5-6-Lecture-distante

La composition et l'analyse automatisée d'un corpus peut mener à une quantité importante de données, qu'il devient difficile d'analyser manuellement. Il devient alors nécessaire de faire appel à des outils statistiques ainsi qu'à des outils de visualisation pour procéder à une lecture distante (Moretti, 2013) du corpus. Ces analyses permettent d'identifier les récurrences dans les données, et peuvent ainsi aider à cerner des points à étudier en particulier selon une lecture proche.

Dans ce cours, nous allons employer deux outils accessibles en ligne permettant une lecture distante d'un corpus : Palladio et RAWGraphs

Palladio

http://hdlab.stanford.edu/palladio/

Palladio est une application de visualisation dédiées aux humanités numériques, développée par le laboratoire Humanities + Design (http://hdlab.stanford.edu) de l'université de Stanford. Cette application est accessible sur le web, et ne nécessite donc pas d'être installée sur un ordinateur.

Palladio peut lire des fichiers tabulaires, idéalement au format .csv, ou tout du moins tous fichiers dont les valeurs sont délimitées par un séparateur (virgule, deux-points, tabulation).

Palladio propose plusieurs modules pour étudier les données, et notamment des modules de visualisation permettant une lecture distante du corpus :

- **Data** : un module donnant un aperçu des données
- **Map** : un module pour observer les données au travers d'une carte
- **Graph** : un module pour observer les relations entre données au travers de graphe
- **Table** : un module pour étudier les données sous forme de tableau
- **Gallery** : un module pour afficher les données sous forme de galerie

Cliquez sur le lien suivant, puis cliquez sur `Start` : http://hdlab.stanford.edu/palladio/ .
Pour découvrir l'interface, nous allons employer le jeu de données proposé par l'application. Pour cela, cliquez sur `Try with sample data`.

Data

L'interface `Data` donne un aperçu des données qui ont été chargées, et de manière générale, donne un aperçu du projet.

Le champ `Provide a title to this project` permet de nommer le projet. Cliquez sur ce champ et nommez le projet `DecouvertePalladio`.

L'option `Show details` permet de remplir des informations complémentaires sur le projet, tel que le nom de.s auteur.s, la date et une description du projet.

Il est possible d'importer plusieurs fichiers dans un même projet.

Le module `Data` affiche un tableau différent pour chaque fichier du projet. Par exemple, dans les données de tests, on peut trouver un tableau `People` et un tableau `Places`. Chaque ligne du tableau correspond à une colonne du fichier original. En cliquant sur le nom de la colonne, on peut afficher les données qui correspondent.

Il y a plusieurs types de données possible :

- `Text`
- `Number`
- `Date`
- `Coordinates`
- `URL`

Un point rouge avec le message `Please review` apparaît sur une des dimensions du tableau s'il y a une erreur dans les données (par exemple, un caractère interdit)

Les dates doivent suivre le format suivant : Année-Mois-Jour (ex: 2024-09-11).

Les coordonnées doivent être contenues dans une même colonne, en suivant l'ordre Latitude-Longitude. Les valeurs doivent être séparées par une virgule (ex: 41.95, 12.5). Des outils en ligne, ainsi que des LLMs, peuvent vous aider à obtenir les coordonnées d'un lieu.

Map

Le module `Map` permet d'observer les données géographiques. Pour cela, il faut que celles-ci soit associées à des coordonnées, par exemple dans une colonne `Coordinates` qui contient la latitude et la longitude d'un lieu.

Par défaut, le module `Map` affiche un fond de carte vierge. Pour ajouter des points à la carte, il faut y ajouter des couches de données à l'aide du men sur la droite de l'interface.

Cliquez sur `New Layer`. Dans la fenêtre qui s'ouvre, entrez `Birthplace` dans le `Name`, puis choisissez `Birthplace` dans le champ `Places`. Si vous le souhaitez, changez la couleur dans le champ `Color`, ainsi que la taille des points en cliquant sur `Size point`. Enfin, cliquez sur `Add layer`. Observez les points qui s'ajoutent à la carte, et répétez l'opération en utilisant d'autres colonnes.

Il est également possible de changer le fond de carte. Pour cela, cliquez sur `New layer` puis sur `Tiles`. Choisissez un des éléments proposés (par exemple Terrain) puis cliquez sur `Add layer`.

Graph

Le module `Graph` permet de visualiser les relations entre deux colonnes du jeu de données, par exemple pour visualiser la relation entre des personnes et des lieux.

Cliquez sur le menu à droite de l'interface. Dans la fenêtre qui s'ouvre, sélectionnez `Name` dans le champ `Source` puis `Birthplace` dans le champ `Target`. Cliquez ensuite sur `Show links` ainsi que `Size nodes`. Vous verrez apparaître des points, certains reliés. Répétez l'opération en changeant les valeurs pour `Source` et `Target`. Essayez également d'identifier des relations potentiellement intéressantes dans le jeu de données.

Table

Le module `Table` permet d'observer les données sous forme de tableau. Il faut d'abord choisir un colonne qui servira de référence, avant de pouvoir sélectionner les colonnes que l'on souhaite observer.

Cliquez sur le menu à droite de l'interface. Dans la fenêtre qui s'ouvre, choisissez `Name` dans le champ `Row dimension`. Dans le champ `Dimension` choisissez au moins trois colonnes qui vous semblent pertinentes. Prenez le temps d'observer les données qui apparaissent. Répétez l'opération en sélectionnant d'autres colonnes, et observez les résultats.

Gallery

Le module `Gallery` permet d'observer chaque point du jeu de données sous forme de galerie d'images. Chaque carte de la galerie peut afficher une des informations suivantes :

- **Title** : le titre de la fiche
- **Subtitle** : le sous-titre de la fiche
- **Link** : un lien vers une page Web externe

- **Image URL** : un lien vers une image

L'option `Sort by` permet de trier les fiches en fonction d'une colonne.

Dans le menu à droite, appliquez les valeurs suivants aux champs des fiches :

- **Title** : Name
- **Subtitle** : Gender
- **Link** : Pic
- **Image URL** : Pic (Pic est le seul champ avec une URL)

Répétez l'opération en essayant d'autres valeurs. N'hésitez pas à essayer différents valeurs pour `Sort by`

Facet

Comme dans OpenRefine, le menu `Facet` permet d'observer le corpus selon la perspective d'un point de donnée en particulier. Il est possible de sélectionner plusieurs colonnes dans la même facette. Un champ `Description` permet de donner un titre à la facette.

Allez dans le module `Table` puis cliquez sur l'onglet `Facet` en bas de l'interface. Assurez vous que le champ `Name` est utilisé comme valeur du champ `Row dimension`. Dans le champ `Dimensions` à droite, sélectionnez `Birthplace`. Vous verrez un tableau apparaître à gauche du menu `Facet`, listant différents lieux. Cliquez sur l'un deux pour voir les données dans la table se mettre à jour. Répétez l'opération en ajoutant et en changeant les colonnes sélectionnées comme facettes, et observez les résultats. N'hésitez pas non plus à observer les résultats dans les modules `Map, Graph, Gallery`.

Timeline

Le menu `Timeline` permet de visualiser la distribution des données dans le temps. Plutôt que comme une frise chronologique, les données apparaissent sous forme de diagramme à barres. L'abscisse, déterminée par le champ `Dates`, doit être une valeur de type `Date`, tandis que l'ordonnée, déterminée par le champ `Height`, correspondra au nombre de points de données (par exemple, nombre de lignes dans la table `People` ou la table `Place`).

Dans le champ `Dates` choisissez `Birthdate` puis choisissez `Number of People` dans le champ `Height`. Dans le champ `Group by`, choisissez `Name`, puis observez les résultats. Survolez le graphique avec la souris pour voir à quoi correspondent les couleurs colonnes, déterminées par le champ `Group by`. Vous pouvez aussi sélectionner plusieurs colonnes avec la souris, ce qui mettra à jour les données apparaissant dans le module au dessus.

Timespan

Le menu `Timespan` permet également de visualiser l'aspect temporel des données, en se focalisant sur la durée dans le temps. Ici, les deux axes (en haut et en bas) affichent une date de départ et une date de fin. Le menu `Layout` permet de changer l'aspect de la durée, qui est soit représentée sous forme de barre (`Bars` , `Grouped bars`), soit sous forme de lien entre deux dates (`Parallels`).

Dans le champ `Start date` choisissez `Birthdate` , puis dans le champ `End date` choisissez `Date of Death` . Dans le champ `Label` , choisissez `Name` , et observez les résultats. N'hésitez pas à changer la valeur de `Label` et de `Layout` . Comme pour le menu `Timeline` , il est possible de sélectionner des points du graphique avec la souris, mettant ainsi à jour les données affichées dans le module.

Importer des données

Si Palladio fournit des données tests, il est tout à fait possible de travailler avec ses propres données.

Dans le menu `Create a new project` , copiez-collez le contenu du fichier `People.csv` dans le champ texte, puis cliquez sur `Load` .

Une fois dans le menu `Data` , renommez la table en `People` . Cliquez les points rouges puis sur `Verify special characters` pour corriger les éventuelles erreurs liées à des caractères spéciaux.

Par défaut, Palladio ne permet que de chercher un seul fichier. Il est cependant possible d'associer les données à d'autres tables, par exemple pour associer des mentions de lieux à des coordonnées.

Cliquez sur `Birthplace` puis sur `Add new table` . Copiez-collez ensuite le contenu du fichier `Places.csv` dans le champ texte, et cliquez sur `Load` . Reproduisez la même procédure pour la colonne `Place of death` et `Arrival point` .

Sauvegarder le projet

Le bouton `Download` en haut à droite de l'interface permet de sauvegarder le projet sur le disque. Ce projet est normalement sauvegardé au format JSON, et peut être réimporté plus tard pour reprendre le travail.

Télécharger votre projet en cliquant sur le bouton `Download` .

RAWGraphs

Les cartes et frises chronologiques sont des outils très puissants pour accéder aux données selon différentes perspectives et pour étudier un corpus. Cependant, cela nécessite une préparation importante des données (identifier les coordonnées des lieux, associer des dates aux points de données...). Il est pourtant possible de procéder à une lecture distante efficace du corpus à l'aide d'outil de visualisation simples, tels que des graphiques en barres ou des graphes.

Il existe aujourd'hui de très nombreux outils pour générer de telles visualisations (Excel, Numbers, Adobe...). Pour ce cours, nous allons employer l'outil accessible en ligne RAWGraphs (https://app.rawgraphs.io).

RAWGraphs est un framework de visualisation de données open source conçu dans le but de faciliter la représentation visuelle de données complexes pour tous.

Chargement des données

RAWGraphs permet de fournir les données en les copiant-collant (`Paste your data`)

Cliquez sur le menu `Try our sample data` puis choisissez le jeu de données `Iris flowers`. Ce jeu de donnée décrit les propriétés de sépales et pétales de trois variétés d'Iris.

Sélection d'un type de graphique

Le menu `Choose a chart` de l'interface propose une grande variétés de graphiques. Tous les graphiques ne peuvent pas être employés pour toutes les données. Il faut donc bien choisir le type de visualisation en fonction des données dont on dispose.

Le jeu de donnée `Iris flowers` décrit les propriétés de sépales et pétales de trois variétés d'Iris. Il est donc adapté à des graphiques permettant de visualiser des valeurs numériques selon des catégories (par exemple, la taille des pétales par variété d'iris).

Dans le second menu, choisissez `Bar chart`.

Sélection des données à visualiser

Le menu `Mapping` permet de choisir quelles colonnes du jeu de données employer pour la visualisation. Les colonnes apparaissent sur le côté gauche dans l'onglet `Dimensions`. Ces colonnes peuvent être saisie avec la souris. Les paramètres du graphique apparaissent à droite

dans l'onglet `Chart variables`. Ces paramètres dépendent des types de graphiques sélectionnés.

Par exemple, le graphique `Bar chart` a comme paramètres :

- `Bars` : les valeurs représentée par les barres
- `Size` : la taille des barres
- `Color` : la couleur de chaque barre
- `Series` : génère un graphique séparé pour chaque catégorie différente

Un astérisque rouge apparaît pour chaque paramètre obligatoire. Par exemple ici, seul le paramètre `Bars` est obligatoire. De plus, chaque paramètre indique les types de données acceptés par un symbole :

- `#` : données numériques
- `Aa` : données textuelles
- Symbole de l'horloge : données temporelles

Glissez la colonne `Species` dans la case `Bars`. Glissez ensuite la colonne `Sepal length` dans la case `Size`, puis la colonne `Species` dans la case `Color`.

Le graphique apparaîtra dans le menu `Customize` plus bas. Ce menu propose également plusieurs options pour modifier le graphique (largeur, hauteur, couleurs, légende...). Prenez le temps de tester ces paramètres, ainsi que de produits des graphiques en barres avec d'autres colonnes.

Reproduisez les étapes précédentes, en essayant les types de graphiques marqués `Correlations` ou `Proportions`, tels que :

- Alluvial diagram
- Multi-set bar chart
- Bubble chart
- Contour plot

Exporter les résultats

Le menu `Export` permet d'exporter le graphique, soit sous forme d'image (.svg, .png, .jpg), ainsi qu'au format .rawgraph. Ce dernier format permet de sauvegarder le projet, et de le recharger dans RAWGraphs plus tard.

Après avoir expérimenté avec les paramètres du graphiques à barres, exportez le au format .png.

Visualiser les relations

RAWGraphs propose plusieurs graphiques permettant de visualiser les relations entre points de données, par exemple les relations entre personnages d'un roman.

Dans le menu `Try our sample data`, choisissez le jeu de données `Lannister vs Starck relationships`. Choisissez ensuite un des types de graphiques marqué pour `Networks`, comme :

- Arc diagram
- Chord diagram
- Sankey diagram
  Dans le menu `Mapping`, glissez la colonne `Source` dans le paramètre `Source`, `Target` dans le paramètre `Target` et la colonne `weight` dans le paramètre `Size`. Observez ensuite les graphiques générés, et essayez d'en tirer des conclusions des relations entre personnages.

Visualiser les données temporelles

RAWGraphs propose plusieurs graphiques permettant de visualiser les données temporelles, dont certains à la manière de Palladio.

Dans le menu `Try our sample data`, choisissez le jeu de données `Italians PMs and President`. Choisissez ensuite un des types de graphiques marqué pour `Time Series`, comme :

- Gantt chart
- Bump chart
- Calendar heatmap
- Matrix plot
- Streamgraph
  Dans le menu `Mapping`, glissez la colonne `Start date` dans le paramètre `Start Date`, `End Date` dans le paramètre `End date`, `Politician` dans le paramètre `Groups` et `Role` dans le paramètre `Color`. Observez ensuite les graphiques générés.

Visualiser la hiérarchie des données

RAWGraphs propose plusieurs graphiques permettant de visualiser les données hiérarchiques, par exemple les taxonomies.

Dans le menu `Try our sample data`, choisissez le jeu de données `Orchestras by musical instruments`. Choisissez ensuite un des types de graphiques marqués pour `Hierarchies`, comme :

- Circle packing
- Circular dendogram
- Sunburst diagram
- Treemap
- Treemap (Voronoi)
  Dans le menu `Mapping`, glissez les colonnes `Orchestra Type, Group, Instrument` dans le paramètre `Hierarchy`, la colonne `Number` dans le paramètre `Size`, et la colonne `Instrument` dans les paramètres `Color` et `Label`. Observez ensuite les graphiques générés, ainsi que les catégories qui s'en dégagent.

Ressources

Tutoriel Palladio :

- https://hdlab.stanford.edu/palladio/help/

Autres logiciels de visualisation :

- Gephi : https://gephi.org
- Voyant Tools : https://voyant-tools.org

Bibliographie :

- Lecture distante : Moretti, F. (2013). *Distant reading* (Vol. 93). Verso.