**Notes du 28/02 (RDZ No.2)**

Le lien de visio pour notre RDV : <https://rendez-vous.renater.fr/lejeunegael_0bc213-a8a9f4-dd9659>

Les conseils de prof:

1. Après pour pour aller charger trois puces sur Google Colock, vous mettez aussi, du coup là je ne peux pas vous le copier coller, mais sur internet vous le trouverez un import, c'est quoi ces drive montent, quelque chose comme ça. En gros, il faut indiquer à Google Colop que vous allez monter des journées depuis le. Et il faudra accepter toutes les conditions de Google parce que sinon vous ne pourrez pas, vous ne pourrez pas les tous les, enfin Google collabore, vous ne pourrez pas accéder à vos données dans. Google quoi enfin de de votre compte ça va ça va ça va faire la déconnexion donc il faudra chaque fois ré-accepter les conditions. Faites un partage d'écran.

然后，要在 Google Colab 上加载三个项目，您也需要这样做。不过我现在无法复制粘贴给您，但您可以在网上找到一个‘import’，类似于‘drive mount’之类的东西。基本上，您需要告诉 Google Colab，您将从 Google Drive 挂载文件。而且，您必须接受 Google 的所有条款，否则您将无法访问您的数据。Google Colab 也无法访问您的 Google 账户数据，这会导致自动注销，所以每次都需要重新接受条款。请进行屏幕共享。

1. on ne peut pas être nombreux pour accéder dans overleaf
2. Si vous avez besoin de quelques lignes de code pour dire pour dire que vous voulez aller dans tel dossier, lire tel fichier ou chercher tel chemin dans tel tellement de fichiers dans tel dossier puis que vous allez lire enfin que la machine va lire le texte qui est dans ce chemin voilà donc oui il faut quelques lignes de code pour enfin je suis pas sûr d'avoir pu rester sûr mais il faut que dans votre programme vous ayez des lignes de code pour aller lire les fichiers point c x t du corpus. Et comme vous allez passer par Google Colv, enfin si vous passez par Google Colv, il faut mettre les les lignes que le

如果您需要一些代码来指定要进入某个文件夹、读取某个文件，或者搜索某个路径中包含的多个文件，并且机器需要读取该路径中的文本，那么是的，您确实需要一些代码。我不确定是否表达清楚，但在您的程序中，必须有代码来读取语料库中的 .txt 文件。而且，如果您是通过 Google Colab 运行的话，就需要添加相应的代码……

4)Vous avez plusieurs choses à mettre dans votre programme. Vous avez déjà des lignes pour dire que vous définissez le choix vers votre fichier, la ligne pour dire que vous allez lire le fichier et ensuite les lignes pour dire que vous allez faire la reconnaissance d'entités données sur votre fichier. Puis comme l'a souligné monsieur Lejeune tout à l'heure, il faut stocker les données quelque part donc ça veut dire aussi des lignes de code pour indiquer ce que vous ce que vous stockez ou dans quelle forme

您的程序中需要包含多个部分。首先，您需要一行代码来定义文件路径，然后一行代码用于读取该文件，接着是执行命名实体识别（NER）分析的代码。然后，正如 Lejeune 先生 刚才所强调的，数据需要存储在某个地方，因此还需要代码来指定存储内容及其格式。

1. Ensuite je me suis permis de commenter votre Google doc là parce que je vois que pour la réunion la dernière fois que vous aviez pris notre travail, faire des annotations manuelles et automatiques, donc des annotations manuelles, parce que j'ai trop trop souligné, des annotations manuelles non, ce n'est pas du tout ça dans un premier temps, ça c'était une perspective pour l'évaluation. Dans un premier temps, c'est déjà faire tourner le le code et faire des annotations automatiques, d'accord Parce que je vois que vous l'avez écrit en premier, donc je pense que c'est là où il y a une incompréhension entre nous. C'est que moi je vous avais dans un premier temps dit voilà on fait la notation automatique et caetera et pour l'évaluation on pourra on pourra annoter manuellement des des sorties. Et donc pour dans un premier temps, il y a vraiment la notation automatique.

然后，我在您的 Google 文档上做了一些评论，因为我注意到，在上次会议时，您记录了我们的工作，包括手动和自动注释。然而，手动注释并不是当前阶段的重点，那只是评估时的一个展望。

当前的首要任务是运行代码并进行\*\*自动注释\*\*，明白吗？我看到您在文档中首先提到了手动注释，所以我认为这可能导致了我们的某些误解。最初的计划是\*\*先进行自动注释\*\*，然后在评估阶段，我们才可以对输出结果进行\*\*手动注释\*\*。因此，在第一阶段，主要任务是自动注释。

1. Et surtout commencer par des choses simples. C'est pour ça que j'insiste sur ça, je suis chiant. Commencez par des choses simples parce que si vous partez sur des trucs et alors des choses fonctionnent pas, on ne va pas avancer. On réussit à faire fonctionner ou pas les quatre modèles, le premier, le deuxième, le premier, le premier, le premier, le quatrième. C'est très bien d'avoir à partir de l'entreprise de comment tester de l'école et le but après ça va être de les comprendre vous allez voir que les formats sont pas tout à fait les mêmes après on va pouvoir commencer à discuter des choses et d'abord on va pouvoir faire fonctionner tout seul changer l'entrée et c'est

最重要的是从简单的事情开始。这就是为什么我一直在强调这一点，可能会显得有点烦人。但请先做简单的事情，因为如果一开始就处理复杂的内容，而它们又无法正常运行，我们就会停滞不前。我们需要先确保四个模型是否能运行，分别是第一个、第二个、第一个、呃，第一个，还有第四个（可能是口误或强调第一个模型的重要性）。能够测试这些模型是很好的，无论是在企业环境还是在学校环境中，关键是理解它们。您会发现，不同模型的格式不完全相同，接下来我们才会开始讨论这些差异。首先，我们要确保它们能够单独运行，然后调整输入，看看它们的表现如何……

1. Et donc n'hésitez pas à nous envoyer des messages pour poser des questions sans attendre la prochaine réunion qui donc aura lieu le vendredi quatorze à neuf heures, on est tous d'accord là-dessus, c'était quatorze mars à neuf heures. Moi Oui.

下一次会议定在 星期五，3月14 号，上午 9 点

7）Surtout si vous avez des sites comme ça,

des caravines assez importantes, on ne peut pas trouver des solutions. Les solutions sont simples, des fois elles sont marquées. Surtout ne perdez pas de temps sur ces trucs là. Vous testez, vous prenez bien la documentation que vous avez devant vous.

尤其是当你们遇到这样的问题时，

如果问题比较复杂，我们可能找不到解决方案。其实解决方案往往很简单，有时候就在文档里清楚地写着。所以千万不要在这些问题上浪费时间，请先测试，充分利用你们手头的文档。

1. Parfois les gens prennent les choses dans les tutoriels en disant, tiens, c'est super simple, etc. En fait, ce n'est pas exactement le cas à vous. Mais en gros, un truc de reconnaissance entité donnée, ce n'est rien de plus simple que quelque chose qui prend une chaîne de caractères random et qui vous donne un petit scénario qui t'envoie.

有时候，人们会按照教程的方式去做，以为很简单，其实情况并不完全一样。

但基本上，命名实体识别（NER）其实并不复杂，它只是一个处理随机字符串的工具，并返回一个带有标注的小场景。

1. Donc voilà, aucune question de question. Mais d'abord, je teste sur une chaîne de caractères random. Vous avez vu, sur le code que vous envoyez, c'est avec un texte, mais complètement vidéo. Pas captif. Et je crois que pour le prochain rendez-vous, avant le rendez-vous, on va voir en maillet par mail, avec des liens ou bien des documents nécessaires pour ne pas perdre de leçon. Oui, si vous avez des choses dans le document, c'est bien, vous avez créé le journaliste, c'est tout. vraiment l'entrée la sortie des arrières vous aurez une séance sur le spacy

所以，没有特别的问题。但首先，我建议先在随机字符串上进行测试。

你们应该也看到了，在你们发送的代码里，文本是完全空的，并不是一个具体的示例。

我认为，在下次会议之前，我们应该通过邮件提前沟通，发送必要的链接或文档，这样就不会遗漏任何关键信息。

是的，如果你们在文档里已经记录了一些内容，那很好，你们已经创建了一个清晰的记录，这就足够了。

真正的重点是理解输入和输出的逻辑。之后，你们还会有一场关于spacy的会议。

Les liens données par les profs:

* <https://huggingface.co/Jean-Baptiste/camembert-ner>
* <https://huggingface.co/docs/transformers/model_doc/camembert>
* <https://stackoverflow.com/questions/48376580/how-to-read-data-in-google-colab-from-my-google-drive>
* <https://docs.google.com/document/d/1Dx9f14aHbNwJ4rDPU-n5lUsvFpbJ5gYqj9MefSqKBtQ/edit?tab=t.0>

Les codes donnés par les prof :

Code 1:

from google.colab import drive drive.mount('/content/drive')

import spacy !python3 -m spacy download fr\_core\_news\_sm nlp = spacy.load('fr\_core\_news\_sm') doc = nlp("Je teste spacy à Paris") for word in doc.ents: print(word.text, word.label\_)

Code 2:

import spacy

!python3 -m spacy download fr\_core\_news\_sm

nlp = spacy.load('fr\_core\_news\_sm')

doc = nlp("Je teste spacy à Paris")

for word in doc.ents:

print(word.text, word.label\_)

from transformers import AutoTokenizer, AutoModelForTokenClassification

from transformers import pipeline

tokenizer = AutoTokenizer.from\_pretrained("Jean-Baptiste/camembert-ner")

model = AutoModelForTokenClassification.from\_pretrained("Jean-Baptiste/camembert-ner")

nlp = pipeline('ner', model=model, tokenizer=tokenizer, aggregation\_strategy="simple")

nlp("Je teste spacy à Paris")

Pour le moment:

- faire l’annotatiojn automatique et faire tourner le code (commencer et avancer avec les choses simples )

-- faire tourner le code pr chaque modèle,tester le code avec le corpus de poèmes et sauvgarder le résultat

-- tester le même code avec un corpus

- l’évaluation sera après= annotation manuelle

Le travail donné par les profs pour le deuxième rdz:

- familiarisez vous avec les données

- Créez vous un template latex de mémoire et créez le sur overleaf

- créez un document collaboratif où vous noter vos idées, les choses à faire et les compte rendus de réunion

- Il faut commencer à appliquer spaCy au corpus.

- Vous pouvez déjà vous entraîner en faisant tourner votre programme sur les textes présents dans le corpus.

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

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

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

**Rappels du professeur avant le début du projet**

| 1. Gaël Lejeune   Site web: [http://www.lejeunegael.fr](http://www.lejeunegael.fr/) | C'est un sujet qui nécessitera de la programmation de votre part  Une sélection des termes importants :  - Les **systèmes de Reconnaissance d’Entités Nommées**[...] semblent très sensibles à la variation des genres traités (presse, littérature…) [...](Lejeune 2023). - **modèles contextuels** (Luoma & Pysalo 2020). - On parle de **True casing**, Lita 2003)  - l'objectif est d’**évaluer la performance** de systèmes état de l’art de REN - mesurer à quel point la **recherche d’un meilleur découpage en phrase** et d’une **casse plus fidèle** [...] contribuerait à améliorer les performances.  Séminaire "Variation Linguistique et Linguistique Computationnelle" : http://stih-sorbonne-universite.fr/dokuwiki/doku.php?id=seminaire\_lc |
| --- | --- |

| Mlle.Ljudmila Petkovic | À toutes fins utiles, voici **le glossaire des humanités numériques** conçu par M. Arthur Perret qui pourrait vous être utile  [https://digithum.humanum.fr/ressources/glossaire/](https://digithum.huma-num.fr/ressources/glossaire/) |
| --- | --- |

| Mme. Caroline Koudoro-Parfait | Le contrat de Marie-Sophie Desmarest s'est terminé récemment, il n'est donc plus possible qu'elle vous apporte son aide.  **Il faudrait que vous consultiez la documentation de spaCy pour la REN ：**<https://spacy.io/usage/spacy-101>   et que vous **l'appliquiez au corpus** mis à votre disposition.  Vous trouverez via se lien le corpus, pour le moment il y a 9 recueils, il en maque 1.  [https://github.com/TheseSCAI2023/CORPUS\_POEMES](https://github.com/These-SCAI2023/CORPUS_POEMES)  Il faut **commencer à appliquer spaCy au corpus**.  Vous pouvez déjà **vous entraîner en faisant tourner votre programme sur les textes présents dans le corpus**. Je vous tiens informé quand j'ai ajouté le dernier texte.  Séminaire NER for Historical Documents : <https://ner-for-historical-docs.github.io/> |
| --- | --- |