaient au gouvernement de faire montre d'une certaine faiblesse, il est venu rendre confiance dans l'avenir aux hésitants, il a dissipé certains doutes ou certaines appréhensions et a rendu à beaucoup une conscience plus nette de la tâche à remplir. On observe toutefois des réactions différentes selon les classes sociales.

<u>Les éléments conservateurs</u> ont appris avec satisfaction les mesures prises contre les sociétés secrètes[^3]. La bourgeoisie s'est montrée satisfaite des déclarations contre le communisme.
```

Script Python pour éditer les métadonnées embarquées de plusieurs fichiers PDF

Contexte :

- Dépôt de fichiers PDF dans Nakala.
- Métadonnées embarquées comme documentation supplémentaire des fichiers déposés
- Traitement d'environ 200 fichiers
- Transformation automatisée des fichiers PDF ⟵
- Utilisation de Python pour réaliser l'opération ⟵

Prompt : « J’ai besoin d’un script pour éditer le titre, l'auteur, le sujet et les mots-clés d'un PDF. Les nouvelles métadonnées sont inclues dans le script Python »



Les limites de ChatGPT en version gratuite

Les mises à jour de librairies Python après 2022

Source : documentation PyPDF, https://pypdf.readthedocs.io/en/latest/meta/history.html

Finalement, le script fonctionne après un paramétrage manuel

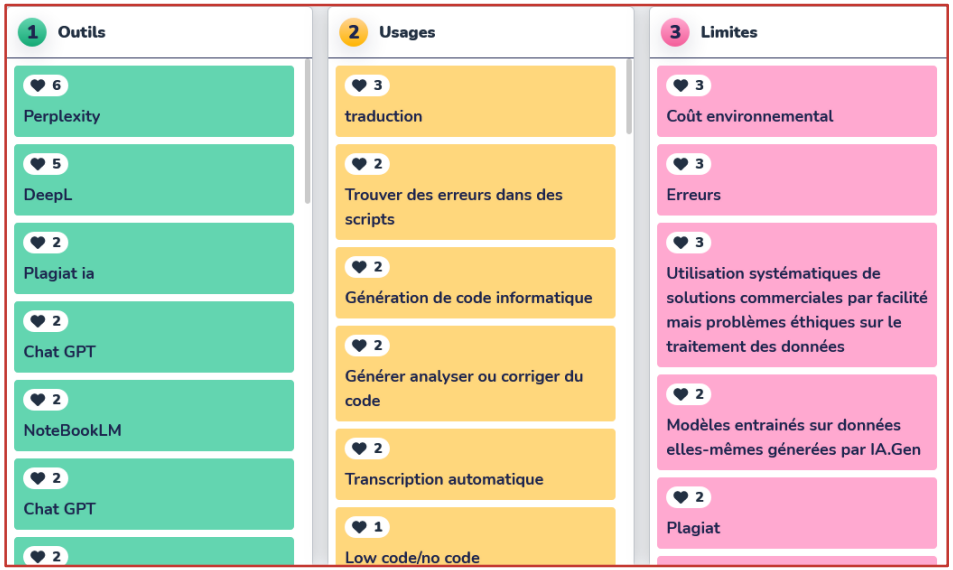
```python
import os
import re
from pypdf import PdfReader, PdfWriter

def extract_number(filename):
    """Extract the numeric part of the filename for sorting."""
    match = re.search(r'(\d+)', filename)
    return int(match.group(1)) if match else float('inf')  # Mettre les fichiers

def change_pdf_metadata(directory, new_metadata):
    # List all PDF files in the directory
    pdf_files = [f for f in os.listdir(directory) if f.endswith('.pdf')]
    # Sort files by their numeric part
    pdf_files.sort(key=extract_number)

    if len(pdf_files) != len(new_metadata):
        print("The number of PDF files and new metadata entries does not match.")
        return
```

Modification de la librairie appelée :
pypdf au lieu de PyPDF2

Résultat obtenu :
le script fonctionne

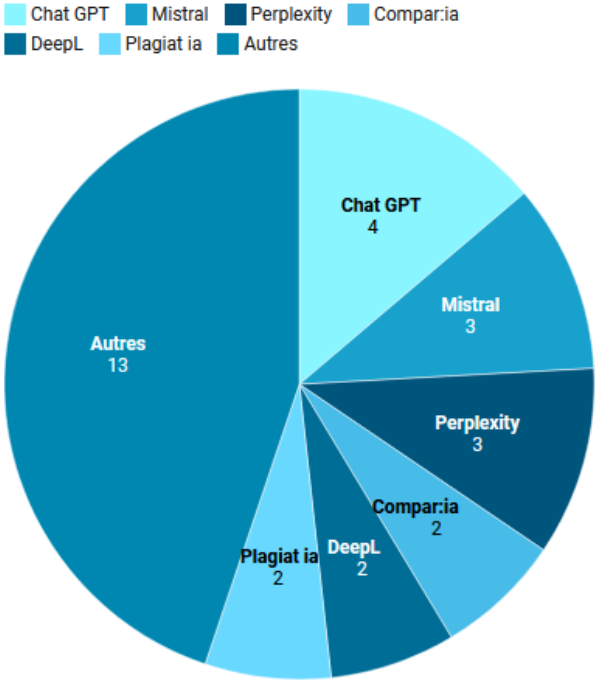
```
Changed metadata of REVE_FMA_A_P1.pdf to Title: Brahim, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P2.pdf to Title: Amanar, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P3.pdf to Title: Kadda, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P4.pdf to Title: Abdelrazak, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P5_6.pdf to Title: Mustafa, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P7.pdf to Title: Taher, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P8.pdf to Title: Mebarek, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P9.pdf to Title: Tahar, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P10.pdf to Title: Rabah, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P11.pdf to Title: Achour, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P12  Amar.pdf to Title: Amar, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P13 Miloud.pdf to Title: Miloud, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P14 Arezki.pdf to Title: Arezki, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P15 Rahman.pdf to Title: Rahman, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P16 Hamlaoui.pdf to Title: Hamlaoui, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P17 Lamine.pdf to Title: Lamine, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P18 Youcef.pdf to Title: Youcef, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P19 Boumediene.pdf to Title: Boumediene, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P20 Haadi.pdf to Title: Haadi, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P21 Boualem.pdf to Title: Boualem, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P22 Belkacem.pdf to Title: Belkacem, parcours biographique, Author: Mathias Gardet
Changed metadata of REVE_FMA_A_P55 Mohand.pdf to Title: Mohand, parcours biographique, Author: Mathias Gardet
```

Réponses uniformisées Wooclap (avril 2025)



- 29 réponses « Outils »
- 54 réponses « Usages »
- 27 réponses « Limites »
- = 110 réponses au total

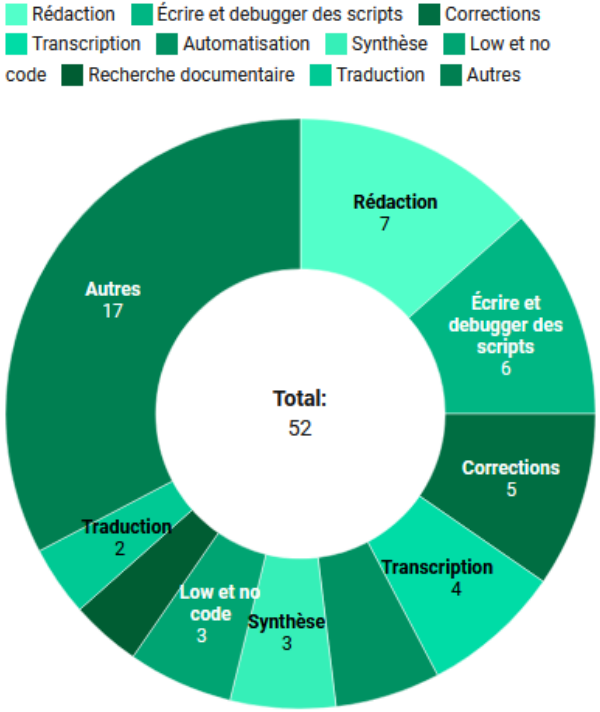
Outils



Réponses homogénéisées de la catégorie "Usages" (Wooclap atelier-garage, avril 2025)

Source: Mate-SHS Condorcet, cycle IA • Récupérer les données • Créé avec Datawrapper

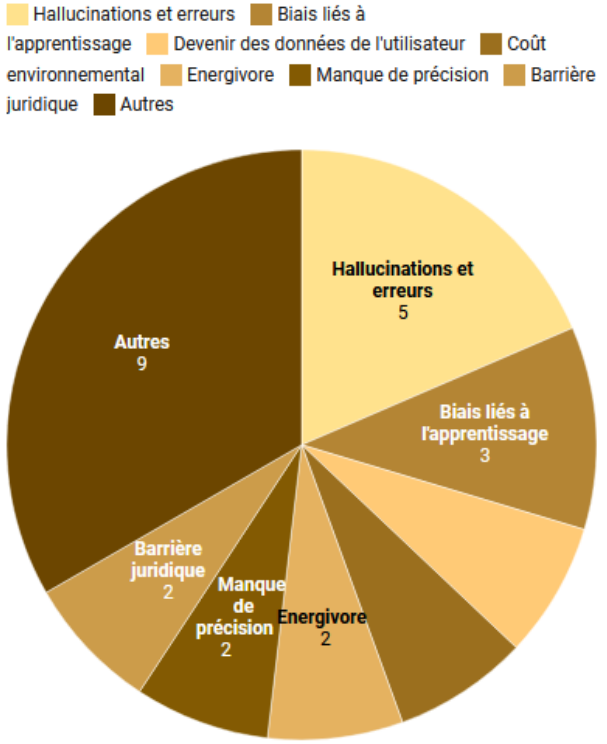
Usages



Réponses homogénéisées de la catégorie "Usages" (Wooclap atelier-garage, avril 2025)

Source: Mate-SHS Condorcet, cycle IA • Récupérer les données • Créé avec Datawrapper

Limites



Réponses homogénéisées de la catégorie "Limites" (Wooclap atelier-garage, avril 2025)

Source: Mate-SHS Condorcet, cycle IA • Récupérer les données • Créé avec Datawrapper

https://www.datawrapper.de/_/9vjsC https://www.datawrapper.de/_/YCtPO https://www.datawrapper.de/_/xN0J8