**Notes du 14/03 (RDZ No.3)**

N'hésitez pas à rédiger quelques lignes pour expliquer le truc de l'installation.

请随时写几行文字来解释安装的过程

il faut qu'on ait les 4 mêmes formats, la même structure, si on peut les comparer à personne. Et ça vous pouvez le montrer, comme je disais tout à l'heure, en vous mettant en avant le côté, le point d'abonnement pour faire ça, c'est ce qu'on a fait en cours. On va avoir la même entrée, la même sortie. Donc la même entrée c'est vos textes

我们需要确保四个系统有相同的格式和结构，这样我们才能进行比较。你们可以像我之前说的那样展示，突出重点，就是我们在课堂上做的那些。我们会有相同的输入，和相同的输出。这里相同的输入就是你们的文本 .

Ensuite il a fallu trouver le moyen de donner à l'outil de reconnaissance entité nommée le truc sous la bonne forme et de pouvoir après récupérer le résultat. Il peut y avoir des petites différences, ça ne marche pas pareil pour Stormform par exemple, comme ce qu'on a ici, et pour les deux modèles Transformer. Donc vous montrez ce truc là, et là on va juste chiffrer. On va dire tel modèle il en trouve tant dans接下来我们需要找到一种方法，把正确格式的文本提供给命名实体识别工具，并且能够之后获取结果。可能会有一些小的差异，像Stormform和这边的两个Transformer模型的工作方式就不一样。所以你们可以展示这个内容，然后我们就可以进行计数。我们会说某个模型在每个类别中找到了多少个实体，等等。 chaque caisse, etc.

Donc en gros, on sort de tableau, ici. Ça c'est un premier résultat où on va dire ici j'ai les oeuvres, d'accord, donc oeuvre 1, oeuvre 2 avec les titres, d'accord? Et puis on va mettre ici le nombre d'entités qu'on a trouvé dans ce pays-ci, qu'on en a 23 ici, 8 là, j'en sais rien, avec Flaubert, j'en ai 5, j'en ai 6, d'accord? Juste un truc qui montre les résultats de manière globale, d'accord? Et ensuite on va regarder ce qu'il y a en commun.

基本上，输出是一个表格。这个是一个初步的结果，我们会说这是我的作品，对吧？作品1，作品2，带上标题。然后我们会在这里写出我们在每个地方找到了多少实体，比如这里是23个，那里是8个，我不知道，Flaubert是5个，6个，对吧？只是一个展示结果的表格。然后我们接着看哪些部分是相同的。

Pour vous rappeler, on a fait en cours ce truc-là, donc Caroline pourra vous indiquer une chouette librairie en Python qui permet de faire ce qu'on appelle les diagrammes de Rémi. En fait, on va pouvoir voir qu'est-ce qui est commun. Il y a une partie des entités qui sont trouvées par tous les systèmes, d'accord? S'il y a quatre sources différentes qui disent la même chose, probablement que l'entité est intéressante, d'accord?

为了提醒你们，我们在课堂上做了这个，所以Caroline可以给你们推荐一个很好的Python库，能帮助你们做所谓的Rémi图。其实，我们可以看到哪些部分是相同的。有一部分实体会被所有系统识别出来，对吧？如果四个不同的来源都说同样的事情，那么这个实体可能是重要的，对吧？

Et parfois, on va avoir des entités qui vont être présentes dans un seul... enfin, dans la haute couche, si je vais y arriver, d'un seul modèle. Potentiellement, c'est peut-être du bruit, etc. Donc, on va vouloir juste avoir ces résultats. Donc là, après on a un truc, là on peut commencer à rédiger des choses. Donc un tableau récapitulatif de ça. Ensuite on peut faire la même chose en distinguant par type d'entité.

On a parlé de ça hier. La reconnaissance identifiée de nommer, c'est un principe de double classification. 2 titins, est-ce que cette séquence de mots est ou n'est pas une entité? Et si c'est une entité, à quelle catégorie elle appartient? Donc ce truc là global, on peut aussi le faire pour les quatre types d'entités. C'est compliqué de tout mettre dans le même tableau parce que vous voyez qu'en fait ici on a trois dimensions. On a la dimension des oeuvres, la dimension des modèles et la dimension des types d'entités.

有时候，我们会遇到一些实体，这些实体只会出现在一个……如果我能理解的话，是在一个高层次的模型中。可能这些实体只是噪声，等等。所以我们只会关注这些结果。之后，我们就可以开始写一些内容了。做一个总结表格。接下来，我们可以按实体类型做同样的事情。昨天我们谈到过这个问题。命名实体识别是一种双重分类原则。首先，判断这段词序列是不是实体？如果是实体，那么它属于哪个类别？因此，这个整体过程，我们也可以为四种类型的实体做同样的处理。把所有这些内容放在同一个表格里比较复杂，因为你们会看到，这里实际上有三个维度：作品的维度，模型的维度，和实体类型的维度。

Un tableau il y a deux dimensions. Donc si je mets trois dimensions c'est le bordel. Donc je pense que c'est intéressant de séparer. Et là vous avez déjà deux pages de mémoire. En gros, ici on est dans la partie... C'est l'expérience 1? Le modèle de mémoire, c'est le modèle IMRAN.

一个表格有两个维度。如果我加上第三个维度，那就会变得非常复杂。所以我认为分开处理会更有意义。你们现在已经有了两页论文了。基本上，这部分内容...这是第一部分的实验吗？

论文的模型是IMRAN模型。

Introduction, Mathematical Methods, Results and Discussion. Ce que vous avez là, c'est dans les résultats. C'est les tout premiers résultats. Ce dont j'ai parlé avant sur le fait d'expliquer comment fonctionnent les modèles, etc. C'est là-dedans. C'est la méthode. Donc en gros, vous allez pouvoir commencer à rédiger des trucs, on se dit, ben tiens, voilà, on a des critiques, allez vous faire demander, etc.

引言、数学方法、结果和讨论。你们现在看到的是结果部分。最初的结果，之前我提到的关于解释模型如何工作的部分，等等。这是方法部分。所以基本上，你们可以开始写一些内容，像是，嗯，我们有了一些批评意见，去请求反馈等。

Peut-être qu'il y a des éléments d'introduction qui vont vous venir. C'est qu'en fait, ici, c'est un truc dont vous ne vous rendez pas compte, mais donc, pardon, le mien est perdu, parce que c'est le truc que je dis à chaque fois, mais vous ne vous rendez pas compte de l'expertise que vous avez par rapport à vous d'il y a un mois. D'accord? Si on revenait dans le passé par rapport à il y a un mois, vous savez 100 fois plus de trucs sur le sujet. Donc, il faut commencer à écrire des trucs, même les trucs qui vous sont...

也许一些引言的内容会浮现在你们的脑海中。实际上，你们并没有意识到，但相比一个月前，你们现在已经拥有了很多专业知识。对吧？如果我们回到过去的一个月，你们已经学会了100倍的内容。所以现在就开始写一些东西，即使是你们认为的琐碎部分……

Pourquoi on fait la réconnaissance? Inspirez-vous de ce que j'ai raconté hier et dites-le en mieux. Et on dit, tiens, ça serait bien d'avoir une référence qui fait ça. L'introduction, en fait, même si elle est au début du mémoire, on la rédige en fait à la fin. Ou plutôt, on met des notes au fur et à mesure et on finit de la rédiger à la fin. Parce qu'on sait ce qu'on a raconté.

为什么要做实体识别？可以参考我昨天说的内容，并用更好的方式表达出来。然后可以说，嗯，也许可以找到一个相关的引用来支持这个论点。引言部分，虽然它位于论文的开头，实际上我们是在最后才完成撰写的。或者说，我们会在写作过程中做一些笔记，并在最后把它们整理好完成。因为我们已经知道自己写了什么内容。

C'est-à-dire qu'un mémoire, ça ne s'écrit pas du début à la fin, il s'écrit d'abord par le milieu, puis après on écrit le début, puis après on écrit la fin. Donc ici, hop, la partie descriptive, on a utilisé Pelletru, vous avez été voir des tutos, ben regardez, j'ai utilisé Pelletuto pour faire ça, et il y a des problèmes. D'accord? Quel est le point commun entre ces trucs-là? Bon ben par exemple, si je regarde votre code, Flaubert et Calenbert, en vrai, c'est un peu la même structure.

也就是说，一篇论文并不是从头到尾写的，而是先写中间部分，然后再写开头，最后写结尾。所以在这里，嗯，描述部分，我们使用了Pelletuto，你们去看了教程，嗯，看看我用了Pelletuto做这个，确实有些问题。对吧？这些东西有什么共同点呢？嗯，比如说，如果我看你们的代码，Flaubert和Calenbert，实际上，它们的结构有点相似。

Effectivement, ça fait l'essence que vous allez regarder Marina sur Spacey, vous allez dire, ça serait cool que Spacey puisse charger ça. Je sais pas pourquoi, mais ça ne fonctionne pas comme ça, Hugging Face et Spacey, c'est fait différemment. Il y a des raisons et tout mais ce qui fait que de mon point de vue en tout cas Spacey est plus facile à utiliser je trouve que les modèles transformés etc. D'accord? Et dire le modèle TRF et bien on l'a pour l'anglais, on l'a pas pour le français. D'accord? C'est à dire que j'ai le modèle small, medium, large, donc ça on a vu ça hier, et le modèle standford je l'ai pour l'anglais,peut-être pour d'autres langues, dans ce cas je l'ai pas pour le français. Pourquoi pour la reconcentrité de l'art. Il a existé à un moment, puis ça a disparu. Voilà. Ne vous enouyez, on est dans ce truc-là. Et donc, ce que je vous raconte ici, ça va vous donner des idées. Ah oui, il faudra dans l'introduction qu'on ait introduit ce truc-là. Pourquoi on s'intéresse aux entités nommées dans les couettes,ça va falloir l'expliquer dans l'introduction. Et ça, ça va être plus clair au fur et à mesure. Et là, vous pouvez prendre des notes, et en fait, votre mémoire, il se peuple comme ça d'un petit truc et tout ça. Et là on a vraiment du contenu qu'elle a. Donc matériel et méthode, vous pouvez décrire le corpus.

实际上，等你们看看Marina在Spacey上做的，你们可能会说，嗯，如果Spacey能够加载这个就好了。我不知道为什么，但它不像这样工作，Hugging Face和Spacey的工作方式不同。有一些原因，但从我的角度来看，我认为Spacey比使用变换器模型等更容易使用。对吧？然后说到standford模型，我们有英文版的，但没有法文版。对吧？就是说，我有small、medium、large模型，昨天我们已经看过了，standford模型我有英文版，可能还有其他语言的，但没有法文的。为什么呢？因为它曾经存在过，但后来就消失了。就这样，不要担心，我们处在这种情况中。所以，我给你们讲这些内容，应该会给你们一些灵感。哦，对了，在引言中我们得提到这一点。为什么我们要在文本中识别命名实体？这部分需要在引言中解释清楚。随着时间的推移，这部分会越来越清晰。你们可以做一些笔记，实际上，你们的论文就是这样一点一点地填充内容的。现在我们已经有了真实的内容。接下来是“材料与方法”部分，你们可以描述一下语料库。

On a vu ça en cours, comment est-ce qu'on décrit un corpus? Quel est le nombre d'oeuvres? Quelle est la taille en mots? Quelle est la taille en vers? Vous faites des tableaux. Vous pouvez regarder dans les articles que publie Caroni et Loup de Minac. Dans leurs articles, à chaque fois au début, il y a un tableau qui décrit les oeuvres. C'est-à-dire en gros c'est ça, sauf que là ça va être oeuvre 1, quel est l'auteur, combien de mots, combien de lignes, la date, etc. Savoir ce que j'ai dans le corpus.

我们在课程中看过，如何描述一个语料库？作品的数量是多少？字数是多少？行数是多少？你们可以做表格。你们可以查看Caroni和Loup de Minac发表的文章。在他们的文章中，每次一开始都会有一个表格描述作品。也就是说，大致是这样的，除了这次它会是作品1，作者是谁，字数是多少，行数是多少，日期等等。了解语料库中的内容。

Ça prend un peu de temps à faire, mais vous en aurez besoin à un moment. Et tous ces trucs là, hop, ils se remplissent au fur et à mesure. Et ça c'était intéressant pour nous, on a un résultat tangible. Aujourd'hui on peut montrer des trucs. Pardon, il faut que j'arrive là dessus, on assiste un peu brutalement, etc. Je suis vraiment désolé

这需要一些时间来完成，但最终你们会在某个时刻用得上。而且所有这些内容，嗯，它们会随着时间的推移逐渐填写。对于我们来说，这是一个有意义的结果。今天我们可以展示一些成果。抱歉，我必须调整一下，稍微有些突兀。我真的很抱歉。.

L'idée c'est que moi ce que je veux c'est avoir un résultat. Ce que je veux c'est que là, demain après demain, vous soyez à l'aise pour programmer des trucs, etc. et démystifier ça. J'ai conscience que ça peut être dur, etc.

我的想法是，我想要一个结果。我希望明天或后天你们能轻松地编程，等等，去揭开这些事情的神秘面纱。我知道这可能会很难，但目标是要有输出。我们不会一直跟你们说“嗯，假如我们做这个呢？”，不是的，我们要做这个。我们需要的是具体的成果。现在，我们有了成果。像是切好的胡萝卜。

Ah, oui, voilà, exactement. On a des résultats tangibles, etc. Et là, maintenant, on peut se dire, à partir de ça, qu'est-ce qu'on aura besoin de raconter en amont? Tiens, ça sera bien d'avoir une section qui définit la reconnaissance qu'il avait, ça sera bien de faire un état d'émerge.

啊，对了，没错。我们现在有了可触及的成果，等等。现在，我们可以思考，从这一点开始，我们需要提前讲些什么内容？嗯，可能需要有一部分来定义什么是命名实体识别，做一个文献综述也是很有必要的。

Mais moi je considérais dans ce mémoire qu'il ne faut pas partir de ça, il faut d'abord partir des trucs forgés. Parce que maintenant vous vous rendez compte de ce que c'est. Et encore une fois, c'est toujours pareil, c'est un système d'entrée et de sortie. En entrée j'ai une chaîne de caractère, en sortie on catégorise des éléments. Ces éléments sont des suites de mots,il y a une gamme de mots qui peuvent être de longueur 1, 2, 3, etc. On a des trucs. Mettez des exemples. Vous avez une super chouette bibliothèque, c'est display ici, c'est ça? Oui, vous avez vu ça? Qui permet de présenter les visuels et les mots.Qui permet de présenter les visuels et les mots, et c'est comme si vous aviez mis un coup de stabilo sur les mots, c'est très joli. C'est vrai? Oui, c'est vrai. Oui, pardon, stabilo, c'est ça? Oui, oui, c'est ça. Non mais je parle pas, je parle pas, ok, toi je sais que tu sais mettre, c'est avec des faux leci, des mots.Et donc, ça sort tout de suite, vous mettez des fiches capture d'écran. Ah ben tiens, voilà un exemple de ça, voici ce qu'on a en sortie, etc. Le display ici, il prouve une sortie sur display ici, d'accord? Donc vous pouvez le faire sur un bout de poème, vous pouvez même prendre un poème en entier. Alors on n'en met pas 50, on ne fait pas du remplissage, mais un poème en entier, où il y aurait plusieurs entités, et on va se faire ça.

但我在这篇论文中认为，不应该从那开始。首先应该从你们已经做好的工作开始。因为现在你们已经知道这些是什么了。再说一次，这总是一样的，它是一个输入输出的系统。输入是字符链，输出是对元素进行分类。这些元素是由单词组成的序列，单词的长度可以是1、2、3个字等。我们有这些东西，举些例子。你们有一个非常棒的库，叫做display，对吧？你们看过它吗？它可以呈现视觉效果和单词。就像你用荧光笔标出那些单词，非常漂亮，对吧？是的，没错。对，是的，荧光笔，就是这样。不是的，我不是在说你，哈哈，我知道你能做的，像是用一些假的颜色，标记那些词语。所以它会立刻显示出来，你可以截图，给它做一些截图。嗯，比如说，这是一个例子，这是我们的输出等等。display在这里展示了输出结果，明白吗？你们可以在一首诗上操作，甚至可以拿整首诗来展示。我们不会放50个例子，不能做填充，但如果是一整首诗，里面有多个实体，我们就可以这样做。

Vous avez vu que dans les poèmes, il y a aussi le titre qui est présent, l'éditeur, etc. Ça fait partie du document, d'accord? Donc c'est intéressant de montrer ça. Nous, on va s'intéresser à ce qu'il y a dans le texte, mais dans un premier temps, on prend tout. Et là, on remplit. Et là, vous avez gagné la troisième partie. Vous pouvez passer en haut, il faut que j'aille sur le public pour le temps de l'interview.

你们也看到了，诗歌里也有标题、编辑者等等，这些也算是文档的一部分，明白吗？所以展示这些也很有意思。我们会专注于文本内容，但首先，我们会先拿到所有的内容。然后填充这些部分。到这个阶段，你们就完成了第三部分。

Vous envoyez un article qu'on a signé avec Edmila, dans lequel vous avez cette présentation d'image, en cette manière de protéger, et vous avez les tableaux qu'on a fait pour... Je vous demande, ça va si, prochainement, on va comparer aussi le temps qui s'amuse... Oui, c'est l'idée que... Vous utilisez, on a vu hier, TQDM, d'accord?

你们已经看过Edmila和我签署的那篇文章，在那篇文章中，你们会看到如何展示这些图像，保护这些信息的方式，并且有我们为此制作的表格。你们可以参考这些。接下来我们还会比较一些时间和其他的情况。是的，这正是我们昨天讨论的那个问题。你们使用过TQDM库吗？

C'est un TQDM, vous allez voir. C'est parfait, voilà. Comme ça, vous avez le temps, vous mettez le temps global. On s'en fiche de savoir la seconde traite. Ce qui nous intéresse, c'est ce qu'on a dit tout à l'heure, d'accord? Vous dites c'est beaucoup dans la mesure où, ce que je disais hier, on est en train d'attendre devant le notre bout. Ce qui est intéressant c'est les ordres de grandeur. Est-ce que quand j'ai un texte, et donc pour comprendre je dois connaître sa longueur，d'où les statistiques sur le corpus dont on parlait, ça prend 15 minutes, ça prend 3 minutes. Et en gros, sur le corpus global, on a un corpus de 15 minutes, ça prend 3 minutes. Et en gros, sur le corpus global, Stone Farm va être plus long que Space Hill. Je ne me jure pas bien qu'Empereur, Camembert, Flaubert vont être plus longs eux aussi que Space Hill. Mais ce qu'ils attendent, c'est les emplois. Est-ce que c'est 2 fois, 5 fois, 8 fois plus long?

这是TQDM，你们会看到的。它非常完美。这样你们就能看到总时间了。我们不关心秒级的精确时间。我们关注的是我们刚才说的内容，对吧？你们说这很重要，因为，正如我昨天说的，我们是在等待自己的那部分。重要的是量级。当我有一段文本时，我需要了解它的长度，了解统计数据，从而能够知道它需要多少时间。比如说，这个统计对整个语料库很重要，它花15分钟，花3分钟等等。整体语料库上，Stone Farm会比Spacey慢。我的意思是，Camembert和Flaubert这两个模型也会比Spacey慢。但是我们想知道的是它们的延迟，具体是两倍、五倍、八倍之差？

Vous êtes d'accord? Je dis que c'est d'accord, mais je ne sais pas si c'est d'accord. Mais tu vas être d'accord. Vous pouvez reprendre les templates des tableaux et les remplir avec vos propres données. Vous pouvez demander ça, sinon à la chaîne JPT aussi, vous mettez vos données dans la chaîne JPT et vous lui dites, fais-moi un tableau, il faut décrire dans le compte ce que vous voulez mettre en programme, en ligne, etc.Ça marche plutôt bien. Pour revérifier après, des fois, il peut inventer des données. Oui, oui, il y a de la réussite. 你们可以从表格的模板开始，填入你们自己的数据。你们可以要求JPT帮你们做这件事，或者你们直接将数据放进JPT里，然后告诉它：“给我做一个表格，描述我想要的内容，比如在代码中用程序行数等等。这运作得相当好。之后为了核实，有时它可能会编造数据。

C'est à dire là on a fait par corpus et là c'est tous les trois corpus ensemble. Il y a différentes manières de présenter selon le grain auquel vous voulez. Vous comme vous avez un corpus de poèmes, je vous conseille plutôt de faire un truc avec les différents noms d'ouvrages, les auteurs.

也就是说，我们按语料库来做，然后这是所有三个语料库的合并。根据你们需要的细节，可以有不同的呈现方式。你们既然有一个诗歌的语料库，我建议你们做一个表格，列出不同的作品名称、作者。

Il ne faut pas voulu mettre les années de parution mais il faut préciser quelque part qu'il s'agit de poèmes de la fin du 19e, début du 20e. Il faut un peu situer. Après, juste pour finir ce qu'on a dit dans la d'autres activités, mais pourquoi est-ce que... Est-ce que vous avez un peu réfléchi à la question de pourquoi récupérer les antiphénomènes dans des poèmes? Parce que M. Lejeune, il a encore drôlé cette question-là.

你不必想要加入出版年份，但必须在某处注明这些是19世纪末到20世纪初的诗歌。需要稍微定位一下。然后，最后，关于我们在其他活动中讨论的内容，为什么... 你们有没有想过为什么要在诗歌中提取命名实体？因为Lejeune先生曾经问过这个问题。

On parle du fait que le modèle SM va récupérer plus d'entités nommées qui ne sont pas forcément des vraies entités nommées, qui vont être des faux positifs. Parce qu'il ne connaît pas, il va le récupérer comme une antithémologie. Oui. Alors que le spécifilard va moins faire ça. Donc du coup, c'est par rapport au fait qu'on va avoir des mots inconnus sur lesquels il n'a pas été entraîné et il va les récupérer, mais ce n'est pas forcément des antithémologies.我们讨论了SM模型会收集更多的命名实体，这些实体并不一定是正确的命名实体，可能是误报的，因为它并不认识，它会把这些当作反义现象来处理。是的。而特定的模型就不会这样做。所以说，问题在于我们会遇到一些模型未曾训练过的生词，它会收集这些生词，但这不一定是反义现象。

Ça peut être des adjectifs, ça peut être vraiment... ou du bruit de l'OCR. Là, en l'occurrence, comme j'ai pas regardé exactement tous les textes, je sais pas s'il y a des textes qui sont très bruités, d'autres moins, mais vous bien-sûr, si SM, il va avoir tendance à récupérer des chaînes de caractères qui sont même pas des mots, qui sont juste du bruit de l'OCR. Du moment qu'il y a des majuscules, par exemple, il va récupérer des trucs avec des majuscules qui ne sont pas des antiphénomènes et les étiqueter comme des antiphénomènes.它们可能是形容词，甚至可能只是OCR的噪音。实际上，因为我并没有完全检查所有的文本，我不确定是否有一些文本的噪音较多，而有些则较少，但你们肯定知道，如果是SM模型，它更容易捕捉到那些甚至不是词语的字符，只是OCR的噪音。只要是有大写字母的，比如说，它就会捕捉到一些带有大写字母的东西，把它们当作反义现象来标记。

Ce que fait moins SpaceyLG. En fait, SpaceyLG, comme il est plus... il est entraîné sur un plus grand corpus, il a plus de vocabulaire, plus de... Chez AnglaisDing, ils ont été entraînés sur plus de contexte et plus de vocabulaire, du coup. Et du coup, il va faire peut-être moins d'erreurs. Par contre, en contexte de bruité, SpatialSem est plus intéressant ，parce que comme entre guillemets il prend plus de risques, c'est-à-dire qu'il va récupérer des trucs, mais pas d'entités, il va avoir tendance à récupérer plus de formes contaminées des entités nommées. Donc les formes contaminées des entités nommées, pour rappel,ça va être écrit Ancestor. Ce PICSM va avoir tendance à récupérer plus ce type de termes, comme étant une entité nommée, que ce PICSG va passer plus à côté des variances, des entités contaminées. Donc oui, ça nous indique un peu le fait que finalement, plus un modèle est entraîné et spécialisé, plus sur données brutées, il devient un peu limitant. Alors que plus un modèle est peu entraîné ou pauvrement entraîné,  
moins il va être limitant sur la donnée brutée. Mais du coup, il va rapporter plus de bruit. Donc, il y a du filtrage à faire. Alors qu'avec PXIFG, il y a peut-être un peu moins de filtrage. Mais par contre, en contexte bruté, il y a peut-être des formes contaminées qui vont plus monter. Mais ce n'est pas tout le temps le cas.

而SpaceyLG就做得较少。事实上，因为SpaceyLG在一个更大的语料库上训练，所以它的词汇量更大，知识面也更广。对于AnglaisDing来说，它们在更多的上下文和词汇上进行训练，因此它会减少错误。相反，在噪音较多的情境下，SpatialSem就显得更有用一些。

因为它，姑且可以说，它更愿意冒险——也就是说，它会收集一些东西，但不是命名实体，它会更倾向于收集那些变形的命名实体。因此，变形的命名实体，比如说，写作Ancestor这样的词，SM模型会更倾向于把这些当作命名实体来处理，而PICSG则会更倾向于忽略这些变形的实体。所以，这让我们得出一个结论：其实，一个模型训练得越多、越专业，它在处理噪声数据时可能会变得有些局限。

然而，一个模型训练得较少或者训练得不够充分，它在噪音数据下的限制较少，但会带来更多的噪音。因此，需要进行更多的过滤处理。而PXIFG模型可能需要的过滤就较少。但在噪音较多的情况下，可能会有更多的变形命名实体被识别出来。但这并不是一直都发生的。

En fait, ce qu'il y a, c'est que c'est le cas à grande échelle. Si on regarde sur des textes comme ça, on va voir qu'il y a des variations. Et ça, ça dépend beaucoup du contexte. Et ça, c'est quelque chose que moi, je n'ai pas pu analyser en profondeur. Ça m'aurait demandé encore deux années de thèse, je pense, pour vraiment regarder, phrase par phrase, pourquoi dans ce contexte là? Moi j'ai surtout fait des études un peu globales sur des textes et globales sur des corpus et sur des langues. Mais c'est intéressant comme question. En tout cas j'espère que j'ai répondu. Donc je pense que dans cet article là on en parle un peu. Sinon dans l'article que vous envoyez... Désolée je sais pas que je veux que m'auto référencer, c'est pas ça, mais c'est que je suis une des personnes qui a écrit la piste. Et dans les bibliographies, des articles-là, il y a d'autres gens qui ont écrit sur les deux niveaux, donc vous pouvez aller les récupérer dans les bibliographies. Normalement, j'ai essayé de faire bien mon travail, et dans la partie état de l'artiste. Donc vous pouvez retrouver les artistes. Alors le géo... Après je ne dis pas que c'est... Je ne dis pas qu'il n'y a que moi qui fais ça, c'est pas ça, mais je suis une des personnes qui a le plus écrit là-dessus ces 2-3 dernières années.

实际上，这是在大规模数据上才会出现的现象。如果我们查看这样的文本，就会看到有很多变化。而这些变化在很大程度上依赖于上下文。这是我没有能够深入分析的内容。我觉得如果要真正逐句分析并了解为什么在特定上下文中会出现这种情况，可能还需要再花两年时间做论文研究。我主要做的是对文本、语料库和语言的整体研究。不过这个问题很有意思。无论如何，希望我有回答到你的问题。所以我认为这篇文章中有提到这一点。如果没有，你可以查阅你们发送的文章... 抱歉，我并不是想自我引用，不是这个意思，而是因为我也是写这篇文章的作者之一。而且在这篇文章的参考文献中，还有其他人也写了关于这两个层面的内容，你可以从参考文献中找出来。一般来说，我尽力做好了我的工作，尤其是在文献综述部分。所以你可以找到相关的作者。然后，我并不是说只有我在做这些研究，并不是这样，而是过去两三年中，我是最活跃的研究者之一。

Avec ma tête, parce que... C'est pas ça. C'est pas pour m'autociter, genre, je suis la meilleure, c'est pas ça. C'est ça où j'essaie de vous dire que c'est pas juste que je suis genre, ah oui, c'est juste... Et en fait, l'explication que vous m'avez donnée, ça me fait penser un peu à la question que vous avez posée sur pourquoi exprès des antiphénomes du pèle parce que j'avais pensé à ça, dans le sens que le poème, lui-même, il a conçu une structure que, par théorie, a une syntaxe qui est différente de la langue traditionnelle, on va dire,我说这不是为了自我引用，也不是想说我最好，而是想告诉你们，并不是我只是说“哦，这很简单...”。实际上，你给我的解释让我想到你之前问过的问题——为什么要特别提取诗歌中的“反现象”。我曾经想过这个问题，因为诗歌本身就有一种结构，而这种结构理论上是与传统语言的语法不同的。

Je pense qu'on attend qu'elles soient plus difficiles à être extraites. On n'a pas le contexte. Si on n'entraîne pas, si on n'entraîne pas la modélisation des textes, les textes ne sont pas de poèmes, et la chantatine n'est pas la même. Je me demande si ce n'est pas une idée reçue en fait. Moi c'est une réquestion que je pose.

我认为我们预期这些反现象会更难提取。因为我们没有上下文。如果我们不训练模型，文本本身就不是诗歌，语调也不一样。我在想这是不是一种固有的观念。其实我在提出这个问题的时候，并不是在评判你写的东西，而真的是一个我自己在思考的问题。真的，是否需要专门训练一个模型来识别诗歌中的元素？

Je me demande si ce n'est pas une fausse vrille d'essayer. Oui, oui, c'est à ça que je pense. Non, mais moi, vraiment, c'est une vraie question que je me pose. C'est pas un truc en jugement sur ce que vous écrivez. C'est plus vraiment une vraie question que je me pose. Est-ce que vraiment, il y a besoin d'entraîner un modèle sur des poèmes pour faire de la reconnaissance sur les poèmes. Ce que vous m'avez dit, que de plus, c'est entraîner moins l'intérêt de changement de langue dans les sciences de prononciation, ça m'a fait penser un peu à ça en vrai. Donc vraiment, je pense que ça c'est une vraie question. En tout cas, ça c'est la question la plus en vogue actuellement. Est-ce qu'on a besoin d'entraîner des modèles sur tout un tas de données et ensuite on va leur rajouter des couches de co-finance, de fine tuning pour les spécialisés.

我在想，这是不是一种错误的尝试。是的，我就是在想这个。不是说“这是我认为的”，而是真的在问自己这个问题。你告诉我，确实，有时候训练模型的目的是为了适应语言的变化，这让我想到了这个问题。所以我真心觉得这真的是一个问题。现在这个问题也很流行。我们是否需要针对大量数据训练模型，然后再通过一些微调层来专门化它？

Ça c'est des vraies questions maintenant, qu'est-ce qu'on ne peut pas fournir des larges modèles de langues entraînés sur tout un tas de trucs Et ensuite, avec vos collègues Camille et Yéline, on a commencé l'année dernière un travail avec Camille pour l'entraînement d'un modèle sur données brutées, un modèle de SPECI, et on s'est rendu compte qu'on n'avait pas à noter assez de données pour que ça ait un impact. Donc en fait, ces larges modèles de langues, ils sont bien. Après on dit oui on peut les fine-tuner avec des données spécifiques. Oui mais quelle quantité de données spécifiques est suffisante pour vraiment fine-tuner parce qu'ils disent oui il n'y a pas besoin de beaucoup de données. Donc nous on a dit ok, il n'y a pas besoin de beaucoup de données. Mais en fait nous on n'a pas beaucoup de données. Pour nous c'était déjà 20 heures d'annotation, 7 annotateurs et ça n'a apparemment pas été vraiment suffisant. Pour l'instant on n'a pas encore la réponse définitive. On va peut-être la voir dans les semaines à venir. Mais pour l'instant l'intuition qu'on a c'est que finalement on n'a pas annoté assez de données alors que quand même on a annoté je ne sais plus 13 000 ou 15 000 tokens. Et c'est écrit dans les tutos qu'elle n'avait pas besoin de beaucoup de données. Donc ça veut dire quoi exactement? Ça veut dire quoi? Tu as beaucoup de données? En plus, même si tu avais dit qu'il fallait ajouter 15 000 paquets, peut-être que dans ton cas particulier, ça ne serait pas appliqué. C'est ça la difficulté.

这些现在是很现实的问题。我们是否不能用大型的、已经训练过的大型语言模型，去解决一切问题，然后通过你和你的同事Camille和Yéline的工作，我们去年和Camille一起开始做的一个任务就是在原始数据上训练一个模型，一个SPEC模型，结果我们意识到我们没有足够的数据来产生实际的影响。所以，大型语言模型很好。但后来我们说，“是的，我们可以用特定数据进行微调。”但问题是，究竟需要多少数据才能进行有效微调？他们说“数据不需要太多”。于是我们说，“好吧，数据不需要太多”，但实际上我们没有足够的数据。对我们来说，20小时的标注，7个标注员，似乎还是不够。现在我们还没有得到最终答案，可能会在接下来的几周里看到。但我们目前的直觉是，实际上我们标注的数据量不够，尽管我们标注了大约13,000到15,000个token。这些都写在教程里说不需要太多数据。那么，真正的意思是什么呢？那到底意味着什么？你说需要15,000个数据包，或许在你特定的案例下，它可能不适用。这就是困难所在。

En tout cas, c'est des questions assez justes. Après, alors, moi, pourquoi j'avais proposé ce sujet sur les poèmes? C'est qu'Alina Alco, quand j'ai travaillé à Alina Alco il y a deux ans maintenant, on avait commencé à travailler sur des poèmes d'Alco Linaire où il n'y a pas de ponctuation. Et on s'était aperçus, et donc moi je m'en suis réaperçue après sur d'autres textes, que Spacey, par exemple, avait tendance à récupérer les phrases entières comme étant des entités de mots. Et en l'absence de ponctuation, à ce moment-là, donc il s'agissait du Spacey 3.5, je crois. Donc là, vous travaillez avec, je ne sais pas si c'est Spacey 3.8 que vous avez installé, voilà. Peut-être que c'est différent, mais à l'époque, il y avait des modèles 3.5 pour le français, ils étaient un peu promis quand il n'y avait pas de ponctuation. C'est-à-dire, ils récupéraient vraiment des morceaux de verbe, enfin des verbes presque complets, qui n'étaient pas du tout des entités nommées. Du coup, c'est aussi ça ma question, c'était par rapport au roman où il y a de la ponctuation, enfin, au roman du 19e, 20e, où il y a une ponctuation qui est quand même assez respectée.

无论如何，这些问题都很有道理。接下来，我为什么会提出这个关于诗歌的主题呢？是因为在两年前我在Alina Alco工作时，我们开始研究Alco Linaire的诗歌，这些诗歌没有标点符号。我们意识到，实际上我在其他文本上也发现了这一点，Spacey，比如说，会倾向于把整句句子当作一个词实体来提取。在没有标点符号的情况下，那个时候，使用的是Spacey 3.5版本。我不知道你们是否安装了Spacey 3.8版本，可能有些不同，但当时对于法语，Spacey 3.5的模型在没有标点的情况下就有点问题。它们会把动词的片段，甚至几乎完整的动词片段，作为实体来提取，而这些根本就不是命名实体。我的问题就在这里：对于有标点符号的19世纪、20世纪的小说，它们的标点符号还是比较规范的。

On n'est pas dans le roman 20e, 20e, donc des fois il y a aussi des romans qui sont des formes un peu, comment dire, particulières, oui, des formes un peu en prose, finalement, de notre poème, et où il n'y aurait pas de ponctuation, c'est-à-dire, est-ce que finalement le contexte trastique, parce que la ponctuation c'est un peu ça, c'est un contexte trastique, comment on dit, du coup pas lexical mais typographique, ou je ne sais pas comment ça s'appelle. Trastique ça veut dire sur la phrase. Sur la phrase, sur le segment, que nous, humains, on appelle une phrase-fois. Parce que s'il y a des points, des virgules, c'est une phrase, c'est une syntaxe,c'est comme les gens disent. Et du coup, est-ce que la syntaxe impacte la reconnaissance de ce qu'est le méprendre? Après le méprendre, c'est celui qui fait les poèmes, les oeuvres, les dessins, les matos... Mais là, je crois que dans les poèmes, il n'y a pas l'océan qui n'a pas marché sur les éléphants. C'est un autre sujet que j'aimerais bien aborder ailleurs, c'est comment on fait pour oscillériser des textes dessinés des galénames. En tout cas, c'était, quel est le paysage, dans ma thèse en gros, l'ambition qui n'a pas été réalisée du tout. D'autres choses ont été réalisées, mais ça c'est une perspective encore maintenant.

我们不是在谈论20世纪的小说，所以有时候，也有一些小说是一些特别的形式，嗯，是一些有点像散文的形式，实际上是我们的诗歌形式，而这些形式中可能没有标点符号。也就是说，最终标点符号的上下文是否很重要？因为标点符号其实就是上下文的一部分，它不是词汇的上下文，而是排版的上下文，或者我不知道应该怎么称呼它。标点符号实际上是在句子上，或者在我们作为人类所称之为“句子”的部分。因为如果有句号、逗号，那就是一个句子，这是语法，就像人们所说的那样。那么，这种语法是否会影响到对所提取的命名实体的识别？至于“命名实体”，就是那些创造诗歌、作品、图画和材质的人……不过，我认为在这些诗歌中并没有“海洋走在大象上”的情况。这是我想在其他地方讨论的一个话题，那就是如何将绘制的文本或图形进行数字化处理。总的来说，这就是我在我的论文中提到的，尽管这些目标没有完全实现，其他方面倒是完成了，但这仍然是我当前的一种视角。

C'était, est-ce qu'on peut arriver à dessiner, à présenter un paysage de la littérature du XIIe et XXe siècle alors qu'en France, au Portugal et en Angleterre. Et pour l'instant, les résultats que j'avais vus, pour le français par exemple, c'était qu'on avait quand même un paysage très européen. Après pour l'anglais et le portugais, j'ai pas encore de résultats, mais pour l'instant, même s'il y a des évocations de l'Orient oudes Amériques, etc., ça reste quand même très européen. Et l'Europe antique, c'est-à-dire le Grèce-Italie sont beaucoup citées dans les romans français. Après, évidemment, pour tout ce qui est de romans réalistes, on a beaucoup de choses qui se passent dans les campagnes françaises et après à Paris. énormément de diversité dans la multiplicité des citations de lieux. Parce que moi je parle vraiment des entités nommées de lieux. Alors là c'était génératif si je viens de comprendre. Ah bah là c'était avec les mêmes modèles, c'était Spacey, Calendar, Flair, Stanza.

这个问题是，我们能否描绘出12至20世纪文学的景观，尤其是在法国、葡萄牙和英格兰的文学中。到目前为止，我所看到的结果，比如法语文学，确实还是一个非常欧洲化的景观。至于英语和葡萄牙语的部分，我还没有得到结果。但目前，即使有关于东方或美洲等的提及，整体上仍然非常欧洲化。而古代欧洲，尤其是希腊和意大利，在法国的小说中被大量引用。之后，显然对于所有现实主义小说来说，我们有很多情节发生在法国的乡村，之后是在巴黎。关于这些地方的引用，表现出了极大的多样性，因为我确实是指命名实体中的地点。然后，这是生成性的，若我理解得对。啊，那时候使用的模型包括Spacey、Calendar、Flair和Stanza。

J'avais pas utilisé Flaubert mais... C'est passé pour... Bah je faisais la reconnaissance d'entités nommées sur les lieux. D'accord, mais j'ai pas compris. Vous dessinez des paysages? Oui, après vous faites des cartes.我没有使用Flaubert，

但...这是为... 我是在进行地点的命名实体识别。明白了，但我不太明白。你们是在绘制景观吗？是的，之后你们可以做地图吗？

Quand vous récupérez les noms de lieux, après vous pouvez utiliser une librairie de géolocalisation. Ok. Là, c'est dessiné dans un sens, pas dessiné, mais représenté sur des cartes les lieux qui sont évoqués dans la littérature, donc un paysage littéraire de la littérature française au départ et après comme on avait deux autres corpus. C'est assez bien intéressant. Voilà, donc après on peut recalquer la même chose dans la poésie. Est-ce que la poésie du 19e-20e siècle est une poésie qui va être plutôt réaliste, c'est-à-dire qui va plutôt son créer dans les campagnes françaises, etc. Ou est-ce qu'elle est très orientaliste,c'est-à-dire, est-ce qu'elle va évoquer le Maghreb, l'Asie mineure, tout ce genre de choses. Ça, ça peut être des pistes. Mais ça, c'est vraiment plus, par exemple, pour l'introduction à ce que je vous dis, ce que disait M. Lejeune, c'est plus des pistes à ouvrir dans l'introduction. Vous n'allez pas l'explorer forcément, mais en perspective, en gros, vous pouvez dire, voilà, ça peut servir à ça, pour l'introduction.

当你们提取地点名称后，你们可以使用地理定位库。好的。所以，这是用某种方式表示出来的，而不是绘制，是将文学中提到的地点在地图上展示出来，因此这是法国文学的一个文学景观。然后，我们有其他两个语料库。这样是很有意思的。之后我们可以在诗歌中做同样的事。19至20世纪的诗歌是现实主义的诗歌吗？也就是说，它是否更倾向于描写法国的乡村等？还是它非常东方主义？也就是，它是否会提到马格里布、小亚细亚，或者这些地方。这些可以是一些思路。但这其实更像是我刚才所说的开头部分，正如M. Lejeune所提到的，更多是开篇的一些思路，可能不一定需要在文中展开，但是从整体来看，你们可以说，这些内容可以为后续研究提供帮助。

Et après, en perspective, bon, ben, nous, maintenant, on a fait ça, mais désormais, on pourrait aller vers ça. Vous voyez ce que je veux dire? Donc, ça, c'est vraiment important de bien voir qu'on ne va pas faire ça, là. Mais c'est plutôt pour vous donner des idées d'intro et de perspective. Parce que, si je viens complet, ça répond à la question pourquoi les espèces de notre cité n'ont pas les déformes, de façon que le traçage des paysages est défini. Voilà, c'est ça. Tout au but de ce que j'ai dit, c'est que j'ai envie de poser cette question que M. Bézaine a évoquée sur qu'est-ce qu'on peut dire en introduction et après, moi, je vous dis, même en perspective, en fait,parce que, vous, votre travail, là, c'est plutôt, on démarre, quoi. Donc, au démarrage, il faut regarder, est-ce que ça marche? Déjà, comment ça marche? Donc ça, vous l'avez fait. C'est une étape de passé. Vous pouvez déjà, comme avec M. Lejeune, l'a décrété. Et ensuite, maintenant, c'est plus comment ça marche, c'est est-ce que ça marche? Donc on passe à l'étape de l'évaluation manuelle, quand on vous parle des tableaux.

然后，作为展望，好的，我们现在已经完成了这个，但未来我们可以朝这个方向发展。你明白我的意思吗？所以，这一点非常重要，要清楚我们不会马上做这些，但更多的是给你们一些开篇和展望的思路。因为，如果我完全回答的话，实际上也是为了回答M. Bézaine所提到的问题——我们可以在引言部分谈些什么，然后，我会告诉你们，甚至在展望部分，实际上。因为你们的工作，实际上是一个起步的过程。所以，在起步阶段，首先需要看的是，这是否有效？首先，它是如何工作的？这一点你们已经做到了，这是一个已完成的步骤。就像M. Lejeune已经宣布的那样。接下来，就是看它是否真的有效。所以，接下来就是手动评估的阶段，我们会讨论表格时。

L'article « Spatial Name Entity Recognition in Literary Texts », c'est mon tout premier article. En gros, déjà, vous voyez, j'avais fait des petits tableaux pour présenter les différents résultats. Alors moi, c'était par rapport aux OCR et avec SpaceyHM, par exemple.

我发给你们的第二封邮件中的文章《Spatial Name Entity Recognition in Literary Texts》是我的第一篇文章。我做了一些小表格来展示不同的结果。我的工作是关于OCR和SpaceyHM之类的。

Mais vous, vous pouvez faire du coup, ici, vous n'avez pas d'OCR à même, vous pouvez mettre les noms des outils de reconnaissance entité nommé. Et puis vous mettez vos exemples. Voilà, vous dites tiens, Spacey, lui, il a trouvé cette entité sur la même phrase. Camembert, il ne l'a pas trouvée ou il l'a trouvé un périmètre plus large，parce que des fois, l'outil Spacey va prendre la graisse tout court. La graisse et puis il va prendre la graisse tout court. Et puis Camembert, il va avoir pris la grèce et puis Flaubert va peut-être pris la grèce bleue. Vous voyez ce que je veux dire? Donc ils ne déterminent pas le même périmètre pour l'entité donnée. Donc ça c'est important aussi de bien voir ça.但是你们现在不需要OCR，直接可以放入你们的命名实体识别工具名称，然后展示你们的例子。你们可以说，Spacey识别了这个实体，在同一句话里。Camembert没有找到它，或者找到的范围更广。因为有时，Spacey工具会直接识别“grèce”（希腊），而Camembert可能会识别“Grèce”，Flaubert可能会识别“grèce bleue”（蓝色希腊）。你们明白我的意思吗？所以它们在确定给定实体的范围时是不同的。所以这点也很重要，大家要注意这一点。

Donc ça c'est des tableaux. Donc franchement n'hésitez pas à vous inspirer des tableaux, des choses, des graphiques. Pour faire les diagrammes de veines, j'ai utilisé veine 2. Je vais vous envoyer un gros paquet de veine. Vous pouvez aller couler dedans. Donc ça c'est tous mes travaux de tête.

这些都是表格。所以，真的不要犹豫，尽管参考这些表格、图表。为了做维恩图，我用的是Veine 2。我会给你们发一个Veine的完整包，大家可以在其中探索。这些是我所有的工作内容。

Je vais faire un Jupyter Notebook qui résume les différents programmes que j'ai utilisé, ils datent un peu parce que j'ai fait évoluer les programmes depuis, mais vous pouvez récupérer les différents programmes qui ne sont peut-être pas parfaits, vous pouvez les améliorer, les facturiser, ce que vous voulez, qui me permettait de réaliser les différentes étapes de l'émission de ta liste. Juste pour mon côté, je ne sais pas si tu es d'accord, mais est-ce que vous pouvez éventuellement corriger le code FlaVert pour ajuster le chemin? Quitte à commenter et donner des commentaires sur quel chemin je dois indiquer, parce que même quand je télécharge un local et quand j'essaie de mettre dans mon drive, j'arrive pas à charger pour un drive, pour une raison. Je pense que c'est à cause du réseau. Parce qu'il faut se mettre dans Collab aussi, je pense que c'est... Oui, oui, je suis sur Collab et j'avais essayé de... Non mais parce qu'il faut se mettre dans un dossier qui s'appelle Collab. Ah ok. Parce qu'une fois l'année dernière j'avais travaillé avec une étudiante et elle ne travaillait pas dans Collab et ça ne marchait pas. Jusqu'à ce que je lui dise mettez-vous dans Collab et ça a marché. Je ne sais pas pourquoi, peut-être que ce n'est pas comme ça chez tout le monde, mais elle ça avait été ça, ça ne marchait pas si elle n'était pas dans le dossier. Oui, donc là, je n'arrive pas à accéder à votre dossier à vous. Donc, ça, c'est mon grain d'argent.

我会做一个Jupyter Notebook，总结我用过的不同程序，它们有些已经有点过时了，因为我已经对这些程序做了一些更新，但你们可以获取这些程序，可能它们并不完美，你们可以改进它们，优化它们，随你们怎么做。这个Notebook帮助我完成了实现你列表中不同步骤的过程。只是从我的角度来看，不知道你们是否同意，但你们能不能修改一下FlaVert的代码，调整路径？

如果能的话，麻烦加点注释，告诉我该输入什么路径，因为即使我下载到本地，然后尝试上传到我的云端，我也无法加载到云盘中，原因不明。我觉得可能是网络的问题。因为你们也得进入Collab，我觉得是这个问题。是的，我现在就在Collab上，我试过了...不，因为你们得进入一个叫做Collab的文件夹。哦，明白了。

因为去年我曾经和一位学生一起工作，她没有在Collab上工作，然后就不行了。直到我告诉她要进入Collab，才解决了问题。我不知道为什么，可能不是每个人都会遇到这个问题，但她就是这样，如果不在那个文件夹里就不行。是的，所以我现在无法访问你们的文件夹。这就是我想说的。

Vous avez travaillé avec GitHub, à l'époque? GitHub? Oui, très bien. Vous connaissez la plateforme? Là, je vais leur envoyer un email. Oui, mais comme... Je vais directement conduire l'école avec... ça je ne sais pas comment vous le disiez. Non, non, non, non, c'était pas pour tourner directement, c'était pour déposer. Ah oui, j'ai un bitaba, moi je peux déposer aussi.

对了，你们之前用过 GitHub 吗？

GitHub？是的，很熟悉。你们了解这个平台吗？我现在准备发邮件给他们。是的，不过我直接用学校的邮箱联系... 不太确定怎么做。

不不不，不是直接给他们发，是上传到 GitHub 的意思。哦对，我也可以上传的。

Super, c'est toujours mieux parce que quand vous rédigez le mémoire, vous aurez besoin de dire voici tous les matériels qu'on a besoin. Utilisez l'école, le corpus, etc. sur le vitel. Donc c'est mieux si on commence déjà à faire ça avec les choses qui marchent? Ou en tout cas, ça vous sert vraiment à... En chacune ou en vitel pour les deux? Ça peut être en vitel pour les deux.En chacune, c'est mieux. Faut pas contrater. Pour faire des poules chez des poules, il faut bien se coordonner quand on travaille sur le même dépôt. Sinon ça fait plein d'erreurs à chaque fois. Ça peut arriver de temps en temps, mais c'est pas grave.

因为当你们写论文时，需要说明所有需要的材料。比如使用学校的资源、语料库等。这样做最好是从开始就处理好，确保每件事情都有效。如果有问题，可以在版本控制系统里解决。对于每个人来说，最好单独操作。我们在同一个仓库里工作时需要协调，否则容易出错。虽然有时出错不可避免，但没有关系。

En fait, il faut communiquer quand on utilise GitHub et qu'on n'a pas envie de faire du versionnage. Parce qu'après, on peut faire des branches et des nationales, mais bon c'est un peu chiant. Donc juste écrire un petit message pour dire ah j'ai fait un pouce.

其实，当我们使用 GitHub 并且不想做版本控制时，需要进行沟通。因为之后我们可以创建分支和本地仓库，但这有点麻烦。所以只需写一条简短的消息说“我做了推送”。

Et normalement dedans vous avez faire des déclins de veines. Et vous pouvez regarder, il y a d'autres choses que j'ai mises. Que vous n'êtes pas obligé de faire. Après comme je disais, les programmes peuvent être améliorés, parce que c'est au moins nécessaire. D'ailleurs, les programmes dont je me sers maintenant ne sont plus tout à fait ceux-là. Mais voilà, il y a plein d'améliorations à faire, je pense, sur ce programme. Mais ils sont utilisables en l'état normalement. Peut-être qu'il faudra juste adapter, parce que moi j'avais une entrée où c'était une sortie. C'était un fichier, un auteur. Alors que vous, vous avez un dictionnaire avec le nom des auteurs en clé. Donc ça, il faut peut-être adapter la manière de lire les fichiers. Parce que moi, mes fichiers, c'était des listes d'entités.

所以关于维恩图，我之前发给你们的就是最新的，那个稳定版本的代码，它就在程序里。你们可以看不同的 notebook，但你们也可以看看其他的。我认为 Atala Bâtelier 的 BNS 应该是最新的。很快就有一年了。 所以我认为这个是最新的。而且通常它里面有做维恩图的功能。你们可以看看，我也放了一些其他的内容，你们不必做这些内容。就像我之前说的，这些程序可以改进的，因为它至少是必要的。其实我现在用的程序已经不完全是这些了。但是呢，我认为这个程序还有很多改进空间。

Et donc pour les diagrammes de Venn, vous allez récupérer le programme dans la partie que ça s'appelle... Les intersections. Évaluation de la reconnaissance de notre filmé sur des données de qualité limite des intersections. Et là, vous avez le programme qui sert à faire les programmes de même.

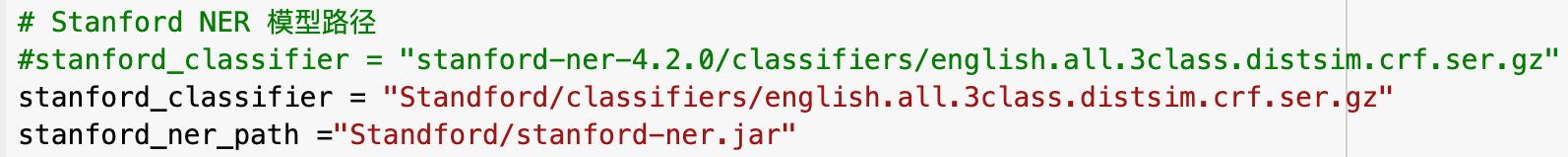
所以关于维恩图，你们需要从“交集”部分获取程序。交集的评估是关于在质量有限的数据上进行命名实体识别的程序。

Je vais en prendre encore plus. Programme. Kanda. Vous avez le modèle français. L'autre Kanda modèle langage STR-FR. et sinon pour lancer l'installation.c'est quoi? Download langue login de la serveur donc là nous on veut lancer le français donc vous voyez la langue vous mettez français. En gros, normalement, enfin, on va dire... Non, ça va pas être dans ce PC. C'est dans Kacus...

Kanda。你们有法语模型。另一个Kanda模型是STR-FR。然后，如果要启动安装，是什么？下载语言登录到服务器， 所以我们要运行法语，你们可以看到语言，选择法语。大致来说，嗯，我们说... 不，这不会在这个PC上。它在Kacus里。

Les conseils de prof:

* Le test du code prend 15 minutes, c'est normal.
* Il est préférable de montrer les résultats séparés pour chaque texte dans Camembert et Stanford.
* Téléchargez à nouveau les éléments du test Stanford en français.

[https://github.com/These-SCAI2023/CORPUS/blob/master/prog/Use\_stanza.py](https://github.com/These-SCAI2023/CORPUS/blob/master/prog/Use_stanza.py)

* Apprendre de ses camarades dans la collaboration d’un projet.
* Écrire le code tout en commençant à rédiger le brouillon du mémoire, en se basant sur ce que nous avons déjà fait. Sinon, une fois que tout le code sera terminé, on risque de ne plus s’en souvenir.
* Prendre des notes à chaque rendez-vous est très important, et l'organisation après le rendez-vous l'est tout autant. Cela nous aidera pour la rédaction du mémoire.
* Lire les articles du professeur en lien avec ces quatre modèles
* Lire les articles du professeur pour voir comment les deux tableaux ont été réalisés
* S’inspirer des articles du professeur pour les tableaux, graphiques, etc.
* À l’avenir, essayer d’écrire le code sur Google Colab autant que possible, car c’est plus pratique pour nous et plus accessible pour les enseignants.

Les questions posés par prof:

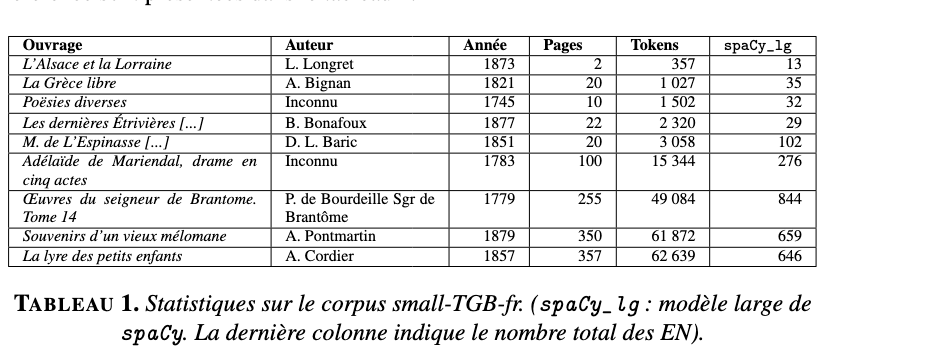
* Pourquoi devons-nous chercher EN ?

Les liens données par les profs:

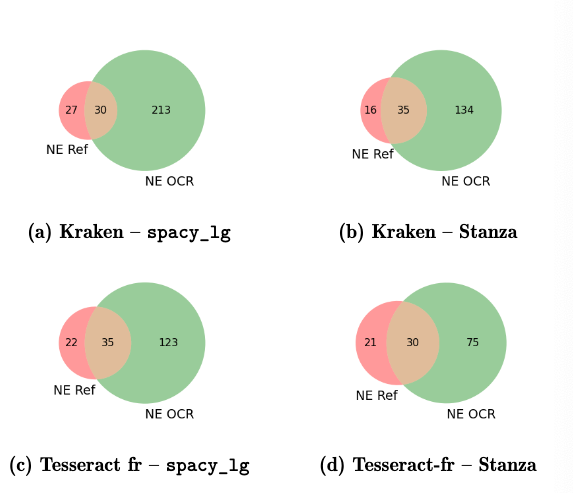
* <https://huggingface.co/Jean-Baptiste/camembert-ner>
* <https://huggingface.co/docs/transformers/model_doc/camembert>

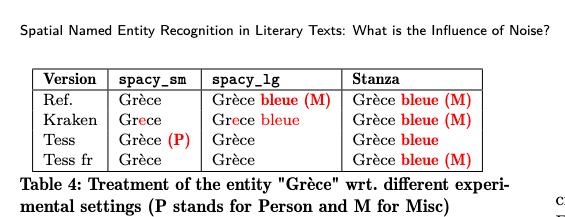
Les deux liens ci-dessus peuvent être utilisés pour explorer CamemBERT.

* <https://www.overleaf.com/read/ygxdrfxztghn#59d511>

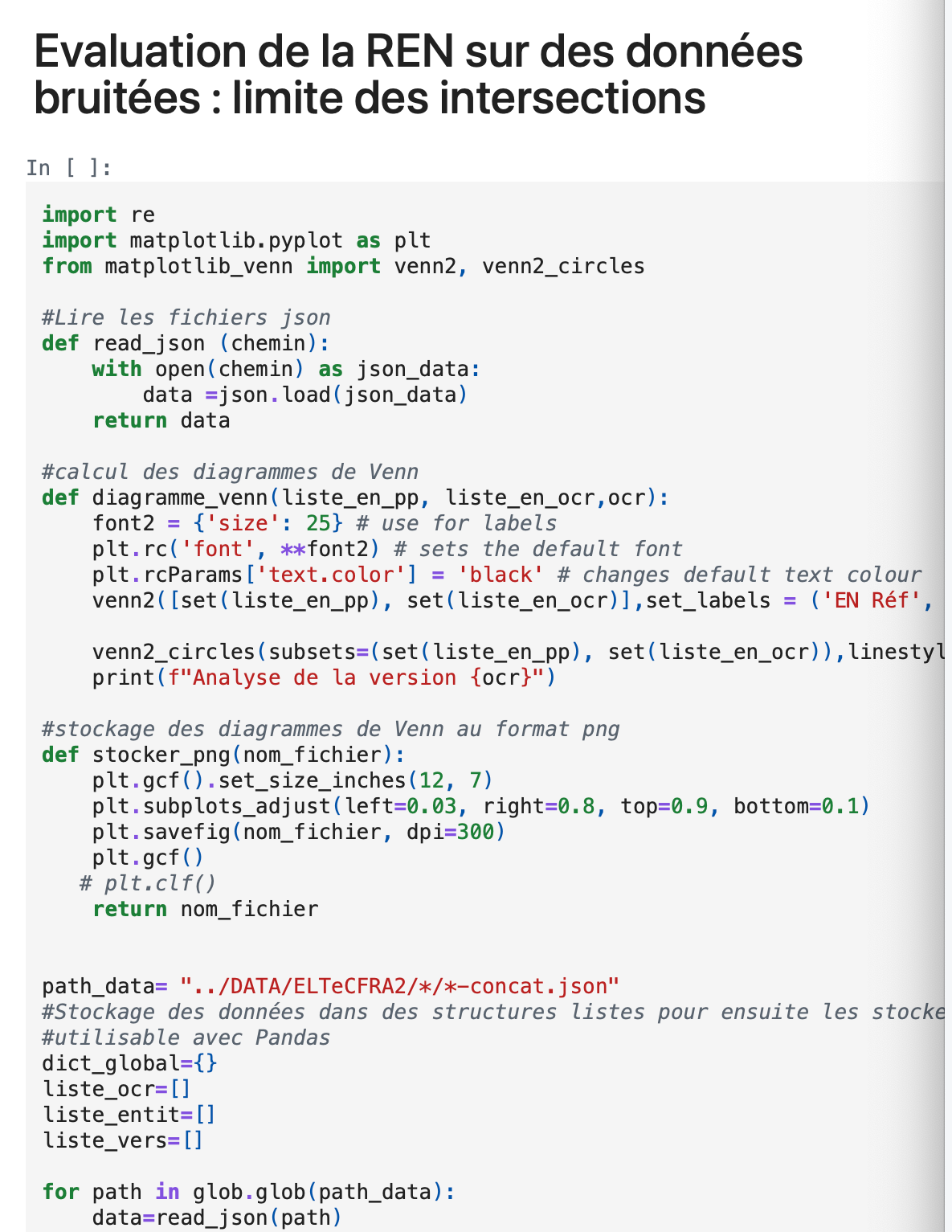


* [https://www.overleaf.com/read/rjnbybvcrvkm#3086d1](https://www.overleaf.com/read/tmkbtqwrfthk#239d62)



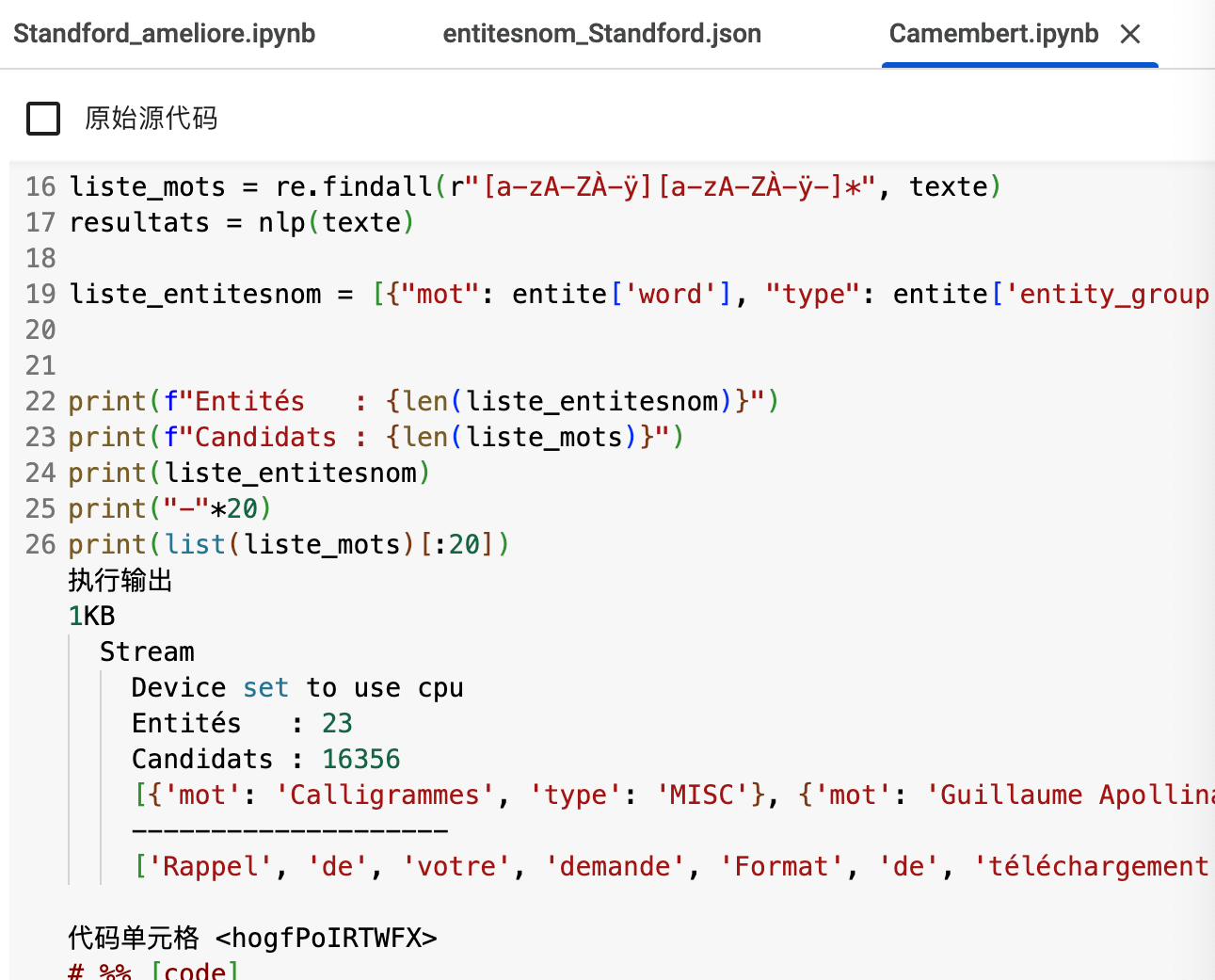


* <https://www.overleaf.com/read/tmkbtqwrfthk#239d62>
* La page GitHub du professeur : <https://github.com/These-SCAI2023/EXPE39_DHENS-2023_13092023/tree/master>
* <https://github.com/These-SCAI2023/EXPE39_DHENS-2023_13092023/blob/master/DATA/ELTeCFRA/ADAM/ADAM_REF/ADAM_Mon-village_REF.txt>
* <https://github.com/These-SCAI2023/EXPE39_DHENS-2023_13092023/blob/master/prog/BnF-datalab-atelier_presentation_.ipynb>



Nos Idées:

* Il faut éviter d'utiliser des commentaires en chinois dans le code autant que possible.
* Quand j'ouvre à nouveau le code que j'ai écrit sur mon drive, je peux le revoir.



* Le code doit absolument être écrit par nous-mêmes afin de pouvoir l'expliquer. Par exemple, nous pouvons utiliser le code et les concepts appris en cours, bien les maîtriser et les intégrer dans notre projet.
* Avant chaque rendez-vous, il faut partir en avance et, avant de sortir ou avant le rendez-vous, envoyer par e-mail au professeur notre travail (code, résultats) ainsi que les notes du précédent rendez-vous. Cela évitera de perdre du temps sur Google Drive pour obtenir les droits d'accès aux fichiers et permettra au professeur de donner directement ses retours.
* Si un code pose problème et que nous ne le résolvons pas nous-mêmes, nous perdrons du temps lors du rendez-vous avec le professeur. Il est donc essentiel d’y consacrer du temps en amont.
* Le code doit toujours avoir une sortie (sortie des résultats).
* Chaque fois qu’un point n’est pas compris, il faut poser des questions au professeur immédiatement, au lieu de perdre du temps à chercher seul.
* Conserver soigneusement tous les articles ou liens fournis par le professeur lors des rendez-vous.
* Le code doit être compris et écrit par nous-mêmes.

Les codes donnés par les prof :

la bibliothèque display pour voir l’exemple

utiliser le tqdm pr enregistrer le temps

Pour le moment:

Le travail donné par les profs pour le troisième rdz:

* **faire un tableau pour montrer quels sont les choses communes des 4 modèles ?**
* dans le code, faire l’explication de l’installation de modèles
* On a la même entrée maintenant, et on modifie le code pr avoir les mêmes résultats en utilisant les 4 modèles
* on peut écrire déjà trois pages de mémoire
* L'annotation manuelle est-elle toujours prévue pour plus tard ? oui
* Continuer à améliorer le code de SpaCy et Flaubert.
* Commencer à rédiger la thèse sur Overleaf, tout en lisant des articles pour mieux comprendre le sujet du papier et les 4 modèles utilisés, ce qui nous aidera aussi à rédiger l'état de l'art.

L’explication de prof sur le sujet :

pq les poèmes et la langue fr?

c’est quoi EN, pq on travaille sur ça?

résumer manuellement les résultats des quatre modèles dans un tableau：

|  | spacy | flaubert | camembert | standford |
| --- | --- | --- | --- | --- |
| entité |  |  |  |  |
| mots |  |  |  |  |
|  |  |  |  |  |

end discussion

INRAD

créer un tableau qui résume les résultats des quatre modèles pour chaque fichier .txt ：

|  | spacy | flaubert | camembert | standford |
| --- | --- | --- | --- | --- |
| HUGO\_Contemplations-T2 |  |  |  |  |
| LOISEAU\_Fleurs-d-avril |  |  |  |  |
|  |  |  |  |  |

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 |  |
| 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).