## Journal de suivi

Contenu

[Journal de suivi 1](#_Toc500781321)

[Jour 1 (01/11/17) - Durée travail env. 8h : 3](#_Toc500781322)

[Jour 2 (02/11/17) – Durée travail env. 13h: 3](#_Toc500781323)

[Jour 3 (03/11/17) – Durée travail env. 13h: 4](#_Toc500781324)

[Jour 4 (04/11/17) – Durée travail env. 10h: 4](#_Toc500781325)

[Jour 5 (05/11/17) – Durée travail env. 2h: 5](#_Toc500781326)

[Jour 6 (06/11/17) – Durée travail env. 4h: 5](#_Toc500781327)

[Jour 7 (07/11/17) – Durée travail env. 3-4h: 5](#_Toc500781328)

[Jour 8 (08/11/17) – Durée travail < 2h: 5](#_Toc500781329)

[Jour 9 (09/11/17) – Durée travail 4h: 5](#_Toc500781330)

[Jour 10 (10/11/17) – Durée travail 6h: 6](#_Toc500781331)

[Jour 11 (11/11/17) – Durée travail 6h: 6](#_Toc500781332)

[Jour 12 (12/11/17) – Durée travail 8h: 6](#_Toc500781333)

[Jour 13 (13/11/17) – Durée travail 3h: 6](#_Toc500781334)

[Jour 14 (14/11/17) – Durée travail 2h: 7](#_Toc500781335)

[Jour 15 (15/11/17) – Durée travail 3h: 7](#_Toc500781336)

[Jour 16 (16/11/17) – Durée travail 2h30: 7](#_Toc500781337)

[Jour 17 (17/11/17) – Durée travail 3h: 7](#_Toc500781338)

[Jour 18 (18/11/17) – Durée travail 9h: 7](#_Toc500781339)

[Jour 19 (19/11/17) – Durée travail 6h: 8](#_Toc500781340)

[Jour 20 (20/11/17) – Durée travail 1h: 8](#_Toc500781341)

[Jour 21 (21/11/17) – Durée travail 3h: 8](#_Toc500781342)

[Jour 22 (22/11/17) – Durée travail 2h: 8](#_Toc500781343)

[Jour 23 (23/11/17) – Durée travail 2h: 9](#_Toc500781344)

[Jour 24 (24/11/17) – Durée travail 2h: 9](#_Toc500781345)

[Jour 25 (25/11/17) – Durée travail 7h: 9](#_Toc500781346)

[Jour 26 (26/11/17) – Durée travail 8h: 9](#_Toc500781347)

[Jour 27 (27/11/17) – Durée travail 4h: 10](#_Toc500781348)

[Jour 28 (28/11/17) – Durée travail 4h: 10](#_Toc500781349)

[Jour 29 (29/11/17) – Durée travail 4h: 10](#_Toc500781350)

[Jour 30 (30/11/17) – Durée travail 3h30: 11](#_Toc500781351)

[Jour 31 (01/12/17) – Durée travail 2h: 11](#_Toc500781352)

[Jour 32 (02/12/17) – Durée travail 10h30: 11](#_Toc500781353)

[Jour 33 (03/12/17) – Durée travail 8h: 11](#_Toc500781354)

[Jour 34 (04/12/17) – Durée travail 2h: 12](#_Toc500781355)

[Jour 35 (05/12/17) – Durée travail 2h: 12](#_Toc500781356)

[Jour 36 (06/12/17) – Durée travail 3h: 12](#_Toc500781357)

[Jour 37 (07/12/17) – Durée travail 2h: 12](#_Toc500781358)

[Jour 38 (08/12/17) – Durée travail 3h: 12](#_Toc500781359)

[Jour 39 (09/12/17) – Durée travail 8h: 13](#_Toc500781360)

[Jour 40 (10/12/17) – Durée travail 8h: 13](#_Toc500781361)

[Jour 41 (11/12/17) – Durée travail 3h: 13](#_Toc500781362)

[Temps de travail : 200h 13](#_Toc500781363)

### Jour 1 (01/11/17) - Durée travail env. 8h :

1er jour de formation sur le cours de Data Analyst. Celui-ci commence par un premier cours pour apprendre à apprendre. N’ayant jamais « arrêté » d’apprendre de manière autodidacte, celui-ci s’est plutôt bien passé. Un peu de temps pour mettre à jour mes CVs (j’ai 2 versions différentes), une petite demi-heure pour la mise à jour de mon profil [Linkedin](https://www.linkedin.com/in/nicolas-m-b5a66174/) incluant le début de la formation mais aussi l’ajout des 19 certificats déjà faits… En effet, j’ai commencé les cours depuis pas mal de temps sur OpenClassrooms mais j’ai décidé de faire les Quiz/Activités que récemment en ayant pris goût à l’apprentissage de la Data Science. Le passage en Premium m’a donc permis d’avoir accès au Certificats.

J’ai aussi rapidement mis à jour mon [site personnel](http://www.nicolasmine.com/) (sur lequel j’ai mis mon CV ainsi que différents scripts fait avec des Notebooks, des news et divers liens concernant l’apprentissage du langage Python ou du Machine Learning) afin de mentionner le début de la formation sur celui-ci aussi !

Finalement, sachant bien pourquoi je me lançais dans cette formation, les objectifs d’apprentissages ont aussi été simples à mettre en place. Je passe donc au projet n°1 et là, 1ère petite frustration… En effet, je commence après le Projet 1 celui-ci qui, comme je m’en doutais un peu, s’oriente sur le nettoyage de données. J’avais déjà eu affaire à du traitement de dataset pas clean mais généralement c’est très léger (ajouter/supprimer des lignes sans valeurs, Encoder/OneHotEncoder les Labels) mais là je me rends compte que c’est un tout autre chantier… Un dataset de 800mo, 320 000 lignes et 162 colonnes. Le tout sans trop savoir ce que l’on cherche. Bref, il faut prendre le temps de comprendre chaque donnée uni variée et de les nettoyer.

### Jour 2 (02/11/17) – Durée travail env. 13h:

Etant en congé 2 jours, je me réveille tôt pour avancer (tout en sachant que la tâche sera longue et fastidieuse). Le data cleaning avance par à-coup. Notamment lors de colonne complexe à traiter (par exemple les traces, additifs, qui sont des concaténations de données). Cependant cela avance. A la fin du 1er nettoyage il reste 84 features mais je décide de commencer en parallèle l’exploration en espérant découvrir d’autres possibilités de nettoyage. Celle-ci-se solde par une réduction importante de dimensions ainsi que d'autres possibilités de Nettoyages de Données (notamment pour les données /100g. Certaines ont des valeurs négatives, d'autres supérieur à 100 ou même extrêmement grande pour ce type de données, par exemple 30g de vitamines).

Commençant à voir un nettoyage correct ainsi que certains patterns arriver, je commence aussi à bloquer sur la réduction de dimensions ainsi que l'objectif final de l'exercice un peu flou. Je décide de passer au projet 3 en attendant de discuter des blocages avec mon mentor quand je l'aurais en rendez-vous.

Le Projet n°3 me semble aussi assez sombre à première vue car on doit effectuer une Clustering mais sans avoir de labels particuliers. Du coup j'ai du mal à voir comment j'évaluerais par la suite les modèles. Mais bon il faut commencer par le commencement, place au nettoyage du dataset.

Le dataset étant beaucoup plus petit, celui-ci va aussi beaucoup plus vite (connaissant aussi certaines commandes pandas grâce au précédent nettoyage). Afin de réduire la taille du notebook, j'opte pour un générateur qui me permet de parcourir chaque colonne et d'effectuer la même analyse sur chaque feature. La majorité sont très simples, il faut juste remplir les valeurs manquantes par des données logique (dans certains cas 0, la moyenne ou couleur dans le cas de la couleur du film). Certaines sont plus complexe à traiter, c'est le cas du type de film, des mots clé et des acteurs. Je décide d'abandonner les mots clé ayant les types de film et cela ferait un peu doublons. La fatigue commençant à se faire sentir, je décide de laisser faire le traitement de ces colonnes le lendemain.

### Jour 3 (03/11/17) – Durée travail env. 13h:

Pour les 2 colonnes à problème, je décide dans un premier temps d'utiliser le LabelEncoder. L'objectif étant d'avoir "un premier jet" pour tester les modèles. Je l'applique donc sur chaque colonne de type texte et teste le Kmeans, PCA et TSNE. Peu importe le nombre de cluster, le résultat du kmeans ne me semble pas terrible visuellement. Si je regarde la liste des films pour chaque cluster, je trouve difficilement du sens au classement. Quant au PCA, si je réduis trop les dimensions, je n'explique qu'environ 30% de la variance, le résultat est donc d'autant plus mauvais. Je teste donc le TSNE. Je regarde par distance euclidienne les points les plus proches du film choisi et je commence à voir une corrélation (par exemple si je visualise un des Pirate des Caraïbes, il me propose de voir une partie des autres). Je décide donc d'améliorer mon nettoyage de données. Pour ce faire, je décide de compter les occurrences de chaque type de film ainsi que de chaque acteur et les encoder par leur nombre d'occurrence. Mon objectif étant d'avoir les acteurs qui jouent beaucoup assez proche. En effet le LabelEncoder pour donner le Label 1 à Johnny Depp et 55 à Orlando Bloom. De ce fait la distance entre ces 2 acteurs sera grand et donc illogique. Par contre tous l'acteur ayant joué dans 1 seul films, auront le même label et seront donc proche (cela pour permettre au fan de film peu connu d'avoir des choix de film peu connus aussi). J'applique la même méthode sur le type de film.

Là les résultats sont meilleurs, le PCA peut expliquer 60% de la variance en 10 dimensions, le K-means commence à classer les films par catégorie mais ses résultats ne sont pas fameux. Quant au TSNE, il réussit très bien. Je reste donc sur ce modèle, génère un dataset de production avec juste les données utilise et leurs positions afin de faire le site.

Ayant déjà codé des sites, j'ai juste à reprendre le cours de Flask sur OC et cela avance plutôt bien. J'ai fait une petite page d'accueil ou l'on peut entrer l'ID du film ou une liste déroulante avec tous les titres dans l'ordre, lors de la validation, celui-ci affiche 5 autres film similaire avec diverses info (date de sortie, acteurs, etc...). Je m'arrête donc là pour aujourd'hui.

Au passage, j'ai contacté ce jour mon mentor pour avoir un premier rendez-vous.

### Jour 4 (04/11/17) – Durée travail env. 10h:

Etant en attentes d'informations avec mon tuteur sur le projet n°2 et le projet n°3 étant démarré, je commence plus doucement le projet 4. Celui-ci s'appuie sur 2.8 go de données assez cleané avec pour objectif de mettre en place un modèle linéaire. Les données sur ce dataset sont propres, il ne manque pas de points, il n'y a pas outliers donc on passe assez rapidement à de l'exploration. Avant cela il faut juste se débarrasser apriori des features inutiles/ imprévisible. Pour certaines cela est évident à la description de la feature sur le site pour d'autres, cela a été décidé basé sur l'exploration. Au départ tout s'est fait sur un seul moins pour accélérer le processus et une fois fini, un script propre a été fait pour fusionner les dataset et faire le cleaning d'un coup. A cause de bug sur le dataset d'avril un peu de temps a été perdu lors de l'exploration mais globalement tout s'est bien passé. Le passage au modèle quant à lui pose beaucoup plus de soucis.

En effet, si on veut pouvoir prédire le retard, on a besoin des données prédictible (date, aéroport de départ, aéroport d'arrivé, compagnie, …) le problème c'est que le retard n'est aucunement linéaire avec ces features (il y a plus de retard en journée, plus le lundi et vendredi que le reste de la semaine, plus en été que le reste de l'année et variable en fonction des aéroports). Une solution connue pour ce problème est de faire un One Hot Encoder sur celle-ci. Bien que l'on ait 5m de points de données, on a peu de risque du fléau de la dimensionnalité mais le problème c'est que cela ne passe pas en mémoire. Je vais donc devoir en discuter avec mon mentor car je sèche.

Je profite donc de ce blocage pour commencer la veille thématique. Dans 1 premier temps, je cherche un sujet. Je pensais parler des Enpsulated networks ou du Synthetic Gradient mais il sera difficile de faire 3 exemples dessus. Ne voulant pas parcourir des sujets déjà vu dans les cours, je décide de partir sur les Recurrent Neural Network. Ceux-ci sont d'actualité avec Siri/Google Home et tous les outils de reconnaissance vocale donc je trouve que c'est un bon sujet. Cela commence par de la recherche sur les différentes cellules (Simple RNN, LSTM, GRU, QRNN). C'est là-dessus que ma journée va s'achever.

### Jour 5 (05/11/17) – Durée travail env. 2h:

Cette journée n'a pas été très productive. J'ai travaillé le matin (6h-14h) et ensuite j'ai dû changer mes pneus. Additionné à un début de maladie, j'ai juste passé quelques heures à commencer l'étude et la rédaction de la veille technologique.

### Jour 6 (06/11/17) – Durée travail env. 4h:

Malade, je ne peux donc pas aller au travail et essaye un peu de continuer les projets. Cependant, avec la tête prise ce n'est pas simple. Je continue donc la veille technologique que j'alterne avec quelques petites modifications sur les différents projets déjà commencés. Rien de fabuleux mais tout de même un peu d'avancement.

### Jour 7 (07/11/17) – Durée travail env. 3-4h:

Toujours malade (conjonctivite), je reste à la maison mais difficile de travailler. Je continue donc tranquillement la veille technologique en passant en revue le modèle du LSTM. Fatigué de lire, je décide de reprendre la compétition Kaggle sur la classification de race de chien que j'avais commencé car elle est utile pour le Projet 8. Je reprends donc la partie de préparation des données ou j'avais bloqué. En fois préparé, je me sers du VGG19 pre-trained dans Keras pour entrainer mon modèle. Je laisse le training 4-5h pour atteindre un résultat de 15% de classification correcte.

### Jour 8 (08/11/17) – Durée travail < 2h:

Ce jour, je me rends compte que mon pre-trained modèle n'a pas été sauvegardé malgré la mise en place de checkpoints si le résultat est meilleur. Mon training n'a donc servi à rien et s'est fait écrasé en fin de la 1ere Epoch… Je cherche donc comment réutiliser mon check point quand je relance le modèle mais sans succès pour l'instant.

### Jour 9 (09/11/17) – Durée travail 4h:

Pour ce 9eme jour, l'objectif était d'arriver à sauvegarder le modèle. Etant donné qu'il y a 5 min entre chaque epoch, je décide donc de le faire sur des scripts fait précédemment plus rapide à entrainer. Une fois le problème résolu, je l'ai implémenté sur le Projet 8 et lancé le training. En parallèle j'ai continué l'avancement du site pour la Classification de film. Le résultat visuel est maintenant correct et j'attends donc mon mentor pour discuter de quelques points (l'évaluation des modèles sans labels). Je lance donc l'entrainement du projet 8 pendant la nuit et cela clos ma journée.

### Jour 10 (10/11/17) – Durée travail 6h:

L'entrainement du Projet 8 ayant durée toute la nuit, j'étais content du résultat mais en arrêtant le process de training, la sauvegarde a été corrompue car j'ai interrompu le process pendant l'écriture de la sauvegarde suivante. J'ai donc modifié le code pour générer des sauvegardes différentes à chaque fois avec une incrémentation pour éviter ce souci par la suite.

Une fois cela fait, je repasse sur le projet 4 (retard des avions) et commence à faire des modèles. Vu la taille du dataset, c'est assez restreint (il n'y a que le SGD Regressor). Je lance un Grid Search exhaustif pour trouver les meilleurs paramètres et pendant le calcul, met aussi en place un petit NN pour faire aussi la régression Histoire de voir s'il performe vraiment mieux. Le résultat des 2 ne me plait pas plus que ça, je réfléchis donc à utiliser les ensembles pour faire le training mais je laisse ça pour le lendemain. En attendant, je commence aussi à préparer les données pour le site (le système de sauvegarde des modèles (scaler et modèle), les données nécessaire à l'UI, etc …)

### Jour 11 (11/11/17) – Durée travail 6h:

Un peu fatigué avec la maladie, j'ai décidé de me changer les idées et partir sur quelque chose de complètement nouveau avec le Projet 5. En suivant les cours j'ai mimé le principe du RBF en essayant d'inclure plus de paramètres (notamment la localité). Ce projet a bien commencé mais j'aimerais aller plus dans le détail par la suite. Cependant certaines techniques vues dans le cours vont m'être utile dans les précédents projets (notamment le Projet 2 avec le classement des aliments)

### Jour 12 (12/11/17) – Durée travail 8h:

Durant cette journée, l'objectif était d'avancer le projet 4 avec la prédiction du retard des avions. Le problème majeur sur ce dataset était le volume de donnés qui faisait crasher le notebook. En conséquence j'ai repris le cleaning sur un script appart et gérer 2 dataset (un d'exploration et un de prod). J'en ai profité pour faire des essais de "simplification de données" (par semaine / par mois et date du mois, agrégation par quart d'heure, …). Cela m'a permis de trouver une configuration permettant d'obtenir de meilleurs résultats (MAE réduite de 14min à 9). J'en ai aussi profité pour avancer le site pour pouvoir faire les premiers tests.

### Jour 13 (13/11/17) – Durée travail 3h:

L'objectif de cette journée est de reprendre le Projet 2 avec les connaissances vues lors du projet 5 (notamment le groupby pour transformer la dataframe). Malheureusement après 1h de recherche pour essayer d'afficher la moyenne de chaque colonne pour les bons et mauvais produits, impossible de trouver une façon. Je décide donc de faire un pause, essayer de déterminer le meilleur format de dataset pour le modèle de M/L concernant le retard des vols. D'un côté je décompose par semaines (de 0 à 52) et par jour (de 1 à 7) et de l'autre je garde par mois (de 1 à 12), jour du mois (de 1 à 31) et le jour de la semaine (de 1 à 7). Pour ce faire je génère les 2 datasets évalue chaque modèle dessus. Le modèle par semaine est légèrement plus précis je reste donc la dessus. Maintenant je nettoie un peu les code pour faire un script qui pilote les modèles et stocke les données dans un fichier texte.

### Jour 14 (14/11/17) – Durée travail 2h:

Comme commencé la veille, cette soirée n'a pas eu d'avancement incroyable. J'ai juste décomposé le code du projet 4 pour avoir 1 script pour alléger les dataset mensuelle, un autre pour les fusionner en 1 seul et générer le dataset d'exploration et de production. Un dernier script permet de scaler les données, préparer les matrices et les sauvegarder pour ne pas le refaire à chaque lancement de modèle. Ensuite je créer différents scripts pour chaque modèle qui écrivent leur résultat dans un fichier texte.

### Jour 15 (15/11/17) – Durée travail 3h:

Ayant tous les scripts prêts, j'ai appliqué aussi la méthode de la partie 4 avec l'affectation de note sur chaque aéroport basé sur les valeurs médianes. L'objectif étant d'avoir que 4 types d'aéroports au lieu de 331 aéroports distinct. L'intérêt était de pouvoir ensuite pourquoi ne pas utiliser un OHE sans avoir d'explosion de dimensions ayant déjà 4.5m de lignes. J'ai ensuite testé différentes façon de générer les features (Aéroport avec et sans OHE, gestion des dates par mois/jour ou par semaine) et j'ai appliqué" le GridSearch sur le SGDRegressor, Adaboost et un simple ANN. Comparé aux premiers modèles testés avec les données "brutes", une augmentation significative a eu lieu permettant de passer d'une MAE de 14 min à 10 min. Il sera malheureusement difficile de passer en dessous car chaque feature est désormais assez propre. C'est un point à discuter avec mon mentor lors d'une session de mentorat

### Jour 16 (16/11/17) – Durée travail 2h30:

Durant ce jour, j'ai repris le projet 3 avec pour objectif de prendre en compte quelques classes que j'avais passé pour l'instant (par exemple le rating). Les modèles ont ensuite été re-testé et donne des résultat légèrement meilleur (par exemple avec Spiderman 1, la 1ere prédiction est Spiderman 2 au lieu d'être en 2nd). Le kl\_divergent a aussi été réduit de 0.80 à 0.75. Cela a pris environ 1h30.

Une petite période a aussi été prise pour regarder le nouveau modèle de reconnaissance d'images (You Only Look Once - YOLO) au cas où le sujet sur les RNN ne soit pas valide comme Veille Technologique.

La seconde partie était la 1ère session de mentorat. Faite pour la présentation du parcours, tout s'est bien passé. Le projet 1 a été envoyé au mentor pour validation et le projet 2 a été aussi abordé afin d'en savoir plus sur les exigences et objectifs. Au final, le rapport ainsi que la présentation doit maintenant être préparé pour la prochaine session de mardi.

### Jour 17 (17/11/17) – Durée travail 3h:

Ce jour, l'objectif a été de commencer une trame graphique pour les différentes présentations. Pour ce faire, le cours d'OC a été repris et une charte graphique mise en place. Par la suite la présentation a été commencé mais devant régénérer certains graphes au fur a à mesure de la rédaction celle-ci prends un peu de temps. Par la suite, cette présentation me servira de support pour le rapport. A ce stade, la présentation est faite à environ 40%.

### Jour 18 (18/11/17) – Durée travail 9h:

L'objectif de la journée était de finir la présentation, tester la durée sur des fausse présentations à blanc et rédiger le rapport. La présentation a pris le plus de temps car il a fallu générer les images, la structure etc… Une fois fini le rapport a été assez rapide bien qu'il fasse 10+ pages. Une relecture sera cependant nécessaire à tête reposée.

J'en profite donc pour "finaliser" le projet 3. Les résultats de tests étant très satisfaisant à mon gout, j'ai décidé de fixer les derniers bugs sur la génération du dataset de prod. Une fois terminé, j'ai mis en ligne le site sur Pythonanywhere (http://coni57.pythonanywhere.com/) .

Une fois cela fait, j'ai commencé la présentation tranquillement. Mon objectif est d'avoir quelque chose de correcte lors de la prochaine session de mentorat (présentation + rapport + site) pour voir pour une seconde soutenance.

### Jour 19 (19/11/17) – Durée travail 6h:

Ce jour, j'ai continué la présentation du Projet 3 ainsi que rédigé le rapport. Le contenu est à voir avec mon mentor Mardi car je ne vois pas comment évaluer numériquement les modèles. Cependant, cela a bien avancé car il ne reste qu'à finaliser le rapport et la présentation s'il y a des changements à apporter.

Ayant pris de l'avance en attendant la 1ere session de mentorat, je me retrouve a devoir faire les rapports et présentation à la suite ce qui est un peu frustrant car cela prend beaucoup de temps et est fastidieux. Du coup je vais faire une pause pour cet fin d'après-midi après avoir passé près de 15 h en 2 jours à rédiger des rapports. L'objectif demain sera de commencer la présentation sur le retard des avions ayant un modèle plutôt satisfaisant aussi.

### Jour 20 (20/11/17) – Durée travail 1h:

Ce jour je n'ai pas eu le temps en particulier d'avancer étant pris sur un autre projet à part. J'ai juste pu faire quelques diapos du Projet 4.

### Jour 21 (21/11/17) – Durée travail 3h:

Entre la fin du travail et la session de mentorat, j'ai continué la présentation du P4 (présentation des modèles).

Lors de la session de mentorat, j'ai malheureusement eu l'information que je vais devoir changer de mentor pour les futurs projets. De ce faits, cela ralenti encore la progression car lors des sessions hebdomadaires, je ne profite même pas des sessions pour résoudre les petites informations nécessaires (lors de la 1ere session, le mentor ne s'est pas présenté. La seconde s'est bien passée mais est fait pour la présentation du parcours et le 3ème est celle où l'on a regardé pour une soutenance et un changement de mentor). Malgré cela, on a revu rapidement le contenu du rapport du P2 pour voir quelques points à rajouter que j'ai fait par la suite pour finaliser le rapport et la présentation.

### Jour 22 (22/11/17) – Durée travail 2h:

La soutenance de P2 a lieu ce jour, je passe donc 1 petite heure à relire la présentation, m'entrainer à blanc pendant environ 45 min, ensuite en attendant la soutenance je continue rapidement la présentation du P4.

La soutenance du P2 s'est bien passée et est validé. Les discussions ON et OFF avec l'évaluateur ont été très intéressantes.

### Jour 23 (23/11/17) – Durée travail 2h:

Ce jour je n'ai pas trop eu le temps d'avancer ayant participé à une conférence sur le Machine Learning (<https://www.meetup.com/fr-FR/LuxRgroup/events/244398875/>). Cette conférence abordait les Generative Adversarial Network et l'utilisation du maximum likelihood pour générer de la musique.

Du coup, je n'ai eu le temps que d'essayer de corriger une erreur (mauvais approche de classification) que j'avais faite sur le code de mon P8. Cette correction a mené à un autre problème avec le Softmax qui "vanish" sur le LSTM. Un post sur SO a été créé mais sans avoir aucunes pistes de résolution (https://stackoverflow.com/questions/47460383/keras-lstm-categorical-cross-entropy-falls-to-0)

### Jour 24 (24/11/17) – Durée travail 2h:

Une alternative a été trouvée pour le problème du Softmax. Au lieu de l'appliquer le Softmax après les LSTM, une couche fully-connected a été ajoutée entre eux. Cela a fonctionné mais le souci qu'il pose c'est que sur la Grammaire de Reber simple, tous les modèles ont 100% de réussite en génération. Du coup j'ai modifié le code pour faire de même avec la version embedded. De ce changement on peut remarquer que les modèles performent de façon similaire à environ 50-60 % de réussite. Du coup je vais utiliser cette version pour le rapport. Je devrais donc reprendre ce rapport ce week-end.

### Jour 25 (25/11/17) – Durée travail 7h:

La grande partie du travail de ce jour a été d'essayer de finaliser le P3. Essayer de trouver des métriques pour les algorithmes de Clustering et la récupération des 5 films les plus similaires avec le clustering hiérarchique. L'optimisation de Kmeans a aussi été essayée avec un nb de Cluster variable.

La seconde partie de la journée a été sur la rédaction du rapport. Beaucoup de changements ont été apporté suite à une discussion avec mon mentor. Celui-ci n'est pas encore fini mais devrait l'être demain pour une préparation à une soutenance la semaine prochaine.

### Jour 26 (26/11/17) – Durée travail 8h:

La 1ere heure a été le mentorat #3. Tout s'est bien passé avec la finalisation de points sur le P3, l'inscription en soutenance. Par la suite le P4 a été abordé avec pas mal de choses a faire. L'objectif étant une soutenance dans une dizaine de jours.

Par la suite j'ai donc mis à jour la fin du rapport du P3, apporté les dernières modifications suite au mentorat avant de passer dans l'après-midi sur le P4.

Lors de cette après-midi, j'ai mis en place un second modèle de prédiction uniquement basé sur le temps. L'idée était de prédire le retard par jour uniquement (car il manque parfois des heures dans le dataset). De ce fait, un nouveau dataset a été fait avec juste date et retard. Pour l'évaluation, le dataset de test a été pris sur le site en utilisant les données de 2017. En effet, le principal défaut de mon précédent modèle linéaire était que j'utilise certains points aléatoires du dataset comme test set. Or ceux-ci sont très liés aux points précédents/suivants. De ce fait l'utilisation d'un dataset de 2017 peut être pas mal bien que un peu hors sujet. Une alternative aurait été de prendre le mois de décembre mais ce mois-ci est très différents en terme de tendance des autres car il y a Noel.

En fin de journée, il reste à avancer le ARIMA, l'optimiser un peu plus proprement que fait actuellement et tester la prédiction sur 2017. Le modèle linéaire va aussi être évalué sur la donnée de 2017 en comparaison simple.

### Jour 27 (27/11/17) – Durée travail 4h:

L'objectif de la journée était de travailler sur le second modèle (ARIMA). Apres avoir cherché des paramètres optimaux pour celui-ci il s'avère que le GridSearch a donné de meilleurs résultats. Je suis donc resté sur celui-ci. Un des bémols de ce modèle est que la prédiction ne peut se faire que sur les 7 jours suivants. De ce fait, il est difficile de l'estimer sur 1 semaine uniquement de 2017.

Par la suite j'ai tenté de mettre à jour l'API suite aux changements de modèle (OHE + week format). Cela a nécessité la mise en place de plusieurs autres scripts pour générer les pickles, fitter et sauvegarder les modèles etc… Une fois ceci fait et l'API mise à jour…. Grosse déception car la prédiction sont totalement incohérentes (le modèle ne prédit que des avions en avance)… Du coup, je me demande s'il ne faudrait pas refaire tout de zéro… script/rapport/présentation car je suis complètement à côté.

### Jour 28 (28/11/17) – Durée travail 4h:

Etant un peu perdu sur le P4, je décide donc en attendant le mentorat du lendemain de continuer sur le P5 ou je ne suis pas bloqué. Pendant environ 2h je modifie le code basé sur les idée du RBM mais afin d'avoir plus de feature interessante pour le clustering (moyenne entre chaque visite, nb de visites, panier moyen par visite, nb d'objets par visite, etc …)

Par la suite, j'ai la soutenance du P3 qui a été un peu stressant avec un faux départ et des problèmes pour partager l'écran. Cependant, une fois ce stresse passé, la soutenance se passe bien et le P3 est validé !

Dans la soirée, une idée m'est venu pour le P4. Partir du principe que les retards sont dûes a un nombre de vol au décollage plus important que possible. L'idée étant de regarder par aéroport, le nb de vol moyen par jour et le retard pour ceux-ci. De même on peut regarder l'évolution du nb de vol en fonction de la semaine et du jour. C'est ce que je me décide de regarder le lendemain

### Jour 29 (29/11/17) – Durée travail 4h:

Après le travail, je me décide donc de fouiller du côté du nb de vols par jour et la corrélation avec le retard, de la date du jour etc … Le problème c'est que cela donne "des matrices" de facteurs mais rien d'exploitable via un modèle classique. Cependant je garde cette idée en tête.

Par la suite, une session de mentorat a lieu où je discute de ces points avec mon mentor (le blocage et cette nouvelle idée). On trouve une des raisons du mauvais fonctionnement du modèle et ouvrons des idées d'améliorations/de changement à apporter. C'est ce que je vais tenter par la suite

* Clipper les vols en avance par un retard nul
* Ne pas supprimer les vols avec retard semi-prévisible et non prévisible
* Regarder quand le weather delay a lieu,
* Regarder du côté des features interactions
* Si possible ajouter nb de vol dans le dataset

Au final, le moral est un peu remonté, j'ai des pistes à explorer et justifier lors du rapport/présentation si le modèle ne répond pas comme imaginé au problème.

### Jour 30 (30/11/17) – Durée travail 3h30:

L'objectif de la soirée était de mettre en place un dataset prenant en compte le nb de vols. Du coup avec diverses agrégations, du cleanup, j'arrive à avoir le dataset que je désire. Qui plus est, il est plus léger par la concaténation. Du coup je peux tester les modèles directement dans le notebook sans crash. Je peux aussi faire l'optimisation dedans et même aller jusque comparer 2 encodages de dataset différents. Du coup, ça avance bien et je commence à comprendre pk le précédent modèle ne marchait pas. L'objectif du lendemain à ce stade est de terminer l'optimisation avec du boosting, terminer la partie visualisation et si possible mettre à jour l'API. Comme cela, je pourrais rédiger le rapport et mettre à jour le power point durant le week-end !

Pour finir la soirée, étant fatigué, je repasse sur la veille technologique afin de tester une autre façon d'évaluer les RNN (le fonctionnement en générateur fonctionne trop bien pour tous les modèles et ne permet pas de voir les avantages/inconvénients des LSTM/GRU)

### Jour 31 (01/12/17) – Durée travail 2h:

Le modèle de boosting a été aussi évalué et perfome un peu mieux. Diverses visualisations ont été ajoutées pour la préparation du rapport/présentation. De plus un README a été mis en place dû aux divers fichiers présents pour mettre en place le modèle. Sans celui-ci, il sera difficile à l'examinateur de ne pas être perdu.

### Jour 32 (02/12/17) – Durée travail 10h30:

Ce jour, la rédaction du rapport a été faite. Celle-ci a été assez pénible car le rapport est long (26 pages) et s'appuie sur divers modèles faits dont il fallait retrouver les informations dans les divers notebooks. Celui-ci a mis près de 5h à être rédigé. Par la suite, le ppt a été repris pour suivre la même trame et tenir dans 30 min de présentation (à blanc cela fait 27 minutes).

Par la suite, j'ai mis à jour le site suite au changement de format du dataset. Celui-ci n'a pas encore été mis en ligne car je ne sais pas s'il faut supprimer le P3.

Pour finir la journée, j'ai décidé de reprendre la veille technologique pour modifier le contenu du rapport ayant fait un modèle de présentation qui ne convenait pas.

### Jour 33 (03/12/17) – Durée travail 8h:

Durant la nuit, j'ai pensé à un problème possible dans les visualisations du P4 à cause d'un shift entre 2 dataset. Du coup durant la matinée, je reprends les graphes qui confirment mon idée et aussi me rassure sur l'analyse des résultats qui sont plus cohérents.

Par la suite, je reprends la relecture du rapport où j'en profite pour corriger ces points. Ensuite, je modifie la présentation, corrige quelques fautes et refait un passage à blanc (29 min 47 ☺).

Durant l'après-midi, je reprends la structure de l'API pour rendre tous les projets visible sur une page d'accueil pour le jury final.

Une fois ceci fait, je décide de me renseigner sur les Virtual environments. Cependant pour éviter de faire planter qqch durant la formation, je testerai cela qu'une fois tous les projets finis (les diverses installations de CUDA / CudNN etc ayant été difficile initialement, je ne veux pas tout faire planter).

Pour finir, je décide donc de continuer sur le P5 en nettoyage un peu le notebook afin de le rendre plus lisible et commence à regarder pour la partie Clustering des Clients.

### Jour 34 (04/12/17) – Durée travail 2h:

Après avoir repassé la présentation du P5, je décide d'ajouter une analyse des résidus pour avoir un intervalle de confiance. Ne connaissant pas trop comment faire, cela a pris pas mal de temps. Une fois fait, le rapport ainsi que la présentation ont été updaté pour la soutenance le lendemain.

### Jour 35 (05/12/17) – Durée travail 2h:

Je repasse une dernière fois la présentation pour éviter de ne pas trouver un mot ou bafouiller. A 19h la soutenance a lieu. Malgré quelques hésitations et erreurs, celui-ci est validé. Il n'y a qu'un seul point un peu curieux sur la Cross Validation que je dois revoir avec mon mentor le lendemain.

### Jour 36 (06/12/17) – Durée travail 3h:

Ayant validé le P4 et étant en attente d'info pour le P5, je décide de regarder le P6 qui aborde un sujet complètement inconnu. Je télécharge juste un dataset. Je vois que je dois cleaner pas mal le texte (qui a des balises HTML). Avant de commencer tout cela, je regardé du coté de la lib nltk sur des videos de Sentdex.

A 18h30, la session de mentorat a lieu où l'on a repris les points du P4 qui posaient des questions, par la suite on aborde le P5 ou le cleaning a été un peu trop violent. J'ai donc des pistes a explorer (OHE sur les articles notamment). Une fois la session terminée, je teste ca mais une fois de plus "Memory Error"… Je trouve peut-être une façon de regrouper les articles par sous-groupes via un traitement par mot clé à faire.

### Jour 37 (07/12/17) – Durée travail 2h:

L'objectif était de trouver comment réduire le nb d'objet sans perdre le sens des articles. Du coup je mets en place quelques Regexp pour grouper les articles entre eux mais cela n'est pas évident. Malgré une dizaine de Regexp, la réduction n'est pas énorme. Une alternative reste à générer via un script python ce dataset et fusionner dans un 1er temps par Invoice Number. Cela va être à tester ce weekend end. Le reste des datas était plutôt bien géré (gestion par jour et moment de la journée, nb d'achats, moyenne de visites etc…). Du coup après ce traitement, le clustering va pouvoir être mis en place pour labéliser les clients et il restera à faire un classifieur.

### Jour 38 (08/12/17) – Durée travail 3h:

Le problème majeur étant les articles, au lieu de travailler sur 400 000 lignes regroupant 4000 article, une autre possibilité est de travailler à part sur un second dataset avec pour objectif de faire le OHE dessus. Par contre au lieu, d'utiliser les articles dans le OHE, il faut trouver une autre façon de les classer. Le fonctionnement par Regexp n'étant pas faisable, je pense à faire un 1er clustering et remplacer chaque article par son cluster. Ensuite il sera plus facile de le OHE et grouper le dataset par invoice et par Client au final. L'objectif de la soirée a été de mettre en place ce clustering.

### Jour 39 (09/12/17) – Durée travail 8h:

Le clustering étant bien commencé, j'ai fait l'agrégation au dataset initial et préparé le dataset par clients. Sur ce nouveau dataset, un second clustering va être appliqué pour labéliser les clients en fonction de ce qu'ils achètent mais aussi en fonction de leur fréquence d'achat, panier moyen, localisation etc.

Le résultat de ce second clustering est un peu curieux car j'ai 2 principaux Cluster avec 3300 clients, 1 de 120 et un de 4… Un premier test de classifier donne des résultats correct mais il faut encore en tester certain et fine-tune les meilleurs. Une fois cale fait, il restera a préparer le code pour tester un client lambda.

### Jour 40 (10/12/17) – Durée travail 8h:

Pas mal de temps ont été nécessaire pour rendre le notebook un peu plus lisible comme il n'y a pas de rapport à fournir. Lors de la mise en place de visualisations, je me suis rendu compte aussi d'erreurs que j'ai fixé. Au final il reste des points assez curieux au niveau du clustering. Ce sont des points à voir avec mon mentor lors de la prochaine session. Mais globalement, le projet à bien avancé, il reste à faire le code de classification sur les données d'un client lambda ainsi que la présentation.

Etant un peu lassé de ce projet, je fais une petite pause passe sur le P8 pour la fin de la journée.

### Jour 41 (11/12/17) – Durée travail 3h:

Lors de l'investigation sur l'échec de clustering, je me suis rendu compte que cela était du à une mauvais interprétation du Score de Silhouette. J'utilisais toujours le nb de cluster minimisant ce score (comme s'il évaluait une entropie). J'ai vu cela suite à un autre test de clustering ou le Score était négatif. De ce fait, j'ai repris les étapes du clustering. Pour aider le modèle, j'ai retiré les keywords très fréquents qui ne décrivent pas les objets (set, Christmas, Vintage, …). Additionné à cela, j'ai utilisé un discount factor sur les colonnes pour réduire l'impact sur le clustering d'un mot très peu utilisé. Cela m'a permis d'avoir un bien meilleur clustering (Silhouette Score x 10 par rapport au précédent).

Ensuite, j'ai repris le Clustering des clients. Au lieu d'utiliser un Standard Scaler qui réduit le Score de Silhouette, je l'ai repris avec le MinMax et là, j'ai eu un clustering bien plus pertinent.

Pour finir j'ai repris la partie d'entrainement des classifieurs car le meilleur classifieur a aussi changé (passé de SVC à Random Forest). Le Fine-Tuning a été fait mais pas encore de visualisations.

Pour finir la soirée, j'ai essayé de continuer la rédaction du P8 qui me donne du fil à retordre plus que je ne l'avais imaginé.

## Temps de travail : 200h