Inspirathèque

Pour vous aider à vous préparer au mieux – que vous soyez novice ou plus expérimenté – cette inspirathèque est créée pour vous.

Contraction des mots « inspiration » et « bibliothèque », l'inspirathèque est une bibliothèque d'inspiration. Elle vous permet de découvrir/redécouvrir des ressources qui pourront vous servir pour préparer vos outils de traitement de la donnée pour le week-end de prototypage qui vous attend.

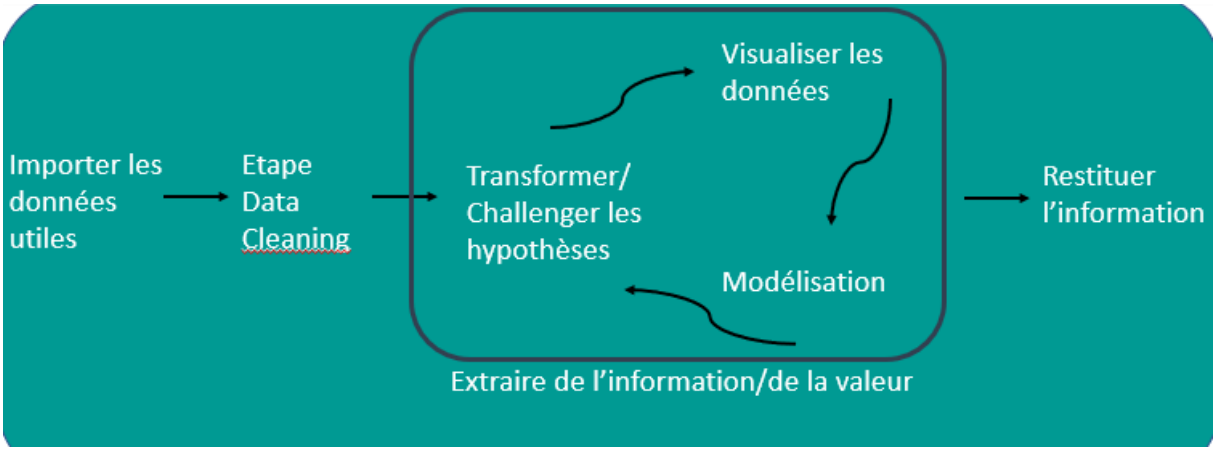
*Les deux langages principalement présentés sont **R** et **Python**, néanmoins **cela n'empêche pas d'utiliser d'autres langages d'analyse lors du hackathon**.*

Les rubriques sont alimentées au fur et à mesure, vous pouvez ainsi venir consulter régulièrement le site.

Les informations réunies ici proviennent de ressources de la communauté du libre. L'intégralité des liens vers les principaux contributeurs sont précisés dans la rubrique « crédits ».

Objectifs de cette rubrique :

- Vous permettre de comprendre les thématiques qu'il faut approfondir pour le cas d'usage que vous choisissez,
- Vous permettre de préparer au plus tôt vos outils pour le week-end intense de prototypage,
- Vous permettre d'orienter vos travaux vers des développements innovants (en ayant en tête un panorama des outils déjà existants évitant ainsi d'explorer des pistes déjà couvertes).



Source : R4DS – Hadley Wickham / Garett Grolemund

Rubrique – Import de données

Langage R

- **Importer des fichiers plats avec le package `utils`**

Le package `utils` est chargé par défaut lorsque vous démarrez R.

Deux fonctions sont utiles pour importer des fichiers csv, `read.csv()` et `read.csv2()` (ces fonctions ont des utilités similaires, néanmoins si vous devez charger des fichiers avec des variables alphanumériques, vous utiliserez la fonction suffixée d'un « 2 »).
Pour charger un fichier .txt, les fonctions `read.delim()` et `read.delim2()` sont utiles (ces fonctions ont des utilités similaires, néanmoins si vous devez charger des fichiers avec des variables alphanumériques, vous utiliserez la fonction suffixée d'un 2).
Pour charger un fichier .txt avec un séparateur autre qu'un espace, la fonction `read.table()` est utile.

Deux packages spécifiques pour charger vos données.

- **`readr`** → la fonction `read_csv()` de ce package permet également d'importer un fichier csv. La table en sortie sera aux formats `tbl_df`, `tibble` ou `dataframe` (`tble_df` et `tibble` étant des formats de tables R optimisés)
- **`data.table`** → ce package se distingue par la rapidité d'exécution – cette fonction gère automatiquement dans un fichier csv la présence d'entête de colonnes ou leur abscence. Il reste toujours la possibilité de spécifier manuellement des options pour gérer le type des variables, le nom des colonnes etc. Il est également possible en utilisant cette fonction de sélectionner des variables par exemple. Les options de cette fonction sont nombreuses et valent la peine d'être investiguées.

- **Importer des données du Web**

Le package **`Rvest`** : permet d'extraire le contenu d'une page web à l'aide d'`Xpath` ou de CSS.

Langage Python

- **Importer des fichiers plats**

`Pandas` : c'est une importante bibliothèque Python pour la manipulation des données, le querelle et l'analyse. Elle fonctionne comme un ensemble intuitif et facile à utiliser pour effectuer des opérations sur n'importe quel type de données. Pour l'installer 'pip install pandas'.

(Javascript Object Notation) : permet de stocker des données textuelles. Json est installée par défaut.

- **Importer des données du Web**

`Selenium` : Lance et contrôle un navigateur web. Sélénium est capable de remplir des formulaires et simuler des clics de souris dans ce navigateur.

`Beautifulsoup` : pour parser du HTML, le format dans lequel les pages Web sont écrites.

`Webbrowser` : il permet de télécharger les données d'une page. Il est installé avec Pyhton.

`Requests` : télécharge des fichiers et des pages Web à partir d'Internet.

Rubrique – Manipuler et mettre en forme vos données

Langage R

La suite `tidyverse` : ensemble de packages pour manipuler vos données sous R – Cours Julien Barnier : https://juba.github.io/tidyverse/index.html

Focus sur le package `dplyr` – un des packages du `tidyverse` http://www.rpubs.com/Marylene/test

Le package `data.table`

Le package `Stringr` : package pour la gestion de caractères.

Quelques références utiles ressources R pour manipuler vos données :

`Dplyr,` Introduction : https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

`Dplyr`, manipulation de deux tables : https://cran.r-project.org/web/packages/dplyr/vignettes/two-table.html

`Tidyr` : https://cran.r-project.org/web/packages/tidyr/tidyr.pdf

Aide mémoire de Rstudio sur `dplyr` et `tidyr` : https://www.rstudio.com/wp-content/uploads/2016/01/data-wrangling-french.pdf

Si vous préférez vous mettre à `data.table` https://s3.amazonaws.com/assets.datacamp.com/img/blog/data+table+cheat+sheet.pdf

Créer un package sous R : https://thinkr.fr/creer-package-r-quelques-minutes/ – ressource ThinkR

Langage Python

Voir site https://hackr.io/

Voir site https://www.pythoncheatsheet.org/

Voir site https://sinxloud.com/fr/python-cheat-sheet-beginner-advanced/

Rubrique – Visualiser ses données (datavisualisation)

Je n'ai aucune compétence en programmation ou en graphisme, y-a-t-il des outils de datavisualisation qui peuvent me servir ?

Le site **`datawrapper`** https://academy.datawrapper.de/ permet à la fois de vous initier aux principes et aux concepts de la datavisualisation mais également de réaliser des visualisations de données. De nombreux tutoriels courts de prise en main existent.

Les logiciels de datavisualisation

- Tableau
- Qlick
- PowerBi

Langage R : visualiser vos données

Package `ggplot2` : initiation à l'utilisation du packages – principaux éléments http://r-toulouse.netlify.com/diapos/2018-09-25_ggplot2_initiation.pdf

Package `esquisse` de DreamRs : pour réaliser des graphiques ggplot2 de manière interactive

Package `shiny` (et son écosystème de packages associés). Pour maitriser les bases de Shiny (langage R) : https://mastering-shiny.org/preface par Hadley Wickham

Package `flexdasboard` : pour créer des tableaux de bord interactifs avec R https://rmarkdown.rstudio.com/flexdashboard/

Package `Golem` ThinkR – un outil pour la construction d'applications Shiny pour la production : https://rtask.thinkr.fr/fr/demarrer-avec-golem/

Ressource RStudio : plusieurs packages de visualisation des données

https://rviews.rstudio.com/2019/10/29/sept-2019-top-40-new-r-packages/?utm_content=buffer7af42&utm_medium=social&utm_source=linkedin&utm_campaign=buffer

Langage Python : visualiser vos données

`Matplotlib` : cette bibliothèque permet de faire de la visualisation.

`Seaborn` : cette bibliothèque fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.

Des « helpers » pour personnaliser vos graphiques R ou Python

Créer des palettes de couleurs utiles pour vos graphiques sous R : https://neocarto.hypotheses.org/1458

Générateur de palette de couleur : https://coolors.co/

Retrouver une couleur (à partir du nom d'une couleur, retrouver son code hex ou RGB) : http://chir.ag/projects/name-that-color/#714693

Choix d'une couleur : https://flatuicolors.com/

Rubrique – Modéliser vos données

Les thématiques importantes en lien avec les cas d'usage de l'idéathon

- **Analyse textuelle**

En python :

JSON (Javascript Object Notation) : permet de stocker des données textuelles. Json est installée par défaut.

String : permet de faire des opérations usuelles sur des chaines de caractères. String est installée par défaut.

NLP (Natural Language Processing)

`NLTK` (Natural Langage Toolkit) : cette bibliothèque permet la création de programme pour l'analyse de texte. Pour l'installée 'pip install NLTK'.

`Gensim` : cette bibliothèque permet d'extraire automatiquement les sujets sémantiques des documents. Les algorithmes dans `Gensim` sont Word2Vec et Fastext. Pour l'installer 'pip install gensim'

`Spacy` : cette bibliothèque permet de construire des systèmes d'extraction d'information ou de compréhension en langage naturel. Elle est conçue pour une utilisation en production et fournit une API concise et conviviale.

- **Les techniques d'océrisation (La reconnaissance optique de caractère (ROC) ou optical character recognition (OCR))**

Package `Tesseract` : c'est bibliothèque OCR parrainée par google. Tesseract est connue pour être le meilleur système de ROC open source et le plus précis qui soit. Pour l'installer 'pip install tesseract'.

Package `pytesseract` : une fois Tesseract installée. Vous serez prêt à installer pytesseract, qui utilise votre installation tesseract existante pour lire les fichiers image, les sorties strings et objets pouvant être utilisés dans les scripts Python. Pour l'installer 'pip install pytesseract'.

- **Datavisualisation**

`Infogram` : pour créer des graphiques, des cartes et des infographies

`Visme` : pour créer des graphes et des infographies

- **UX/UI design**

Pour une liste d'outils UX : https://uxtools.co/

Pour prototypes web et mobile : InVision (permet de hiérarchiser des projets / écrans de manière simple et claire.)

- **Séries temporelles / analyses de données / méthodes supervisées / méthodes non supervisées /deep learning**

Le site https://hackr.io/ vous permet d'accéder à des ressources et programmes en différents langages.

Rubrique - Ressources pour approfondissement

Traitement des données

Livre interactif gratuit « R for DataScience – Hadley Wickham, Garett Grolemund » : https://r4ds.had.co.nz/index.html

Ce livre permet d'apprendre à traiter des données sous R, les importer, les mettre en forme, les visualiser, les modéliser et les restituer.

Accéder aux API de l'Etat

Le site www.api.gouv.fr permet un accès unique aux API de l'Etat. N'hésitez pas à y jeter un œil. Certaines API sont en accès direct et totalement libre, d'autres nécessitent une mise à disposition en contactant leur propriétaire. Si vous êtes d'ores et déjà intéressé par un cas d'usage, vous pouvez contacter le support de certaines API afin d'obtenir l'accès aux outils qui vous intéressent.

Travailler en mode projet sous RStudio

Organiser son projet : https://nicercode.github.io/blog/2013-04-05-projects/

Autres ressources

Etalab lance https://code.etalab.gouv.fr/ , les données sous-jacentes permettent de filtrer sur le langage principal de chaque dépôt : https://www.data.gouv.fr/fr/datasets/inventaire-des-depots-de-code-source-des-organismes-publics/

Par exemple, on trouve 27 projets R produits par l'administration, pour la communauté #rstats française. https://pbs.twimg.com/media/EGbSNTEXkAIMzJq.png

Compléter les données / trouver des données complémentaires pour traiter le cas d'usage que vous choisissez

Rubrique : quels développements/travaux déjà effectués pouvant vous inspirer ?

- Etude sur les contractuels de la fonction publique

  https://www.collectivites-locales.gouv.fr/files/files/statistiques/bis_138_contractuels_fpt_bs2017_1.pdf

- Des travaux conduits en 2017 / 2018 par la Caisse des dépôts, en lien avec les employeurs publics, sur l'usure professionnelle et le maintien dans l'emploi, qui ont donné lieu à un « référentiel » à l'usage des employeurs. (À lire pour vous informer de ces questions RH très spécifiques).

  https://www.cnracl.retraites.fr/employeur/actualites/recueil-du-maintien-dans-lemploi-des-fonctionnaires-territoriaux-et-hospitaliers

Les grandes règles de la retraite d'un agent de la fonction publique (titulaire et non titulaire)

Les informations disponibles à partir de cette page détaillent les principales règles concernant la retraite des fonctionnaires civils et contractuels des 3 fonctions publiques. Ces éléments ne suffisent pas à estimer la date exacte d'un départ à la retraite, ni le montant d'une future pension mais sont d'excellents outils à des fins de prototypage.

https://www.service-public.fr/particuliers/vosdroits/N379