**Notes du 11/04 (RDZ No.5)**

**Chaque fois, avant la prochaine réunion RDZ, consacrez deux à trois jours par semaine pour accomplir le travail et l'améliorer.**

**Pour la mise en commun:**

* **Mardi 20h-22h**
* **Mercredi 14h-16h**
* **Vendredi 9-12h, si nécessaire**

**Avant chaque mise en commun, c’est le moment de travailler séparément**

**Après chaque RDV, il faut discuter ensemble pendant 30mn-1h pour décider ce qu’on va faire prochainement.**

**Les nouveaux conseils donnés par prof:**

* **可能要更细致地分配和完成任务，我要在overleaf上开始写东西了**
* **这次RDZ 结束以后，要为12/05的RDZ做准备，开完估计就要等到考完期末考试的一周，5.24-5.31弄了**
* **tout est sur github et envoyer tous les documents en avant**
* **论文和代码都要往前**
* **Caroline 2周以后有事**
* **faire un plan de rédaction**
* **contacter Monsieur Gael pour le rdz 0425**

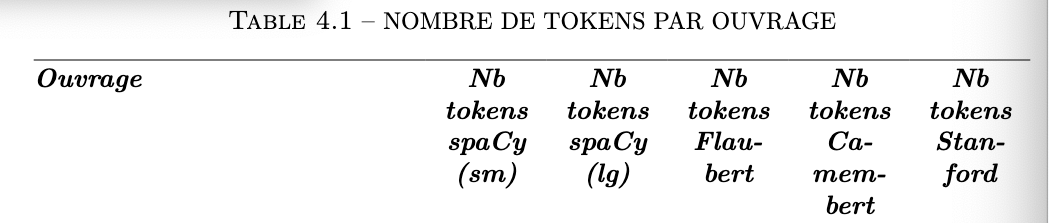
**Tableau:**

faire des opérations dans le python

**avec l’aide de GPT:**

dictionnaire qui enregistre clé-valeur, mettre tout dans un dictionnaire

le dictionnaire est plus pratique, faire le json dictionnaire,et à partir de ça, faire des calculs



“Na” pour tester

laisser les chiffres de faux, c’est faux

garder le nm de tokens par ouvrage n’est pas nécessaire pour chaque modèle

gérer des tableaux mais pas des résultats dans la sortie, i y a la librairie pour faire ça

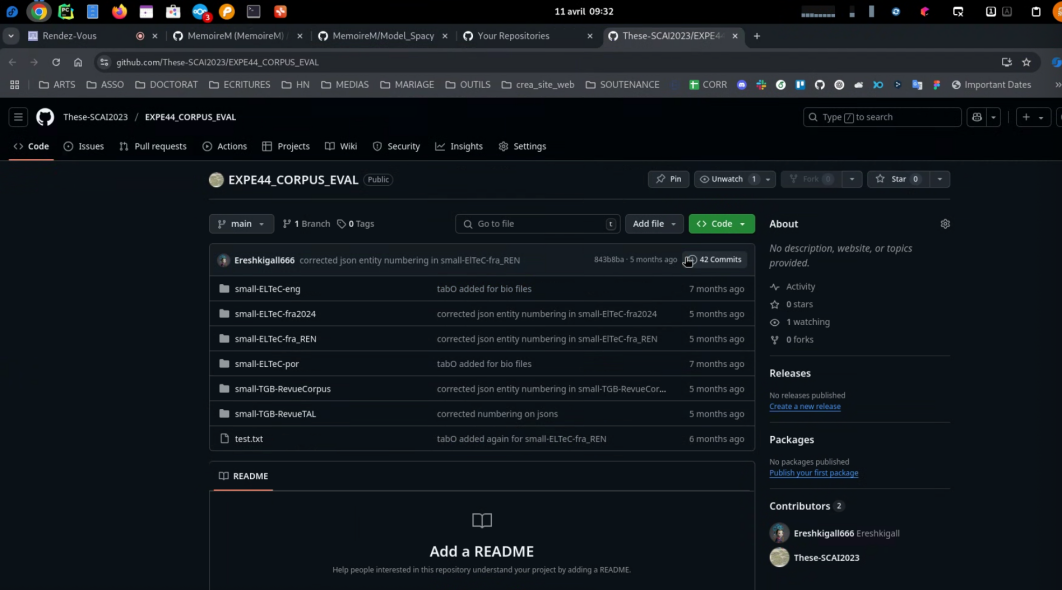
**Github:**

il faut être accessible

ce dépôt est très nécessaire chaque fois

ce grand dépôt: code plus mémoire

github的命名



“prog” pour porgramme

partage le compte de github avec une autre personne

plein de données et de dossiers, se concentrer dans le github

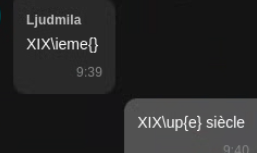
quel nom donné pour le dépôt: dossier,sous-dossier

**Mémoire :下一阶段的重点**

Le lien d’overlearf : **<https://www.overleaf.com/project/67b47266b94c349f24135b3d>**

il est temps de présenter les outils

XIX\up{e} siècle

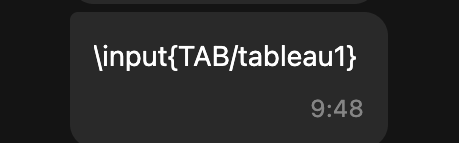


régler le problème d’affichage des bibliographies sur overleaf

论文的结构参考Gael在课上讲的东西

\Tableau

\Image



**Les choses importantes:**

* En fait, votre bibliographie, vous ne la rendez pas, enfin, elle ne s'affiche pas du tout, parce que justement, il y a ce problème d'import. Pour l'instant, il n'existe pas. Attendez, le défi, c'est que le ... commenter la bibliographie ou mettre en commentaire... Oui voilà, pour l'instant, pour que vous puissiez au moins afficher correctement et après vous devez nommer bien votre fichier de bibliographie que vous appelez et ensuite ça va fonctionner. 其实，你们的参考文献并没有被呈现出来，完全没有显示出来，因为正是存在一个导入的问题。目前它还不存在。等一下，问题是……你们可以选择把参考文献注释掉，或者把那部分内容用注释符号标起来……对，就是这样。这样至少你们现在可以让页面正常显示。之后，你们需要把你们调用的参考文献文件命名正确，那样它才会正常运行。
* Il faut créer, si vous voulez mettre une bibliographie, il faut créer le fichier qui s'appelle bibliographie.com. Mais normalement, il n'existe pas, c'est pour ça que ça m'est satiré en fait. Je me permets, parce que là, il vous a... Ok, c'est bon maintenant pour moi.如果你们想要添加参考文献，那就必须创建一个名为 bibliographie.com 的文件。不过现在它其实并不存在，这也是为什么会报错的原因。
* Bien, bien, bien. Je pense que ce n'est pas ça que je vous ai demandé, ce n'est pas de faire des opérations entre tableau et la tec, c'est de faire des opérations sous Python, ou des listes de chiffres, ou entre des dictionnaires de chiffres, mais déjà, nous, on ne travaille jamais en faisant des opérations entre tableau et la tec, je ne sais pas si ça me paraît bizarre de faire comme ça. Oui, et j'ai cherché après sur Excel, c'est possible entre tableau différent sur la tec, je ne sais pas. Là, on n'utilise pas Excel, on utilise Excel en anglais.我觉得这不是我让你做的，我并没有让你用 LaTeX 里的表格之间做运算。而是让你在 Python 里进行操作 —— 比如用数字列表或者字典来做运算。我们从来不会在 LaTeX 的表格之间进行操作，这样做听起来就很奇怪。”“对，我后来查了一下，Excel 的确可以在不同表格之间做计算，但 LaTeX 我就不知道了。”“可问题是我们现在并不是在用 Excel，而是用 Python，而且我们这个课是‘信息开发课程’。”
* On est dans un cours de développement informatique. On est dans le cadre du développement informatique avec Python. On n'est pas dans le cadre de... on utilise... il faut passer à l'étape où maintenant on fait du développement informatique et on travaille en faisant de la programmation, de l'algorithmique, en utilisant l'algorithmique et la programmation Python. Il n'y a pas de solution du côté d'Excel en fait.“我们现在是在做信息开发，用 Python 做开发。我们不是在做办公自动化。你需要进入一个新阶段，那就是开始进行程序开发和算法设计，用 Python 的编程和算法来处理问题。Excel 这边是没有解决方案的
* Ça c'est peut-être pour d'autres cours, mais pour le mémoire, on est dans un mémoire de traitement automatique du langage avec de l'algorithme et de la programmation. Donc votre problématique, en fait, elle n'est pas pertinente dans le cadre de vos travaux. Donc ce qu'il faut faire là, c'est que vous ayez en entrée des listes ou des dictionnaires, ce serait mieux de garder des dictionnaires, parce que dans un dictionnaire, vous avez la possibilité de mettre des clés qui deviennent les métazonées de vos données, et donc ensuite de faire les calculs et de faire en sous ces données, soit dans chaque dictionnaire, soit dans un même dictionnaire,vous avez trois clés primaires, une avec le nombre de tokens par ouvrage, une avec le nombre d'entités par ouvrage. Vous faites deux clés primaires, et ensuite, vous avez des clés qui seraient les noms des auteurs, ce serait plus simple. et en bas vous avez les chiffres... Après pour considérer le nombre de tokens, je trouve ça pas forcément super utile. Alors là par exemple déjà il y a des chiffres qui sont bizarres dans votre tableau 4.1,vous n'avez pas remarqué que c'est bizarre les chiffres que vous obtenez comme nombre de tokens avec camembert et Stanford ?这可能是其他课程的内容，但对于您的论文来说，我们是在进行自动语言处理的研究，涉及算法和编程。所以，您的问题在这方面其实并不相关。现在需要做的是，您输入的数据应该是列表或字典，最好是使用字典，因为在字典中，您可以设置键值，这些键可以作为数据的元数据，然后您可以在这些数据中进行计算，可能是在每个字典中进行，或者将它们放在一个字典中进行。您有三个主键：一个是每个作品的token数量，一个是每个作品的命名实体数量。您可以设置两个主键，然后再添加作者的名称作为键，这样会更简单。然后，在底部，您将有数字。至于token的数量，我觉得这并不是特别有用。比如，在您的表格4.1中，已经有一些奇怪的数字。你没有注意到使用 Camembert 和 Stanford 模型时，得到的 token 数字很奇怪吗？
* Je pense que là c'est un peu ce qu'on a dit la dernière fois, qu'il fallait que vous regardiez si ça prenait bien tout le texte en entrée. Là clairement on voit que ça ne prend pas tout le texte en entrée, c'est pas possible qu'on ait que 17 tokens d'un côté et de l'autre côté 55 910. Ou alors d'un côté ça a calculé les caractères et de l'autre côté les tokens, je ne sais pas.Mais même dans 17 tokens on n'ai pas complété ta bande d'artistes.我认为这有点像我们上次说的，应该检查一下是否正确地处理了所有的输入文本。显然，我们可以看到它没有处理全部的文本，17个 token 和 55,910个 token 之间明显不可能是这样。要么一方计算的是字符数，另一方计算的是 token 数量，我也不确定。
* Il faut bien mettre dans ces cas là, vous mettez N A en mathuscule.Voilà, quand vous voulez juste tester ou quoi, vous mettez N si vous êtes sûr que vous avez mis A. Parce que là, c'est un risque de 1, que vous montrez ça, donc moi, je m'appuie là-dessus, je me mets là aussi, je pense.在这种情况下，应该将其写为大写的"N A"。当你只是测试时，应该写"N"如果你确定已经写了"A"。因为这样的话，可能会有一个风险，所以我也在这里依赖这个，我想我也会这么做。
* Et de 2, c'est un risque de laisser ces chiffres qui sont faux. Alors que si vous avez juste pas de chiffres, il faut mettre NA. Et en plus j'allais venir au fait que, je ne sais pas si pour le nombre de tokens par ouvrage, c'est vraiment très très nécessaire de garder pour chaque outil en fait. On va garder SpaceySM ou SpaceyG, peu importe. Vous en gardez un ou Flaubert si vous préférez. Et puis voilà, parce qu'en fait, je ne vais pas pour le calcul de la densité.而且第二点是，留下这些错误的数字是有风险的。相反，如果没有数字，应该写“NA”。另外，我还想提到，对于每个工具计算的每个作品的token数量，是否真的有必要保留每个工具的结果。我建议保留SpaceySM或者SpaceyG，随便选择一个，或者如果你更喜欢Flaubert的话，也可以保留。就这样，因为实际上，我并不需要所有工具的数据来计算密度。
* Quoique non, après vous faites densité par outils.Ouais, non, ok, je vais vous dire ce que j'ai dit.Si en vrai, on s'avait dit qu'on devrait faire ça, parce que comme Flaubert, on n'était pas sûr s'il prenait tous les œuvres. C'était pour vérifier ça.虽然，之后你们可以根据工具来计算密度。是的，没问题，我会告诉你们我说的。如果实际上，我们之前说过应该这么做，因为Flaubert我们不确定它是否涵盖了所有的作品。那是为了验证这一点。
* Ok, ça marche. Après, sur la question du calcul, comme je vous disais, ben non, il ne faut pas prendre une solution avec Excel. En tout cas, ça ne sert à rien de conserver le format tableau LaTeX aussi. Faites un dictionnaire dans lequel vous avez vous donné et vous verrez, vous pourrez très bien, à vos décrits francs, faire des calculs.关于计算问题，正如我之前所说的，确实不应该使用Excel解决这个问题。至少，保留LaTeX表格格式没有意义。你们可以创建一个字典，其中包含你们的数据，然后就可以很方便地用Python进行计算。
* Oui, oui, bien sûr. Peut-être que j'ai mal compris, mais j'avais compris avec monsieur Léjeune qu'il fallait faire le calcul avec le latex.是的，当然。也许我理解错了，但我之前理解的是，Léjeune先生说过需要用LaTeX来进行计算。
* Et c'est ça que je n de toute façon, ces calculs-là que j'ai déjà, je ne les ai pas calculés manuellement. Ce que vous avez peut-être demandé, c'était de générer les tableaux, les formats tableaux automatiquement, c'est-à-dire pas de générer les calculs avec la tech, mais de générer les tableaux.而正是因为这些计算，我实际上并没有手动计算这些内容。你们可能要求的是自动生成表格的格式，而不是用LaTeX进行计算，而是生成表格。
* Oui, je pense que monsieur Lejeune, il avait pensé, en fait quand vous cherchez Gizont ou la tech, je sais pas si vous avez un logiciel très, très, très spécifique mais... voilà.是的，我认为Léjeune先生的意思是，当你们使用Gizont或者LaTeX时，我不确定你们是否使用了非常非常特定的软件，但就是这样。
* Tu sais qu'il y a des librairies qui permettent de faire ça ou sinon moi je passais par TLGPT. Oui oui. Je donnais mes dictionnaires, j'avais des dictionnaires stockés au format JSON. Et je donnais à TLGPT en lui disant les clés primaires du dictionnaire deviennent les noms des lignes, des colonnes.你知道有一些库可以做到这一点，或者我也可以通过GPT来实现。是的，没错。我把我的字典存储为JSON格式，然后将字典传给GPT，告诉它字典中的主键应该成为行和列的名称。
* Les clés secondaires deviennent les noms des lignes ou des colonnes, selon ce que je voulais faire. Et classe-moi les résultats dans un tableau en latin. Ça me fait très bien aussi si vous avez du mal à générer les tableaux. Voilà, je pense que ce que vous savez comprendre, c'est générer les tableaux en latin, mais pas générer les résultats.字典中的次级键会变成行名或列名，取决于我想做什么。然后把结果整理成拉丁文的表格。如果你在生成表格方面遇到困难，我也觉得通过这种方式可以很好地解决。总的来说，我认为你需要理解的是生成表格，而不是生成结果。
* tu peux juste répéter, comme tu disais, par rapport à la clé primaire et secondaire, dans le contexte de ce travail.关于主键和次键，在这个工作的背景下

Dans le dictionnaire, tu peux avoir plusieurs niveaux de clés, tu peux avoir un dictionnaire de dictionnaire.

在字典中，你可以有多个层级的键，可以有字典的字典。

Le côté primaire, c'est les premières clés.

主键是指第一个层级的键。  
Tu as un dictionnaire dans un dictionnaire.

你有一个字典嵌套在另一个字典里面

Et après, les clés secondaires. Mais je n'ai jamais lu ça, c'est moi qui appelle ça comme ça. Parce que je m'appuie sur la manière de parler en base de données. Ah oui, oui. C'est juste pour ça, parce que je ne sais pas, après, vraiment, j'ai pas vu un vocabulaire pour appeler ça. Alors je me suis dit, je vais aller peut-être le vocabulaire là. Et le JPT a l'air de comprendre.次键是第二个层级的键。但我从来没有读过这样的用法，这是我自己这么叫的，因为我是基于数据库的术语来理解的。哦，是的，没错。这只是因为我不知道该怎么称呼它，所以我用了这个词汇。而且 GPT 似乎能理解。

* Quand je lui ai dit le dictionnaire, il comprend que c'est le premier niveau de clé. C'est les clés secondaires, il comprend que j'avais l'impression que la GPT comprenait ça et que ça marchait.当我说到字典时，他明白那是第一层级的键。而次键，他理解我想表达的是 GPT 理解了这个概念，而且它有效。
* Mais tu voulais dire qu'ici, donc en l'oc'on avait des clés primaires qui seraient les noms des auteurs. Donc ça serait le baron d'auteur. Et après à l'intérieur de chaque auteur, on aurait un nouveau dictionnaire avec comme clé à chaque fois, qui seraient donc les clés secondaires, nombre de tokens pour Space CLG, nombre de tokens pour Space SM, nombre de tokens for Ber, nombre de tokens calendaires,nombre de tokens stem.在这里，我们有主键，那就是作者的名字。所以它会是“作者名”。然后在每个作者内部，我们会有一个新的字典，里面每个键就是次键，例如 Space CLG 的 token 数量，Space SM 的 token 数量，Flaubert 的 token 数量，以及 Camembert 的 token 数量。
* Donc là, ça fait un dictionnaire de dictionnaires qui est plutôt facile à gérer après, que ce soit pour faire des calculs, ou pour faire... De plus, si on prend les clés par hauteur, après on peut ranger toutes les données concernant les tokens et les entités qu'on met. C'est-à-dire, on peut avoir token spacy, token flobert, token calombert, et en spacy, en flobert, en calombert. Et ça nous fait les... Attends, 3, 4, 5... ça nous fait 10 clés par hauteur. Et après on peut demander à faire le calcul entre token space CSM et entité space CSM pour avoir la densité.这样就形成了一个字典中的字典结构，之后会更容易处理，无论是进行计算，还是进行其他操作。此外，如果我们按层次提取键值，我们可以把所有关于 token 和实体的数据都整理好。也就是说，我们可以有 spacy 的 token，Flaubert 的 token，Camembert 的 token，分别为 spacy、Flaubert 和 Camembert 的实体数据。这样我们就会有 10 个键，每一层都有这些数据。之后，我们可以请求计算 spacy CSM 的 token 和 spacy CSM 的实体数目来得到密度。
* Et tu vois Marina, ça simplifie quand mêmeil faudrait complètement changer la mise en page de ce tableau parce que... Non, pas changer la mise en page du tableau, là je parle d'un dictionnaire. Là je parle du dictionnaire. Il n'y a pas besoin de changer la mise en page de ce tableau, parce qu'après dans le dictionnaire tu mets ce que tu veux. Tu lui dis, pour le tableau nombre de tokens, tu lui dis, prendre en ligne les clés primaires pour chaque auteur, et en clé secondaire pour chaque colonne,tu leur dis dans le cadre du nombre de tokens, les clés dans lesquelles il y a un token. Et comme ça il y a les cartes saines dans lesquelles il y a entité. C'est juste qu'il faut réfléchir en amont, il faut juste réfléchir aux données qu'on veut et bien mettre les bons noms de métadonnées en gros, parce que les clés c'est les métadonnées des données. Et après c'est hyper simple à utiliser, puisque si on a tout bien nommé, et qu'on a bien nos données rangées dans un dictionnaire,c'est pour ça qu'en liste c'est plus compliqué d'obtenir ce genre de chiffres. Le dictionnaire c'est plus compliqué d'obtenir ce genre de chiffres. Le dictionnaire, c'est vraiment plus pratique. Mais parce que pour l'instant, votre format de données, c'est quoi concrètement ? Oui, c'est des dictionnaires JSON. Mais même si on envoie les tableaux, la texte, si le chat JPP comprend, il fait la densité.你看，Marina，这样确实简化了很多。其实不需要完全改变这个表格的布局，我在说的是字典的结构。你不需要更改表格的布局，因为在字典里，你可以存放你想要的东西。你只需告诉它，对于“tokens数量”这个表格，你让每个作者的主键（keys）作为行，然后让每列的二级键是每个 token 对应的列。这样就可以得出哪些键包含 token，以及哪些键包含实体。关键是提前思考清楚你想要的数据，并且确保正确命名元数据（metadata），因为键本身就是数据的元数据。只要我们正确命名，并且把数据整理在字典中，使用起来就非常简单。因为如果我们使用列表，它会更复杂，无法很容易地获取这些数字。字典方式更方便。但现在，你的数据格式是什么呢？对，确实是 JSON 格式的字典。即使我们将表格转换成 LaTeX 格式，如果使用 Chat GPT，它也能计算出密度。
* Oui, les calculs dans le chat GPT, ça ne montre pas que l'algorithme vous pourrait servir. Oui, oui, oui. sur le usage de l'algorithme. C'est-à-dire que faire le fichier Gisane, le dictionnaire, comme je l'ai décrit tout à l'heure, puis faire les calculs à partir de ce dictionnaire de manière algorithmique, en utilisant Python,puis après demander à l'GPT de générer le tableau avec vos résultats aux bons endroits, ce n'est pas la même chose que de donner à ChatsGPT les tableaux LaTeX ou Excel et lui dire calcule-moi le truc. Oui, complètement. Là, en fait, vous êtes vraiment dans le cadre de je dois montrer que je sais programmer des choses.是的，使用 ChatGPT 进行计算并不会展示算法的实际用途。是的，是的。关于算法的使用，意思是你先按照我之前描述的方式制作 Gisane 文件和字典，然后通过 Python 编写算法来进行计算，接着让 ChatGPT 生成带有你计算结果的表格，这与直接将 LaTeX 或 Excel 表格交给 ChatGPT 并要求它计算是完全不同的。完全正确。实际上，在这种情况下，你是在展示自己能够编写程序进行操作。
* Donc, ça n'empêche qu'on peut utiliser le chat GPT pour des tâches un peu répétitives annexes qui ne sont pas facilement généralisables. Après c'est faisable de faire des tableaux en mathèque à partir d'un script Pythagore. C'est juste que ça demande des fois, moi c'est pour ça que je ne l'ai pas fait pour ma thèse par exemple, parce qu'à chaque fois ça demandait de réadapter un peu le code parce que je ne faisais pas exactement les mêmes tableaux, etc. Si on sait qu'on fait à chaque fois exactement le même tableau,c'est ce qu'il y a de mieux.尽管如此，我们仍然可以利用 ChatGPT 来处理一些重复性的任务，这些任务不容易被普遍化。之后，通过 Python 脚本生成 LaTeX 表格是可行的。只不过有时候这会需要调整代码，因为每次生成的表格不完全相同。如果我们知道每次都要生成完全相同的表格，那确实是最好的方式。
* En parlant de la programmation, vous, Ruchi, j'ai vu que vous aviez envoyé par mail les codes et le drive, donc encore une fois, c'est ok, mais il faut quand même penser à passer sur GitHub, parce que sur GitHub, vous allez déposer toutes vos données, plus vos scripts. Donc je pense que ça rejoint ce que tu dis Karoline, dans le sens où vous devez montrer que vous savez programmer. Donc si on voit le code sur GitHub, alors ça c'est déjà valorisant pour vous, et ça donne le poids à votre travail. Si vous faites hypothétiquement, si vous faites vos calculs dans le chat GPT,alors ça vous coupe les pieds, entre guillemets. Donc il faut vraiment faire un effort pour faire de la programmation et ensuite mettre en avant, mettre en avant ses résultats, ses codes je veux dire. Donc pas seulement les résultats et le mémoire, mais on doit savoir que derrière le mémoire, il y a une base très solide qui est par exemple votre dépôt GitHub. La plupart du temps c'est ça ce que les chercheurs utilisent. Ils déposent toutes les données, tout le matériel nécessaire pour reproduire les expériences,puisque vous travaillez dans le domaine du tag. Et donc ça c'est tout votre écosystème, on peut dire. Ensuite, j'aimerais bien savoir où est votre guitare ? Vous avez envoyé un URL ?关于编程的部分，Ruchi，我看到你通过邮件发送了代码和驱动链接，虽然这没问题，但你还是应该考虑使用 GitHub。因为在 GitHub 上，你可以上传所有的数据和脚本。我觉得这和 Karoline 说的有点关联，也就是说，你们需要展示你们具备编程能力。如果我们在 GitHub 上看到你的代码，这对你们来说是有价值的，也能增加你们工作的权重。如果你们仅仅是用 ChatGPT 来做计算，那就相当于削弱了你们的能力展示。真正的关键是要做出编程的努力，并且展示出自己的代码。所以，展示的不能只是结果和论文，还必须要让大家看到背后有一个非常坚实的基础，比如你们的 GitHub 存储库。大部分研究人员都这样做，他们会上传所有的数据、材料，确保别人可以重复他们的实验，因为你们是从事标签领域的工作，这就形成了你们的整个生态系统。
* Alors merci. Donc ça c'est une guitare que vous partagez toutes les deux ? Oui, partagé. C'est quoi le... C'est du top de Marina ou de Lutti ? Je crois que c'est de moi. Mais en fait vous n'avez pas partagé, vous avez partagé votre profil. Il faut savoir parce que là on voit plusieurs dépôts, Model Spacey, Model Flobber.那谢谢你。所以这是你们两个一起共享的一个 Git 仓库吗？是的，共享的。那这个是……是 Marina 的项目还是 Lutti 的？我觉得是我的。但实际上你们并没有共享仓库，而是共享了你们的个人资料。这个需要搞清楚，因为我们现在看到的是好几个仓库，比如 Model Spacey、Model Flobber。
* Ce qu'il faut faire, je pense qu'il faut rassembler tous vos dépôts parce que vous avez divisé tout simplement modèle spécifique, modèle Flaubert, etc. Alors que, idéalement, vous devez comprendre qu'il s'agit d'un seul projet qui est votre mémoire. Donc c'est OK, vous l'avez appelé Memorem, mais vous pouvez spécifier plus précisément ce que c'est votre mémoire. Et à l'intérieur de ce grand dépôt, vous mettez vos codes, vos notes si vous le souhaitez, mais c'est pas obligé, votre mémoire, ou la tech, parce que si c'est dispersé dans plusieurs dépôts, c'est pas centralisé, alors que nous ce qu'on veut c'est...我觉得你们现在需要做的是把所有的仓库整合到一起。因为你们现在是把项目拆分成了“Specific 模型”、“Flaubert 模型”等等。但理想情况下，你们应该明白，这其实是一个完整的项目——也就是你们的毕业论文。你们是把它叫做 Memorem，这个没问题，但你们可以更明确地说明这份论文的具体内容。在这个主仓库里，你们可以放入你们的代码、（可选的）笔记、论文、或者 LaTeX 文件。因为如果资料分散在多个仓库中，就不集中。而我们希望的，是一个集中统一的项目结构。
* Ok, je vous montre un dépôt de travail partagé avec une collègue. A quoi ça ressemble ? Donc j'ai mon compte, d'accord, qui s'appelle Tez Sky 2023 et ma collègue ici elle s'appelle LHKigal66. On ouvre son compte, on voit qu'elle a 5 propos qui sont publics. Moi j'en ai 44 publics, parce que je vais toujours en public. Et en fait ce qui se passe c'est que moi j'ai fait ça sur mon compte à moi, donc j'ai fait le déco XP44 sur mon compte à moi, puis dans settings, collaborators, là il faut taper sur le mot de passe,好的，我现在给你们看一个和我同事共享的工作仓库是什么样子的。所以我有一个自己的账户，叫做 Tez Sky 2023，然后我同事的账户是 LHKigal66。我们打开她的账户，可以看到她有 5 个公开的仓库。而我这边有 44 个公开的仓库，因为我一向都是公开的。实际上，我做这个项目是用我自己的账户做的，比如我在我自己账户里创建了 deco XP44 这个仓库。然后，在“设置”（Settings）里，进入“协作者”（Collaborators）部分，就需要输入密码。
* est-ce que je me rappelle ? Je tente ça, on va voir si c'est bon, oui c'est bon, voilà, là vous voyez je l'ai Je vais ajouter le nom de mon collaborateur. P E T K O V I C Ah oui c'est parce que, attends, il faut que je fasse Help people L J P E T Voilà, alors là,我还记得密码吗？我试试看……好啦，成功了。现在你们可以看到，我在这里添加了我的协作者的名字：P E T K O V I C。啊对了，等等，我需要点击“帮助添加人员”（Help people）……L J P E T……好了，现在……
* le mien, vous voyez donc c'est comme ça qu'on partage un dépôt et donc là où je rejoins Ludmilla, c'est que moi ce que je vois là dans votre c'est que vous avez fait un mode L2 spécif, je sais pas ce que ça veut dire, il y a juste un readme dans lequel il n'y a rien. Donc ce n'est pas très utile.我这边的仓库，你们看，就是这样来共享一个仓库的。在这一点上我同意 Ludmilla 的看法，因为我在你们的项目里看到的，是你们创建了一个叫 “mode L2 spécif” 的仓库——我不太清楚这是什么意思。而且里面只有一个 README 文件，但这个 README 是空的。所以这样其实没有什么参考价值。
* Donc là, ce serait mieux d'avoir le nom d'une des deux personnes. Donc soit Marina, soit les autres filles. De manière un peu plus identifiée. Parce que comme ça, pour nous, les enseignantsants ça nous permet de voir les choses. Donc je ne sais pas si ça peut être changé à un nom de compte mais peut-être. Et puis quand on va par exemple sur modèle SPACI, de voir ici les collaborateurs en fait.这时候最好能在仓库名称或者描述中加上你们其中一个人的名字。比如 Marina，或者其他几位女生中的一个名字。这样我们老师在查看项目的时候会更清楚谁是谁。我不确定是否可以直接更改账户名，但仓库名是可以改的。另外，当我们进入像 “modèle SPACI” 这样的项目时，也希望能看到谁是协作者。
* C'est plus ça qui nous intéresse, vous voyez. Vous pouvez voir les contributeurs ici. Est-ce que ça vous paraît clair ?这才是我们真正关心的地方，你们看。你们可以在这里看到项目的贡献者（contributors）。这样说清楚了吗？
* Et après, bon, là, moi, c'est pas forcément un très bon exemple, mais je peux vous montrer un où j'ai travaillé toute seule sur ma thèse par exemple je sais pas où il est, lxp42, il y a pas de contributeur, il y a que moi voilà et là vous voyez j'ai mis plein de trucs dedans c'est peut-être un peu trop, il faut que je trie, mais j'ai plusieurs modèles par exemple, tac et tac, si je vais dans, j'ai plusieurs expériences, par exemple là c'était des expériences sur l'archéologie de Spacey, où j'ai testé plein de modèles de Spacey,trois modèles de Spacey différents, et vous, je n'ai pas fait un dépôt pour ça particulièrement. Je l'ai mis dans le même dépôt. Donc en fait, on peut faire vraiment plein de niveaux de dossiers à l'intérieur. Jusqu'à arriver au résultat qui est là. Le résultat, c'est des graphiques.然后呢，其实我现在展示的这个例子可能不是最好的。我可以给你们看一个我自己独立完成的论文项目，比如……我不知道它在哪儿了……lxp42。这个项目没有协作者，只有我一个人。你们看，我把很多东西都放进去了，可能有点太多了，之后我还得整理一下。比如说我在里面有多个模型，像这样、这样（点击示意）。我还有多个实验，比如这些是我在研究 SpaCy 的考古学应用时做的实验，我测试了多个不同的 SpaCy 模型。三个不同的 SpaCy 模型。我并没有为每一个实验都单独创建一个仓库，而是都放在了同一个仓库里面。其实我们可以在一个仓库内部创建很多层的文件夹结构，一层层组织，最后就能得到这样的成果。成果是什么呢？成果就是这些图表。
* Et je dois avoir le programme. Par contre, tous mes programmes sont rangés dans un... Je n'ai qu'un seul endroit où j'ai tous les programmes, qui s'appelle Programme. Et là, il y a tous les programmes. Et après, tous les autres dossiers, c'est que les résultats générés à partir des programmes qui sont dans le programme. Donc en fait, ce qui est important à retenir, je dirais, c'est déjà que chacune vous faites un cours d'eTab, ensuite il y en a une de vous deux qui crée un dépôt d'eTab et qui le partage avec l'autre et ensuite comme le disait madame Petkovic la dernière fois, enfin la dernière fois qu'on s'est vu je me souviens qu'elle a dit ça. Et ensuite, vous collaborer en faisant les poules et les couches pour pouvoir faire que le dossier se complète.

而且我应该还有一个叫“Programme”的文件夹，所有的程序代码都放在这里。

我只有这一个文件夹，里面集中存放了所有的程序。

然后，其他的文件夹里都是运行这些程序后生成的结果。

所以，其实最重要的一点我想说的是：

首先，你们每个人都要在 github上完成一节课程；

接着，你们两人中有一个人来创建一个 github 的仓库，然后把它共享给另一个人；

然后，就像 Petkovic 老师上次说的（我记得我们上次见面的时候她是这么说的），

接下来你们要一起协作，像“搭积木”一样，一层一层往上搭建，把整个项目的文件夹慢慢完善起来。

* Donc ça, c'est vraiment important de procéder comme ça. Elle a vraiment raison, Mme Pécovit, tu t'as insisté là-dessus, parce que, de manière générale, que ce soit pour la rédaction d'un article, pour la rédaction d'une thèse, ou peu que ce soit dans cette mode de méthode de langage, on a besoin de pouvoir réutiliser les programmes sans parler de l'évaluation du mémoire, je veux dire, dans l'absolu. d'avoir un endroit où on peut réaccéder au programme et pouvoir aussi peut-être avoir des données présentes.

所以按照这种方式来操作真的非常重要。

Pécovit 老师说得很对，她特别强调了这一点。

因为通常来说，不管是写文章、写论文，还是任何跟语言建模相关的方法，我们都需要能够重复使用我们的程序——这甚至不光是为了最后论文的评估，我是说，从更广义的角度来看，

我们都需要有一个地方，能够重新访问这些程序，

也能够把一些相关的数据保留下来。

* Parce que vous avez plein de données, plein de matériel que vous mettez un peu à gauche et à droite alors que vraiment vous devez vous concentrer sur votre dépôt guita par exemple. Donc c'est une bonne chose quand vous disposez de pas mal de choses mais juste sachez qu'il ne faut pas trop se disperser parce que vous risquez de vous perdre dans une multitude de données. Donc non seulement vous avez besoin de tout mettre dans un seul endroit mais voilà, commeame Coulon-Parfait a dit, vous devez penser aussi aux conventions de nommage c'est à dire que vous devez décider quel nom vous allez donner à votre dépôt pour que ça soit parlant pour vous par exemple et non seulement le nom du dépôt mais les noms des dossiers, des sous-dossiers, des fichiers, pour que vous-même soyez complètement certain de ce que vous mettez dedans. Parce que plus le temps passe, plus vous êtes plongé dans le travail et il se peut qu'à un moment vous oubliez ce que vous avez fait. Donc plus vous explicitez ça pour vous-même et par conséquent par nous, ça facilite un peu le travail de tout le monde.

因为你们现在有很多数据、很多资料，放得到处都是，左一点右一点。

其实你们真的应该专注于一个主仓库，比如你们的 GitHub 仓库。

有很多资料当然是一件好事，但你们要知道，不要太分散，

否则很容易在这些海量的数据中迷失自己。

所以你们不仅需要把一切都集中在一个地方，

就像 Coulon-Parfait 老师说的那样，

你们还需要认真思考命名规范的问题。

也就是说，你们要决定如何给你们的仓库命名，

这个名字要对你们自己来说是有意义的、清晰的。

不仅是仓库的名字，

还包括文件夹、子文件夹、具体文件的命名——

这样你们自己在使用的时候，才会清楚每一项内容是什么。

因为，随着时间的推移，你们会越来越深入项目，

可能有一天就会忘记当初放了什么、做了什么。

所以越早把这些东西表达清楚、命名明确，

不仅对你们自己有帮助，

对我们老师来说，也更容易理解和评估你们的工作。

* Après, par rapport au temps de passé, au mémoire en tant que tel, pour l'instant il est un peu vide, enfin, il n'y a pas de... donc ça serait peut-être bien de commencer à rédiger certaines choses, comme par exemple présenter les outils. Donc déjà, on remarque pour l'instant qu'il n'y a pas de titre.Donc dans l'introduction, vous présentez en gros un peu ce qu'on a dit l'autre fois sur à quoi ça sert de faire de la reconnaissance des entités nommées sur des poèmes du XIXe siècle. Alors comme je ne peux pas corriger, je voulais vous dire, ce serait mieux dans votre titre d'écrire 19e en chiffres romains. Et en français en fait on n'écrit pas comme ça. Moi aussi j'ai fait l'erreur. tu peux l'écrire comme ça aussi. Euh oui. Alors attends, E-M-I-E-M.

另外，关于你们在这个项目上花的时间，以及论文本身的内容，

目前看起来论文还是有些空——也就是说，还没有什么实质性的内容。

所以现在可能是时候开始撰写一些东西了，比如说介绍你们所使用的工具。

我们目前注意到，连论文的标题都还没有。在引言部分，你们可以大致介绍一下我们上次讲过的内容，

比如：为什么要对19世纪的诗歌进行命名实体识别，这个任务的意义是什么。还有，我不能直接帮你们修改论文内容，不过想提醒一下，

你们的标题里最好把“19世纪”用罗马数字表示。

用阿拉伯数字写“19e”在法语里其实不太规范，我自己以前也犯过这个错。你也可以这样写：E-M-I-E-M（可能是在打论文标题或词组的首字母）。

* J'avais un trou de mémoire. Et donc, l'introduction, ok. Ensuite, le chapitre 1, ça serait plutôt, il faudrait l'appeler directement là, vous pouvez l'appeler, comment... littérature sur la reconnaissance d'entités nommées dans du texte ancien par exemple.

我刚刚一时想不起来。

所以，引言部分就照我们刚说的来写。

然后，第一章，你们最好直接命名为，比如说：

关于古文文本中命名实体识别的文献综述。

* En tout cas c'est l'état de l'art. Voilà. Et puis le chapitre 2 ça va être sur les résultats de ces maths et... matériaux et méthodes, en général ça s'appelle mathémètes.

也就是我们说的“研究现状”部分（l'état de l’art），对，就是这个意思。

接下来，第二章，就应该是关于你们的实验结果，还有——

材料与方法。

在学术写作中，这一部分通常会被简称为 "matémètes"（来自“matériaux et méthodes”的拼读）。

* C'est matériaux et méthodes, en général ça s'appelle mathémètes, c'est matériaux et méthodes, et là vous pouvez mettre du coup, comment ça s'appelle, les données, la description des données, la description des outils, donc tout ça c'est des choses que vous devez commencer déjà à faire. Donc déjà, avoir un tableau qui récapitule les...

是的，材料与方法，在学术写作中通常简称为“mathémètes”。

在这一章里，你们可以开始描述数据、数据的来源，以及工具的描述，这些都是你们现在需要开始着手写的内容。

所以，你们应该先准备一个表格，把这些内容概括出来——

比如列出使用的数据集、所用的工具、每个工具的简要介绍等等。

* Comment dire ?Comme le tableau 4. Voilà.Ce tableau là, par exemple, c'est déjà pas mal. Et garder les deux autres tableaux pour l'analyse et résultats plus tard peut-être. En tout cas, il faut commencer à mettre des titres à bout de chaque livre et à bout de tout ce qui est écrit. Mais des titres qui indiquent un peu, ce n'est pas forcément des titres définitifs, mais c'est qui indique ce que vous allez faire dedans.

就像表格 4 那样。对对，没错。比如说，这个表格已经做得不错了。

而其他两个表格，可以暂时保留，用于分析和结果部分，以后再加上去。总之，你们需要开始为每一章、每一部分加上标题。这些标题不一定是最终的定稿，但至少要能表明这一部分是做什么的——也就是预示着你们在这一部分会做什么内容。

* Et l'état de l'art de la reconnaissance d'antithénomène à partir de corpus des poèmes, très spécifiquement, où c'est corpus ?  
  Ça peut être un coup large, un coup bien. Après si vous avez des chances dans les corpus de poèmes, parfait, je vous y arrive. Et comme je ne sais pas si c'est pas un truc de niche un peu, où il n'y a pas grand monde... Oui, c'est pour ça queil y a du monde sur du texte ancien. Mais après si vous trouvez sur des textes d'écore plus poétiques, même dans d'autres langues que le français, c'est très intéressant.  
  En gros un débat de l'art, il faut que ça soit, ça part souvent en quelque chose d'un peu large, parce que quand on travaille sur quelque chose d'un peu doux, si tu n'as pas grand monde à travailler, généralement c'est bon signe quand on a du mal à commencer son état de l'art en partant de son sujet à soi. Parce que ça veut dire que son sujet à soi,il n'a pas trop écrit de travailler, donc que c'est quelque chose d'intéressant pour peut-être fermer des portes ou ouvrir des nouvelles portes. Alors que si on a déjà 15 000 états de l'art tout fait sur sa recherche particulière, ça veut dire que notre recherche elle est plus compliquée, dans le sens où il va falloir ouvrir des nouvelles portes ou fermer des nouvelles portes sur des choses que beaucoup de gens ont déjà... Je ne sais pas si c'était clair ce que je viens de dire...

关于从诗歌语料库中识别“反现象”的研究现状，特别是那种“到底哪些语料库”可以用的问题。这个问题可以是个大话题，也可以是个好话题。如果你在诗歌语料库方面有一些可以利用的资源，那就太好了，我也正是要去接触这一块。因为我也不确定这是不是一个有点小众的领域，可能研究它的人不是特别多……这也是为什么在处理一些“古老文本”方面的研究，人会多一些。但如果你能找到哪怕是其他语言中的“更具诗意的文本”，那也非常值得关注。大体来说，一个“研究现状综述”（state of the art），往往从一个比较宽泛的角度开始。因为当我们研究的是一个比较冷门的领域时，如果我们在围绕自己的研究方向写“研究综述”时很难找到开头，那其实是个好迹象：说明这个方向还没有太多人研究过，也许是个值得探索的空白点，有机会去打开新问题，或是封闭一些旧问题。反过来说，如果我们的研究主题已经有一堆现成的研究综述，比如已经有一万五千篇文章都写过差不多的东西，那我们的工作就会更难一些：不是说不能做，而是意味着你需要在已有研究的基础上进一步开辟新的视角或角度，这反而更具挑战。我不知道我刚刚说的有没有讲清楚……

* Ok, en tout cas, je dirais que là je suis tout à fait d'accord avec Péter Kovic. Il faut que votre guitare soit accessible et claire. Il faut qu'il y ait des choses dedans, pas juste des noms de dépôt avec des rythmes qui ne sont pas écrits. Et il faut que vous avanciez aussi sur la rédaction de votre mémoire. Donc peut-être que, comme la dernière fois on en a parlé, peut-être que chacune d'entre vous a écrit ailleurs dans un autre fichier, c'est le moment de rassembler ça là, parce que nous c'est moi je vais plus aller regarder les autres fichiers. Voilà. Maintenant, je pense que Manuel Bétho et toi aussi, moi je vais me concentrer sur la lecture du 8 mois.

好的，总之我想说，我完全同意Péter Kovic的意见。你的“吉他”必须是清晰而易于理解的。里面要真的有内容，而不只是一些仓库的名字，或者是没有写清楚节奏的东西。你们也要继续推进论文的写作进度。

也许，就像我们上次聊到的那样，你们每个人可能都在别的文件里单独写了一些内容——那现在可能是该把这些内容集中起来的时候了。因为我这边不会再去看其他的零散文件。

就这样吧。现在，我想Manuel Bétho 和你，还有我自己，我们都要集中精力去读你们的“第八个月”的进展。

* Parce que là, il faut arrêter de s'éparpiller. bien l'ultime que vous nous envoyez vos codes deux minutes avant la réunion. Je veux dire que moi je ne vais pas aller lire vos codes deux minutes avant la réunion. À ce moment-là, il faut les envoyer en avant. Mais en tout cas, il faut que ça soit dans un GitHub. Si je veux y aller jeudi, j'ai tout par exemple. Et j'ai aussi ce que Marina avait fait en fait. Et en plus, on sait qui a contribué à quoi, puisque les contributeurs sont désignés sur Github. Et puis pareil, dans le cliffier Overleaf, si je vais dans History, je vois très bien qui a fait quoi.

因为现在真的不能再分散精力了。你们把代码在会议前两分钟才发过来这种事，真的该停了。我是说，我不会在会议前两分钟还去看你们发来的代码。所以，代码要提前发过来。

不管怎样，所有东西必须放在GitHub上。这样的话，如果我周四想看的时候，我可以一口气看到所有内容，比如也能看到Marina做了哪些东西。而且，通过GitHub我们也能清楚地知道每个人的贡献，因为贡献者在GitHub上是可以看得一清二楚的。

同样地，在Overleaf上也是一样的道理，如果我去看“历史记录”，我也能非常清楚地看到每个人具体写了什么。

* Pardon, je vois très bien pour l'instant qu'en effet, il n'y a que Mme Lopez qui a apparemment mis des choses dedans. justement pour Overleaf, étant donné que vous serez amené à rédiger votre établi prochainement, il faut que vous régliez ce problème d'affichage de la bibliographie. C'est très facile de faire, il faut juste trouver, bien nommer le fichier.Il faut juste créer un fichier qui s'appelle bibliographe.mille.

不好意思，我现在确实看得很清楚，目前好像就只有Lopez女士在Overleaf里上传了东西。

正因为你们接下来要开始撰写你们的论文了，所以你们现在必须把参考文献的显示问题解决掉。

其实这很简单，你们只需要：

* 找到参考文献的文件，
* 给它正确命名，
* 创建一个叫做 bibliographie.bib 的文件就可以了。
* Et il faut bien l'appeler parce que là, quand vous l'avez appelé, on avait vu qu'il y avait des espaces et deux colonnes, donc forcément, Overleaf avait du mal à l'interpréter. Donc, quand vous réglez ce problème, à partir de ce moment-là, vous pouvez citer autant de fois que vous voulez et comme ça, ça va vous aider parce que vous aurez déjà des références que vous employez. On va voir que d'accord, vous avez vu que tel auteur et telle autrice travaillait sur ça, c'est bien. Donc vous avez couvert un domaine de votre état d'alarme et ensuite vous pouvez passer à autre chose. Donc c'est pour ça que quand vous réglez ce problème et que vous commencez vraiment à citer des travaux, ça ira plus vite.

你们一定要正确命名文件，因为之前你们命名的时候，我们看到文件名里有空格和两列内容，所以Overleaf无法正确解析。

所以，当你们解决了这个问题后，从那时起，你们就可以随时引用文献了，这样会很有帮助，因为你们已经开始使用已有的参考文献。

这样我们就能看到，你们已经阅读了某个作者或某个女性学者的研究成果，这就表明你们已经覆盖了你们研究领域的一部分，这很好。

然后，你们就可以继续展开其他方面的研究了。

因此，当你们解决了这个问题并开始真正引用文献后，写作进度会更快。

* Par contre on ne peut pas mettre des commentaires dans le Overview.  
  Comment ça ?J'essaye de mettre un commentaire dans le fichier. Mais je ne vois pas commentaire parce que quand je vais dans Review, je n'ai pas d'espace pour écrire. Ah oui c'est vrai, moi non plus.Moi je ne peux pas vous mettre parce que... Je pense que c'est parce qu'on est en mode électrice. Je ne peux pas vous mettre éditeur parce qu'il n'y a que deux possibles au maximum. Donc j'ai déjà mis Lucine.

不过，我们不能在概述（Overview）中添加评论。

怎么回事？我试着在文件里加一个评论，但我没有看到评论，因为当我进入审阅（Review）模式时，根本没有可以写的地方。

哦，对，确实，我也不能加。我不能给你们加评论，因为……我想是因为我们现在处于查看模式（lectrice）。我不能把你们设置为编辑者（éditeur），因为每个文档最多只能有两个人。所以我已经把Lucine设置为编辑者了。

* Donc il est d'accord. Là c'est hyper restrictif parce que du coup, le mieux c'est'on puisse mettre des commentaires. Ce que je voulais vous mettre comme commentaire pour les tableaux, c'était de les sortir du mail. C'est-à-dire de faire un dossier que vous appelez tab. Et dans ce dossier, vous ouvrez des fichiers. Par exemple, pour le tableau 4.1, vous allez l'appeler NombreToken, et puis dedans vous mettez les lignes de précision. Vous copiez dedans le tableau qui est de la ligne 83 à la ligne 115.Et puis vous faites pareil pour le tableau suivant. Et du coup vous aurez un dossier TAB, avec dedans plusieurs fichiers qui contiendront les tableaux. Et ensuite vous mettez dans le main, input. Et puis le chemin sera TAB, puisque vous l'appeleelez table. Et puis ensuite le nom du tableau. Vous appuyez pas sur tableau 1 ou tableau 2, on laisse juste pour les gens. Vous mettez des noms vraiment explicites.

所以他同意了。现在这种方式很限制，因为最好能有办法添加评论。我想给你们的评论是，把表格从邮件中分离出来。也就是说，创建一个名为 tab 的文件夹，然后在这个文件夹里打开一些文件。比如，对于表格4.1，你可以把它命名为 NombreToken，然后把第83行到第115行的内容复制到这个文件里。接着，对下一个表格做同样的操作。

这样你们就会有一个TAB文件夹，里面有多个文件，每个文件包含一个表格。接下来，你们只需要在主文件（main）中做一个 input，路径指向TAB文件夹，因为你叫它tab，然后接着就是表格的名字。不要再使用“表格1”或者“表格2”这种名字了，我们保持这种命名方式对其他人更清晰。要确保文件名是很明确的。

* Ça, je peux vous montrer comment on a fait par exemple sur l'article ?Après accessoirement, vous pouvez appliquer la même logique même aux images. Vous pouvez créer un dossier images, vous pouvez stocker vos fichiers qui contiennent des images et comme ça vous pouvez appeler uniquement des noms de fichiers, en tout cas des chemins vers ces fichiers. Par exemple, pour la revue STAL, on a notre main avec le NLP de Twitch. Vous voyez notre main, il n'y a pas grand chose dessus. Ce qu'il y a dessus, c' dessus le titre du papier, les résumés, les noms des auteurs, les mots-clés. Et puis après, on a nos différents inputs, donc l'introduction, l'état de l'art, voilà. Et après on a même nous des fichiers qui correspondent à l'introduction, à l'état de l'art, à la présentation du corpus.Donc là c'est les inputs. Et à l'intérieur on va faire appel soit aux images, soit aux tableaux. Vous voyez, nos noms de tableaux, on essaie d'être assez descriptifs.

我可以给你们展示一下我们在文章中是怎么做的。例如，你们也可以把同样的逻辑应用到图像上。你们可以创建一个图片文件夹，把包含图片的文件存放在里面，这样你们只需要调用文件名或者文件路径就可以了。

例如，对于STAL期刊的论文，我们有一个主文件，里面包含了Twitch的NLP部分。你们可以看到我们主文件里内容不多，里面有文章的标题、摘要、作者的名字和关键词。然后我们有不同的输入文件，比如介绍部分、文献综述部分等。接着我们还有一些文件分别对应于介绍部分、文献综述部分和语料库介绍。这些就是我们的输入文件。

在这些文件内部，我们会引用图片或者表格。你们看，我们的表格名称，我们尽量做到描述性强、清晰明确。

* Et pour chaque fichier, vous compris l'extension .tech, soit vous donnez juste le nom effectif du fichier, parce que normalement Overleaf va reconnaître comme ça.On avait même plusieurs niveaux de déco, on a pris le dossier images, puis le dossier de ce qu'on voulait mettre, c'était les boîtes à moustache par exemple, et à l'intérieur on avait les différents dossiers avec les images. Nous c'était classé selon la langue du corpus. Donc soit le corpus anglais, soit le corpus français, soit le corpus portugais, soit le corpus de la très grande bibliothèque de la TGV.

对于每个文件，你们要么包含 .tex 后缀，要么只给出文件的实际名称，因为正常情况下，Overleaf会自动识别这种格式。

我们甚至有多个文件夹层级，首先是图片文件夹，然后是我们需要的内容文件夹，比如箱线图，然后在里面我们有不同的文件夹来存放图片。

我们按语料库的语言进行分类。比如，有英语语料库、法语语料库、葡萄牙语语料库，还有TGV大图书馆的语料库。

* En fait, vous ne serez pas forcément obligés d'avoir autant de niveaux d'arborescence ou peut-être que vous allez juste avoir...C'était pour vous montrer que c'est possible, c'est-à-dire que c'est pas tout le temps obligé de tout réorganiser.Vous préférez qu'on fasse ça juste pour les images, pour les tableaux, ou on fait en main exactement comme vous avez fait, même avec introduction, conclusion ?Alors idéalement, justement, vous avez vu que l'ossature du fichier de l'overleaf de Madame Coudoura Parfait, contient la préambule, donc là vous appelez tous vos packages qui s'occupent de tel ou tel fonctionnement, et après votre préambule, c'est là où il y a le begin document et end document. Donc ça c'est votre corps du texte. Et à l'intérieur, vous n'écrivez pas le texte manuellement, enfin je veux dire, vous appelez tous vos matériels, c'est-à-dire les fichiers qui contiennent des chapitres, des sections, des sous-sections, des images, des tableaux. Donc là, on a un exemple dans le chat InputTable qui est le dossier slash qui contient le fichier tableau 1. Pour une image vous allez peut-être avoir un dossier qui s'appelle image slash image1 ou image2, ou évidemment vous l'appelez comme vous voulez pour que ce soit clair pour vous et pour le fichier tech qui va représenter votre chapitre ou votre section ou sous-section vous allez l'appeler input peut-être votre dossier nom du dossier slash votre nom du fichier .tech d'accord avant de terminer ce qui va faire le gilet je voudrais revenir sur le tableau 4.2. Je vois que pour la colonne Flaubert, c'est le nombre de tokens et pas le nombre d'entités nommées qui figurent. Donc là c'est pareil, il faut mettre NA si vous n'avez rien à mettre. Si vous n'avez rien à mettre, ne mettez pas des chiffres, mettez NA. Ah non mais c'est bon, si vous recompilez j'ai mis NA pour Flaubert, c'est bon, les données sont justes. Ah ok d'accord. Même pour le tableau 4.2 du moment ?

其实，你们不一定需要有那么多的文件夹层级，或者你们可能只是……

这是为了给你们展示一下，这种做法是可行的，也就是说，不一定每次都要重新整理所有内容。

你们希望我们只针对图片和表格做这种整理，还是像你们现在在主文件（main）中做的那样，包括引言、结论等部分都一起整理？

理想情况下，正如你们所见，Madame Coudoura Parfait的Overleaf文件框架包含了前言（préambule），在这里你们调用所有处理特定功能的包（packages），然后在前言之后就是 begin{document} 和 end{document}，这部分是你们的正文。

在正文内部，你们不是手动写文本，而是调用所有的素材文件，也就是那些包含章节、节、子节、图片和表格的文件。

比如，我们在聊天中看到一个InputTable的示例，它是一个文件夹，里面包含了表格1的文件。对于图片，你们可能会有一个文件夹，叫做 image/image1 或 image2，当然，你们可以根据需要自己命名，只要对你们来说清晰就好。而对于表示你们章节或小节的文件，你们可以用 input 来命名，比如文件夹名称就是 input/文件名.tex。

好了，最后在结束前，我想回到表格4.2。我看到Flaubert列里是tokens的数量，而不是命名实体的数量。这也是一样的，如果你们没有数据，应该填写NA。如果没有数据，就不要填数字，填NA。

哦，没问题，我已经重新编译过了，Flaubert那一列我已经填了NA，数据是正确的。

好吧，明白了。那么表格4.2也可以这么处理吗？

* Ok ça marche. Je recompile, norme, baromètre. Alors moi ça prend du temps pour... Non, je viens de recompiler, dans le tableau 4.2 pour Flaubert.  
  Oui, pour Camembert et pour Stanford.Pour les nombres de tokens.  
  Non, mais Flaubert c'est juste, c'est juste Camembert et Stanford qui s'est...  
  Non, non, non, je me dis que Flaubert c'est pas juste. Dans le tableau 4.2, il peut pas y avoir 60 937 entités nommées pour 60 937 tokens. Vous avez 60 000 et quelques tokens, mais par contre pour les entités nommées, là il va falloir changer les numéros.  
  Ah, j'ai vu, ils sont mélangés. Ok, je vais mettre une A et après une I.

好的，明白了。我重新编译了，规范，条形图。

其实这需要一些时间来……不，我刚刚已经重新编译了，表格4.2中的Flaubert。

是的，Camembert和Stanford也是一样的。

对于tokens的数量。

不，Flaubert是对的，问题出在Camembert和Stanford。

不不不，我是说Flaubert的部分不对。表格4.2中，命名实体的数量不可能等于tokens的数量。你们的tokens有大约60,000个，但对于命名实体，这些数字需要调整。

哦，我看到数字是混在一起了。好的，我会把它们调整成A和I。

* Et même ici pour les entités nommées qui ont un B restant fort, là vous avez laissé 5 et 28. Donc ça vous pouvez aussi changer en moyenne A. Ok, et pour les densités du coup, vous mettez si vous n'avez pas de... enfin si c'était faux résultat, il faut mettre N-A aussi. En même temps que vous n'avez pas des résultats que vous jugez pertinents, enfin que... ils sont bons, il ne faut pas les mettre dans le tas de mots. Oui. Alors, du coup, avant de partir, pour la prochaine fois, donc ça serait bien que tu as à peu près tout ce qu'on a demandé figure dans le... Avance.  
  Dans le Overleaf, là, dans le fichier Overleaf. Plus donc les calculs d'intersection, avec Chine, ce serait bien que vous fassiez donc des diagrammes de Venn. J'avais partagé avec vous normalement comment faire les diagrammes de Venn, c'est la numérique de Venn 2, voilà. Donc ce serait bien qu'il y ait des images, des diagrammes de Venn. Ce serait bien que pour la prochaine fois, on ait quelque chose à analyser de Diable, quoi. Puis s'analyser tous ensemble avec Monsieur Lejeune. Je pense que là, il y a un petit coup d'accélérateur à mettre dans la rédaction aussi des différentes parties.

甚至这里，关于命名实体的部分，你们在剩余部分留下了5和28。所以你们也可以把这些改成平均A。好的，至于密度部分，如果你们没有数据，或者如果结果是错误的，也应该填写NA。同时，如果你们认为某些结果不重要，或者说它们并不准确，那么不要将它们算入总数。是的。那么，接下来的任务是，下一次会议之前，你们应该确保所有我们要求的内容都已经写入Overleaf文件中。另外，还有交集的计算，结合中国，希望你们能做一些维恩图。我之前应该已经和你们分享过，如何做维恩图，可以用Venn 2的数字工具。所以最好是有一些图片，例如维恩图。

希望下一次我们能看到一些可以分析的内容，然后大家一起分析，包括Lejeune先生。我认为现在是时候给不同部分的写作加把劲了。

* Et puis, comme le soulignait Madame Petkovic, il faudrait faire le GitHub de manière efficace. Peut-être, je ne sais pas si la priorité à ce moment-là, ce serait la rédaction. Moi, je pense que la priorité, c'est tout. C'est tout. C'est tout. Oui mais je veux dire, il faut avancer sur tout parce que là en deux semaines, moi je me dis bon c'est un peu... il faut y aller. Donc je ne m'attends pas à ce que dans deux semaines tout soit fini, c'est pas ça,mais à ce qu'il y ait les diagrammes de veines dans le mémoire et qu'ils soient aussi accessibles sur le GitHub. Je pense que c'est pas hyper compliqué de mettre en place le GitHub, donc si madame Pekoitch a raison, s'il y a une priorité à mettre, ce serait faire le GitHub, parce que je pense que c'est pas très long et même pas très difficile.

正如Petkovic女士所强调的，GitHub的管理需要高效。也许，我不确定此时是否应该优先进行写作。我个人认为，一切都很重要，都是一样的。是的，但是我的意思是，我们必须在所有方面都推进，因为两周的时间，我觉得有点紧迫……我们得加把劲。所以，我并不指望两周内一切都完成，并不是这个意思，而是希望能够在文档中看到维恩图，并且它们也能在GitHub上获取。我觉得设置GitHub并不复杂，所以如果Petkovic女士说得对，如果真的要设定一个优先级，那就是先做GitHub，因为我认为这不需要太长时间，也不难做。

* Et puis après, s'imaginer vraiment sur la rédaction, c'est-à-dire vous faire un plan de rédaction, qui réquille la présentation des corpus, voilà. Enfin je sais pas, moi c'est comme ça que... C'est pour ça que je sais pas s'il y a vraiment une priorité là, mais plutôt il faut avancer sur les deux conjointement, bien que moi je m'attende pas à ce que ça soit fini, je veux dire c'est pas ça, c'est plus là il faut...
* Voilà, parce que ça fait déjà deux fois je crois qu'on parle du GitHub et que c'est toujours pas clairement mis en place. Donc là je pense qu'il faut vraiment clairement le mettre en place. Et tu avais dit le nom de la librairie pour les Diablards de Vinz, tu peux me rappeler juste ? C'est Ven2, c'est ma bibliothèque, Ven2.

然后，再想清楚写作部分的内容，也就是说，制定一个写作计划，包括语料库的展示，就是这样的。我不知道，反正我就是这么理解的……所以我不知道是否有特别明确的优先级，但我认为两者应该同时推进，尽管我并不指望它们会完全完成，我的意思是，这并不是重点，重点是要继续推动工作。

因为我觉得我们已经讨论过两次GitHub，但它始终没有明确的落实。所以现在我认为我们真的需要把它明确地建立起来。而且你之前提到过Vinz的Diablards的库名，你能再告诉我一遍吗？是Ven2，就是我的库，Ven2。

* Après, je sais qu'avec M. Lejeune, on avait trouvé une autre librairie. C'était peut-être 23, je vais vérifier avant de dire.

之后，我知道和Lejeune先生一起我们找到了另一个库。可能是23，我会在说之前再确认一下。

* Ah oui, 22 22 fonctions. 2, 3, 4, 5 groupes ensemble plutôt déjà faire des diagrammes de veines avec deux ensembles c'est bien pour commencer.

哦，是的，22，22个函数。2、3、4、5组一起，最好是先做带有两个集合的维恩图，这样开始会更好。

* Ok. C'est juste une question, c'est que pour le modèle Stanford j'ai changé pour Stanza. C'est votre adresse mail, c'est mon adresse mail. J'ai pu utiliser Stanford mais c'est en zap, qui est aussi créé par l'université de Stanford, pour faire les programmes, les entités nommées et les candidats. Parce que j'ai vérifié le modèle pour la langue anglaise, il y a des erreurs pour les entités nommées.

好的。只是有个问题，就是关于Stanford模型，我已经改用Stanza了。您的电子邮件地址是我的电子邮件地址。我原本可以使用Stanford，但它在zap上，也由斯坦福大学开发，用于做程序、实体识别和候选项。因为我检查了英语模型，发现实体识别存在错误。

* Oui, parce qu'il n'y a pas de modèle pour le français. Oui, il n'y a pas un modèle pertinent pour traiter le corpus. Donc j'ai changé pour standard.  
  Ok. Et donc ça c'est quel modèle ? C'est le modèle français ?C'est le modèle français. Ok. Et aussi pour les deux dossiers de code, je ne sais pas maintenant, est-ce qu'on doit améliorer la structure de code ?

是的，因为没有适用于法语的模型。

所以我换成了标准模型。好的。那么这是哪个模型？是法语模型吗？

是法语模型。好的。那么关于这两个代码文件，我现在不确定，是否我们需要改善代码结构？

* Est-ce que vous devez améliorer la structure du code ? Alors, il faut s'assurer que votre code produise les résultats qui sont corrects. Donc je reviens sur cette histoire de Kernel-Berr et Stanford, je pense que ça c'est le plus important de régler. Donc de s'assurer que vous avez vraiment les nombres d'entités nommées, les tokens, pour ces deux modèles qui posent problème ici pour qu'on puisse y compléter les tableaux parce que pour l'instant on voit que pour spacey smlg et flubber vous avez des données qui passent correct mais pas pour scamandr.info donc ça c'est au niveau du contenu ce que vous devez changer pour l'instant on s'intéresse pas trop à la forme. Je ne sais pas, Kéonis, ce que tu en penses, mais là, je pense que... Excuse-moi, je ne vais pas...

你们需要改善代码结构吗？那么，首先要确保你们的代码能够产生正确的结果。回到这个Kernel-Berr和Stanford的问题，我认为这是最重要的，首先需要解决。也就是说，要确保你们对于这两个模型（当前存在问题的模型）正确地计算了命名实体和tokens的数量，这样我们才能完善表格。因为现在我们看到，对于spacey smlg和flubber，你们的数据是正确的，但对于scamandr.info模型，数据存在问题。所以，目前我们要关注的是内容的正确性，而不是格式。Kéonis，你怎么看？但是我觉得...抱歉，我不能继续说下去了

* Si vous cliquez le lien dans le chat, c'est le dossier de code pour Gamembert. Je crois qu'on a déjà les résultats pertinents, parce que j'ai classé au niveau du corpus et au niveau de chaque texte. On a déjà le résultat pertinent, le problème c'est qu'on n'a pas encore le mettre dans.Et donc ça, ça produit les entités nommées... Parfois, toujours. Vous parlez de quelle cellule déjà ? Peut-être si vous partagez l'écran, ce sera plus facile. Comme ça on peut tout regarder.Là, on voit l'auteur, mais je pense qu'il affiche pas tous les auteurs. Il affiche juste certains. On a Sauvage, Rimbaud, Loiseau.

如果你点击聊天中的链接，那是Gamembert代码的文件夹。我想我们已经得到了相关的结果，因为我已经按语料库和每个文本分类了。我们已经有了相关的结果，问题是我们还没有将其放入文件中。所以这部分会生成命名实体...有时会，始终如此。你们说的是哪个单元格？也许如果你们共享屏幕，会更容易一些。这样我们就可以一起查看了。现在我们看到了作者，但我认为它没有显示所有的作者。它只显示了部分，比如Sauvage，Rimbaud，Loiseau。

* Ça sera ça. Vous avez 9 fichiers. 10 parce qu'on commence avec 0. Enfin 10 pardon. En fait il vous faut encore, encore combien ? Non c'est bon je pense que c'est bon. Ah d'accord.Ouais je pense qu'il faut juste mettre sur le tableau.Et pour le modèle camembert, il existe aussi la limite de traitement, c'est que chaque fois c'est 512, mais en bas de ce dossier de code, j'ai vérifié la longueur maximale de ton code pour débarquer le modèle. Est- parce que ça veut dire que vous allez rester sur les résultats que vous avez déjà. On ne voit pas le nombre de tokens et d'entités nommées effectif. Donc il faut trouver un moyen pour continuer à traiter. Donc disons que vous traitez un premier bout de texte qui contient 512 tokens, il faut continuer à traiter les restes. Oui, c'est dans... Je crois que c'est dans là. C'est ici ?

就是这样。你们有9个文件，实际上是10个，因为从0开始。嗯，10个，抱歉。其实你们还需要多少个？不，应该够了，我觉得够了。哦，明白了。我认为只需要将结果放到表格中。而对于Camembert模型，它确实有处理的限制，每次处理的最大长度是512，但在这个代码文件的底部，我已经检查了你代码中处理模型的最大长度。那意味着你们将只能使用已经获得的结果。我们看不到实际的tokens数量和命名实体的数量。所以你们需要找到一种方法来继续处理。也就是说，如果你们处理的第一部分文本包含512个tokens，那你们需要继续处理剩余的部分。嗯，是在这里吗？

* Oui, pour exclure l'erreur d'utiliser la méthode de fenêtre glissante.Donc vous avez testé ça ou pas ?On a testé.D'accord, et qu'est-ce qui...Le nombre de troquons est le même que le nombre de tentatives mais pour les indénommés il y a la répétition. Et les indénommés, les informations ne sont pas complètes. Donc il y a des erreurs pour la méthode fenêtre glissante. Sans redoublant c'est à peu près le même nombre de dénommés.  
  L'idée de la fenêtre glissante était de s'assurer que vous traitez tout le texte. Et là, ce que vous me dites, c'est que vous avez le même nombre d'entités nommées que de tokens ? C'est avec la méthode fenêtre glissante, le nombre de tokens et le nombre d'entités nommées c'est le même. Alors ça c'est problématique, c'est pas possible justement qu'il y ait le même nombre. En théorie vous aurez beaucoup beaucoup plus de tokens que d'entités nommées.

是的，为了排除使用滑动窗口方法的错误。你们测试过这个吗？

我们测试过了。

好的，那么结果如何？

tokens的数量和尝试的次数是相同的，但对于命名实体来说，有重复的情况。而且对于命名实体，信息不完整。因此，滑动窗口方法存在错误。如果没有重复的情况，命名实体的数量大致相同。

滑动窗口方法的目的是确保你们处理了整个文本。而你们告诉我的是，使用滑动窗口方法后，tokens的数量和命名实体的数量是相同的？

是的，使用滑动窗口方法后，tokens的数量和命名实体的数量是相同的。

这就有问题了，这不可能。理论上，你们应该有比命名实体多得多的tokens。

* Donc cette méthode du coup, on en a tout.Ouais, ça c'est pas normal. Je pense qu'il y a un problème dans le socours gardien.Si la fenêtre glissante vous stockez pas...Est-ce que vous faites bien de la reconnaissance en s'entendant des dates d'organisation ? C'est quand même ma première question.

所以这方法现在没问题了。

对，这不正常。我觉得滑动窗口方法中可能有问题。如果在滑动窗口中你们没有存储...

你们是否确实做了实体识别，特别是日期和组织名的识别？这是我的第一个问题。

* Oui, c'est là où ça parle de l'utilisation de la méthode fenêtre glissante. C'est quasiment tout en bas. Alors, traiter au niveau de chaque texte... En fait, cherche dans le sommaire, exclure l'erreur due à la limitation du nombre de toquets. Exclure l'erreur due à la limitation du nombre de toquets.

是的，这里提到了滑动窗口方法的使用。 几乎在最下面。然后，处理每个文本的内容... 实际上，查找目录，排除由于令牌数量限制而导致的错误。排除由于令牌数量限制而导致的错误。

* Il y a une colonne. Oui, mais sur ça, regardez dans votre méthode de présentation, il n'y a pas une erreur dans le programme. Parce que ça ne vient pas de Camembert, ça. Vous voyez ce que je veux dire ? Ça vient de vous, de ce que vous écrivez, je pense. Parce que si à un moment donné, vous voyez bien que vous Vous avez dit c'était où ? C'était au début ?

有一个列。是的，但你们看看你们的显示方法，程序里其实没有错误。这不是 Camembert 的问题，你明白我的意思吗？我觉得这是你们写的代码或方法的问题。因为如果某个时候……你们之前说是在哪儿来的？是在开头？

* C'est ça ? Il y a 23 entités. Donc si là, Calambar, il vous dit... Il y a 21 297 tokens. Puis il y a 23 entités sur les 20 562 candidats. Je ne sais pas trop ce que c'est les candidats. Et que là, tout en bas, vous faites à peu près la même chose que ce qu'il vous dit. Il y a 84 morceaux et 84... C'est qu'il y a un domande où vous mettez la même chose dans les deux variables. Donc c'est ça en fait, c'est ça que j'essayais soi. Moi je vous avoue que pour travailler avec Camembert, je ne me souviens pas qu'on soit passé par tout ça avec le comédie. Vous avez regardé un peu le programme que je vous ai montré l'autre fois ?C'était dans Corpus... Je suis désolée, il va falloir qu'on vous laisse, parce qu'il est déjà dix heures et quart. Donc c'était là-dedans. Voilà.

是这个吗？有 23 个实体。那么，如果 Camembert 告诉你……有 21,297 个 token，但只有 23 个命名实体，在 20,562 个“候选项”中。我不太确定“候选项”是什么意思。

然后，在最下面的地方，你们做的几乎跟上面一样：有 84 个片段（chunks），也有 84 个……这说明在某个地方，你们把相同的内容放进了两个不同的变量里。

其实，这就是我一直想说的。我坦白讲，在使用 Camembert 的时候，我不记得我们之前在处理 Comédie（喜剧作品集）的时候有这么复杂的流程。

你们有看过我上次给你们展示的那个程序吗？是在 Corpus 文件夹里……对不起，现在已经十点一刻了，我们得结束了。就是在那里面。

* Et normalement, vous pouvez peut-être avoir le plus bas.  
  Je crois que pour le standard ça va. Normalement, vous pouvez avoir le plus bas, mais dans Prog... Pour le standard, ça va...  
  Dans Prog, vous avez UseCalember et je crois pas que ma collègue utilise mal ça. Excuse-moi, donc sur le GitHub, tu parles de... Vous allez dans Prog, dans le bouton, plus dans Prog, il y a UseCalember. Ah oui, ok. Je vais vous montrer le programme BIO Calendars. C'est un peu comme ma collègue avait utilisé ce programme. Je ne sais pas trop exactement comment elle a fait, mais si elle a l'air d'utiliser, c'est peut-être un peu mieux. Non, c'est pour le BIO Calendars qu'elle utilise ça. ...

正常情况下，你们也许可以得到更低的结果。我觉得对于 Stanford 模型来说是可以的。正常来说你们可以得到更低的值，但在 Prog 里面……对于 Stanford 模型应该是可以的。

在 Prog 文件夹里你们有个叫 UseCalember 的脚本，我不觉得我那位同事在使用它时哪里错了。不好意思，你说的是 GitHub 上的……你们去 Prog 文件夹，然后在里面有个 UseCalember。啊，好，我明白了。我会给你们展示那个 BIO Calendars 程序。

这跟我那位同事使用的方式差不多。我不太清楚她具体是怎么做的，但她看起来确实是用了这个程序，也许她的处理方式会更好一些。

不，其实她是用这个程序来处理 BIO Calendars 的

* Enfin, on va les évaluer. Donc il y aura des faux négatifs, des choses comme ça.  
  On a parlé des faux résultats dans le sens où Camembert et Robert, je pense, ils présentaient des résultats au niveau des entités nommées qui sont beaucoup plus bas par rapport au nombre de tokens. Donc c'est ça où on avait dit, il faut regarder comment dire à Camembert de traiter tout le texte et par conséquent de sortir le nombre de tokens raisonnable et le nombre d'entités nommées. Après les résultats qui vont différer en fonction de modèle, ça c'est ok. On est bien d'accord, il faut qu'on voit la différence entre chaque modèle parce que vous allez évaluer combien d'entités nommées Canberra a sorti sortie, SPC ou autre. Donc explique ce que tu m'as dit sur les résultats, il ne faut pas qu'on ait les résultats faux pour faire l'intersection, tu peux l'expliquer. On doit garder ce genre d'erreur, protoplan, l'un, le fausse antidéoms par exemple, le mot c'est IE et le type c'est organisation.  
  Est-ce qu'on doit garder ce genre d'erreur pour

最后，我们会评估它们。所以会有假阴性，类似这样的情况。

我们谈到过错误的结果，Camembert和Robert的结果，我认为，他们展示的命名实体的结果比token的数量低得多。

所以就是这个问题，我们之前提到过，必须看看怎么告诉Camembert处理完整个文本，从而输出合理的token数量和命名实体的数量。

然后，结果会根据模型有所不同，这点没问题。我们需要看到每个模型之间的差异，因为你们会评估Camembert输出了多少命名实体，SPC或者其他模型也是。

所以解释一下你之前跟我说的关于结果的部分，我们不能接受错误的结果去做交集，你可以解释一下吗？我们是不是应该保留这种错误，比如protoplan，假阴性，例如一个词是“IE”，类型是“组织”。

* Dans le JSON ?Oui.Alors, oui, ce que votre modèle génère dans le fichier JSON, ça c'est le résultat du modèle. Après c'est à vous d'évaluer, ah si il y a des faux positifs ou faux négatifs, il ne faut pas modifier le dictionnaire, il ne faut pas modifier les résultats que le modèle vous sort.Car on va évaluer leurs performances respectives après. Oui, nous on avait juste dit au niveau du code que vous utilisez pour générer les résultats, il faut que vous regardiez dans quel endroit votre code génère le même nombre d'entités nommées que le token. Parce que madame Coudoro-Papel avait dit que vous allez avoir probablement des mêmes valeurs qui sont stockées dans les deux variables différentes, donc entités nommées et tokens en gros. Donc il faut voir pourquoi le code vous génère cette erreur. Et dès que vous réglez ce problème au niveau du code,pour l'instant je ne sais pas encore, il faut que vous vous regardiez. Ça peut être juste une erreur syntaxique ou bien de la structuration, où vous avez mal inventé une ligne ou vous avez mal placé votre ligne de code. Donc ça, c'est plutôt à vous de travailler dessus. Et quand vous vous assurez que votre code vous sort des résultats corrects désormais, après vous passez à l'analyse des résultats et à l'évaluation des modèles. Mais ça il faut faire au plus vite, c'est ce que madame Kurofaki a dit.

在 JSON 文件中？

是的，您模型在 JSON 文件中生成的内容就是模型的输出结果。之后，您需要自行评估，如果有假阳性或假阴性，您不应该修改字典，也不应该修改模型输出的结果。因为我们将在之后评估它们的性能。

是的，我们只是说，在您用来生成结果的代码部分，您需要查看您的代码在哪个地方生成了相同数量的实体命名和令牌。因为 Coudoro-Papel 女士说，您可能会在两个不同的变量中存储相同的值，分别是实体命名和令牌。您需要找出为什么代码会生成这个错误。一旦您解决了代码中的问题，我现在还不确定，您需要自行检查。可能只是一个语法错误，或者是结构问题，您可能在某行代码上出错，或者代码行的位置不对。

这些都需要您自己来解决。当您确保代码生成正确的结果后，您就可以开始分析结果并评估模型了。但这需要尽快完成，正如 Kurofaki 女士所说的那样。

* Il faut régler ce problème de code, il faut ouvrir le dépôt GitHub, mettre tout dedans, commencer la documentation, continuer la réduction. Donc vous avez déjà du matériel à écrire. Donc il faut se concentrer sur plusieurs endroits, sur plusieurs domaines, mais il faut avancer dans la mesure du possible. Nous gardons ce genre d'erreur dans des fichiers JSON pour évaluer les performances respectives des 4 modèles. Mais je crois que les données sont aussi nécessaires pour faire l'analyse et faire l'analyse. Si les données sont tabourettes, qui crée des problèmes.Je n'ai pas compris la dernière partie de votre phrase, on garde les erreurs dans le JSON ? Oui, pour faire faux négatifs, pour évaluer.D'accord. Est-ce que ce genre d'erreur influence l'intersection ? Oui bien sûr, parce que dans les graphiques d'intersection vous allez voir, ah tiens, on a beaucoup plus de faux négatifs par exemple que de faux positifs, ce qui donne une représentation visuelle de la performance du modèle.

首先，您需要解决代码中的问题，打开 GitHub 仓库，将所有内容上传到那里，并开始编写文档，继续对数据进行处理。您已经有一些材料可以写了，因此需要在多个方面进行集中工作，并尽量推进进度。

关于错误的保存，您提到需要在 JSON 文件中保留错误数据，以便之后进行四个模型的性能评估。这样做是为了能够计算出假阳性和假阴性，并在图表中反映出来，以便通过视觉展示模型的表现。您关心的是，这种错误是否会影响交集（intersection）分析。答案是肯定的：假阴性和假阳性会影响交集图表的结果。例如，如果您发现假阴性明显多于假阳性，那么在交集图表中这一差异会显现出来，从而能够帮助您理解模型的表现。

* Dans un autre cas de figure, vous allez avoir un modèle qui sera plus performant dans le sens où votre cercle qui représente des faux positifs va être plus petit que par exemple le cercle qui représente des faux négatifs.Des cercles de faux négatifs ça veut dire ? Pas de faux négatifs, vous allez avoir faux positifs. Je ne sais pas si j'ai un... Je crois que l'intersection c'est juste un fold Carambère et on a l'intersection, mais les mots en commun et les mots différents entre deux cèpes. Ah oui, donc vous allez avoir l'intersection entre Stemfold et Carambère. Oui, mais ça, ça va être l'intersection des entités nommées.Oui, on peut faire l'analysation entre deux modèles et aussi entre plusieurs modèles.

在另一种情况中，您可能会有一个更高效的模型，这意味着表示假阳性的圆圈会比表示假阴性的圆圈要小。假阴性的圆圈是什么意思？没有假阴性，您将只看到假阳性。我不确定我是否理解...我认为交集就是Carambère模型的一个fold，包含交集以及两个模型之间的共同词汇和不同词汇。哦，是的，您将会看到Stemfold和Carambère模型之间的交集。

* Ce qu'elle m'a dit, c'est qu'avec les résultats qu'on a par exemple pour Hispécie, qui ne posaient pas de problème en tant que nombre de Tolkien et d'entités nommées, c'était qu'on n'aurait pas à avoir des résultats pour l'intersection, que ça aurait posé de problème. C'est ça que je pense qu'elle n'a pas compris ou que j'ai compris différemment. Et moi, je pense que le fait d'avoir des doublants, des choses comme ça, ça ne pose pas de problème pour créer une intersection.Mais quand vous dites doublant, vous vous référez à quoi concrètement ? Quand vous dites doublant, je ne pense pas à 84 quantités nommées, 84... Le étudiant de Jason pour le texte abolinaire.D'accord. On a le même mot, prévu de deux voix. Et il, pour l'organisation. Par exemple, il a mis organisation, une erreur comme ça.  
  Alors l'erreur, oui, ça on est d'accord qu'on le garde parce que de toute façon ça donne une information sur l'évaluation du modèle. Après il faut que vous compreniez pourquoi vous obtenez de deux blancs. Est-ce que c'est parce que... est-ce que c'est normal en fait ? Peut-être normalement... normalement vous êtes censé avoir une seule occurrence de votre identité nommée qui est labellisée comme loc. Peut-être, en l'occurrence, ça commence par le B en majuscule, peut-être, j'invente un peu l'exemple, peut-être il y aura un autre exemple où bibliothèque c'est écrit en minuscule et vous aurez une autre occurrence de bibliothèque nationale de France qui était interprétée comme ça.

她告诉我的是，以我们对Hispécie的结果为例，其中的Token数量和命名实体没有问题，这样我们就不需要为交集结果进行处理，因为那样会引发问题。我认为她没有理解，或者我理解得不一样。而我认为，重复项之类的东西并不会对创建交集产生问题。但是当你说重复项时，具体指的是什么呢？当你说重复项时，我并不是指84个命名实体，84个......比如学生在处理json文件时使用了相同的词语，像"abolinaire"这样的文本。我们有相同的词，重复了两次。例如，他把组织名称标记成了重复项，这就是一个错误。是的，关于错误，我们是同意保留的，因为这本身提供了对模型评估的信息。之后，您需要理解为什么会有重复项。是因为......这是不是正常的呢？也许，通常您应该只得到一次命名实体的出现，它应该被标记为"loc"（地点）。也许在这个例子中，"Bibliothèque"以大写字母B开头，另一个例子中，"bibliothèque nationale de France"可能被解析为另一个命名实体（即一个不同的识别）。

* Mais si vous tokenisez et si vous mettez tout ça en lowercase, ou même si vous ne faites pas comme ça, vous gardez les résultats. Mais encore une fois, je ne vous dis pas, juste en regardant votre exemple, je ne peux pas dire que c'est correct, parce que je ne sais pas si ça vient véritablement de votre code que vous nous avez montré ou pas.Donc, dès que vous êtes sûr que votre code est propre...C'est ici.Alors, ça c'est le résultat JSON, d'accord. Oui, et pour SPACIE... Ce résultat vient de quel code, en fait ?

但是如果你进行分词，并且将所有内容转换为小写，或者即使你不这样做，只要保留结果是可以的。但再次强调，仅仅通过查看你的例子，我不能确定它是否正确，因为我不清楚这些问题是否真正出自你给我们展示的代码。所以，当你确认你的代码没有问题之后...... 这就是要做的。

* Ça vient de Carambè ou ça vient de... de Spacey de Spacey ok. c'est le deuxième résultat pour Marina il met le résultat JSON de 10 ouvrages dans un même JSON et pour le code c'est ici et le premier résultat, ça vous écartez? Non, le premier résultat ça veut dire SM, le deuxième résultat ça veut dire LG. C'est juste ça. Mais je peux modifier les noms pour que ça soit plus clair.

好的，明白了。如果第一个结果是SM，第二个结果是LG，这样的命名方式可能确实会让人有些困惑。您可以将这些名称修改得更加清晰，以便于理解。您提到的JSON文件包含10个作品的结果，确实需要确保它们被正确地分开处理。如果将所有结果放在一个JSON文件中，需要特别小心确保每个作品的输出不会混淆，避免重复的实体或错误的数据聚合。

* Modifier les noms, c'est bon.Mais je vais mettre SM et LG, comme ça c'est plus clair que le premier et le deuxième. Alors, euh... Mais je vais mettre ça au github de toute façon. Est-ce que vous avez repéré d'autres doublons à part Bibliothèque Nationale de France dans le résultat Gisane? C'est le seul résultat. Qu'est-ce qui vous pose problème? C'est ça ce que je vais comprendre. Donc ça c'est un exemple de doublon. Est-ce que vous avez d'autres répétitions ou c'est la même chose? En fait, la question c'est est-ce que ces genres de... On va essayer de chercher, mais est-ce que ça pose des problèmes pour faire l'intersection? Parce que c'est ça qui hallucine Magie et je ne suis pas sûre d'être

修改名称是可以的。您打算将SM和LG用作名称，这样会更加清晰，区分第一个和第二个结果。好的，您也会将这些更新上传到GitHub。关于Gisane结果中的其他重复项，您提到的“Bibliothèque Nationale de France”是唯一的重复项吗？这是唯一的问题吗？我理解的是，您想确定这种重复是否会影响交集的计算。是否存在其他的重复项，或者只是“Bibliothèque Nationale de France”重复？关键问题是这些重复项会不会对计算交集产生影响，因为Magie提出了这个问题，我不确定我完全理解这个问题。

* Et aussi pour faire la même chose. Oui. Ou finalement on choisit seulement un modèle pour faire des choses. Encore une fois, si dans votre JSON, toutes les entités nommées ou tous les tokens qui ont été labellisés comme entités nommées... Oui, qui ont été labellisés, parce que dans le JSON, vous n'avez que des entités nommées, on est d'accord ? Oui. Ok, donc dans ce JSON, il faut regarder si vous avez d'autres doublons par exemple, parce que je ne peux pas savoir pourquoi vous avez ce doublon il faut voir dans le code il faut voir même dans le texte d'entrée ce qui se passe véritablement parce qu'a priori ce sont des textes correctement orthographiés ce ne sont pas des textes gruilletés où vous avez une bibliothèque écrite de manière telle ou telle, de façon aléatoire. Donc, si vous pouvez me dire si vous avez d'autres doublants, ça serait utile pour moi parce que maintenant je ne peux pas vous dire exactement si ça va fausser les résultats ou pas. Vous pouvez aller directement dans G-Zone pour voir si vous avez d'autres...

并且也要做同样的事情。是的，或者最后我们选择只使用一个模型来进行操作。再说一遍，如果在您的JSON中，所有标注为命名实体的实体或所有标记为命名实体的tokens... 是的，所有标注的实体，因为在JSON中，您只会有命名实体，是吗？是的，好的。那么，在这个JSON中，您需要检查是否有其他重复项，例如，因为我无法知道为什么会有这个重复，您需要查看代码，甚至要查看输入文本的内容，看看到底发生了什么。因为按理说，这些文本应该是正确拼写的，并不是一些乱七八糟的文本，在其中“bibliothèque”可能会以不同的方式随机拼写。所以，如果您能告诉我是否有其他重复项，对我来说会非常有帮助，因为现在我不能确切地告诉您它是否会影响结果。您可以直接去G-Zone查看是否有其他重复项吗？

* Est-ce que c'était le seul exemple que vous avez référé comme problématique ou il y avait d'autres... Je vois qu'il existe d'autres. Il existe d'autres? Pour faire un gros exemple, quatre modèles et pour chaque texte, ce genre de... Ce genre de...Vous pouvez me montrer d'autres exemples juste pour voir? J'essaie de trouver ici aussi. Parce que, enfin, oui. C'est pour spécies moe, c'est pour spécies larges.D'accord, ok. Bon, pour l'instant, alors, je vous dirais, même si pour l'instant, si vous avez des doublants, ça ne veut pas forcément dire qu'il s'agit des erreurs. Quoique ça c'est quelque chose qu'il faut prendre avec un peu de réserve. Je vous ai dit que, et même on s'était mis d'accord que...Ce genre avec des caractères.Oui parce que ça veut dire qu'il s'agit de nouvelles lignes. Oui.

是的，您提到的这个例子是唯一一个有问题的例子，还是说还有其他的重复项？我看到确实有其他的重复项吗？要做一个大的例子，四个模型，每个文本都要做这样的检查...您能展示一下其他的例子吗，看看是什么情况吗？我也在这里尝试寻找一些。因为，嗯，嗯。是针对特定物种的例子吗，还是针对更广泛的物种？好的，明白了。那么，暂时来说，如果您有重复项，并不一定意味着它们是错误的，尽管这点需要谨慎对待。我之前提到过，而且我们也达成共识了...这些带有字符的重复项。是的，因为这意味着它们是新的行。

* L avec apostrophe org, il a écrit comme org.Oui parce que ça c'est dans la préambule ou dans l'introduction qu'ils vous disent... Des trucs avec apostrophe, ils reconnaissent avec des choses différentes, ils repèrent des choses avec apostrophe. Oui, non mais ok, si vous avez d'entités distinctes, c'est ok. Oui, le problème c'est les doublons, il faut pas trouver les doublons. Alors les doublons, on les voit dans la description, mais en tout cas c'est pas votre corps du texte, ça c'est juste l'avertissement de l'éditeur, c'est pas vraiment le corpus que vous traitez, mais dans le corpus... Il y a plusieurs L' par exemple, il a sorti plusieurs L'.Oui, oui, oui, tout est en or. Tout est en or. Alors, encore une fois, vous pouvez poser la question même aux autres enseignants. Moi, ce que je peux vous dire pour l'instant, c'est que pour l'instant, c'est pas très grave, enfin, c'est pas problématique parce que si jamais vous devez écarter les doublants, ça encore une fois, vous pouvez faire dans le code.

带撇号的L，他写作org。是的，因为这是在前言或介绍中提到的...带撇号的东西，他们把这些东西识别为不同的东西。是的，没错，如果你有不同的实体，那就没问题。问题在于重复项，不应该有重复项。重复项我们在描述中看到，但这并不是你处理的正文部分，这只是编辑的警告，并不是真正的语料库内容。但在语料库中... 比如说有多个L'，他输出了多个L'。是的，是的，所有都没问题。一切都可以。所以再次强调，你可以向其他老师询问这个问题。我目前能告诉你的是，暂时这不是一个大问题，或者说不是什么问题，因为如果你真的需要去除重复项，那你可以在代码中处理这个问题。

* Vous pouvez clairement dire, ok, on va ignorer les doublons. Voilà, encore une fois, ça c'est à mon avis moins prioritaire. Je dis pas qu'on va le garder plus tard parce que il va falloir peut-être rassembler tout ça dans un seul, dans une seule antithéorémie, d'autant plus que toutes les antithéorimèques portent la même labelle. Voilà donc je vais pas dire plus que ça. On avait vraiment dit aujourd'hui ce que vous devrez faire et on va pas vous... en tout cas moi je vais pas vous rajouter de nouvelles informations jusqu'à... avant que le travail principal soit fait. Ce sera plus facile pour vous parce que vous allez dégager les queues qui restent. Et ensuite vous pouvez réfléchir à... Ensuite vous pouvez analyser vraiment vos résultats dans le JSON et surtout à partir des graphiques que vous allez générer. Maintenant, il existe des erreurs pour le nombre de nommés pour gammon père et aussi pour flotter parce que c'est à peu près les mêmes pour un dénommé de chaque tournement. C'est à peu près la même.

你可以明确地说，好吧，我们将忽略重复项。再说一次，在我看来，这个问题的优先级较低。我不是说我们以后会保留它，因为可能最终我们需要将所有这些合并为一个单独的实体，尤其是因为所有的实体都有相同的标签。所以，我不会说更多。今天我们确实已经讲了你们需要做的事情，直到主要工作完成之前，我不会再给你们添加新的信息。这样对你们来说会更容易，因为你们将能够清理剩余的部分。然后，你们可以开始分析你们在JSON中的结果，尤其是从你们将要生成的图表中来分析。现在，存在一些关于Camembert和Flotte命名实体数量的错误，因为每次处理的命名实体数量差不多，差不多是一样的。

* Attendez, je comprends pas. C'est quoi l'erreur?  
  Nombre total de dénommés.D'accord, par auteur. Donc ça ça c'est 75 pour Apollinaire.Ça c'est pour Flaubert et ça c'est pour Gammonbert.Et vous avez donc le nombre qui est différent, on est d'accord ? Parce que c'est pas le même modèle qui traite les données. C'est par rapport à une autre entité nommée de spécies, small et large. C'est quelque 1000, mais pour Flaubert, c'est...Oui, ça, il faut le compiler. J'ai mal écrit. Il faut que je refasse. Mais j'ai déjà mis NA. Il faut juste le compiler.C'est aussi le nombre de trucs, mais pasest ici pour Flaubert. Et c'est un peu comme Camembert, c'est par rapport au résultat de spécifiement...Donc maintenant, ça ce sont des bons résultats, c'est ça ce que vous voulez me dire ?

等一下，我不明白。是什么错误？

命名实体的总数。好的，按作者来。这个是75个，针对Apollinaire的。这个是针对Flaubert的，那个是针对Gammonbert的。你们的数字是不一样的，对吗？因为处理数据的模型不同。它是基于另一种命名实体模型，比如specie、small和large的。这个是1000个，但对于Flaubert，是……是的，这个需要编译。我写错了，我需要重新做。但我已经标记了NA。只是需要编译一下。这也是数量，但不是Flaubert这里的。它有点像Camembert，它是针对specifiement的结果……所以现在，这些是正确的结果，是你们想告诉我的吗？

* Là, c'est pas les bons résultats ici. Il faut le recompiler, ça va être N1, parce qu'il faut que je mette ça ici. D'accord, ok. C'est ça, je pense que c'est ça, je sais pas si c'est ça que tu veux dire. Non, ce que je veux dire c'est par rapport au résultat dans le dénommé de spécies, moins spécies larges, le nombre dans le dénommé pour flaubert et aussi pour camembert. Ah, beaucoup plus petit, c'est ça ?  
  Ouais. Oui, mais je pense que c'est parce qu'ils fonctionnent moins...  
  Ça c'est le problème avec le découpage, ça c'était pour Camembert. Et pour Flaubert aussi vous aviez le problème de...Le même problème, je crois que c'est le problème de traitement.Chaque fois c'est 100, 12...Ah, pour les limites, c'est pour ça qu'on vous a dit, réglez d'abord le problème de code. Il faut que le tableau soit complet.Oui, pour qu'on puisse comparer bien.

好的，这里不是正确的结果。需要重新编译，这应该是N1，因为我需要把它放在这里。明白了，好的。我觉得就是这样，我不知道这是不是你想表达的意思。

不，我是说，关于specie和large-specie中的命名实体结果，Flaubert和Camembert的命名实体数量。哦，是的，明显更小，对吧？

是的，没错。我觉得这可能是因为它们的处理方式不同......

这就是Camembert的切分问题。对于Flaubert来说，你们也遇到了类似的问题。我认为这是处理上的问题。

每次是100，12……哦，关于限制，这就是为什么我们告诉你们，先解决代码问题。这样表格才能完整，方便我们进行比较。

* Alors, ce sont des résultats importants, il faut se concentrer dessus. Ensuite, on regarde, ah ah, le JSON, on a des doublants. Après, encore une fois, c'est ce que Caroline vous a dit, que vous devez penser aussi aux intersections. Donc, vous pouvez aller par étapes, par étapes. Vous réglez d'abord ça, et ensuite vous regardez Gisonne et intersection.  
  Ou bien, vous faites maintenant le travail à la réflexion sur Gisonne et intersection, et puis vous revenez ici. Mais c'est un peu à vous de voir ce qui vous arrange le plus. Voilà donc ok c'est une question légitime je suis bien d'accord. Donc encore une fois ce que je pense aussi il faut vérifier aussi c'est que vous allez devoir dire ah ah on a des doublons on va juste  
  on va réduire tout ça en une seule entité nommée.

好的，这些结果很重要，需要集中精力处理。然后，看看JSON，看看有没有重复项。再者，正如Caroline所说，你们还需要考虑交集的问题。所以，你们可以一步一步来。先解决这个问题，再看看Gisonne和交集的问题。或者，大家现在可以开始思考Gisonne和交集的问题，然后再回到这个问题。总之，怎么做取决于你们自己，哪种方式对你们来说最合适。好，明白了，这是一个合理的问题，我完全同意。再说一遍，我认为你们还需要确认的是，你们会需要明确地说：“哦，原来有重复项，我们把它们合并成一个命名实体”。

* Peut-être comme ça, ça sera plus équilibré. Mais c'est ça, il faut qu'on ait un tableau pour comparer. C'est ça qui... si c'est un peu le bazar, on ne peut pas faire ça en fait.Oui, ça c'est le plus important, effectivement, parce qu'après, oui, vos graphiques, vos entités nommées. Si vous n'avez pas le nombre d'entités nommées exactes, ça vous bloque pour plus tard. On remplit ce tableau et on commence à se poser des questions.

也许这样会更平衡。但确实，我们需要一个表格来进行比较。如果数据有些杂乱，那我们就无法进行比较了。是的，这确实是最重要的，因为之后，如果没有准确的命名实体数量，后续的图表和分析会受到阻碍。我们先填好这个表格，然后开始提出问题。

* Et peut-être améliorer les choses.Et n'oubliez surtout pas la rédaction, donc ça vous rajoute encore du travail. Vous avez un peu de retard, mais c'est rattrapable parce que vous avez déjà du matériel concernant l'introduction. Déjà il faut expliquer dans le mémoire ce que vous faites, quel est l'intérêt de ce que vous faites. Donc voilà. Mais aussi audio et canals pour préserver le projet en demi ?Alors déjà oui, ça c'est une bonne question parce que j'attends le retour de M. Lejeune. Si M. Lejeune dit qu'il sera là ou en vision en présentiel le 25, je serai là. Sinon on va reporter la réunion. Parce que ça dépend en fait de lui. C'est vraiment en attendance, il me paraît qu'il va devoir rebondir sur la conversation de madame Coudreau Parfait, peut-être il va se manifester, il va dire qu'il sera là ou pas là, donc ça dépend de sa réponse. Et ensuite je vais, je vais, je vais dire si je vais être là en fonction de sa réponse.

而且可能会有所改进。别忘了写作部分，这会增加你们的工作量。虽然有些进度延迟，但可以赶上，因为你们已经有了关于引言的素材。首先，需要在论文中解释你们在做什么，以及你们所做工作的意义。关于音频和会议部分的问题，首先是的，这是一个好问题，因为我在等待Lejeune先生的回复。如果Lejeune先生说他会在25号现场或线上出席会议，我会在那里。否则我们会推迟会议。因为这取决于他。如果他在Madame Coudreau Parfait的讨论后有回应，说他能否出席，那就要看他的回答了。然后，我会根据他的回答决定是否出席。

* Donc je pense pas que vous aurez besoin de poser la question si on aura une réunion. Si jamais j'oublie ou si voilà il y a quelque chose qui arrive, vous pouvez m'envoyer un mail ou à tout le monde, pas seulement à moi, pour demander si on aura une réunion. Mais encore une fois, attention, si monsieur le jeune dit qu'il sera pas là, moi, pour demander s'il y aura une réunion. Mais encore une fois, attention, si Monsieur Lejeune dit qu'il ne sera pas là, moi je vais dire, ok, on reporte la réunion.C'est en ligne ou c'est...Non, si il dit qu'il n'est pas disponible du tout, que ce soit en présentiel ou en visio, on va reporter après les vacances. Parce qu'en fait, que dans deux semaines c'est toujours des vacances. Donc il ne faut pas s'attendre à ce que le monde sera disponible pour faire des réunions.

所以我觉得你们不需要再问是否会有会议。如果我忘记了或者有什么事情发生，你们可以发邮件给我，或者发给大家，询问是否会有会议。不过再说一次，如果Lejeune先生说他不能来，无论是现场还是线上，我会决定推迟会议。是在线上还是...？不，如果他说他完全无法参加，无论是现场还是线上，我们就会推迟会议到假期后。因为实际上在两周后也是假期。所以大家不应该期望大家会有时间来开会。

* Et peut-être que ce n'est pas institutionnel non plus.C'est un peu, comme vous le constatez, la maison de la recherche reste ouverte même pendant les vacances. Oui, oui, pour nous ce n'est pas grave. Mais comme vous avez constaté aussi, quasiment tout le monde sera en vacances.Oui, non, c'est un problème. Nous on est exposés en visio, c'est juste moi le matin. Non, non, c'est juste pour dire qu'attends justement la réponse de M. Lejeune parce que je préfère qu'on soit plus nombreux, donc qu'au moins deux personnes pour qu'on puisse avoir un peu la même vision des choses. Là, on avait fait un bilan pour M. Lejeune parce qu'il n'est pas là. Mais si on n'est qu'une seule personne avec vous deux, alors il va falloir mettre beaucoup plus de travail pour faire un bilan, pour envoyer ça aux autres directeurs et co-encadrants.Donc c'est mieux si on est tous sur la même page à peu près, quitte à reporter pour quelques jours plus tard, vous ne perdez pas grand chose. Si on ne fait pas une réunion le 25, bon ça va, on va reporter peut-être pour une semaine plus tard, mais ça ne va pas vous... vous ne serez pas beaucoup plus en retard. Au moins vous aurez un peu plus de temps pour avancer sur certaines choses que vous n'avez peut-être pas réussi à faire jusqu'à ce moment là. Parce qu'il y aura toujours du travail à faire. Mais deux semaines c'est un peu... c'est bien de voir ce délai parce que ça vous oblige de travailler plus et comme ça vous ne perdez pas beaucoup de temps. Sur ça.

对，这个确实是一个问题。如果我们只有你们两个人，那就得投入更多的工作，做一个总结报告，然后发送给其他的主管和共同导师。这样的话会更好，至少我们能保持一致的视角。如果我们要推迟几天，那也没关系，你们不会因此丢失太多时间。反而你们可能会有更多时间来推进一些可能在这个时刻没有完成的任务。因为工作总是有做不完的事情。但是两周的延期还是有意义的，它能迫使你们更集中精力工作，这样你们就不会浪费太多时间。所以如果25号的会议不能按计划举行，我们可以推迟一周，这不会让你们特别落后。最重要的是，我们能在同一页上工作，这样进展会更顺利。

**Les conseils donnés par prof:**

* donner les images directements mais pas les montrer dans le code
* Dans un cas, la tokenization n'a pas été effectuée au préalable, tandis que dans un autre cas, elle l'a été. Professeur : Cela peut entraîner une "sur-tokenization" (over-tokenization). Cela me rappelle les tests que j’ai réalisés au début de ma thèse. À l'époque, j'effectuais d'abord la tokenization, puis j'appliquais la reconnaissance des entités nommées (NER). Mais plus tard, en relisant la documentation, j'ai découvert que la tokenization n'était en réalité pas nécessaire, car le modèle l'intégrait déjà. Par exemple, dans SpaCy, il existe un mécanisme de traitement intégré qui gère déjà cette étape.在一个情况下，没有事先进行 tokenization，而在另一个情况下，进行了 tokenization。老师：这可能会导致“过度 tokenization”（over-tokenization）。这让我想起了我博士研究刚开始时做的测试。当时，我先进行了 tokenization，然后再应用命名实体识别（NER）。但后来我重新阅读了文档，发现其实不需要进行 tokenization，因为模型内部已经处理好了。例如，在 Spacy 里，它已经有内置的处理机制。
* Vous pouvez mentionner dans votre mémoire : "Nous avons précédemment effectué une tokenization, ce qui est manifestement la cause du problème."Vous pouvez illustrer cela en expliquant : "Si une tokenization est réalisée en amont, puis qu’une autre tokenization est appliquée par la suite, cela peut conduire à des résultats différents."Pourquoi ? Parce que l’outil va retraiter le texte, mais il aura déjà perdu certaines informations de contexte.可以在论文里提到：我们之前已经进行过一次 tokenization，这显然是原因所在。你可以展示出来，比如说：“如果在前面进行了 tokenization，而后面又进行了 tokenization，会导致不同的结果。” 为什么？因为工具会重新处理文本，但它已经失去了上下文信息。
* utilisier *match.flip.plt.savefig* pour enregistrer les images sur bureau, on ne sera pas obligé d'actionner le code.chaque fois
* mettre *plt.savefig("filename.png", dpi=300)* à la fin de votre code, et puis vous enregistrez les images au format png. Donc comme ça on pourrait les regarder sans avoir à faire tourner le programme
* Pendant le traitement du corpus, notez les éléments importants et inscrivez-les dans le document Overleaf. Essayez de documenter toutes vos actions dans la mesure de votre travail

**Les points à clarifier :**

* Pour graphique des intersections : des diagrammes de veines ou des upset plots qui sont en bâton
* le diagramme de veine est conseillé par prof,le diagramme de veine fait deux cercles qui se coupent
* les données changent selon les textes. Ce n'est pas les mêmes textes, donc ce n'est pas les mêmes entités nommées, donc ce n'est pas la même quantité. Donc c'est normal que nous n'avons pas les mêmes résultats que le mémoire proposé par prof
* Il est important de commencer par expliquer clairement ce que sont les tokens, les entités nommées (EN) et les mots candidats
* Les entités nommées opposées : Cela pourrait faire référence aux "erreurs" ou "résultats incorrects" générés par la machine lors de l'identification des entités nommées. En général, dans la reconnaissance des entités nommées (NER), la machine tente d'extraire des entités nommées à partir du texte (comme des noms de personnes, de lieux, d'organisations, etc.). Cependant, en raison des limites des algorithmes ou des erreurs de modèle, la machine peut parfois identifier incorrectement certaines entités ou marquer des mots qui ne sont pas des entités nommées comme étant des entités. Ces résultats erronés peuvent être considérés comme des "entités nommées opposées".反义命名实体：“反义命名实体”可能是指机器生成的“错误”或“不准确”的命名实体候选项。通常，在命名实体识别（NER）中，机器会尝试从文本中提取命名实体（例如人名、地点名、组织名等）。然而，由于算法的局限性或模型的误判，机器有时会错误地识别一些实体，或者将非命名实体的词语标记为命名实体。这些错误的识别结果可以被视为“反义命名实体”。
* Si vous souhaitez utiliser certains modèles qui permettent de modifier les paramètres, ceux-ci prioriseront soit la précision, soit l'efficacité. L'efficacité traitera les données plus rapidement, mais peut être moins précise, ne récupérant pas tous les entités nommées, ou risquant d'identifier à tort une entité comme une entité nommée erronée, alors qu'elle ne l'est pas. La précision, en revanche, prendra plus de temps pour traiter les données, mais génère généralement des résultats plus précis.如果你们想使用一些模型，允许改变参数，这些模型会优先考虑精度或效率。效率会更快地处理数据，但可能会不太准确，可能不会回收所有的命名实体，或者可能会出现错误，认为某个实体是反义命名实体，实际上它不是。而精度则相反。精度会花更多时间处理数据，但通常会生成更准确的结果。
* Nous devons répéter le processus pour chaque texte et chaque modèle. Quoi qu'il en soit, cela nous permet de faire des comparaisons. Peut-être que nous ne mettrons pas tous les résultats dans l'article final, mais au moins, cela nous aide à choisir les résultats d'intersection les plus significatifs pour nous. Ainsi, pour chaque texte et pour l'ensemble du corpus (ce qui nous aide à adopter une perspective macro), nous devons effectuer une analyse des intersections des quatre modèles.我们需要对每个文本、每个模型重复执行processus，无论如何，我们可以做比较，也许最后我们不会把所有的内容都放进论文里，但至少这样做能帮助我们选择对我们来说最有意义的交集结果。所以，每个text和整个corpus（可以帮助我们在宏观的角度看待问题），都要做四个模型的交集分析。
* Avec cela, vous pourrez utiliser le tableau pour fournir plus de détails dans votre résumé. Sur la base de ces résultats, vous pouvez créer un autre tableau, par exemple, montrant les résultats des intersections, car actuellement ce tableau présente la distribution des entités nommées, ce qui est simplement une méthode de visualisation, montrant le contenu commun à chaque modèle.相信有了这些，你们可以用表格提供更多的细节来进行总结。基于这些结果，你们可以创建另一个表格，比如显示交集的结果，因为现在这个表格，展示的是命名实体的分布，这只是一种可视化方法，它展示的是每个模型共有的内容

**Les codes donnés par prof:**

plt.savefig(dpi=300) 【il faut modifier pour s'adapter】

plt.savefig("filename.png", dpi=300) 【le nombre de dpi c'est pour la qualité de l'image】



**Les liens donnés par prof:**

<https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.savefig.html>

[https://python-graph-gallery.com/venn-diagram/](https://python-graph-gallery.com/venn-diagram/" \t "/Users/zhengruixing/Documents\\x/_blank)

[https://github.com/These-SCAI2023/CORPUS](https://github.com/These-SCAI2023/CORPUS" \t "/Users/zhengruixing/Documents\\x/_blank)

**Pour le moment:**

**Ne soyez pas pressé d'avancer, ne sautez aucune étape et terminez le travail de manière complète.**

**Tableau:**

* Le tableau sur tout le corpus et chaque text doit parler de la densité, la durée d’exécution, du nombre de tokens et du nomdre d’EN pour chaque modèle.
* Il faut changer le paramétre pour vérifier si Flaubert est bien appliqué sur tout le corpus
* rajouter Spacy-sm et Spacy-lg dans le tableau

**Latex：**

1.Génération automatique de tableaux :

-Vous pouvez utiliser LaTeX pour générer des tableaux au lieu de les remplir manuellement.

-En utilisant un script (par exemple en Python), vous pouvez exporter automatiquement les données au format tableau LaTeX, puis les insérer directement dans votre document LaTeX, ce qui évite la saisie manuelle.

1.自动生成表格：

你可以用 LaTeX 生成表格，而不是手动填写数据。

通过 脚本（比如 Python）来自动输出 LaTeX 格式的表格数据，然后在 LaTeX 文档中直接插入，避免手工输入。

2.Utilisation de la commande \input pour gérer les documents :

-Dans LaTeX, vous pouvez utiliser \input{fichier.tex} pour diviser différentes parties du document, comme les résultats d’analyse de textes, afin de mieux organiser votre travail.

-Cela vous permet de gérer plusieurs fichiers en même temps (résultats pour différentes œuvres) sans que votre document LaTeX ne devienne trop volumineux.

2.使用 input 命令管理文档：

在 LaTeX 中，你可以使用 \input{fichier.tex} 来分割不同部分，比如不同的文本分析结果，方便管理。

这样你可以同时处理多个文件（不同作品的分析），而不会让 LaTeX 文件变得太庞大。

**Discussion sur le tableau:**

-À travers le tableau basé sur l’ensemble du corpus, nous pouvons voir le nombre de tokens, le nombre total d’entités nommées (EN) et observer la performance moyenne des modèles. Cependant, se limiter uniquement à la moyenne peut ne pas suffire, car différentes œuvres (documents) peuvent présenter des variations. Certaines œuvres peuvent contenir plus d’entités nommées (densité plus élevée), tandis que d’autres en ont moins (densité plus faible).Nous devons donc analyser comment la densité varie à l’intérieur du corpus et quelles sont les différences entre les différents modèles.

通过针对整个corpus的表格，我们看到了tokens的数量、命名实体（EN）的总体数量，可以观察到模型的平均表现等。但是，仅仅看平均值可能还不够，因为不同的作品（文献）可能存在差异。有些作品可能包含的命名实体更多（密度更高），而有些则较少（密度较低）。我们需要去观察密度如何在语料库内部进行变化？以及不同模型之间的区别是怎样的？

**Modèle:**

* Confirmer s'il existe une limitation sur le nombre de tokens traités par Flaubert (c'est-à-dire vérifier si Flaubert a traité une partie réduite du texte en raison des paramètres par défaut).Vérifier si les résultats actuels sont exacts :Si les résultats sont corrects, on peut continuer à faire des statistiques et des analyses par œuvre.Si les résultats ne sont pas corrects, il faut préciser que ces résultats ont été obtenus avec les paramètres par défaut, puis approfondir l'étude pour voir s'il existe réellement une limitation du nombre de tokens. Ensuite, ajuster ces paramètres afin d'obtenir des résultats plus cohérents.确认 Flaubert 处理的 token 数量是否存在限制（即检查 Flaubert 是否因为默认参数的限制导致少处理了一部分文本）。确认当前结果是否准确：如果结果是正确的，可以继续按照不同作品进行统计和分析。如果结果不正确，就需要说明：这是在默认参数下得到的结果，并进一步深入研究，看看是否真的存在 token 处理限制，然后进行调整，以获得更合理的结果。

**Mémoire:**

* En observant le nombre d'entités nommées dans le tableau, nous pouvons constater que Flaubert a produit des résultats étranges, car le nombre d'entités nommées générées est extrêmement faible. Vous pouvez tout à fait expliquer ce phénomène en indiquant qu'il est nécessaire d'ajuster certains paramètres afin de garantir que Flaubert traite l'intégralité du corpus, et non seulement une partie des documents. Car il est évident que 77 entités nommées ne correspondent pas à un traitement complet du corpus. C'est un phénomène intéressant, et vous devrez l'expliquer dans votre thèse.Vous pouvez conserver les 77 résultats de Flaubert et les inclure dans le texte principal, mais lorsque vous mettrez à jour le code et découvrirez que Flaubert doit traiter l'intégralité du document, vous utiliserez les résultats mis à jour, c'est-à-dire ceux de la deuxième version. Toutefois, vous pouvez également expliquer dans le texte que les 77 résultats initiaux ont été obtenus pour une raison particulière.
* En créant un tableau montrant combien de noms d'entités ont été extraits de chaque œuvre, nous pouvons voir que dans une œuvre, nous pouvons avoir 100 000 entités nommées, tandis que dans une autre œuvre, il n'y en a que 50 000. C'est une information très intéressante, qui peut être utilisée pour expliquer que dans certains cas, il y a moins d'entités nommées. Vous pouvez même formuler des hypothèses pour expliquer pourquoi cela se produit.
* Assurez-vous que tout le contenu généré dans les différents documents soit importé dans Google Drive. Si lors de la prochaine réunion, quelqu'un vous dit que certaines choses ne devraient pas être incluses ou qu'il y a des modifications à apporter, ce n'est pas grave. Cela ne fera pas perdre de travail potentiellement précieux, et évitera que ces éléments restent simplement dans vos disques durs personnels ou vos notes. Il est donc préférable de rédiger le texte et d'y inclure tout le contenu.
* Cette discussion du tableau ne nécessite pas de longs discours, elle doit simplement expliquer le tableau, car le tableau en soi n'est pas suffisant. Il constitue le cœur de votre travail, mais c'est sur cette base que vous devez expliquer ce que vous avez fait. Ce que nous voyons, c'est le tableau, mais dans le corps du texte, vous pouvez peut-être ajouter quelques commentaires qui proviennent du tableau, mais qui ne sont pas nécessairement évidents, car vous avez déjà travaillé dessus. Par exemple, vous pouvez revenir sur Flaubert. Supposons que vous ayez obtenu des données sur les entités nommées, ces données seront écrites dans le tableau, mais comme les autres lecteurs externes ne savent pas comment vous avez obtenu ce chiffre, vous devez le décrire dans le texte, car dans la version finale du tableau, vous aurez un nouveau chiffre. Dans le texte, vous pouvez dire : au début, nous avons rencontré ce problème car le modèle n'a pas traité l'intégralité du texte, ce qui a donné 77 entités nommées, mais nous avons finalement corrigé les paramètres et obtenu un nouveau nombre d'entités nommées. Pour l'instant, nous parlons de Flaubert, car il avait des problèmes lors du traitement du corpus.
* Pendant le traitement du corpus, notez les éléments importants et inscrivez-les dans le document Overleaf. Essayez de garder un équilibre, car mettre tout dans le document est une bonne chose, mais il faut aussi savoir comment le traiter. Donc, si vous présentez tout dans votre thèse, c'est bien, mais vous devez être conscients qu'il faut sélectionner certaines choses. Il faut être clair sur ce qui est pertinent, et vous pouvez expliquer au professeur : 'Nous avons décidé d'inclure ce contenu, mais nous ne sommes pas sûrs qu'il soit approprié.' Ensuite, le professeur pourra nous dire ce qui est pertinent. Ainsi, ne vous souciez pas trop de la forme, il est préférable d'inclure tout le contenu, puis nous pouvons noter toutes les informations, même si c'est juste pour nous-même, mais les inclure dans le document. Car si nous oublions quelque chose, ce serait un problème. Si vous indiquez 'ceci est une note' et 'voici ce que nous voulons vraiment inclure dans la thèse', cela clarifie les choses et aide le professeur à comprendre l'avancement du projet. Cela permettra aussi au professeur de décider si certains éléments doivent être conservés ou non.

**Nos idées：**

* Essayer de décrire clairement les questions que nous souhaitons poser au professeur
* Après chaque RDZ, assurer d'avoir bien compris ce que les profs ont dit ainsi que les tâches qui nous ont été attribuées

**Attention:**

* accepter les conditions de google pour importer le corpus dans google colab
* il faut bien organiser le temps pour ne pas le gaspiller inutillement et ralentir l’avancement du projet
* n’hésiter pas d’écrire le problème à prof avce la capture d’écran
* Avant chaque rendez-vous, envoyez par email à votre professeur ce que vous avez réalisé.
* Après chaque rendez-vous, confirmez les tâches à réaliser pour la prochaine étape.

**Le calendrier des RDV** （toutes les deux semaines ）

|  |  |  |
| --- | --- | --- |
| Date(à 9h) | Lieu | Temps |
| 14/03 | sur place (Bureau 211,Maison de recherche) | 1h30mn |
| 28/03 | sur place | 1h30mn |
| 11/04 | sur place | 1h30mn |
| 25/04 | sur place |  |
| 9/05(12/05 à 14h30 en remplacemen) | sur place |  |
|  |  |  |

Il est probable que les rendez-vous pendant les vacances d'été soient des réunions en ligne. Si nous souhaitons avancer le plus possible sur le mémoire avant les vacances, il faudra envoyer plus de mails à profs pour poser des questions (code, résultats, sorties, captures d'écran).