Synthèse – 11 février 2021
Séance 2 : BaOIA / MoDOAP. Les manuels scolaires et guides de voyage.

I) Présentation des outils

A) Extraction et classification

- **Script 1** : téléchargement des documents nécessaires disponibles sur Gallica. Cet outil permet de récupérer automatiquement les données issues des corpus (guides et manuels) sous la forme d'un fichier JSON. Ce format de fichier est particulièrement manipulable pour les analyses ultérieures à effectuer et permet de conserver de nombreuses informations : les métadonnées des documents (titre, auteur, date de publication, nombre de pages, etc.), l'océrisation complète (les textes). La position des textes est aussi conservée : à la fois la page dans laquelle ils apparaissent ainsi que leur position sur la page. Il sera ensuite possible de rajouter des informations sur ces fichiers JSON et constituent un document de référence. Le script fonctionne avec deux types d'entrées : l'identifiant ARK d'un document ou un classeur (type Excel) avec une liste des liens des documents sur Gallica. Un fichier JSON est généré par document dans un dossier, sur un drive.

- **Script 2** : classification automatique des par genres des textes. Cet outil permet de classer automatiquement des passages de documents textuels par genre. Après avoir annoté des passages manuellement selon des catégories prédéfinies, l'algorithme classe des nouveaux documents qui lui sont donnés automatiquement dans les classes. Deux types de classements ont été expérimentés :
    - le classement binaire : cet extrait appartient-il à cette classe ?
    - le classement multilabel : à quelles classes peut-on identifier ce texte ?

Le corpus de manuels a été annoté avec plusieurs genres pour l'expérimentation : histoire, description, portrait, leçons de choses, exercices, chronologie, architecture, récit, dialogue, paratextes (contenant titres, notes d'édition, etc.) et morale. Les résultats du premier type de classe sont probants et l'identification se fait très bien (la mesure de précision est élevée, ce qui est signe que le modèle prédit avec précision les catégories), par contre, le classement selon plusieurs labels ne se fait pas encore bien.

B) Reconnaissance des entités nommées et géolocalisation

- **Script** 1 : extraction des fichiers de textes bruts océrisés de Gallica. Ce premier outil permet d'extraire l'océrisation des documents disponibles sur Gallica. Il permet d'obtenir des fichiers TXT de texte brut, utiles ensuite pour la reconnaissance d'entités nommées. Il est possible ici aussi de fournir un fichier Excel avec une liste de lien pour récupérer automatiquement les données et se constituer le corpus nécessaire dans un dossier, sur le drive.

- **Script 2** : reconnaissance des entités nommées (NER). Les entités nommées sont les noms de personnes, de lieux, d'organisations, d'évènements, etc. présents dans un texte ou un corpus de documents. Il est utile de les extraire car ils constituent des références utiles pour l'étude des sources : notamment les personnages historiques cités dans les manuels et les lieux cité dans les guides touristiques. L'extraction des entités permet la représentation et la conservation, sous forme d'index pour ensuite effectuer des calculs statistiques utiles. Cet outil permet de reconnaître les entités nommées dans un texte et de les visualiser (ils sont reconnus d'une couleur particulière dans le texte). Il faut charger un modèle préalablement entraîné selon la langue des textes à analyser (ici, le modèle utilisé est le français). Deux fichiers sont ensuite créés : un fichier TXT avec la liste des lieux reconnus et un fichier TXT avec la liste des personnes reconnues.

- **Script 3** : géolocalisation des lieux et cartographie. Cet outil permet de créer des cartes interactives à partir des lieux repérés lors de la reconnaissance des entités nommées. Il suffit de fournir le fichier TXT avec la liste des lieux. Les coordonnées GPS sont reconnues : la latitude et la longitude, ce qui permet à la fois d'identifier précisément le lieu dont il est question et d'éviter toute ambiguïté ainsi que de le placer ensuite sur une carte. Il est possible de créer 3 types de cartes :

- une carte regroupant tous les points de localisation souhaités pour identifier des tendances et visualiser l'ensemble des lieux cités (d'un ouvrage, chapitre, paragraphe)
- une carte regroupant plusieurs points de localisation et triés en catégories choisies comme monuments / lieux de promenade. On peut penser aussi à choisir les lieux d'un guide ou manuel pour les comparer avec d'autres, ou encore d'effectuer des comparaisons entre dates.
- une carte retraçant un parcours chronologique, dans l'ordre d'identification des lieux contenus dans le fichier TXT d'origine (carte en mouvement).

Pour ces trois types de carte, il est possible d'enregistrer les cartes créées au format HTML qui permet de les ouvrir dans n'importe quel navigateur et de conserver le format interactif. De ce format HTML, il est possible via le navigateur d'enregistrer les cartes au format PDF.

C) <u>Textométrie et Topic modeling</u>

- **Script 1** : calculs statistiques d'occurrences, de cooccurrences, identification de termes pivots et des termes qui leurs sont associés. Cet outil permet d'effectuer les premiers calculs statistiques sur le contenu des textes : nombre d'apparition d'un terme et des termes qui lui sont associés.

- **Script 2 :** Outil de similarité entre plusieurs textes : identifier les passages communs entre documents, les reprises identiques ou des extraits inspirés, ressemblants. Cet outil permet d'identifier les circulations qui peuvent avoir lieu entre des ouvrages différents, ou bien des originaux et des rééditions.

- **Script 3** : Topic modeling : identifications des « topics » (thèmes, sujets principaux) d'un texte ou d'un groupe de textes. Il peut être utilisé préalablement à un outil de classification pour identifier les grandes thématiques. Le nombre de thématiques à identifier est encore difficile à

définir, et dépends des documents à analyser. Il permet, sans connaître le contenu des guides ou manuels de définir les grands axes thématiques traversants.

- Script 3 : plongement de mots et visualisation. Cet outil permet de visualiser les termes qui se rapprochant dans un espace vectoriel d'autres termes utilisés. Certaines thématiques sont particulièrement bien reconnues : le travail minier par exemple dans
Le Tour de France par deux enfants (1878, disponible sur Gallica : https://gallica.bnf.fr/ark:/12148/bpt6k373586p?rk=64378;0). De même, sont bien reconnus dans les manuels les leçons de choses. Il pourrait être intéressant de traiter au préalable tous les corpus, de conserver tous les textes vectorisés sous forme de dictionnaire, d'une base de données pour pouvoir ensuite établir facilement des comparaisons avec un nouveau document.

II) Questionnements et besoins spécifiques

- Identification de labels pour la classification qui correspondent aux besoins. Réfléchir en particulier aux distinctions entre paratextes / thématiques, envisager peut-être des classes spécifiques pour chaque corpus. Il faut réfléchir à de nouvelles classes liées à la morale, à la religion, à définir.
- Identifier la volonté d'un personnage d'aller dans un lieu en particulier en fonction du vocabulaire, des temps verbaux utilisés ? Pistes : utiliser l'outil de plongement de mots, s'intéresser aux outils d'analyses morphosyntaxiques (voir avec *spaCy*). Grâce à la visualisation, il est possible d'identifier les thématiques liées à un lieu, un personnage, etc.
- Concernant les manuels destinés aux malentendants : il pourrait être envisagé d'utiliser les outils de détection de similarité pour identifier des reprises de passages d'autres manuels, certaines thématiques pouvant être seulement abordées comme prétextes, elles auraient pu être repris d'autres ouvrages.
- Comment analyser les aspects diachroniques des documents ?
- Extraction des images pour pouvoir effectuer ensuite des études statistiques, pour quantifier la part d'images dans les manuels. Le rapport texte/image doit aussi être étudié, ainsi que la mise en page.

III) Poursuites envisagées

- Récupération des documents qui ne sont pas disponibles et/ou océrisés sur Gallica.
- Création d'ateliers personnalisés pour identifier les besoins pour chaque corpus et tester les outils.
- Identification automatique de « types » de lieux (villes, rues, monuments, établissements, etc.) pour le classement automatique et la création de cartes plus rigoureuses et plus en accord avec les axes d'analyse, pour pouvoir trier les lieux selon ces critères. L'identification peut être réalisée en récupérant automatiquement les informations sur Wikidata.
- Affinage de l'algorithme de classement automatique en fonction de classes choisies.

- Lier les entités nommées de personnes aux données de Wikidata pour la récupération automatique des dates et lieux de naissance et de mort, et faciliter la recherche d'informations.

Prochaine séance : lors de la prochaine séance, nous travaillerons sur l'extraction des images et sur le rapport entre textes et images dans les guides et manuels, ainsi que sur l'étude des mises en pages et choix éditoriaux.

Les scripts qui permettent d'utiliser les outils présentés seront disponibles sur les sites des deux projets :
https://modoap.huma-num.fr/
https://baoia.huma-num.fr/