**Notes du 28/04 (RDZ No.6)**

**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.**

**我的法语好差，但是这不应该成为不努力表现和严格要求自己的理由**

**Lejeune老师限时完成自己的任务，**

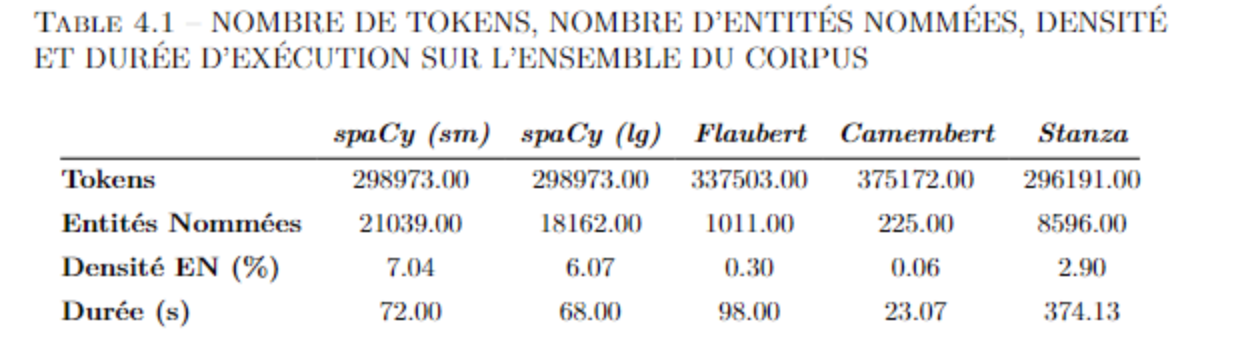
Les suggestions de l'enseignant pour améliorer le code de Camembert sont :

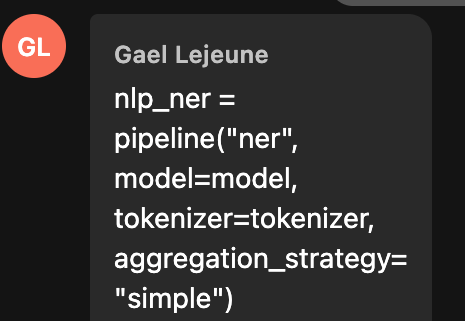
老师建议使用的改进camembert的代码：

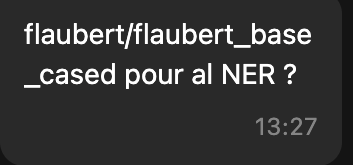
<https://github.com/These-SCAI2023/CORPUS/blob/master/prog/Use_camemBERT.py>

Le problème du code : il faut étudier attentivement les codes de Flaubert et Camembert.

代码的问题：flaubert和camembert的代码要细细研究







Conservez le problème de Flaubert (qui reconnaît 17 entités nommées) et comparez-le avec Spacy (qui en reconnaît plusieurs milliers). L'objectif est de comparer simplement le nombre d'entités nommées (EN) identifiées par les deux modèles.

保留flaubert（识别出17个EN）的问题，做表格和spacy（识别出几千个EN）进行比较。比较EN的数量就可以了

Créez un notebook séparé : sélectionnez une phrase ou un texte poétique et testez-le avec les quatre modèles en même temps, en vous assurant que la taille d'entrée est identique pour tous.

单独的notebook：找一句或者一个诗歌文本，用四种模型一起进行测试【确保taille d’entree是一样的】

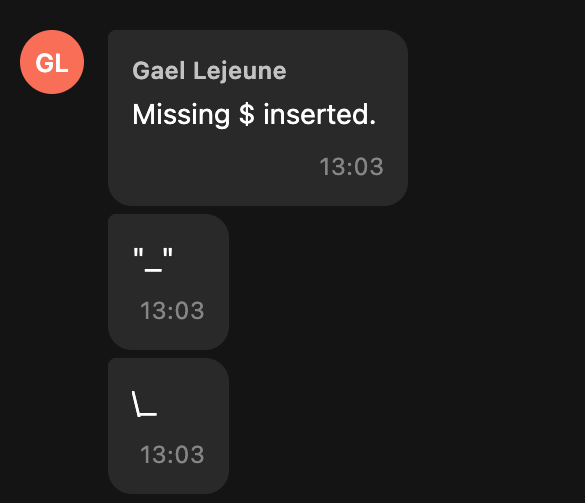
Latex：

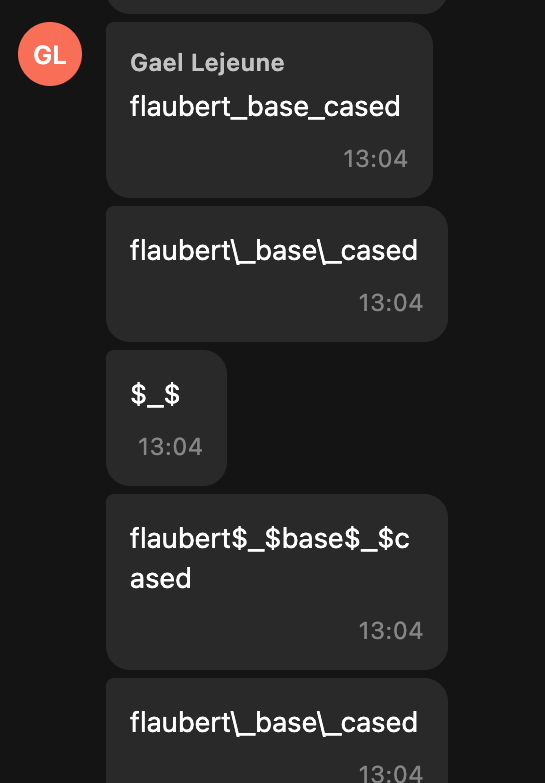
Il vaut mieux continuer à écrire sur Overleaf, car les différentes parties sont séparées, ce qui permet de les modifier indépendamment.

要么继续在overleaf上写，要么在wps里写，然后，保持更新。

En plus, vous pouvez les télécharger sur GitHub, ce qui permet de bien conserver l'historique des modifications de chaque partie.

最好在overleaf上写，因为各个部分是分开的，可以分别进行修改。同时可以上传到github上，各部分的变化轨迹，可以很好得保留。





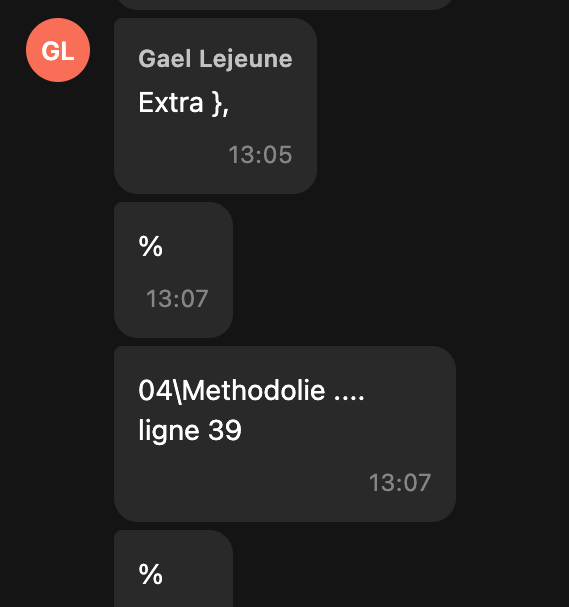


Il faut examiner attentivement les erreurs, car il y en a beaucoup, ce qui entraîne des problèmes de compilation.

要仔细看错误，因为有很多的错误，所以才引起compiliation的问题

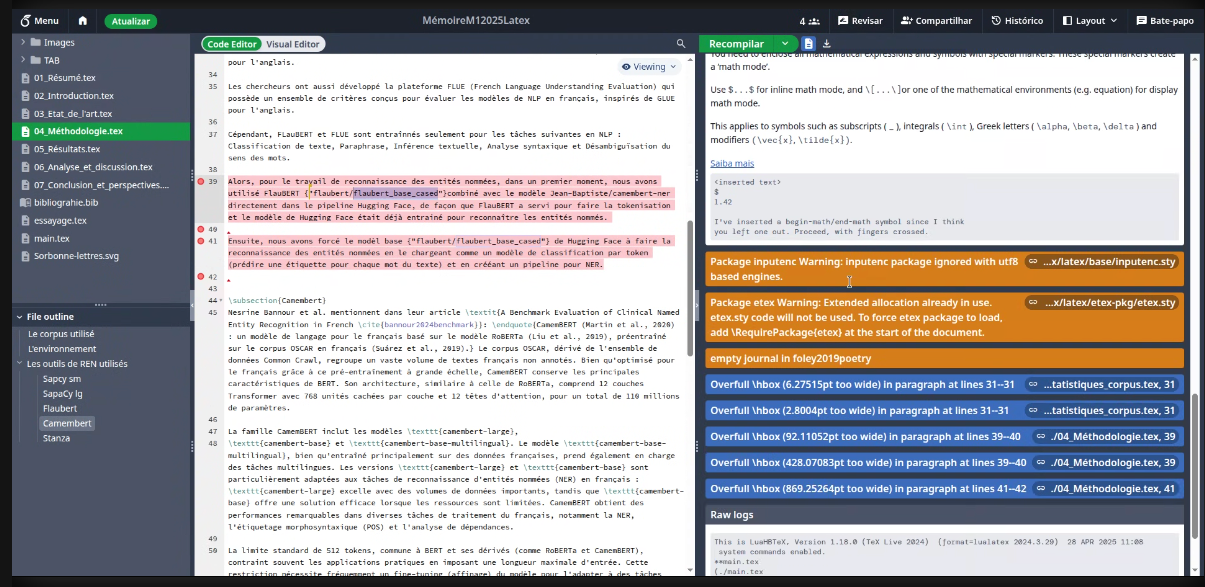
Dans LaTeX, les erreurs sont affichées à gauche, sur quelle ligne elles se trouvent, donc il suffit de les corriger sans que cela n'affecte le travail futur.

latex上有显示错误在左边的哪行，就去修改，不妨碍之后的工作



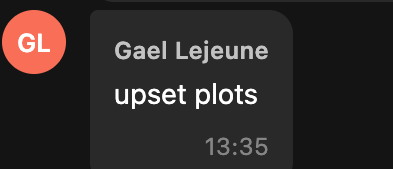
Les éléments en jaune indiquent qu'il manque des informations dans le fichier LaTeX.

黄色的东西，表示latex里面缺少信息



* Le diagramme de Venn peut être utilisé pour montrer l'intersection de 2 à 3 modèles en même temps.venne图可以同时做 2-3个模型的intersection
* Il est possible de réaliser l'intersection de plusieurs modèles en même temps.可以同时做多个模型的intersection

<https://github.com/These-SCAI2023/EXPE42_EVALUATION_ELTeC-fra_09072024/blob/main/prog/upstpt_modif_ljudmila.py>



* [https://github.com/mandalsubhajit/venndata](https://github.com/mandalsubhajit/venndata" \t "/Users/zhengruixing/Documents\\x/_blank)

Si le diagramme de Venn ne fonctionne pas, utilisez une autre bibliographie.如果venn图不行的话，就用其他的bibliographie

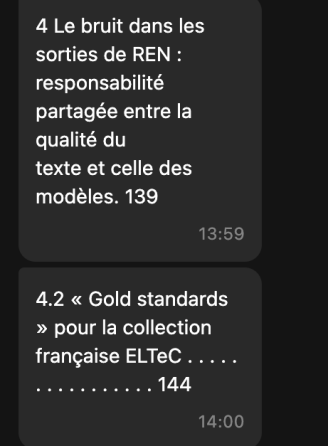
* [https://github.com/These-SCAI2023/EXPE42\_EVALUATION\_ELTeC-fra\_09072024/blob/main/prog/upstpt.py](https://github.com/These-SCAI2023/EXPE42_EVALUATION_ELTeC-fra_09072024/blob/main/prog/upstpt.py" \t "/Users/zhengruixing/Documents\\x/_blank)
* [https://github.com/These-SCAI2023/EXPE42\_EVALUATION\_ELTeC-fra\_09072024/blob/main/Upsetplot\_intersection/GLOBAL/CORPUS-COMPAR\_TAL-ENS2\_spacy-stanza/CORPUS-COMPAR\_TAL-ENS2\_spacy-stanza\_ACCMAJ\_\_PNG/CORPUS-COMPAR\_TAL-ENS2\_spacy-stanza\_ACCMAJ\_ACCMAJ-6000.bio.json\_REF\_upsetplot-size-100.png](https://github.com/These-SCAI2023/EXPE42_EVALUATION_ELTeC-fra_09072024/blob/main/Upsetplot_intersection/GLOBAL/CORPUS-COMPAR_TAL-ENS2_spacy-stanza/CORPUS-COMPAR_TAL-ENS2_spacy-stanza_ACCMAJ__PNG/CORPUS-COMPAR_TAL-ENS2_spacy-stanza_ACCMAJ_ACCMAJ-6000.bio.json_REF_upsetplot-size-100.png" \t "/Users/zhengruixing/Documents\\x/_blank)

entrée：

* [https://github.com/These-SCAI2023/EXPE42\_EVALUATION\_ELTeC-fra\_09072024/blob/main/Upsetplot\_intersection/GLOBAL/small-ELTeC-fra2024\_REN/small-ELTeC-fra2024\_REN\_spacy-lg-3.7.5.json](https://github.com/These-SCAI2023/EXPE42_EVALUATION_ELTeC-fra_09072024/blob/main/Upsetplot_intersection/GLOBAL/small-ELTeC-fra2024_REN/small-ELTeC-fra2024_REN_spacy-lg-3.7.5.json" \t "/Users/zhengruixing/Documents\\x/_blank)
* Tout cela est sur GitHub, n'oubliez pas d'utiliser le code des professeurs.这些都在guthub上，记得利用老师的代码
* [https://upsetplot.readthedocs.io/en/stable/](https://upsetplot.readthedocs.io/en/stable/" \t "/Users/zhengruixing/Documents\\x/_blank)
* [https://www.overleaf.com/learn/latex/Bibtex\_bibliography\_styles](https://www.overleaf.com/learn/latex/Bibtex_bibliography_styles" \t "/Users/zhengruixing/Documents\\x/_blank)

Il y a encore beaucoup de références à ajouter, y compris les références utilisées dans les articles des professeurs.还有很多的reference要加入，包括老师的文章里使用的reference

* [https://theses.hal.science/tel-05042915](https://theses.hal.science/tel-05042915" \t "/Users/zhengruixing/Documents\\x/_blank)



* [https://huggingface.co/asi/gpt-fr-cased-base/blob/main/README.md?code=true#L119](https://huggingface.co/asi/gpt-fr-cased-base/blob/main/README.md?code=true" \l "L119" \t "/Users/zhengruixing/Documents\\x/_blank)
* [https://github.com/bcallonnec/ner\_bert](https://github.com/bcallonnec/ner_bert" \t "/Users/zhengruixing/Documents\\x/_blank)

reLe fichier readme.md doit être terminé avant le prochain rendez-vous.

Avant le prochain rendez-vous, toutes les problématiques rencontrées aujourd'hui doivent être résolues.adme.md

在下次rdz前要写完

下次rdz前，要把今天遇到的问题都解决好

Lors du prochain rendez-vous, il faudra répondre si la soutenance sera faite en juin. N'oubliez pas de tenir compte des informations mentionnées par Lejeune concernant les dates : la date de soumission du mémoire et la date de la soutenance.下次rdz的时候，回复是否在6月进行soutenance。注意lejeune提到的时间，论文提交的时间和进行soutenance的时间

* "spacy\_lg" et "spacy\_sm" semblent avoir des performances similaires en reconnaissance de l'anglais.
* En observant attentivement le fichier JSON généré , nous avons constaté des erreurs de reconnaissance.

nous ne prenons pas ces erreurs en compte, mais nous nous concentrons uniquement sur les différences de performance des cinq modèles lorsqu'ils traitent les mêmes données.

Plus tard, une fois que nous aurons effectué l'annotation manuelle, nous envisagerons le calcul de métriques telles que le F1-Score.

**Pour le moment:**

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

**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(28/04 à 13h en remplacemen | en ligne | 1h30mn |
| 9/05(12/05 à 15h30 en remplacemen) | en visio | 1h |
| 26/05 à 13h | en visio |  |

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).

Alors, comme j'étais sur le document, je voulais juste demander à toutes les deux un petit message dans le chat. Vous avez cette erreur là dans votre latex, c'est parce que le underscore en LaTeX, vous ne pouvez pas l'utiliser comme ça. Il faut le déspécialiser en mettant un petit sable devant. Donc pensez à regarder de temps en temps les messages des marques dans votre LaTeX. Il y en a plusieurs sur ce truc-là.

那么，因为我正在看文件，我想在聊天里给你们俩留个小信息。  
如果你们在 LaTeX 中遇到了这个错误，是因为在 LaTeX 里，下划线（\_）不能直接使用。  
需要在前面加一个反斜杠（\）来转义它。  
所以，请时不时查看一下你们 LaTeX 文件中标记出的错误提示。  
在这个问题上，会有好几个提示

Parce que sinon, votre document ne va pas compiler. Donc vous avez fait ça, je vous montre juste rapidement. Quand vous avez quelque chose comme ça, il va falloir... ...d'écrire comme ça. Il y a quelques caractères en fait là que vous ne pouvez pas utiliser comme ça. L'autre solution étant d'écrire comme ça, vous pouvez aussi faire comme ça.

否则，你们的文档是无法编译成功的。

我简单给你们演示一下，当你们有类似这样的内容时，必须这样写（加反斜杠转义）。

其实在 LaTeX 中，有一些特殊字符是不能直接使用的。

另一种解决方法是使用双美元符号（$$）进入数学模式，也可以这样写。

En le mettant, en fait, l'anti-sabre c'est pour déspécialiser le caractère, comme dans les extensions écrires. Et si vous mettez avec les deux larmes, la deuxième manière, ça permet de mettre en place mathématiques, là le underscore et acheter messages mais ça va faire un truc bizarre. Donc vraiment la méthode à faire c'est celle-ci. Prenez l'habitude de bien lire les messages d'Eva. En tout cas il faut 10$ parce qu'en fait il pense que c'est l'underscore qu'il utilise dans l'environnement math, mais il faut vraiment...

实际上，加反斜杠是为了转义特殊字符，就像在某些扩展环境中一样。

如果你用双美元符号的方法，那么下划线就可以在数学模式中使用，但会引发一些奇怪的显示效果。

因此，正确的方法还是在普通文本中加反斜杠转义。

养成认真阅读 LaTeX 编译器（比如Eva）的提示信息的习惯。

否则 LaTeX 会误以为你是在使用数学环境里的下划线，导致错误。

0:07:21

(Speaker 1)

Et donc c'est pour ça qu'il y a des erreurs. Ah ok, je pensais que ça allait être quelque chose de plus grave. Il y a pas mal d'erreurs. Et il y a un... Une aquale fermée sans aquale au ventre. Voilà, ça c'est la deuxième erreur qu'il y a. Et donc ça probablement, en fait, il cherche à trouver où est-ce qu'il a le plus d'un collage au ventre

还有一个……一个未闭合的大括号（aquale fermée sans aquale ouverte）。

这是第二个错误。

所以，大概是系统在试图寻找有没有未配对的大括号。

Ouais, ouais, ouais. Sur Overleaf, il arrive même pas à compiler le document. Parce que si t'as des erreurs dans le syntaxe, textulier ou ailleurs, il arrive à compiler, mais tu vois dans le document que c'est mal revu. Et bon, il y a un merci pour le petit coup. Ouais, mais ça, ça arrive, ça se compile toujours moins.

在 Overleaf 上，遇到这种错误，甚至文档都编译不了。

如果只是一些小的语法错误，像拼写错误（textulier，可能是笔误，应为 "textuel"），一般还能编译，只是生成的文档显示会有问题。

不过，还是感谢你提醒我们啦。

是的，这种事情经常会发生，但通常文档还是能勉强编译的。

Alors, donc, j'ai répondu à, je pense, en partie, en tout cas à la deuxième question sur la compilation sur OverLife. N'hésitez pas quand vous avez comme ça des parties de votre LaTeX qui ne compilent pas, vous pouvez commenter la partie en mettant ça devant. Par exemple, quand la latex vous met qu'il a trouvé un problème de 1.39, quand vous avez un message qui vous dit que vous pouvez le copier et coller. En gros la tête vous dit j'ai le 0.4 de MERSCODE METALOLOGY blablabl ligne 39 et vous dit à quel moment il a repéré l'erreur. Parfois il suffit de commenter la ligne. Vous avez des points de site en ligne comme ça et puis vous commentez, vous mettez un petit pourcent dedans pour voir si le reste compile.

那么，我已经回答了，至少部分回答了第二个关于在 Overleaf 上编译的问题。

如果你们在 LaTeX 中遇到某些部分无法编译的情况，可以临时注释掉有问题的部分，只需要在前面加一个百分号 %。

比如，当 LaTeX 报错说在第39行（或其它位置）出现了问题，它会给出一个错误提示，比如说什么 MERSCODE METALOLOGY blablabla。它会告诉你出错的具体行数，有时候只需要把那一行注释掉，就可以继续编译其他部分了。

像这样，当你遇到一行代码出问题时，在前面加个百分号 % 注释掉，看剩下的部分是否能顺利编译。

Ok, alors là, je vous partage mon écran. Vous avez pu faire ça dans le cours du premier semestre sur la tech. Donc ici, à côté de l'icône compilée, vous avez l'icône des messages d'erreur. Donc là par exemple, je vais essayer de compiler. Il va chercher, mais il a trop d'erreurs. En fait, il va avoir un dossier qui n' arrive pas, donc il ne va pas pouvoir s'occuper.

好，那我来共享我的屏幕。

这其实在第一学期的 LaTeX 课程里讲过的。

在「编译」按钮旁边有一个「错误信息」图标。比如我现在试着编译一下，它会开始查找错误，但如果错误太多，最终它无法正常生成输出文件，因为存在太多问题。

Donc voilà pour les erreurs de la TEC. Euh... Alors j'ai pas compris votre autre question. Pourquoi parfois l'archive IP et AKBB s'affichent sur GitHub et parfois non ? Je n'ai pas compris. Quelle est la réaction ?

以上就是关于 LaTeX 错误的问题。

呃……然后我没听懂你们的另一个问题。为什么有时候 GitHub 上会显示 IP 和 AKBB 的归档，有时候又不显示？我没太明白。你们想问的是什么？

Je ne peux pas très bien vous piquer la question. Alors ce que j'ai piqué la question, c'est est-ce que vous pouvez entraîner Flaubert pour la réduction de son entité de l'éliminant ? Est-ce que ça te l'ai douté, je crois que ce que j'ai vu, c'est que vous dites que c'est parce que Floubert et Camandale récupèrent moins d'entités nommées que Swayze, LG, ST, PSM, etc. Alors moi ça me paraît pas étonnant que Floubert et Camandale récupèrent moins d'ent statistique sur les 3 modèles de Spleen mais ce qu'on avait vu l'autre fois dans les tableaux et du coup j'ai pas pu voir si les tableaux vous les aviez rectifiés pour mettre les bonnes données parce que ce que vous avez vu, à la fin, il n'y avait pas forcément tout le temps les bonnes données dans les tableaux, il y avait eu des coquilles collées, c'est qu'en fait il y avait vraiment, vraiment, vraiment beaucoup. Donc là je ne me rends pas compte, comme je ne vois pas le tableau, je ne me rends pas compte à quel point il y en a moins. C'est que soit on passait de 100 à beaucoup des fièvres, les proportions étaient comme ça, qui étaient vraiment simples. Après si on passe des proportions de 500 à 200, ça va vraiment être long, enfin en gros c'est des proportions... Voilà, je sais pas si ça vous aide. Et après non, on n'a pas prévu, je pense qu'il y a... Il y a une vidéo où on avait déjà parlé, mais on n'a pas prévu de vous enseigner des modèles, parce que ça impliquerait qu'il fallait qu'on fasse de la notation, enfin voilà. Donc je ne pense pas que ce soit...

没能特别清楚地理解你们的问题。那么我理解到的问题是：你们在问是否可以对 Flaubert（模型）进行训练，以改进它在实体识别（NER）上的表现？关于这一点，你们似乎在怀疑：你们说 Flaubert 和 Camembert 这两个模型，比 Swayze、LG、ST、PSM 等其他模型识别到的命名实体要少。

对我来说，Flaubert 和 Camembert 识别到更少命名实体这件事并不奇怪，因为我们上次在统计表格中也已经看到了。而且我没有确认你们后来有没有把表格里的数据修正，因为最后我看到的表格，并不是每次都包含正确的数据，还有复制粘贴错误（coquilles collées）。实际上，之前真的有很多很多错误。不过现在我没法确认，因为我看不到你们的表格，所以我无法判断到底少了多少。可能是比例的问题，比如说从 100 个减少了很多个，那么比例变化是明显的；如果是从 500 减到 200，那就是非常明显的减少了……总之就是比例变化的问题。我不知道这些解释是否对你们有帮助。另外，关于重新训练模型这一点，我们并没有计划教你们如何训练模型，因为这会涉及到我们需要为数据做标注（annotation）等等工作，这就比较复杂了。因此，我认为目前不会涉及到这方面的教学内容。

Non, non, je ne pense pas complètement étonnant effectivement que le clover et le calendaire soient un peu moins performants, donc potentiellement on a 10 fois moins, mais c'est carrément pas 10 fois moins, puisque les solutions passent décidément sur 3, on va faire le choix ici,et je compte ça pour l'objectif central des OSA. Donc on a des systèmes spécifiques, des systèmes optimistiques, C'est nécessairement, c'est ce qu'on appelle le fait tuning, pour pouvoir plus compter sur un rapport plus ou une tâche par seconde. Alors, ok, je veux voir votre opinion, est-ce que vous avez une impression de l'image réalisée ?

不，不，我觉得并不完全令人惊讶，Flaubert 和 Camembert 的表现稍微差一点，所以可能我们少了 10 倍，但其实并不是少了 10 倍，因为解决方案显然是通过了 3 次，所以我们在这里做了选择，我将其视为 OSA 的核心目标。所以我们有一些特定的系统，优化系统，这是必要的，这就是所谓的微调（fine-tuning），以便能在每秒处理更多的任务或更精确的报告。

Est-ce que tu peux m'appeler ? J'arrive sur votre kit, je vois bien des idées que tu as. J'ai pas l'impression que ce soit un problème spécifique à ce genre de clichés. C'est peut-être sur un autre boucle particulier. Et là, quand je vais dans le nouveau dossier script, je vois bien des PYKB. Donc voilà, c'est la réponse que vous a fait voir.

我看到你们的工具包，我看到你们有的一些想法。我不觉得这是这种陈词滥调的特定问题。也许是在某个特定的循环中有问题。而当我进入新的脚本文件夹时，我确实看到了 PYKB。所以，这是你给我的回答。

Alors après, concernant le tableau, donc là je regarde en effet, est-ce que vous avez vérifié que... que... parce que là, en effet, c'est très étonnant, les résultats de Camembert. Est-ce que vous avez vérifié que tout le texte, que l'algorithme passe bien sur tout le texte, première chose, ensuite que les entités sont bien, comment dire, stockées dans une liste ou dans votre structure de données, c'est la structure de données que vous utilisez, et pas écraser, c'est-à-dire qu'à chaque tour de couple, il n'y a pas un écrasement des données. Là, oui, en effet, c'est très étouffant, mais comme on en a déjà parlé, ça, ça peut être réglé en faisant...

然后关于表格，嗯，我现在看到了，\*\*你们有没有检查过...\*\*因为这里，Camembert 的结果确实很奇怪。\*\*你们有检查过算法是否正确地处理了所有的文本吗？\*\*首先是这个问题。其次，实体是否正确地存储在一个列表或数据结构中，是你们使用的那个数据结构，而不是被覆盖，也就是说，在每次循环中数据不会被覆盖。如果是这样的话，确实会很奇怪，但正如我们之前讨论过的，这个问题可以通过... 来解决。

Donc je vous avais envoyé une plaque que nous avons faite avec madame Zemaré. Est-ce que vous pourriez me dire où vous avez fait ce que vous avez fait ? Donc ce qu'on avait fait, c'est qu'on était placé sur le texte morceau par morceau. On avait fait des morceaux de chacune un caractère... En fait là je ne peux pas, je suis désolée, mon ordinateur, j'arrive pas à découvrir mes scripts ou mes trucs.所以我之前发给你们了一张我们和泽马雷女士一起做的板子。你们能告诉我你们做了什么吗？\*\*我们所做的是将文本分成一部分一部分。我们把每一部分都做成一个字符...事实上，我现在无法操作，非常抱歉，我的电脑，我无法打开我的脚本或其他东西。

sur latex depuis mon... je crois que c'était XP44 sur Giteb depuis XP44, dans le Giteb Corpus Prog, et on avait une musique à vendre. Vous avez vu le lien? Est-ce que vous avez reçu le lien que j'ai envoyé? Il y avait une fonction de changeemploye. Si on n'en employe que pour le BIO, on peut aussi...

关于latex，从我的... 我记得是从GitHub上的XP44，在GitHub的Corpus Prog中，我们有一个音乐要卖。你们看到了链接吗？你们收到了我发的链接吗？里面有一个changeemploye函数。如果我们只为BIO使用，我们也可以...

Ah non, on l' employé pour faire... Oui, donc est-ce que Lucie ou Marina, vous avez essayé d'utiliser là les fonctions ?

你们有试过使用这些功能吗？

Ah, donc en fait on est au même point qu'il y a deux semaines dit c'est d'essayer d'envoyoyer vos canons verts, en tout cas à cette position là, et d'enjoyer vos canons blancs, par exemple, en les adaptant pour l'enjoyer au vert. Moi c'est le costume que je pouvais vous proposer, parce que je sais que ça fonctionne, mais après il faut que vous fassiez l'adaptation.

啊，实际上我们和两周前的情况一样，说的是尝试发送你们的绿色枪，至少在这个位置，然后再发送你们的白色枪，例如，通过调整它们以适应绿色枪。这个是我能提供的建议，因为我知道它有效。

Oui, mais ça n'a pas d'importance Marina, le notebook n'est pas fait pour être ouvert d'ordi, c'est juste pour gérer les versions. Quand on a besoin de le regarder, on le regarde en local, on récupère les erreurs. C'est un peu le mot de quitte, etc.

notebook1或者github并不是为了在电脑上打开的，它只是用来管理版本的。当我们需要查看它时，我们在本地查看，收集错误。这有点像退出命令等等。

Github n'est pas fait pour afficher des notebooks. D'accord ? Donc, ça a marché sur les boutiques, mais dès que vous avez un truc avec des résultats, encore plus avec des images, etc., Donc, vous avez des notebooks simples ou des notebooks avec que du code. Là, ça va fonctionner.

GitHub 并不是用来显示笔记本的。明白吗？所以，它在商店上工作，但一旦你有了带结果的内容，特别是有图片等时，情况就不一样了。所以，如果你有简单的笔记本或仅包含代码的笔记本，它会正常工作。

Quand vous avez des notebooks plus complexes ou des notebooks avec des images, des graphiques, des choses comme ça, là, ça va moins bien fonctionner, moins mal ne pas fonctionner. C'est juste que c'est un problème totalement secondaire. Pas du tout important. Donc, quand on a besoin de regarder vos notebooks, on les regarde dans le cadre. Je peux répondre à la question de Caroline sur Flaubert, qui m'en pose la question de est-ce que c'est le même nombre de top-ends qui sont satés avec Flaubert et avec les autres approches ?Tu as entendu ce qu'il a dit ? Je n'ai pas entendu. Je crois que pour le code de Flaubert et Camembert, Marina a utilisé Camembert pour faire la reconnaissance de notre phénomène, mais pour le Troc, c'est toujours Flaubert. Et le nombre c'est différent entre le modèle de Flaubert et le modèle de Camembert.Attention, je ne parle pas du nombre d'identités, je parle du nombre de tokens. Est-ce que vraiment, vous avez vérifié, vous avez dit que ça fait le bon nombre de tokens. Est-ce qu'il n'y a pas quelque part une limite dans ce que vous donnez à Flober et au lieu de prendre toute la dette, il en prend qu'une partie ?

当你们有更复杂的笔记本或者包含图像、图表之类的东西时，情况就不那么顺利了，虽然它不会完全不起作用。 这只是一个次要的问题，根本不重要。所以，当我们需要查看你们的笔记本时，我们就在框架中查看它们。 我可以回答 Caroline 关于 Flaubert 的问题，她问的是：Flaubert 和其他方法使用的 token 数量是否相同？

你听到他说的了吗？我没听清楚。我觉得在 Flaubert 和 Camembert 的代码中，Marina 使用了 Camembert 来做我们的现象识别，但在 Troc 中，仍然是 Flaubert。Flaubert 模型和 Camembert 模型之间的数字是不同的。注意，我不是在说实体的数量，而是在说 token 的数量。你们真的检查过吗？你们说它们的 token 数量是正确的吗？是不是在你给 Flaubert 的输入中，某个地方有限制，使得它没有处理全部的数据，而只处理了一部分？

C'est ça la question, c'était d'utiliser Flaubert pour l'art case, pour faire la coquinisation. Et après, la seule façon que je trouvais sur Rugged Face, c'était d'utiliser Camembert pour faire les deux combinés. Et Lucie m'a dit que c'était ça qu'on n'était pas d'accord parce que selon Lucie, on ne peut pas utiliser Camembert et Flaubert ensemble. Et donc la première chose que j'ai fait c'était ça. Flaubert, localisation, Camembert, vous pouvez accéder à ces codes, il est disponible sur uTerm. parce que la troisième aurait d'être faire l'entraînement du modèle, ce que vous avez dit qu'on va pas faire et je me rappelais de ça.

这是问题所在，问题是用 Flaubert 来做艺术案例，进行标注。而之后，我在 Rugged Face 上找到的唯一方法是使用 Camembert 来做两个结合。然后，Lucie 告诉我我们不一致，因为根据 Lucie 的说法，不能同时使用 Camembert 和 Flaubert。所以我做的第一件事就是这样。Flaubert 用于定位，Camembert 用于标注。你们可以访问这些代码，它们在 uTerm 上可用，因为第三步应该是训练模型，但你们说我们不做这个，我记得这一点。

J'ai essayé de forcer Flaubert avec cette organisation pour pouvoir faire la reconnaissance d'Antithénomé, mais ça a donné vraiment n'importe quoi parce qu'il n'a rien pour faire. Il a sorti des résultats de façon complètement aléatoire et sans log, sans cursing, parce qu'il n'a rien pour se baser justement. Mais je fais quand même ces deuxièmes.

Donc pour le premier, quand je passe la vérification, tout va bien, il ne met pas d'erreur. Pour le deuxième, c'est celui que j'ai pas utilisé par mon père. J'ai forcé le Flaubert d'organisation. Il donne des erreurs partout.

我尝试强制使用 Flaubert 来做这个组织，以便进行反义术语的识别，但结果真的一团糟，因为它没有任何可供依据的东西。它给出的结果完全是随机的，没有日志，没有错误，因为它没有任何东西可供参考。但我还是做了第二次。对于第一个，当我进行验证时，一切正常，没有错误。对于第二个，我没有按照我的path使用它。我强行使用了 Flaubert 组织，它到处都是错误。

Utilisez un modèle de Flaubert qui s'aide pour la reconnaissance d'entités données. Je ne vous prends pas le feu pour ce que vous voulez dire par forcer la tokenisation, et après, est-ce que vous travaillez bien un modèle qui ait d'arrêt de coup ? Non. Bien sûr, il y a un Frobert, mais je ne suis pas sûr de le comprendre.

使用 Flaubert 模型来进行实体识别。我不太明白你所说的“强制标记化”的意思，然后，你是否正在使用一个停止词的模型？不。当然，确实有 Flaubert，但我不太确定如何理解它。

Là c'est le seul endroit dans le code de Flobber où vous spécifiez que vous voulez faire de la reconnaissance d'entité donnée. Et la question de Caroline c'est, est-ce que le modèle que vous avez utilisé, donc là, au paramètre modèle de donnez le modèle qui s'appelle le monoménaire, est-ce que c'est un variable schlöber qui a spécifiquement été entraîné pour la représentation d'antithénomène ? Puisque la hypothèse c'est que peut-être vous utilisez un modèle générique qui n'a pas été spécialement entraînéue la question de la taille du corpus, effectivement c'est bien sûr les textes de bout en bout qui sont traités là aussi, si ce n'est pas le problème des données en entrée, une autre question peut être le problème du modèle.

这里是 Flaubert 代码中的唯一地方，你指定了想要进行实体识别的部分。Caroline 问的问题是，你使用的模型，在这里模型参数中给出的模型名称，是不是一个专门为表示反义术语而训练的 Flaubert 变体？因为假设是，你可能使用了一个通用模型，而这个模型并没有专门为反义术语进行训练。至于语料库的大小，的确，应该是处理完整文本，如果输入数据没问题，那么另一个问题可能是模型本身的问题。

Donc ici, le modèle que vous chargez, j'ai pris juste ce que j'ai trouvé dans votre code, donc dans votre code, et le modèle qu'on a en amont, C'est donc le Flaubert Baird Case

.所以在这里，你加载的模型，我只是根据我在你们的代码中找到的内容来看，在你们的代码中，前端加载的模型就是 Flaubert Baird Case。

Non, ce modèle là, il n'est pas indiqué et je n'ai pas trouvé aucun modèle de Flaubert qui a été entraîné pour l'année. J'ai trouvé un article, que j'ai mentionné sur le Mémoire, qu'ils ont entraîné leur Flaubert, mais avec des corpus médicaux pour le français clinique. C'est la seule chose que j'ai trouvé liée à Flaubert et les entités nommées. Sinon, il peut accomplir cinq tâches du TAL, mais pas les entités nommées. Je n'ai pas trouvé.

不是的，这个模型没有被提及，我也没找到任何使用 Flaubert 模型训练用于命名实体识别（NER）的案例。我找到一篇文章，我也在论文中提到了，他们用医学语料库训练了 Flaubert，用于法语临床语言处理。这是我找到的唯一与 Flaubert 和命名实体识别相关的内容。除此之外，虽然它能完成自然语言处理中的五项任务，但并不包括命名实体识别。我确实没找到相关内容。

D'accord. Donc, ça, c'était la question qu'on posait il y a aussi quelques semaines. En fait, la réponse existe que non. Apparemment, vous n'aveziste pas que non, il n'existe pas, donc il faut laisser tomber Flaubert.

好的，这其实就是我们几周前问的那个问题。答案显然就是没有。你现在确认了确实没有，所以我们就得放弃 Flaubert。

Ah non, je ne pense pas que ça soit laisser tomber, c'est le mettre de côté. On a quelque chose qui n'est pas entraîné pour la reconsentie, mais on voit qu'il y a beaucoup moins de choses. Là, ce qui manque dans votre comparaison, c'est de savoir ce qu'-ce qui est en commun, qu'est-ce qui n'y a pas en commun, montrer les exemples de la santé etc. Est-ce que Flaubert quand il sort a différemment d'entités nommées que Spacey par exemple, est-ce que les entités nommées sont bien ? Parce que si on a 10 fois mieux mais qu'elles sont bien parce que ça, ça prend du temps, c'est une tâche spécifique. L'idée, c'est de comparer. Donc là, on compare dans les tables. Si vous dites que vous n'avez pas trouvé disponible un modèle dérivé de Flaubert spécialisé pour l'année, très bien.

啊不，我不觉得这是“放弃”，应该说是“暂时搁置”。我们确实有个模型没有针对 NER 任务进行过训练，但它还是有一定的表现。你们对比中缺少的是：哪些是相同的，哪些是不同的，要展示健康语料的例子等等。比如 Flaubert 输出的命名实体和 Spacy 输出的是否不一样？这些实体有没有准确识别？如果一个模型输出的更多，但质量不好，那也是有问题的。这是个具体的任务，需要花时间。我们的目标是进行比较。你们现在在表格里是有比较的。如果你们确认没有找到专为 NER 训练的 Flaubert 模型，那也没关系。

Et ça, voilà, si vous l'avez déroulé, de sorties avec Flaubert qu'avec Spacey. Qu'est-ce que ça veut dire ? Est-ce que ça veut dire que Spacey en sort trop ? Sans doute en partie. Et pourquoi ? Parce que dans le modèle Spacey Small on a beaucoup de choses, on en a moins que là. Est-ce que vraiment il y a une différence ? Moi là j'ai exécuté Voscript par exemple, ils ne trouvent pas du tout le même nombre de tokens dans les mêmes textes. Là quand j'ai exécuté FlowBear, FlowBear trouve sur un appel linéaire je sais plus, 18 000 tokens, là où Spacey en a 26 000. Donc, ça peut être intéressant de montrer ces exemples-là, faire des tableaux comparatifs, là encore, en termes de chiffres. Puisque peut-être que Flaubert sous-tokenise, ou peut-être que Spacey super-tokenise, enfin, il y a plein de questions. Et là, c'est pas très bien. La même question va se poser à Camembert.

是的，比如你已经运行了 Flaubert 和 Spacy 的输出，那这代表什么？这是否意味着 Spacy 输出得太多了？很可能是。为什么？因为在 Spacy Small 这个模型中，它包含了很多内容，可能比别的模型还多。那么这之间是否真的存在差异呢？我自己运行 Voscript 的时候，比如，在同样的文本中，它们找到的 token 数量完全不一样。我用 FlowBERT 执行一个线性调用的时候，它找到了大约 18,000 个 token，而 Spacy 却找出了 26,000 个。所以展示这些例子可能会很有意思，可以用数据来做对比表格。可能 Flaubert 的 token 分得太少，或者是 Spacy 分得太细，总之，这里面有很多问题。而现在还不太清楚。Camembert 模型也会出现类似的问题。

Donc, si on va garder alors le script et les sorties qu'on a eu, parce qu'en vrai c'est ça, les sorties qu'on a à la partie de NERS, elles sont pas Flaubert, elles sont Camembert pour les deux. La seule différence c'est que la tokenisation, une va être Flaubert et l'autre Camembert. Donc si j'ai bien compris, il est intéressant quand même qu'on fasse cette comparaison par exemple, entre la tokenisation Flober et Camembert et après la reconnaissance en titres nommé juste, enfin qui va être juste Camembert. Parce que pour Camembert c'est ça, on a trouvé le modèle entraîné sur Hugging Face, mais pour Flober non. Quand je parle de la comparaison de tokens, c'est juste comparer le nombre. Vous avez des corpus en fait qui, en nombre de tokens, en nombre de caractères, ils sont de la même taille quels que soient les modèles, les fichiers sont les mêmes, la tokenisation change, c'est intéressant, mais c'est pas dans ton tableau. C'est-à-dire que la taille du corpus en token n'est pas la même selon les modèles que vous avez utilisé. En tout cas, c'est ce que j'observe quand je vois votre code. sur une même frame, trouver une phrase qui est tokenisée différemment par les modèles, même phrase ou même un poème. Et ça, c'est tout bon.

那我们就保留现在的脚本和输出吧，因为目前我们在命名实体识别（NER）部分的输出其实都不是用 Flaubert 做的，而是 Camembert。唯一的区别是 token 分词部分，一个用了 Flaubert，一个用了 Camembert。所以如果我理解没错的话，我们应该还是要比较一下 Flaubert 和 Camembert 的分词方式，然后再用 Camembert 来做命名实体识别。因为 Camembert 是我们在 Hugging Face 上找到已经训练好的模型，而 Flaubert 没有。当我说对比 token，是单纯在数量上做比较。你们使用的语料文件，在字符数和内容上是一样的，但用不同的模型分出来的 token 数却不同，这一点很有意思，但在你的表格中并没有体现出来。也就是说，根据你所用的模型，corpus 在 token 的数量上是不一样的。我从你的代码中可以看到这一点。你可以在同一个片段上，找一句被不同模型分词结果不同的话，或者一首诗，展示它们的差异。这就非常有价值了。

Il y a des différences ici. Encore une fois, si on passe de 26 000 à 18 000, c'est un tiers en moins, c'est quand même pas rien du tout. Vous l'avez vu, quand on a fait le cours dictionnaire et linguisme, il y a différentes manières de tokeniser un texte. S'il est suffisamment grand, on va avoir des accords entre les tokeniseurs, y compris on aurait des accords entre humains, mais c'est des accords de l'ordre de 5% ou 10%. Là on parle de 35%. D'accord ? Donc c'est beaucoup plus bassé.

这里确实有差异。再说一次，如果从 26,000 个 token 变成 18,000 个，减少了三分之一，这可不是小事情。你们在“词典与语言学”课程中也看到了，不同的分词方式会带来差异。如果文本足够大，不同的分词器之间会有一定程度的一致性，就算是人工分词，也大概会有 5% 到 10% 的出入。但我们这里讲的是 35% 的差异，对吧？这就大得多了。

Ça peut être intéressant de quantifier ça dans un tableau, de comparer quelle taille ça fait selon Spacey Small, Spacey Large, etc. La taille de l'entrée. D'accord ? montrer des exemples sur ça, un même texte, une strope d'un poème par exemple, donc il disait différemment pour montrer un petit peu, pour éliminer la différence.

这方面做一个数据对比表会很有意思，比如比较 Spacey Small、Spacey Large 等模型下输入文本的长度。可以拿一个例子来演示，比如同样的一段文本、一段诗歌，看看它们的分词方式是如何不同的，以此来消除或解释这些差异。

J'avais un doute concernant ça aussi, puisque j'avais mis un tableau avec les tailles d'entrées, les nombres de pages et tout. Et je ne savais pas si je considérais le nombre de pages qui est marqué sur GAICA, parce que j'ai eu l'impression qu'ils marquent le nombre de pages vraiment des livres physiques, avant qu'ils ont faillu en CR. Le nombre de pages, on s'en fiche.

我对这个也有点疑惑，因为我之前做了一个表格，列出了输入的长度、页数什么的。但我不知道是否该参考 Gallica 上标注的页数，因为我觉得 Gallica 上的页数是根据实体书的页码来的，是书还没被扫描之前的。

Ce qu'on a déjà manipulé, eux, ils n'utilisent pas du grimpage, ils utilisent du grindling, du grintel, si vous voulez. Le nombre de pages, c'est secondaire. Si vous l'avez déjà dans un tableau, vous mettez le nombre de pages de Gallica, ça correspond effectivement, vous avez raison'est l'entrée sortie de Gaëlcar, ça se modifiait, vous n'avez pas besoin de la calculer, on prend ça, il n'y a pas de problème. Par contre c'est après où il y a un intérêt de comparer effectivement est-ce que tous ces outils donnent la même taille. OK. Bonjour. J'ai une lé... du bodysuit c'est pour faire l'intersection. La graphique vivienne c'est pour deux ou trois modèles mais il ne peut pas être utilisé pour cinq modèles en même temps. Est-ce qu'on a un autre moyen pour... Alors, vous avez des librairies pour faire un diagramme de veine. Oui. Pour plus de 3, ça, ça existe. Je ne l'ai pas en tête, je vous demande si Caroline ne l'a pas utilisé. Mais voilà, la librairie de base en python, elle fait pour 2 ou 3 et vous avez une autre livrairie qui permet de faire ça qui permet de faire ça car oui il faut s'impliquer aussi une autre visualisation qui est adaptée qui s'appelle le top 7 plot et bien fichu pour faire faire des comparaisons à plus de 3 ensembles qui va donner quelque chose de différent ça je peut être intéressant de l'évaluer. Et puis ça, je ne sais pas si tu as encore le code du truc de 20 sur plus de 3 sets qu'il y a en ligne. Je pensais que je voulais envoyer... J'en ai trouvé un, donc non testé, et voilà, j'ai simplement fait la requête, 24 moments of free set, et le premier lien c'est ça, je ne dis pas qu'il est bien ou pas, mais voilà, voici le, vous avez au moins trouvé le vocabulaire de ce genre de truc. Si le premier que vous cherchez ne marche pas, n'hésitez pas à regarder d'autres. Il y en a certains qui sont très faciles à utiliser d'autres moins. Le premier lien ne va pas forcément être le plus efficace. Vous testez le machin, si ça a l'air de marcher facilement vous le prenez. Sinon vous pouvez en regarder plusieurs avant d' un bon signe sur l'utilisabilité parce que la dernière mise à jour date d'il y a 6 mois ou un an et pas d'il y a 5 ans. Donc potentiellement c'est du code qui a été remis à jour, modifié, amélioré. Donc là vous allez avoir ça. Donc ça c'est ça les deux solutions. Mais effectivement vous avez raison, le diagramme de VED est fait nativement pour pouvoir fonctionner sur deux ou trois. Là il y a des problèmes de visibilité etc. Mais regardez quand même un petit peu ce qui se fait.

我们之前已经操作过了，他们那边不是用 grimpage（可能是拼写错误或误听），他们用的是 grindling 或 grintel，如果你愿意这么说的话。页数是次要的。如果你已经在表格里有 Gallica 的页数，那你就用 Gallica 给出的页数就行了，确实是对应的。你说得没错，那就是 Gaëlcar（可能是 OCR 或文献项目名）的输入输出信息。那部分是可以变的，你不需要自己去计算，用它们的就可以，没问题。

不过，有意思的是后面这一点：是不是所有这些工具在处理相同的文本时，会给出相同的大小（比如 token 数）？这个对比就很有意义了。

好的，早上好。

我有一个…… bodysuit 是用来做交集的（注：可能是“UpSet plot”或其他交集图工具的误听或误写）。Vivienne 图（graphique vivienne）是用来展示两个或三个模型的，但它不能同时用于五个模型。那我们有没有其他方法可以做到呢？

其实你是有的，有一些专门绘制 Venn 图（diagramme de Venn）的库。对，支持三组以上数据的库是有的。

我现在记不太清了，你可以问问 Caroline 有没有用过。但我知道的是，Python 里的基础库一般是支持 两组或三组集合的 Venn 图。而有另一个库是可以支持更多集合的，确实是可以做到的。

另外你也可以试试一种叫 UpSet plot 的可视化方式，它适合用来比较超过三个集合，做出来的效果也不一样，也许值得评估一下。

还有，你是不是还有那个画三个以上集合 Venn 图的代码？网上是有资源的。我当时只是简单搜了一下，比如输入 "24 moments of free set" 之类的关键字，点进第一个链接就是相关的内容。我不保证它好不好，但至少你可以借此找到这类图形所用的术语。

如果第一个链接不行，也没关系，可以多试几个。有些工具非常好用，有些就比较难搞。第一个结果不一定是最好的。你试一试，如果觉得好上手，那就用它；否则你可以多看几个选项。

你可以看一个判断工具“是否好用”的方法是：看看它最近一次更新是什么时候——如果是 6 个月或 1 年前更新的，那就比 5 年前更新的可靠多了，意味着它可能是最近还在维护、优化的代码。

总之，这就是两种解决方案：

一是尝试扩展 Venn 图支持多个集合的库，

二是改用更适合多集合比较的可视化方式，比如 UpSet plot。

Ce qui est intéressant c'est de voir surtout les ordres de grandeur, ce qui est commun, ce qui n'est pas commun. 有趣的是看数据量级、共同的部分、不共同的部分。是的，关于“输出”的部分（也可能指的是数据处理的结果）

Qu'est- petit programme qui fonctionne. Pour faire ce que disait Gaël, les offset plugs, ce qu'ils vont ressembler du coup. Ça va bien si je me prends un pied. Voilà. Après, le programme que je vous envoie là, c'est un réadapté avec vos données, mais en entrée, ce qu'il faut,entrée, un petit peu. C'est les fichiers qui vont êtrerée, c'est ce type là que les fichiers. Voilà, ce sont des gros exemples. Vous pouvez l'éliminer. Vous avez plus ou moins. C'est bon, normalement on peut l'éliminer.

这是一个可以用的小程序，正好可以实现 Gaël 提到的 offset plots（偏移图/偏移对齐）。比如说处理完之后它们长什么样。这样就好办了。如果我理解正确，这段程序我发给你了，是根据你的数据做过一些修改的。

不过你要做的是把自己的数据作为输入导入进去，文件应该是这个类型……就是这种大文件示例。你可以忽略不相关的部分。差不多就这些了，基本上可以运行了。

Donc moi j'ai travaillé par type d'OCR. J'avais des gros dictionnaires dans lesquels j'avais les entités nommées. Là, en l'occurrence, ça devait être pour Spacey, ce que je vous envoyais. Pour Spacey-LG, il y avait les 3 ou 4 systèmes de CR qui étaient en clé. Et après c'était les listes d'entités. Je ne retrouve pas le programme que j'avais fait pour les diagrammes de veines avec 3,4, 5... Donc tu pouvais explorer ça en regardant les quelques premiers résultats que vous pourriez avoir sur Google, sur ça, sur les diagrammes de Venn pour plus. Et par contre, tu as un code pour les obsèques là, c'est ça ?

我之前是根据 OCR 类型来处理的。我有一些大词典，其中包含了命名实体。这个可能是我之前发给你的，用于 Spacey 模型的。在 Spacey-LG（大型模型）中，我有3或4种 OCR 系统，分别作为不同的 key（键）。然后就是实体列表了。

我没能找到我当时为三、四、五个集合做维恩图的代码。但你可以通过 Google 查一下“Venn diagrams for more than 3 sets”（三个以上集合的维恩图）这样的关键词，看看前几个结果。

另外，你说你现在有 obsèques（偏移数据）部分的代码，是这样吗？

Tout est sur GitHub, vous pouvez faire couiller, normalement c'est public. C'est donc dans XP42, évaluation FRAC, les graphiques en sortie dans UPSET Plus Intersection. ... Si on regarde le programme, Vous n'aurez pas besoin de toucher votre corps de toute façon. Tout ce qui est en commentaire, pour rester en commentaire, tout ce qui est éliminé, c'est ce que j'utilisais pour d'autres données d'entrée. Donc, ce n'est pas la peine de les décommenter. Ensuite, ce qu'on avait faites. On avait trois vérités de terrain, mais vous, vous n'avez pas de vérités de terrain, c'est la référence qu'il faut mettre.

所有内容都在 GitHub 上，大家可以运行（couiller 是“faire tourner”的俚语说法，意思是“运行脚本”），应该是公开的。在仓库 XP42 中，路径是 évaluation FRAC，输出图表的部分在 UPSET Plus Intersection 文件夹里……

如果你查看这个程序，实际上你根本不需要修改主体代码。注释掉的内容就保持注释状态，那些是我用于其他输入数据时的部分，不需要取消注释。

然后我们之前做的是：我们有三份“ground truth”（即标准参考数据），但你这里没有“ground truth”，所以要用你自己的“reference”（参考数据）来替代。

Vous n'avez peut-être pas besoin de la ligne 41. Il faut que vous triez un peu. C'est vrai que je suis désolée, il faudrait que je commente ce problème.

你可能不需要第41行。你需要稍微清理一下代码。是的，确实很抱歉，我应该给这个程序加注释的。

Et après, les lignes 111 à 142, ça c'est pour le visuel du graphique. C'est comment vous voulez marquer, quelle couleur vous voulez il y a quand même pas mal d'aides pour ça. Après c'est pour le stockage. Il y a comment c'est représenté sur la figure. Dans les premières lignes, vous avez comment le fichier de sortie va s'appeler, s'il existe ou pas déjà pour créer ou pas le dossier dans lequel la sortie va être stockée. Après, il y a quelle clé il faut aller lire. Donc ça c'est M81. Jusqu'à M10. Je ne sais pas si vous n'êtes pas obligé de le garder peut-être.

然后是第111到142行，那部分是图表的可视化设置，比如你想怎么标注、用什么颜色等等，这些其实有很多可用的帮助文档。接着是关于图表的存储方式，比如输出图像怎么显示。在前几行代码中，你会看到输出文件的命名方式，以及这个文件是否已经存在，如果不存在，就会自动创建一个用于存储输出结果的文件夹。

然后还有关于读取哪一个键（key）的设置，比如是 M81 到 M10 这些。你也许不需要保留这些设置，可以删掉。

Il faut que vous regardiez un peu, vous n'êtes peut-être pas obligé de tout garder. Vous voyez comment vous structurez vos données en entrée, mais le plus simple est de faire un dictionnaire avec les clés qui vont dans le graphique. Donc quand vous regardez l'exemple que j'était ça ? Enfin, le PNG est là. En gros, les clés dans le dictionnaire d'entrée, c'est ce qui va apparaître après devant les points. Donc là, sur l'image que j'ai envoyée, c'était ACMEA-GRF.

你要看看自己的数据结构，也许不用全部保留原来的结构。最简单的方法是构建一个字典（dictionnaire），里面的键是你想显示在图表上的名称（key），这些键就是你图上的点名。在我发的例子中，PNG 图片上显示的是 ACMEA-GRF，这些就是那个字典的键。

每个键对应的值（value）是该模型识别到的命名实体（entités nommées）列表。这个就是图表构建的基本逻辑。

Donc ça, c'était les clés du dictionnaire, et que pour chacune des clés en valeur, il y avait la liste des entités nommées. Voilà, c'est ça le principe à retenir. Ensuite, comment ça se lit ? Alors là, sur l'exemple que j'ai envoyé, on voit que la barre bleue, c'est les entités nommées qui avaient été récupérées sur la référence. La barre grise entourée de rouge, c'est les entités nommées qui avaient été récupérées par la notation dans la vérité terrain. Et ensuite, la barre bleue entourée de rouge, c'était donc les entités nommées qui sont dans l'intersection entre les deux ensembles. Et donc là, en fait, ce qu'on voit, c'est qu'en gros, on a 5,7% des entités nommées qui sont dans la référence et qu'ensuite, on a 69,8% des entités qui ont été récupérées uniquement sur la référence et 24,5% qui ont été récupérées uniquement sur la vérité terrain. Et ensuite, sur le côté, les chiff entités qui sont présentes dans les sorties de l'accord majoritaire et 75,5% des entités qui sont présentes pour la référence. Pour lire l'intersection, c'est vraiment la dernière colonne, où on a les deux points ACMAS, donc la vérité terrain et les sorties de référence, qui sont liés entre eux. Et oui, en gros, dans Caroline, on constate aussi que 30.2 c'est quand on additionne les actes magiques, les rêves, et un blessant pour le... C'est l'entièreté en fait, c'est l'entièreté de ce qu'il y a Oui, oui, oui Et l'autre c'est l'entièreté de ce qu'il y a ...

所以说，那些就是字典的键（keys），而对于每一个键来说，它对应的值（value）是命名实体（entités nommées）的列表。这个是你需要记住的核心原理。

接下来，怎么读取这个图呢？在我发的那个例子里，我们可以看到：蓝色的柱子表示的是在“参考集”（la référence）中提取到的命名实体；灰色柱子，红色边框的是从“人工标注的真实数据”（la vérité terrain）中提取出来的实体；而蓝色柱子加红色边框的，就是在这两个集合中都出现过的命名实体，也就是交集部分。

所以从图上我们看到，大概有 5.7% 的命名实体同时出现在参考集和真实数据里；有 69.8% 的实体只出现在参考集里；还有 24.5% 的实体只出现在真实数据里。

然后在图的右边，还有一些数字：一个是表示出现在“多数投票输出结果”（sorties de l’accord majoritaire）中的实体数量，另一个是参考集中实体数量的百分比（75.5%）。

至于怎么读“交集”，那就是最后一列了，在那一列里我们可以看到两个点：一个是 ACMAS（真实数据），另一个是参考集的输出结果，这两个点是被连在一起的。

是的，总的来说，在 Caroline 的结果中我们也看到类似的情况。比如那个 30.2% 的值，就是把“magic acts（魔法行为）”、“dreams（梦）”和另一个“blessant（伤害性）”这样的东西加起来之后的总和……意思就是——这是所有内容的总量。对的，是所有的实体数量。

Après moi je me demande une question concernant le mémoire, est-ce que vous pouvez me dire si vous aviez déjà ce genre d'erreur lors de la compilation ou pas ? Parce que de l'autre côté, on n'arrive pas, en tout cas on voit le message d'erreur, ou on voit des erreurs syntaxiques que vous devez régler, mais je me pose la question, même si vous réglez toutes ces erreurs, est-ce que vous recevez une erreur qu'on croit la compilation a expiré, qu'on n'arrive même pas à compiler le projet ?Parfois on peut voir le document PDF et parfois il montre des problèmes de compilation.

然后，后面我有个问题，是关于\*\*论文项目（mémoire）\*\*的：

你们有没有遇到过这种编译错误？

因为我们这边遇到的问题是：要么直接编译失败，要么提示“编译超时”，根本无法运行。有时能看到一些语法错误信息，但我很好奇的是——即使你们把语法错误都修复了，是不是还是可能出现编译器直接报错、项目根本编译不了的情况？有时候我们可以看到 BDF 文档，有时候它却显示编译错误。

Et oui, c'est plus en local que ça fonctionne en faisant les corrections que j'ai dit tout à l'heure. Donc ça peut l'être.

是的，其实在本地运行的时候，按照我刚才说的那几个修改方法，有时确实是可以运行的。所以是有可能的。

Je ne sais pas si j'ai corrigé toutes les erreurs, mais certains, ouais je pense que j'ai pas corrigé tout. Mais d'après ce que j'avais vu, j'ai corrigé et j'arrive pas à compiler encore.

我不确定我是不是已经修正了所有的错误，但有些问题……我觉得我应该没有全部改完。按照我当时看到的错误，我改了一些，但我还是编译不出来

Mais peut-être que j'ai loupé des choses. N'hésitez pas, comme on disait tout à l'heure, à commenter. Commenter tout ce qui pose problème en attendant de trouver une solution. Ou vous pouvez même commenter un input, pour vérifier si le reste fonctionne. Ça permet de répondre à des questions de Mila. Vraiment identifier. Parce que si ça se trouve, vous avez une accolade non fermante à un endroit qui correspond à un truc ouvert il y a très longtemps pour un verseur. ou à la sorte. Des fois ça peut être bien de pouvoir isoler la tech, enfin, Overleap a décidé de vous dire quel est le fichier qui pose problème et quelle est la ligne. Généralement c'est assez précis, mais si vous avez laissé traîner beaucoup d'erreurs, elles peuvent se cumuler en train. Vous regardez ça, si vous avez une erreur que vous n'arrivez pas à corriger, vous nous envoyez la carte à l'écran, juste de l'erreur que vous avez pas fait, vous n'avez pas corrigé. J'avais une autre remarque sur le mémoire. Ce serait bien que vous travailliez un petit peu la bibliographie. Et un petit peu succincte, très succincte. Ceai compilé, il n'y avait que 6 références bibliographiques, pas beaucoup. Et que vous mettiez un style bibliographique un peu plus lisible.

也许你还有一些地方遗漏了。别犹豫，就像我们刚才说的那样，可以先把那些有问题的部分注释掉，等找到解决方案以后再打开。你甚至可以注释掉一个输入 input，看剩下的部分能否正常运行。

这也可以帮助回应 Mila 的问题，真正去定位问题所在。因为有时候，可能只是某处一个没闭合的大括号，而那个括号可能是很早以前开的。

有时候，把错误孤立出来真的很有用。LaTeX 或 Overleaf 通常会告诉你是哪一个文件出错，以及是第几行，通常提示还是蛮准确的。但如果你留下了太多错误，这些提示可能会互相叠加、干扰。

你仔细检查一下，如果实在有某个错误你怎么也搞不定，就把出错信息截图发给我们，只发那一条你搞不定的就好。

我还有另外一个建议是关于你论文的：你最好再补充一下你的参考文献部分。现在那部分太简略了，非常简略。你提交的那个版本只有6条参考文献，真的太少了。还有，建议你采用一个更清晰可读的参考文献格式。

Là le style bibliographique que vous avez mis, ça vient décrocher avec des numéros. C'est pas très difficile parce que du coup, on ne sait pas quand on a un numéro, Je vous ai mis un lien dans le chat sur les styles bibliographiques en latèque. Vous avez des styles beaucoup plus expressifs, où vous pouvez avoir le nom de l'auteur par exemple. Je pense que vous pouvez prendre le style Apalike par exemple, qui est un livre en catholique. Là, ça va vous mettre à chaque fois le nom du premier auteur et la date. Ça va rendre votre document public. Il faut que vous ajoutez de références qui mériteraient d'être ajoutées.

你现在使用的那个参考文献格式会以编号的方式显示，显得有点混乱。这其实不难解决，因为这样就无法明确知道每个编号指的是什么文献。我在聊天里给你发了一个关于LaTeX 参考文献格式的链接，你可以参考一下。

你会发现有很多更简洁直观的样式，你可以像这样展示每篇文献的作者名字。

我建议你可以使用 Apalike 样式，它是一个符合学术标准的样式。使用这个格式后，每次引用文献时，会显示第一个作者的名字和出版年份，这样会使你的文档更规范，更符合标准。 另外，你还需要增加更多的参考文献，因为目前看起来是有些不足的。

Il n'y a pas de chiffre exact, mais clairement, sur un mémoire de la main, 6 références, c'est clair. À la fois sur les outils que vous utilisez, les méthodes, il ne faut pas hésiter aussi à faire référence aux articles dont vous vous appelez, notamment des choses qui ont été écrites par Okvoling et le bilan. Voilà, ensuite il faut, alors on a déjà parlé il y a quelques mois je crois, il va falloir qu'on commence à réfléchir à savoir est-ce qu'on vise une soutenance en fin juin ou pas. C'était bien, donc une soutenance avant l'été, ce qui va changer des trucs un petit peu sur la façon d'organiser le travail. Alors je pense que c'est bien de finir avant l'été parce qu'il va changer des trucs un petit peu sur la façon d'organiser le travail. Alors je pense que c'est bien de finir avant l'été parce qu'ils ont une chanson ou d'autres choses à faire l'été. Vous avez envie de travailler, de vous reposer, c'est bien légitime. Mais voilà, en tout cas il faut toujours réfléchir entre vous, donc vous n'êtes pas obligés de répondre tout de suite maintenant, mais pour la prochaine fois on se voit dire plein de travail en fonction. Il y a un problème pour moi c'est que le 3 mai je vais retourner en Chine pour faire le stage. Oui mais s'il faut qu'on organise une soutenance en hybride parce que vous êtes en Chine pour le stage on le fera. De toute façon si votre stage il démarre le 3 mai, je sais pas à quel point tu finis, mais bon c'est on peut s'organiser pour faire ça. Donc déterminez votre objectif indépendamment de ça. Après vous pouvez très bien dire que vous préférez se tenir en septembre, même avec un mémoire fini fin juin, c'est possible aussi. Mais voilà, séparez les différents trucs.

虽然没有确切的数字，但显然，在手写的论文中，只有6篇参考文献显然不够。无论是你们使用的工具，还是你们的方法，都不要犹豫，应该引用你们使用过的文章，特别是那些Okvoling和Bilan等人的文章。

接下来，我们之前几个月讨论过，可能需要开始考虑是否要在六月底前进行答辩。最好在夏天前完成答辩，这样会稍微改变一些工作的安排方式。我认为在夏天之前完成会比较好，因为夏天他们有其他事情要做，你们也可能希望休息，这样是合理的。但无论如何，你们还是要相互讨论这个问题。

所以，你们不需要现在马上回答，但下次见面时，我们可以基于这个问题来安排后续的工作。

我自己遇到的问题是，我5月3日要回中国去做实习。

说话者1：

没关系，如果你在中国实习，我们可以安排一个混合模式的答辩。反正如果你的实习从5月30号开始，我不确定你什么时候结束，但我们可以根据情况调整。所以，首先确定你们的目标，独立于此之外。

你们也完全可以说，即便论文在6月底完成，你们更倾向于等到9月再答辩，这也是可以的。

但是，记得将这些不同的事情分开考虑。

【无论你们是否选择在6月底之前答辩，论文本身应该在6月底之前完成】

Il me reste deux buddy questions. J'ai envoyé le lien pour la partie d'intersection. En observant les tableaux et les images générés pour Space Large et Space Small, ils semblent avoir des performances similaires en reconnaissance des antidénommés.

我还有两个问题。关于交集部分，我已经发了链接。通过观察为 Space Large 和 Space Small 生成的表格和图像，它们在识别实体命名方面似乎有相似的表现。

D'accord, d'accord. Je dis ça dépend ce que vous appelez similaire, ce que vous voulez dire par similaire.

好的，好的。我说，这取决于你所说的“相似”，你指的是什么呢？

Le nombre de... le nombre d'antidénommés. C'est des tokens, non, qu'il y a le même ? Non, c'est les antidénommés. C'est par exemple pour Svej Silarč et Stanza, il y a par exemple 3000 ou 4000 entités nommées en commun, et pour Svej Smo et Stanza c'est le même nombre. C'est un peu bizarre.

是指……命名实体的数量。是不是这些标记数量相同呢？不，是指命名实体。例如，对于 spacy sm和 Stanza，它们有大约 3000 或 4000 个命名实体是相同的，而对于 spacy large 和 Stanza，它们的数量也是相同的。这有点奇怪。

C'est possible, je sais pas, il y a deux chiffres, il faut aussi comparer le nombre d'entités. Vous avez le nombre d'entités en commun, mais en fonction du nombre d'entités totale. C'est-à-dire que si Spacy, enfin mettons Stanza, a 3 000 entités en commun avec Spacy Small, et Spacy Small a 3 000 entités en commun avec Spacy FG, déjà ce n'est pas forcément les mêmes, c'est la première chose, mais ça dépend de l'entité. C'est que si SpaceySmall détecte 10 000 entités, c'est beaucoup plus facile d'avoir 3 000 entités en commun avec quelque chose où j'ai 10 000 entités. Vous voyez, donc il y a le nombre, il y a la valeur absolue, il y a aussi la valeur relative.Donc là, juste sur une valeur absolue, c'est difficile à juger.

这有可能，我不知道。你有两个数字，也需要比较实体的数量。你有共同的实体数量，但也要看总的实体数。就是说，如果 Stanza 和 Spacy Small 之间有 3000 个共同实体，而 Spacy Small 和 Spacy FG 之间也有 3000 个共同实体，那这些不一定是相同的，首先这一点需要明确。但这也取决于实体的数量。比如说，如果 Spacy Small 检测到 10000 个实体，那就更容易和 10000 个实体的模型有 3000 个共同实体。明白了吗？所以有绝对值，也有相对值。所以，单单看绝对值，是很难判断的。我

Pour les fichiers JSON générés, nous avons constaté des erreurs de reconnaissance. Maintenant, nous ne prenons pas ces erreurs en compte, mais nous nous concentrons plutôt sur les différences de performance des modèles lorsqu'ils traitent les mêmes entrées. Et plus tard, une fois qu'on aura effectué l'anodage manuel, nous envisageons le calcul de métriques tels que F1, score et précision？

我们生成的 JSON 文件中，已经发现了一些识别错误。现在，我们并不考虑这些错误，而是更关注模型在处理相同输入时的表现差异。之后，一旦我们完成手动标注，我们会考虑计算 F1 分数、精度等指标吗？

L'idée c'est d'abord de voir de manière globale ce qu'il y a en commun, est-ce qu'il y a une grande différence, ça permet de hiérarchiser un petit peu les modèles. Vous avez des modèles qui ont beaucoup d'entités, des modèles qui en trouvent peu, etc. D'accord ? Donc déjà, ça permet de donner des tendances. On fait l'hypothèse que les modèles qu'on trouve très peu, on sent beaucoup de silence, que ceux qu'on trouve beaucoup, on sent beaucoup de bruit.

这个想法是首先从整体上看一下共有的部分，是否存在很大的差异，这样可以帮助我们对模型进行一些排序。有些模型识别了很多实体，有些模型识别的很少，等等。明白了吗？所以，首先，这能提供一些趋势。我们假设那些识别得很少的模型，可能会有很多遗漏，而识别得很多的模型，可能会有很多噪音。

Voilà, ça, ça donne des tendances. dans le cours qu'on a eu ensemble le jeudi sur la notation, il va falloir que vous suiviez, donc là je vous invite à regarder ce que Caroline a fait par exemple sur le sujet, un guide d'annotation. Parce que c'est pas si simple de déterminer, voilà, ben tiens là ça, est-ce que c'est une gamme quantité ou pas ? Alors généralement on va faire la différence entre une personne et un lieu, c'est assez facile. Donc une personne et un lieu c'est assez facile. Ensuite, une organisation et un lieu, il peut y avoir des ambiguités, il y a plein de choses à traiter. Donc si vous voyez une référence, alors peut-être que Karine pourra vous indiquer précisément dans quelle section de sa thèse. Donc Karine a fait une étude de ça. Vous allez voir que de la même manière qu'on en a parlé en cours, Donc compliqué à traiter mais en échange on espère avoir un plus grand intérêt à Coratheim de faire l'équipe plus simple où là vous allez avoir moins d'indications ce qui permet d'annoter plus de choses. A mon avis ce serait mieux de prendre un truc plutôt simple. Si vous regardez la partie sur la notation dans la thèse de Caroline elle étudie ça.

对，这样就能得出一些趋势。在我们之前一起上的周四课程中，关于标注的问题，你需要去关注一下，看看 Caroline 在这个主题上做了什么，比如她写的标注指南。因为确定是否某个项是一个范围、数量之类的并不那么简单。一般来说，我们会区分人物和地点，这个比较容易区分。人物和地点是比较简单的，但组织和地点之间可能会有一些模糊不清的情况，还有很多其他需要处理的情况。所以，如果你看到了一个参考，或许 Karine 能具体告诉你在她的论文的哪一部分。Karine 对这方面做过研究。你会发现，就像我们在课上讨论的那样，虽然它很复杂，但通过交换，我们希望能够在 Coratheim 得到更高的兴趣，这样可以做一个更简单的团队，减少一些指引，便于标注更多内容。我的建议是你最好选择一个简单一点的部分。如果你查看 Caroline 论文中关于标注的部分，她对这个做了研究。

Ça vous permet d'avoir des références aussi à faire sur ça. C'est-à-dire pourquoi est-ce qu'on s'intéresse à l'évaluation non-supervisée, aujourd'hui, pour le moment, c'est l'évaluation non-supervisée. Pourquoi ? Parce qu'une réalité de terrain, c'est que beaucoup de systèmes vont nous donner des données, sur des données, Wikil y a une différence entre les deux, il y a les modèles, est-ce qu'il y a des modèles qu'on veut garder à l'occurrence, qu'on veut explorer, c'est ça l'idée de désintersection et tout ça. Et après, ok, on va évaluer cette fois-ci de manière supervisée,

这也能帮助你在做这些事情时获得一些参考资料。也就是说，为什么我们现在关注非监督评估？目前的情况是非监督评估。为什么呢？因为在实际场景中，很多系统会提供数据，对于这些数据，Wikil之间存在差异。我们是否希望保留某些模型，是否有模型我们希望探索，这就是交集（disintersection）等问题的核心。然后，好的，我们接下来将以监督方式来进行评估。

Et pour la soutenance, la dissertation sera finie fin juin, ça veut dire qu'après l'examen final, on n'aura pas le rendez-vous ? Ou on aura...

至于答辩，论文将于六月底完成，这意味着在期末考试之后，我们没有时间安排吗？还是我们会有？

On aura des rendez-vous ensemble jusqu'au rendez-vous du mémoire. Sachant que tous les trois, on aura le coupure cet été, donc il n'y a pas forcément trois ou quatre semaines pendant lesquelles on ne sera pas forcément régal, on ne sera pas forcément tous les 3 exactement les mêmes. C'est une des raisons pour lesquelles nous soutenir fin juin, c'est aussi pour vous, pour l'occasion d'être débarrassé etc. parce que le travail d'été est plus compliqué. C'est un juin de début juillet, et donc avec un rendu du mémoire dix jours avant. Donc si vous voulez soutenir en juin, il faut que le mémoire soit terminé aux alentours du 15 voire 20 juin. Si vous voulez avoir un mémoire terminé le 15 ou 20 juin, il faut qu'on ait une version complète, une première version qu'on puisse rel plus en détail cette espèce de règle sur la ligne, mais voilà, pour vous donner une idée un petit peu des enjeux, on est là sur ça. Mais c'est vraiment totalement à votre question.

我们会一直有会议，直到论文答辩的那天。考虑到我们三个人在夏天都会有休息期，所以可能有三到四周的时间我们不能全都在一起，也不一定是完全同步的。这也是为什么我们建议在六月底进行答辩，这样你们也能尽早完成工作，因为夏天的工作会比较复杂。答辩时间大概是在六月底或七月初，因此论文需要提前十天提交。所以，如果你们想在六月进行答辩，论文必须在大约6月15日或20日之前完成。如果你们希望在6月15日或20日完成论文，我们需要先拿到完整的版本，至少是第一版，这样我们可以更详细地审阅。这些就是一些重要的时间节点和要求，给你们提供一个大致的参考。但最终决定还是完全由你们决定。

Donc dans le chat, je vous ai mis le 4.2, à partir de la section 4.2, enfin je vous ai tout mis dans le chat, donc vous pouvez récupérer ça, où en effet, je t'écris la manière dont j'ai répondu le de l'annotation, comment ont été conduites les complètes d'annotation, et ensuite comment je m'en suis servi dans les évaluations pour la reconnaissance d'entité nommée. Sinon je voulais juste te dire par rapport au document Overleaf que moi là quand je compile ça me fait toujours le time out mais ça ne me met plus les mêmes erreurs. Donc je ne sais pas beaucoup de votre côté mais une des erreurs qu'il y a maintenant c'est par rapport à etex.sti et en fait je sais pas si c'est ça qui pose problème et après j'ai des erreurs sur les citations. Mais j'ai plus aucune erreur en rouge avec des histoires problématiques concernant des acolytes non fermés ou ce que Gaël expliquait quand je suis arrivée sur les underscores, j'ai plus du tout ces deux erreurs. Et donc je sais pas pourquoi ça compile. Oui, c'est ce que je disais aussi, je pense que le problème c'est plutôt au niveau du plan de compilation, en tout cas, parce que ça dit « This project exceeds the compile time outline in our free plan » J'ai peur que ça ait quelque chose à voir avec la taille du document que vous ne pouvez plus compiler parce qu'il est volumineux. Ça, ça arrive avec des présentations. Il y a beaucoup d'images dedans.

我在聊天中给你发了4.2部分，实际上我已经把所有的内容都发给你了，你可以拿去查看。在那里，我写了我如何回答注释部分，以及注释是如何进行的，然后我如何将它应用到命名实体识别的评估中。除此之外，我想告诉你关于Overleaf文档的情况，当我编译时，它还是会出现超时，但不再显示相同的错误。我不知道你那边的情况如何，不过现在我遇到的一个错误是关于etex.sti的，实际上我不知道是不是这个引起的问题，之后还有一些关于引用的错误。不过，我不再看到任何红色错误，也没有关于未闭合大括号的错误，或者像Gaël所说的关于下划线的问题，我完全没有看到这两种错误。因此，我不清楚为什么它能编译通过。是的，我也这么说，我认为问题可能出在编译计划方面，因为它说“这个项目超过了我们免费计划中的编译时间限制”。我担心这可能和文档的大小有关，可能因为它太大而无法编译。这个问题在有很多图片的演示文稿中也经常发生。

Pourquoi il n'y a pas des images ?

为什么没有图片？

Vous avez fait des captures d'écran de choses qui devraient être des tableaux qui découpent du coup, prendrait déjà beaucoup moins de place. Vous l'avez peut-être fait pour gagner du temps et c'est ok. Mais là vous avez mis par exemple une image intersection 1.png qui fait 250 kg, n'hésitez pas à travailler un petit peu sur ça, en tout cas le truc compile sans erreur, donc voilà déjà ça vous permet d'avoir une erreur. Après moi je voulais juste savoir comment on fonctionne pour nous rapporter des corrections, parce que là, on ne peut pas écrire dans le permis, donc je vois qu'on ne peut pas commenter non plus. Enfin, je n'ai pas assez de mots. Du coup, comment on fait ? Je ne sais pas, c'est un peu bête comme question mais... Ah si on peut pas compiler, j'imagine qu'on doit juste télécharger le PDF et... Non non non, je crois que vous pouvez faire des commentaires ou des trucs comme ça.

你们做了截图，本来应该是表格的内容，这样的截图占用了更多空间。你们可能是为了节省时间这样做的，这没问题。但是你们放了比如说“intersection 1.png”这样的一张图片，它有250KB，不妨稍微优化一下这个，反正现在文档没有错误地编译通过了，所以至少你们现在能确保没有编译错误。然后我想知道，我们该怎么报送修改意见，因为我们不能直接在Overleaf中写评论，我看到我们也不能评论，或者是我没找到合适的地方。所以该怎么做呢？我知道这个问题有点傻，但…哦，如果我们不能编译，我猜我们应该下载PDF然后… 不不不，我觉得你们应该可以做评论之类的。

Est-ce qu'on peut mettre sur le commentaire des... Si je trouve que t'as pas le sens. On a des commentaires, mais est-ce que vous voulez faire des couches de nos commentaires sur Github et vous les récupérer ? Est-ce qu'il faut qu'on fasse des... non je sais pas. Gaël, est-ce que tu as une idée simple ? Moi j'ai pas une idée simple.

我们可以在评论中添加... 如果我觉得你没有理解意思的话，我们有评论的功能，但你们是想把我们的评论放到GitHub上然后你们去提取吗？还是我们需要做什么…我不知道。Gaël，你有简单的主意吗？我没有。

Le problème c'est qu'en fait, s'il y a conjointement un truc sur Github et Brewleaf, il faut se faire le lien entre les deux. Vous pouvez le faire, vous pouvez mettre votre mémoire sur GitHub et faire que Overleaf a accès aux données. Ça c'est une possibilité. Après avec toujours ce problème de mise à jour, si on est plusieurs à travailler sur le même document sur GitHub, si les poules et les pouches ne se confarent le bon ordre et tout ça, ça peut être problématique. Donc là on touche potentiellement un peu à la limite de ce qui peut être fait sur Overlim Graphic. Donc je me demande Caroline si on n'a pas demandé à Marceau de nous ouvrir un accès sur ce truc. Oui je pense que... Voilà, parce que vous mettez Marina à queue, donc il y a la limite de 2 personnes pour commenter la édition, voilà, donc ça c'est un premier problème.

问题是，如果同时在GitHub和Overleaf上工作，需要将两者链接起来。你们可以把文档放到GitHub上，并让Overleaf访问这些数据，这是一个可能性。但还是会有更新问题，如果我们几个人在GitHub上同一个文件上工作，文件的顺序不对或者没有同步好，这可能会变得麻烦。所以在这里，我们可能已经触及了Overleaf的功能极限。Caroline，我在想我们是不是没有让Marceau给我们开启这个权限。对，我觉得是… 因为你们把Marina加到队列中，所以每次只能有两个人评论文档编辑问题，这是第一个问题。

Après il y a le problème de la confilation, donc on va voir où de notre côté il y a une solution. C'est pas quelque chose de simple parce que avoir en même temps un GitHub avec Le but c'est pas de faire quelque chose qui soit galère, donc le fait d'utiliser la neige nous permet d'avoir un monument bien formé et tout ça. Malheureusement il y a quelques contraintes en échange, donc nous allons voir, nous allons essayer de vous trouver une solution pour avoir un serveur un peu plus adapté.

接下来是编译的问题，我们会看看我们这边是否有解决方案。这不是一件简单的事，因为同时使用GitHub和Overleaf并不是为了让事情变得复杂。这样做的目的是为了确保文档的格式规范和完整性。不幸的是，过程中有一些限制。因此，我们会尽力找出一个更合适的解决方案，提供一个更适合的服务器。

La première chose c'est sur les tableaux parce que je ne suis pas sûre si j'ai bien fait cette fois-ci. Mais je n'ai pas fait un dictionnaire d'yéison et remporté sur Python pour construire les tableaux parce que je me suis dit que chaque fois ça variait et j'avais beaucoup de choses à mettre, vous pouvez l'accéder sur le guide dans la partie script tab. Là, vous pouvez voir de quoi je parle. Donc, j'ai choisi plutôt d'écrire mon dictionnaire et de construire le tableau comme ça sur Python. Mais si vous pensez que c'est mieux, je peux essayer de faire ça, de faire un dictionnaire dans un JSON, emporter et là, sortir le tableau après le latex.

首先是关于表格的部分，因为我不确定这次是否做得对。我没有在Python中使用字典来构建表格，因为我觉得每次都需要调整，而且我有很多内容要放入其中。你可以在脚本标签的指南部分看到我说的内容。所以，我选择直接写字典，然后在Python中构建表格。如果你觉得这样做更好，我可以尝试改为使用JSON格式的字典，然后导入并在LaTeX中输出表格。

Si votre code fonctionne et qui donne des tableaux bien, c'est ok. Ce qu'on vous disait l'autre fois, c'est que si jamais vous avez de nouveaux résultats et que vous êtes obligés de réexécuter plein de trucs, peut-être que ça aurait été pratique d'avoir stocké des JSON C'est une remarque méthodologique. Ce qui vous amène à le bouger et qui sont déjà générés, c'est très bien. La question qui se pose, c'est quand vous allez avoir le même code qui est réutilisé plein de fois, par exemple votre code qui permet d'utiliser Stanza, plus FlowBear, plus Camembert pour générer tout, et si à chaque fois on doit réexécuter un truc pour faire, le fait de passer par un JSON, ça peut simplifie pas dans votre cas, si votre truc fonctionne, il n'y a pas de problème. La question c'est, si à un moment donné, vous avez besoin de régénérer deux tableaux et que ça vous prend 1h30 de calcul pour un résultat que vous avez déjà obtenu, là peut-être que avoir stocké un JSON aurait été utile. Donc voilà, c'est juste une attaque.

如果你的代码能正常运行并生成正确的表格，那就没问题。我们之前提到过的是，如果你有新的结果，并且需要重新执行很多操作，那么将结果存储为JSON格式可能会更方便。这是一个方法论上的建议。对于已经生成的并且已经可以移动的内容来说，这是非常好的。问题在于，当你需要多次重复使用相同的代码时，例如你用Stanza、FlowBear和Camembert生成所有内容的代码，如果每次都要重新执行一些步骤来做计算的话，使用JSON格式可以简化一些。在你的情况下，如果代码正常运行，那就没有问题。问题是，如果某一时刻你需要重新生成两个表格，而这个计算需要花费1小时30分钟，而这些结果其实已经得到了，那么此时存储JSON格式可能就很有用了。所以，这只是一个建议。

Et si là comme ça, ça fonctionne, c'est parce que je pense qu'on va devoir changer les tableaux et tout, parce que ce que c'est écrit, c'est un peu bizarre, c'est vraiment présenté comme si on était en train d'utiliser Flaubert pour les entités nommées, c'est pas ça. Mais je vais garder les choses et j'aimerais vous demander si je garde, parce que j'avais fait aussi l'archive Flaubert seul, c'est celui qui fait vraiment n'importe quoi parce qu'il n'a pas de référence. Est-ce que je les garde quand même ou non ? Parce que j'en ai en parlé sur le mémoire et il est dans le GitHub.

如果现在这样能正常运行，是因为我觉得我们可能需要更改表格等内容，因为目前写的方式有点奇怪，感觉像是在使用Flaubert模型进行命名实体识别，但实际上并不是这样。但是我会保留这些内容，我想问一下你们，是否我应该保留，因为我也做了Flaubert单独的归档，那个模型其实没什么用，因为它没有参考信息。是否我还是要保留它呢？因为我在论文中提到过它，而且它也在GitHub上。

C'est pas la peine, je supprime.

如果现在这样能正常运行，是因为我觉得我们可能需要更改表格等内容，因为目前写的方式有点奇怪，感觉像是在使用Flaubert模型进行命名实体识别，但实际上并不是这样。但是我会保留这些内容，我想问一下你们，是否我应该保留，因为我也做了Flaubert单独的归档，那个模型其实没什么用，因为它没有参考信息。是否我还是要保留它呢？因为我在论文中提到过它，而且它也在GitHub上。

Ça permet de montrer qu'en fait on ne peut pas juste utiliser un modèle et puis comme ça et qui a besoin d'être spécialisé sur le cage pour donner quelque chose ça me paraît pas inintéressant. Alors est-ce qu'il faut faire toutes les comparaisons avec ? Bon ben ça peut se discuter mais à mi-ma sur les premiers tableaux que vous faites oui. Après s'il y a vraiment un écart trop grand, bon ben là actuellement ça ne regarde pas ce qu'on peut adapter mais on peut bien dire au début, la flemère, ça ne fonctionne pas, je le prouve dans ce cas-là, ça me semble intéressant. Ok. Et j'ai trouvé un modèle préentraîné pour les entités nommées, je ne sais pas si vous avez déjà en parlé, si vous avez déjà entendu en parler, que peut-être si c'est le cas de remplacer par ça, je vais vous envoyer, il s'appelle « camétrietrie verte. Je sais pas... Je suis désolé, il y a un bruit de clavier.

这可以显示出实际上我们不能仅仅使用一个模型就能得出结果，它需要在特定领域进行专业化才能给出有用的结果，这样看起来还是很有意义的。至于是否要做所有的比较，那可以讨论一下，至少在你做的前几个表格中可以考虑。之后，如果真的有太大的差距，嗯，到时候我们可以看看能否做些调整，但我们也可以在开始时就提到Flaubert模型的效果不佳，证明这一点，这样看起来也很有价值。好吧。我还找到一个预训练的命名实体识别模型，不知道你们是否已经讨论过，可能的话可以考虑用它来替换，我会发给你们，叫做“camétrietrie verte”。

Oui, là c'est bon. Donc, je trouve, apparemment, ce modèle-là, il est préentraîné pour les entités nommées. Il est en BERT aussi, pas FLEBERT, mais il est entraîné pour les entités nommées. Ça s'appelle « camp maîtrise. Je voudrais savoir si vous le connaissez et si vous désirez qu'on remplace Flaubert aussi par ce modèle ou non, c'est pas la peine. Je l'ai envoyé sur le chat.

是的，看来这个模型是预训练的，专门用于命名实体识别。它也是基于BERT的，不是Flaubert，但它是专门为命名实体识别训练的。它叫做“camp maîtrise”。我想知道你们是否知道这个模型，并且是否希望我们用这个模型替换Flaubert，或者不需要。这个链接我发在了聊天里。

Le lien que vous avez donné, c'est un lien Github, non?

你给的链接是GitHub的链接吗？

Ah non, c'est pas un lien, j'ai juste mis là. Oui. Il s'appelle Cometri. Apparemment, il est disponible sur Reading Face aussi. Je vais essayer de le trouver.

哦，不，这是我直接写的链接。是的，它叫做Cometri。看起来它在Hugging Face上也有。我会尝试找找看。

le renvoi est pas bien fait quoi, ça renvoie juste sur la Kelt de Hugging Face j'arrive pas à trouver le truc donc je peux pas juger alors la question c'est toujours pareil c'est de savoir qu'est-ce que ça donne. J'ai trouvé ça. Si vous arrivez à le faire tourner, vous pouvez comparer en plus, parce que ça peut être intéressant de comparer en plus. Là encore pour aller dans le sens de montrer qu'un modèle qui est spécialisé sur la tâche fonctionne mieux qu'un modeling qui va juste traiter une tâche spécifique comme une autre.

链接做得不好，它只是链接到Hugging Face的主页，我找不到具体的东西，所以我无法判断。问题依然是，最终还是要看它的效果。如果你们能够运行这个模型，我们可以做对比，因为对比一下可能很有意思。这也有助于证明，专门针对某个任务训练的模型，比起那些只是将任务当作一般任务来处理的模型，更有效

Pourquoi pas.为什么不呢。

Si le truc donne des résultats et que vous pouvez comparer, etc. C'est potentiellement intéressant. Pour remplacer Flaubert, si vous voulez, vous pouvez le conseiller à l'Université d'Ottawa.

如果这个模型能给出结果，且你们能够进行对比，那就很有潜力了。如果你们想用它来替换Flaubert，可以推荐给渥太华大学。

C'est bon ? Je vais essayer. Mais sinon, pour l'instant, il n'y a que le camembert qu'on a trouvé. Ma question concerne de soutenance et tout avec des profs?

明白了，我会试试看。但是目前为止，我们只找到camembert模型。我有一个问题，关于答辩和教授的问题？

Le désencadrant de mémoire disponible le jour D. Donc voilà c'est un jury qui va comporter entre, ça dépend des disponibilités des collègues le jour D, mais entre 4 et 8 personnes. Ah, ça fonctionne. C'est Montassier qui va vous donner des infos précises, mais en gros, vous aurez selon le nombre de personnes qui passent ce jour-là, soit 10, soit 15 000 de présentations, soit 5, soit 10 000 de réponses, vous pourrez faire un truc comme ça.

答辩当天的论文将会有一个可用的时间框架。大致上，评审团会由4到8人组成，具体取决于各位教授的时间安排。啊，成功了。Montassier会给你们提供详细信息，但大体上，根据当天参加答辩的人数，你们将有大约10到15分钟的展示时间，之后是5到10分钟的答辩时间，你们可以做一个这样的安排。

Vous avez un temps de présentation, suivi d'un temps de question.

你们将有一个展示时间，接着是答辩时间。

La note finale est une... Je vous ai donnée à l'issue de la sentence.

最终的评分是根据答辩的表现得出的。

Est-ce que je peux ajouter quelque chose par rapport, Marina, à la consommétrie ? Là, moi, je vous dirais de ne pas trop vous attarder, parce que quand je regarde le GitHub, le rythme n'est même pas écrit.

我能补充一下关于Marina和消耗模型的内容吗？从我查看GitHub的情况来看，实际上并没有写明模型。

OK. quand je regarde le GitHub, l'origine n'est même pas écrite. Ok. Oui, non, je viens de trouver ça comme solution.

好的。根据我查看的GitHub，源代码甚至没有写明。好吧，嗯，我刚刚找到这个解决方案。

Il n'y a pas d'origine non plus.

也没有标明源代码。

Et même le FLEBER qui a été entraîné pour la reconnaissance d'entités nommées des trucs médicaux, je ne l'ai pas trouvé non plus.

即使是FLEBER（用于医学命名实体识别的训练模型），我也没有找到。

Je ne suis pas sûre que... Peut-être que c'est mieux de faire les... Avec ce que vous avez déjà, de bien fignoler ce que vous avez, de bien faire les descriptions et tout ça. Et s'il y a du temps, peut-être d'aller regarder. Mais là, donc... Moi, ce que je vois, c'est que c'est un article qui semble en effet décrire quelque chose qui renvoie vers un github avec un projet source sur lequel il n'y a pas de rythmique, qui date d'il y a 3 ans. Donc déjà ça pour moi c'est, enfin je ne sais pas si c'est pour souligner Hegel mais c'est un truc qu'il faut, je pense que c'est pas équilibré.

我不确定...也许最好先把现有的工作完善好，把描述做得更清楚。如果有时间，再去看看。但是，现在我看到的是，这篇文章似乎确实描述了一个项目，链接到一个GitHub，那个源项目已经三年没有更新了。对于我来说，这是一个问题。就是说，这个项目看起来没有更新，甚至没有详细的描述，我认为这个项目可能不平衡。

Je suis d'accord qu'il y a des choses beaucoup plus importantes en ce moment, surtout, étant donné qu'il y a côté guitar et côté mémoire, donc on doit être à jour sur les deux choses. Donc je pense que, idéalement, je l, j'imagine, le rendez-vous dans deux semaines. Je pense que d'ici là, il faut vraiment régler les erreurs, dans le mémoire, et puis régler les points évoqués lors de la réunion concernant tout ce dont on a parlé. Je dirais juste que quand vous avez un dépôt du TED qui a plus d'un an, sur lequel il n'y a pas de mise à jour depuis plus d'un an, même plus de six mois, et qu'il n'y a pas de rythme vraiment descriptif, il ne faut pas y aller.

对，我同意。现在确实有很多更重要的事情，尤其是考虑到你们有论文和答辩的事情，所以你们必须把这两方面的工作都更新到位。理想情况下，我想在两周后的约会之前，应该先解决论文中的错误，并且处理我们会议中提到的所有问题。我只是想补充说，如果一个GitHub项目已经一年多没有更新，且没有任何更新和详细描述，不应该继续使用它。

Oui, et l'expérience en tout cas, je ne sais pas après, Yann, ce que tu en penses. Non, non, je suis tout à fait d'accord.

是的，经验上来说，至少我不确定，Yann，你怎么看？

Là, je pense que ce qui a été dit, c'est que si c'est facile à faire, le truc fonctionne, les résultats, et que vous pouvez facilement obtenir ce qui vous permet de compter en haut aux autres, pourquoi pas ? D'accord ? Avec toutes les réserves qui ont été dites, effectivement, s'il n'y a pas la documentation, ça va être un peu compliqué. Voilà. En gros, si ça prend une heure de faire le truc,pourquoi pas.

我认为，正如刚才所说的，如果这个东西很容易做，而且它能够正常工作，结果也很好，且你能轻松获得它来跟其他模型进行比较，那就没有问题，对吧？但是，正如我的同事们所说，如果没有文档，那可能会有点麻烦。总的来说，如果这个事情只需要一小时完成，那就可以做。

Mais il y a d'autres priorités, ça a été dit par mes collègues. Si vous me disiez qu'on n'a pas de résultat, peut-être que ça signifie, par rapport à ce que Caroline disait, que le truc n'est pas un jour, il n'y est bien documenté, que la manière de fonctionner est simple, et vous l'avez vu, et ça ça me permet de le dire, c'est un truc qu'il va falloir dire dans le même moment etc. c'est qu'effectivement spaCy c'est peut-être moins bien que d'autres trucs etc. Mais c'est facile à utiliser. Vous voyezinaison entre la qualité intrinsèque d'un modèle et son utilisabilité. Et après le coup, Spacey, en termes d'identité de même, c'est trois lignes de code. C'est du si bon matin.

但是还有其他优先事项，正如我的同事们所说。如果你告诉我我们没有结果，可能就像Caroline说的那样，那个东西没有更新，也没有详细的文档，且使用起来不简单。这就是我想说的，确实像spaCy这样的工具，可能没有其他工具好，但是它很容易用。你们能看到的是模型的内在质量和它的易用性之间的平衡。最后，像spaCy这样，在命名实体识别方面，只需要三行代码，基本上是“早上起来就能用”

J'ai juste une dernière question. On va discuter sur la soutenance en fonction de nos projets, mais est-ce qu'à ce moment-là où on est, vous nous conseillez plutôt d'essayer d'avancer pour la faire en juin, ou vous conseillez mieux en septembre, juste pour qu'on ait un peu de confiance ?

我只有最后一个问题。我们会根据项目讨论答辩的时间，但在目前这个阶段，您建议我们尽量准备六月答辩，还是更好等到九月？这样我们可以有些信心？

Alors, mon avis c'est que vous avez bien avancé, d'accord, par rapport à ce qu'on a vu avant, etc. On a vraiment des trucs concrets, on a des tableaux, des choses comme ça. Je vais pas vous le cacher, du coup, le hack, etc. on a vraiment des trucs concrets. Faut muscler la bibliographie, etc. tout à fait réaliste d'imaginer rendre ça, c'est-à-dire d'avoir une visière de mémoire à la tête des musulmans. Alors, il y a des avantages de se faire accompagner en deux. L'avantage de faire ça fin juin, c'est que vous êtes débarrassé. C'est pas nourrigeable, ça vous permet de passer à autre chose avant clairement, voilà, l'été, vous allez faire vos trucs, vous allez prendre des vacances, vous allez bosser, vous allez faire plein de choses, etc. Ce ne sera pas forcément facile de rester dans l'éclat. Donc en fait, on pense avoir deux mois de plus, on se soutient dans ses temps, en fait on n'a pas vraiment deux mois.En réalité, c'est plutôt avec le stage, etc. Mais ça vous permet aussi de se débarrasser parce que vous qui avez un stage, votre stage il va faire, il va être duré, juin, juillet, peut-être, je n'en sais rien. De toute façon, ça va prendre du temps disponible. Donc le fait de faire tout dans la journée, à mon avis, c'est une bonne solution. le mémoire il est fini au 15 juin et après au 15 juin on a juste la préparation de la présentation, on va la faire, d'accord ? Donc moi je conseille ça personnellement.

那么，我的意见是，考虑到我们之前的进展，你们已经取得了很大的进展。我们确实有一些具体的成果，有了表格、这些东西，我不会骗你们，确实有一些实质性的内容。需要加强的是文献部分等，完全可以在六月前完成，并且交出一个完整的论文。所以，六月底答辩有很多好处。做这件事的好处是，你们可以早点解脱，先把它搞定，然后可以开始做其他事情，去度假，工作，或者做任何事情。这样的话，假设我们给你们多两个月，你们觉得可能会有更多时间准备，其实并没有更多时间，实际上就是要结合你们的实习等安排。你们有实习，可能是六月、七月，甚至不知道具体的安排，反正都会占用你们的一些时间。因此，尽早做完答辩，我认为是一个不错的选择。只要论文能在六月十五号之前完成，之后你们只需要准备答辩就行了。个人来说，我建议尽早完成。

C'est... alors la question que peut vous poser Sémarins en termes de notes est-ce qu'on aurait pas une meilleure note en attendant septembre ? Alors possiblement parce que vous allez pouvoir y passer plus de temps et tout ça, mais d'un certain point de vue le niveau d'exigence d'une soutenance en septembre est un peu plus élevé que pour une soutenance en juillet. Donc le même mémoire, soutenue en juillet et en septembre, au fond, l'emballement est un peu meilleur en juillet. Pas cinq points de mieux, et peut-être un. Donc voilà, moi ça me semble compatible avec une soutenance fin job, comme ça vous vous êtes débarrassé et puis vous arrivez pas trop fatigué à l'entrée en M2 et vous avez le temps de vous occuper de trucs.

这个嘛……你可能会问一个问题：从成绩的角度来看，我们是不是等到九月答辩会拿到更高的分数？答案是：有可能，因为你们可以投入更多时间去完善。但从另一个角度讲，九月答辩的要求其实会比七月答辩高一点。所以，同样的论文，在七月答辩可能反而会得到稍微好一点的评价。当然不会差五分那么多，可能就差一分左右。所以就我看来，七月底答辩是完全可以的，这样你也可以早点完成任务，不至于带着疲惫状态进入硕士二年级（M2），还能有时间处理别的事情。

Donc vous n'avez pas le rendu du PDF mais vous savez que votre truc compile. De toute façon dans un premier temps, la tête qu'a le PDF on s'en fiche, on n'est pas en train de regarder le placement des tableaux, le placement des figures, ce qu'il faut c'est qu'il y ait le texte. Donc en gros, chaque truc que vous dites, il s'appuie soit sur les tableaux de'on en est. Alors donc, voilà, ça c'est la règle générale. Et nous de l'autre côté on va essayer de régler ce problème de lieu de compilation. Mais déjà si vous avez un document avec pas d'erreurs,ça permet déjà d'avancer.

所以你们现在虽然看不到 PDF 的最终效果，但你们知道自己的文档是可以编译通过的。反正，在这个阶段，我们不关心 PDF 的外观，比如表格或图片的位置，我们现在关心的是文字内容是否完整。所以大致上，你写的每一段内容都应该有表格或数据作为支撑。这是我们目前的基本原则。我们这边也会尝试解决编译位置的问题（比如服务器的问题）。但只要你们的文档目前没有报错，就已经很好了。