**Notes du 28/03 (RDZ No.4)**

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

**Pour la mise en commun:**

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

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

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

**Les conseils donnés par prof:**

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

**Les points à clarifier :**

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

**Les codes donnés par prof:**

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

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

**Les liens donnés par prof:**

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

**Pour le moment:**

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

**Tableau:**

Le tableau sur tout le corpus et chaque text doit parler de la densité, la durée d’exécution, du nombre de tokens et du nomdre d’EN pour chaque modèle.

Il faut changer le paramétre pour vérifier si Flaubert est bien appliqué sur tout le corpus

rajouter Spacy-sm et Spacy-lg dans le tableau

**Latex：**

1.Génération automatique de tableaux :

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

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

1.自动生成表格：

你可以用 LaTeX 生成表格，而不是手动填写数据。

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

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

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

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

2.使用 input 命令管理文档：

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

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

**Discussion sur le tableau:**

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

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

**Modèle:**

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

**Mémoire:**

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

**Nos idées：**

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

**Attention:**

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

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

| Date(à 9h) | Lieu | Temps |
| --- | --- | --- |
| 14/03 | sur place (Bureau 211,Maison de recherche) | 1h30mn |
| 28/03 | sur place | 1h30mn |
| 11/04 | sur place |  |
| 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).