

Algorithmique M2 Data Science Projets

Vincent Runge



Laboratoire de
Mathématiques
et Modélisation
d'Évry



université
PARIS-SACLAY

1 Les règles du jeu

2 Les projets

Section 1

Les règles du jeu

Choix optimal d'un projet

Chaque *groupe de 3 étudiants* choisit un *problème d'algorithmique* à traiter. Pour n groupes et n projets, on demande aux étudiants de noter les projets par préférence (ou "satisfaction") **de 1 à n**

C'est un problème algorithmique !!!

Comment répartir au mieux les projets pour maximiser la satisfaction des étudiants?

Remplir ce tableau avec vos choix par équipe

Travail à réaliser

- **Un rapport au format Rmd** (rendu pdf ou html) (**10 points**)
- **Une présentation beamer / Rmd présentation / ppt** (**20 points**)
- **Un package fonctionnel** (**30 points**) du type **M2algorithmique** avec les éléments suivants :
 - ① Solution naïve R (5 points)
 - ② Solution améliorée R (5 points)
 - ③ Evaluation sur des simulations de la complexité (5 points)
 - ④ Le code en C++ et comparaison en temps avec le code R (10 points)
 - ⑤ Un package fonctionnel et bien documenté sur github (5 points)

ça donnera une note sur 60

Les dates

- projets proposés le 10 décembre
 - groupe et vote de 1 à 7 pour le mardi 14 décembre 14h (vous aurez ainsi 1,5 mois pour le projet)
- ① Rendu du **rapport** le **mercredi 26 janvier au plus tard à 23h59** (à vincent.runge@univ-evry.fr)
 - ② **Présentation de 20 min** devant un jury du LaMME (2 ou 3 chercheurs) le **vendredi 28 janvier de 9h à 12h30** et **5-10 min** de questions.
 - ③ Une **étape de suivi** (au minimum une) la **semaine du 10 au 14 janvier** avec moi ou votre tuteur (s'organiser avec lui)

Section 2

Les projets

Projet 1 : Dijkstra's algorithm

Trouve le plus court chemin entre deux noeuds d'un graphe

- ① Algorithme naïf à implémenter : l'algorithme de Bellman-Ford ou une méthode heuristique (algo glouton)
- ② Présenter rapidement des exemples d'application de l'algorithme de Dijkstra

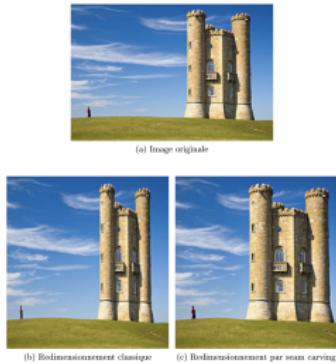
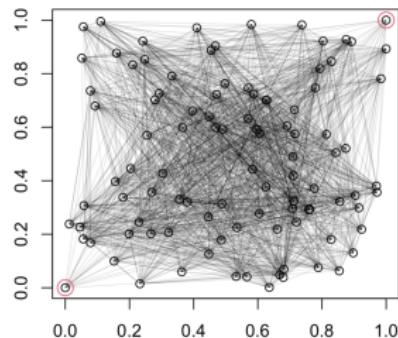
Références :

Un référence pour l'algorithme de Dijkstra

https://en.wikipedia.org/wiki/Seam_carving

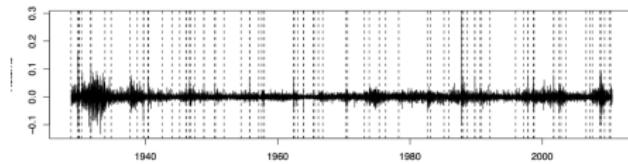
Projet 1 : Dijkstra's algorithm #2

- ① Avec les points du carré $[0, 1] \times [0, 1]$ étudier le nombre de points à traverser et la longueur du chemin (à comparer à $\sqrt{2}$) en fonction du nombre de points et d'arrêtes.
- ② Etudier le problème de la réduction d'image avec cet algorithme



Référent : (doctorant) Kylliann De Santiago
kylliann.desantiago@hotmail.fr

Projet 2 : Variation de variance (volatilité) dans les séries temporelles



Ecrire l'algorithme dit d'*optimal partitioning* (type PELT) qui est un algorithme de programmation dynamique. Il réduit la complexité de $O(2^{n-1})$ à $O(n^2)$ et celui de segmentation binaire de complexité $O(n \log n)$.

La référence

Ce problème a de nombreuses applications à la **finance**. Trouver des données réelles à analyser et comparer le résultat à un algorithme standard de votre choix.

Projet 2 : Variation de variance (volatilité) dans les séries temporelles #2

Références possibles :

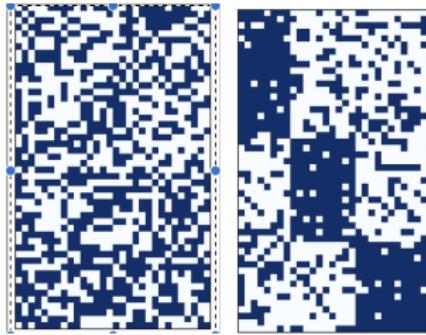
<https://arxiv.org/pdf/1709.03105.pdf>

<https://www.sciencedirect.com/science/article/pii/S0029801810001162>

Référent : (doctorant) Arnaud Liehrmann
arnaud.liehrmann@universite-paris-saclay.fr

Projet 3 : *Biclustering Algorithm*

Les matrices binaires de très grande dimension sont très présentes en **bioinformatique**. Le biclustering consiste à permuter les lignes et les colonnes d'une matrice pour faire apparaître des groupes aux caractéristiques homogènes.



Les algorithmes existants sont cependant assez lents, une approche diviser pour régner existe. L'objectif est de la coder puis de la tester et de la comparer aux méthodes existantes.

Projet 3 : *Biclustering Algorithm #2*

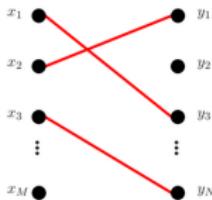
Références :

l'article qui présente l'algo

autres sources: [article1](#) et [article2](#)

Remarque : nous disposons de données de génomique au LaMME pour tester l'algo sur des données réelles

Projet 4 : *Student-project allocation*



Etudier le problème de l'assignation optimale de n étudiants à n projets
(comme présenté en cours)

- ① Trois algorithmes sont à faire : le naïf qui explore toutes les solutions, l'algo de programmation dynamique, l'algorithme génétique
- ② Evaluer la qualité du résultat et le temps de calcul de ces algorithmes
- ③ Présenter une généralisation possible du problème (avec les équations) et une méthode de résolution possible. **Un point de départ**

Projet 5 : *NLP Algorithm*

Il est possible de détection le plagiat par un algorithme de **programmation dynamique**

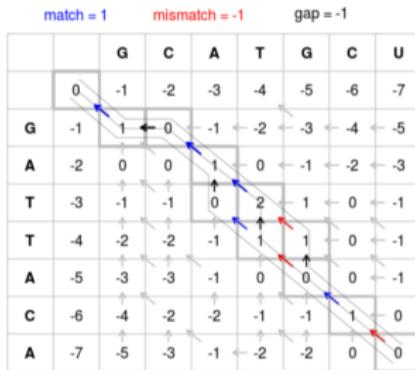
Text 1 : "The First sentence is about Python. The Second: about Django. You can learn Python,Django and Data Ananlysis here."	Text 2 : "The Initial sentence is about Python. The Second: about Flask. You can learn Python,Django and Data manipulation here."
---	---

- ① avec le package `rvest`, obtenir les textes candidats au plagiat (par *web scraping* sur google)
- ② valider les candidats par analyse avec un algorithme de programmation dynamique.

Références: [article1](#) et [article2](#)

Projet 5 : NLP Algorithm #2

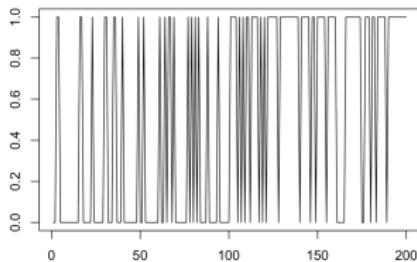
Ce problème s'applique aussi à l'alignement de séquence ADN en bioinformatique par l' algorithme de Needleman-Wunsch



Voir aussi la page wikipédia [sequence alignement](#)

Projet 6 : *change detection in frequencies*

On observe une série de 0 et de 1 (une série binaire) avec un possible changement de probabilité de la loi binomiale dans la série. L'objectif est de détecter **une fraude**



- ① Estimer la probabilité de détecter la fraude correctement pour 2000 données et un point de fraude à 100, 200, 300, ..., 1900 avec différentes valeurs de p_0 et p_1 .

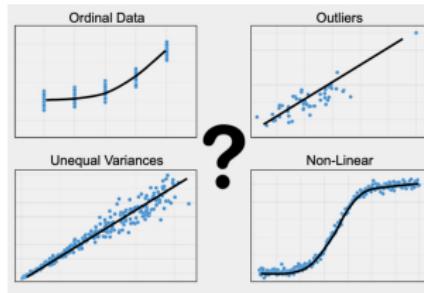
Projet 6 : *change detection in frequencies #2*

- ② Explorer les mesures de score (à partir de la matrice de confusion)
- ③ Que peut-on faire lorsque le nombre de changement de fréquence est inconnu?

On utilisera mclapply et le serveur de calcul du LaMME.

Projet 7 : Mesure de corrélation : le tau de Kendall

Dans les très grandes bases de données, $n > 10^9$ il devient impossible d'évaluer la corrélation par le Tau de Kendall qui est de complexité $O(n^2)$. Pourtant cette mesure de corrélation non linéaire est très utilisée en finance ou bioinformatique



Projet 7 : Mesure de corrélation : le tau de Kendall #2

- ① Implémenter les méthodes naïves en $O(n^2)$ et améliorée en $O(n \log(n))$
- ② On comparera le résultat de ces méthodes à une estimation par *bootstrap* (re-sampling) sur des exemples (données simulées et réelles). On pourra calculer la valeur du tau théorique sur un modèle de simulation simple

Quelques références : [article1](#) et [article2](#)