Université Lumière Lyon 2, Laboratoire ERIC, Lyon

Rapport de Stage

Stage du 09/05/2022 au 08/07/2022

Sujet : Analyse de tweets selon les topics en utilisant la Non-negative Matrix Factorization (NMF)

Abdourahmane Ndiaye
L3 MIASHS IDS (Informatique décisionnelle et statistiques)

Table des matières

Introduction

Sujet de stage

« Le sujet du stage est focalisé sur l'analyse de tweets selon les topics et leur évolution temporelles en utilisant des méthodes de type Non-Negative Matrix Factorization (NMF). Il sera l'occasion de mettre en oeuvre un code sous Python effectuant ces tâches. Le stage sera l'occasion de s'initier et de progresser dans ce langage. Dans l'idéal, le code sera amélioré de manière à détecter automatiquement les données aberrantes au moyen d'une technique très en vogue : les Median of Means. »

Pour la L3, j'ai effectué un stage avec le laboratoire ERIC à Lyon 2 sur le campus de Bron. Mon stage a été encadré par Stéphane Chrétien. L'analyse de tweets via la NMF était le sujet principal de mon stage. C'est un sujet que j'ai choisi parmi 3 autres car c'est quelque chose que j'ai toujours voulu essayer et je voulais progresser en python.

Etat de l'art

La Non-Negative Matrix Factorization est une méthode d'analyse qui permet de trouver les sujets principaux abordés dans des textes.

Dans un texte les mêmes mots vont souvent revenir lorsqu'un sujet est abordé. Lorsqu'on fait de la NMF, on constitut une matrice X ayant sur un axe les mots contenus dans les textes et sur l'autre les documents (les textes). La matrice contient la fréquence d'apparition de chaque mot par document.

Via l'analyse de cette matrice, on va créer des topics (sujets). Un topic est généralement composé de deux-trois mots qui reviennent fréquemment ensemble à travers les différents textes.

A l'aide de ces topics, on génère deux autres matrices W et H. W contient sur un axe les topics et sur l'autre les textes. C'est une matrice qui nous indique la fréquence à laquelle apparait un topic (sujet) pour chaque texte donné. H en revanche contient la fréquence d'apparition d'un mot pour chaque topic.

La suite consiste à calculer la norme de Frobenius, puis à la réduire.

$$\|\mathbf{X} - \mathbf{WH}\|_F^2$$

Pour réduire la norme il faut recréer tour à tour les matrices W et H de sorte que lorsqu'on refait le calcul on ait une norme plus petite. Notez aussi que la matrice que constitue le produit WH est quasiment égal à X.

En dehors de la NMF, nous avons à disposition la LDA, NCPD et l'ONCPD comme méthode pour faire du topic modeling.

La NCPD par exemple est une variante de la NMF dans laquelle on utilise des tenseurs de dimension 3 à la place des matrices. Cela permet d'ajouter une dimension temporelle à l'analyse des textes. Lorsqu'on analyse des tweets, la NCPD est plus performante que la NMF car elle permet de voir plus précisément le moment où un sujet a été le plus actif.

Déroulement du stage

Mes attentes

Je pensais pouvoir analyser des tweets de chez moi et trouver des faits intéressants après ce stage. Je ne connaissais pas les méthodes utilisées. Je pensais m'améliorer en python et en apprendre plus sur les façons de l'utiliser.

Je pensais devoir chercher les données sur tweeter par moi-même afin de les collecter. Il y aurait ensuite eu une étape de nettoyage des données et des prétraitements afin de les rendre utilisable. Je pensais que l'on construirait ensuite un programme complexe qui se chargerait de traiter les données. Je n'avais en revanche pas idée que quel genre de rendue nous allions avoir.

Premiers pas

Durant la première semaine de stage j'ai été dans une phase de recherche. Mr Chrétien m'a donné un article scientifique à lire, intitulé « Detecting short lasting topics using Nonnegative Tensor decomposition ». C'est un article récent (2021) qui compare la NCPD (nonnegative tensor decomposition) aux autres méthodes de topic modeling tout en les décrivant.

Lien de l'article : https://arxiv.org/abs/2010.01600

J'ai créé un document dans lequel j'ai mis par écrit ce que je comprenais ou non à mesure que je lisais l'article. Ce document m'a ensuite servi de journal de bord afin de suivre mon avancement au jour le jour. C'était un bon moyen d'analyser tous les éléments mis à ma disposition et d'avoir un suivi de ce que je devais faire d'un jour à l'autre. Cependant j'ai aussi J'ai fait quelques schémas et pris des notes sur papier.

Figure 1 Sommaire du journal de bord

Dès le début du stage ce sont à travers des échanges que j'ai compris plusieurs des concepts clés. Afin de comprendre les tenseurs, j'ai par exemple questionné Mr Chrétiens et d'autres élèves. J'ai pu échanger avec Loïck qui travaillait sur la LDA et cela nous a permit de mieux visualiser les étapes à suivre pour faire du topic modeling.

Tout le long de mon stage, j'ai aussi recherché des vidéos en lien avec la NMF et autres notions tels que la représentation les topics. Il se trouve dans mon journal de bord la plupart des ressources que j'ai consulté ainsi que des captures d'écran afin d'y illustrer mes réflexions.

Python et git hub

Durant deux semaines nous avons suivi des cours pour à utiliser GitHub ainsi que python afin de mettre tous les stagiaires à la page. Ainsi nous pourrions partager nos programmes et documents pour travailler en équipe.

Les cours étaient encadrés par d'autres stagiaires de Mr Chrétien (Astrid et Benjamin) qui ont pu nous aider tout le long du stage.

Durant la première partie du cours nous avons appris à utiliser GitHub afin de travailler en collaboration et accéder à des ressources mis en commun.

Nous avons ensuite fait des cours de python. Le cours était accompagné d'exercices et avait pour but de nous donner les outils pour faire des analyses statistiques sur python à l'aide des différentes librairies et modules (Numpy, Scipy, Pandas…).

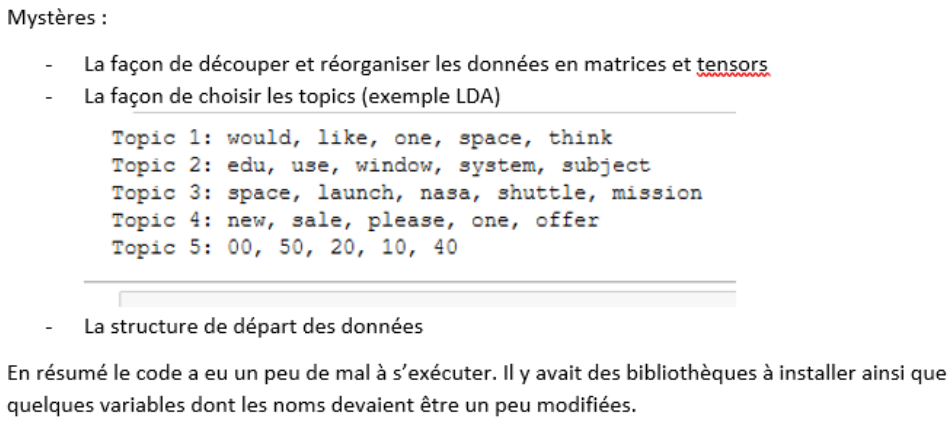
Lien GitHub du cours : https://github.com/StagiairesMIASHS/Introduction

Analyse de tweets

Après les cours de python, nous avons eu à disposition un repo git où se trouvait le travail de Hanbaek lyu et Lara kassab sur le topic modeling. Il y avait aussi des notebooks contenant des codes pour faire de la NMF mais aussi du LDA, NCPD et ONCPD tout à la suite.

L'exécution du premier programme de topic modeling avec les données d'ABC news via jupyter notebook a été compliqué au départ. Il y avait plusieurs choses à installer et paramétrer mais on ne pouvait parfois s'en rendre compte qu'une fois l'erreur affichée.

Le programme a finalement pu s'exécuter cependant je ne comprenais pas encore toute la structure de ce dernier.



Mystères :

- La façon de découper et réorganiser les données en matrices et tensors
- La façon de choisir les topics (exemple LDA)

```
Topic 1: would, like, one, space, think
Topic 2: edu, use, window, system, subject
Topic 3: space, launch, nasa, shuttle, mission
Topic 4: new, sale, please, one, offer
Topic 5: 00, 50, 20, 10, 40
```

- La structure de départ des données

En résumé le code a eu un peu de mal à s'exécuter. Il y avait des bibliothèques à installer ainsi que quelques variables dont les noms devaient être un peu modifiées.

Figure 2 extrait du journal de bord

Il y avait un deuxième notebook que j'ai essayé d'exécuter plusieurs jours de suite cependant celle-ci prenait des heures et finissait par afficher des erreurs. Elle demandait beaucoup trop de ressources à l'ordinateur et je n'ai jamais pu obtenir ses résultats.

Journée des stagiaires

On nous a appris durant le stage que l'on pouvait participer à la journée des stagiaires, cependant la date était très proche.

J'ai décidé d'y participer avec Luan Dechery, un stagiaire de L2 MIASHS qui travaillait sur le même sujet. Nous avons donc préparé un diaporama de 3 pages. Le but était d'expliquer la NMF d'une façon simple avec un diaporama court mais parlant.

Je doute que tout le monde ait vraiment compris ce qu'est la NMF grâce et son fonctionnement grâce à notre présentation. Cependant Ils en auront entendu parler et auront vu les notions principales comme la norme de Frobenius.

Nous avons fait le choix d'illustrations simples mais qui attirent l'attention.

J'ai pu réexpliquer pendant la pose qui a suivi l'exposé le fonctionnement de la NMF à quelques collègues et j'ai répondu à leurs questions.

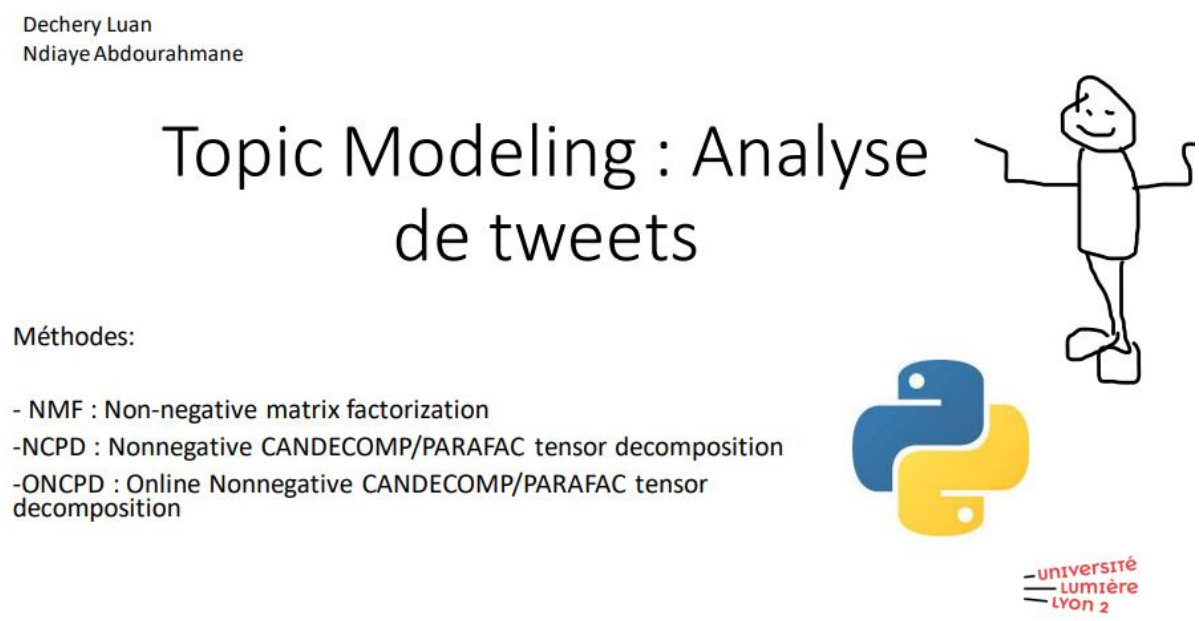
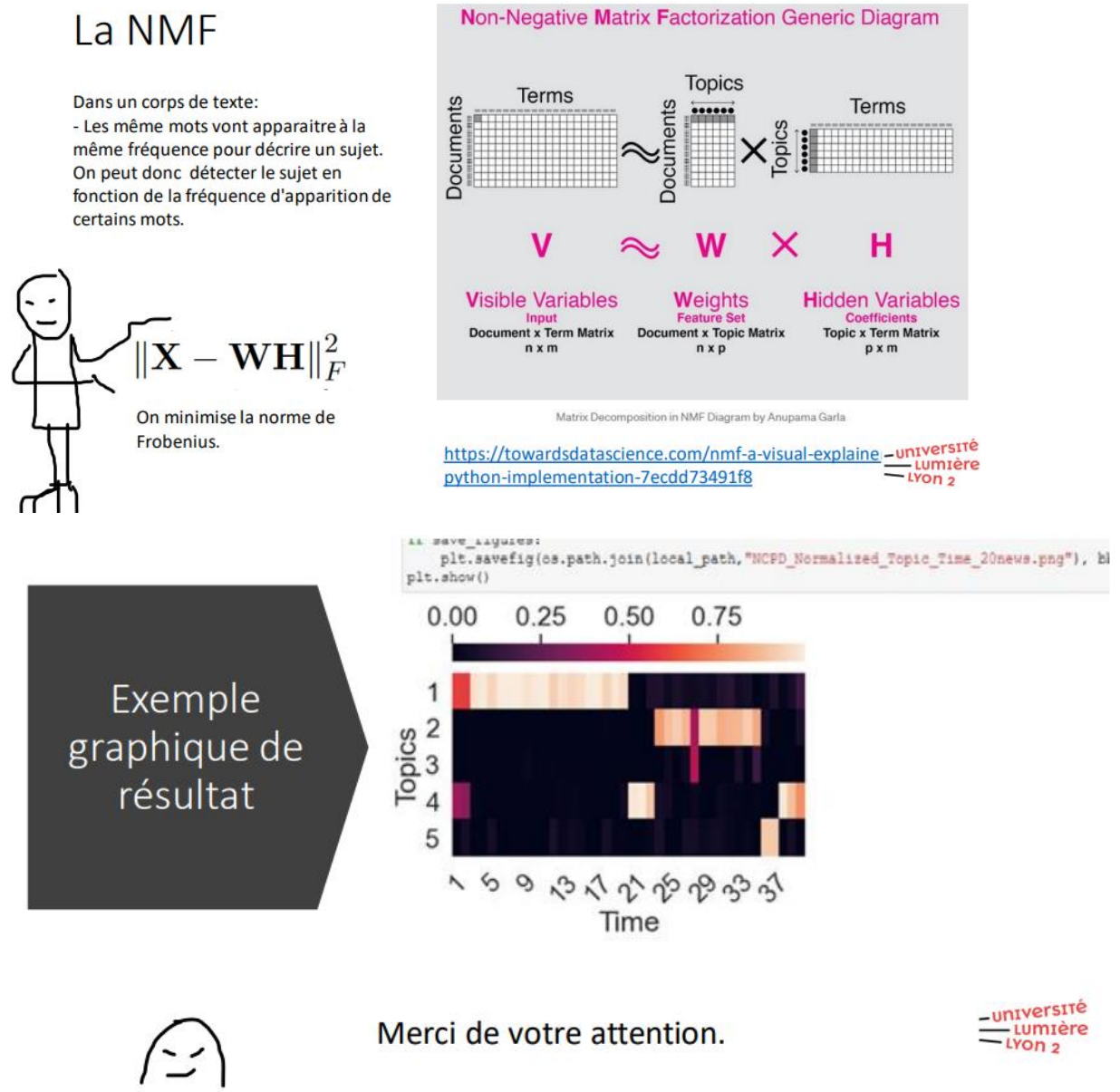


Figure 3 Page 1 de la présentation



Construction d'un algorithme

En recherchant un moyen de mieux visualiser ce qu'il se passe lors d'une NMF, j'ai trouvé le tutoriel d'Anapuma Garla. Elle y explique comment se passe une NMF avec des schémas parlant que j'ai ajouté à mon journal de bord et commenté. C'est de cette façon que j'ai le mieux saisi ce qu'il se passait avec la NMF.

Les explications d'Anapuma sont suivies d'un tutoriel pour faire une NMF avec comme données des discours d'inauguration présidentiel. Je voulais à ce moment-là construire une NMF du début à la fin puis l'appliquer aux données de mon choix.

Tutoriel d'Anapuma Garla : https://towardsdatascience.com/nmf-a-visual-explainer-and-python-implementation-7ecdd73491f8

```
# 3 ---- Isoler les données

"""I'm going to make a dataframe of the President's names and
speeches, and isolate to the first term inauguration speech,
because some President's did not have a second term. This makes
for a corpus of documents with similar lengths."""

"""
Donc Anupama a décidé de s'en tenir aux discours d'inauguration
de premier mandat des présidents. Car ils n'ont pas tous eu de
deuxième mandat.
"""

#select rows that are first term inaugural addresses
df = df.drop_duplicates(subset=['Name'],keep='first')
# comment sait-il que 'first' est dans inaugural addresses

df = df.reset_index() #je ne sais pas ce qu'est l'index

df = df[['Name','text']]
# J'imagine qu'on ne conserve que ces deux colonnes de donnée

df = df.set_index('Name')

print()
print (df.head())

```

Figure 4 Extrait du programme "NMF_tentative_1.py" pour faire une NMF

A mesure que le programme avance, j'ai rencontré quelques erreurs d'exécutions car il manquait des modules et librairies. En cherchant un peu j'ai souvent pu régler le problème. Je n'ai cependant pas réussi à exécuter le programme après l'avoir fini. Il y avait à nouveau quelques erreurs à corriger dans la dernière partie.

Conclusion

Objectifs atteints

J'ai fini par comprendre le fonctionnement de l'analyse de tweets selon les topics et leur évolution temporelles avec la NMF. Cependant je n'ai pas pu exécuter tout le programme que j'ai créé en suivant le tutoriel d'Anapuma Garla.

Progresser en python est un objectif que je pense atteint. Le stage à élargi le champ de ce que je savais faire avec ce langage.

Dans l'idéal, le code aurait pu être amélioré de manière à détecter automatiquement les données aberrantes mais nous ne sommes finalement pas allés jusque-là. J'ai voulu faire du topic modeling sur des données de mon choix en usant toutes les différentes méthodes présentées. C'est peut-être ce qui m'a éloigné de cet objectif ci.

Difficultés rencontrées

Le stage s'est bien déroulé mais n'était pas dépourvu de difficultés. En programmant, il y avait des morceaux de code à modifier et des installations à faire sans que cela ne règle tous les soucis. Il fallait donc être méticuleux et concentré pour ne pas se perdre dans ce que l'on faisait. Mon de bord m'a servi à répertorier plusieurs des erreurs que j'ai rencontrées pour ne pas les voir se reproduire.

La compréhension des articles scientifiques, la compréhension des différentes parties des codes fournis ont aussi été des défis. De même pour l'interprétation des résultats graphiques. Il est arrivé que plusieurs stagiaires exécutent le même programme puis se demandent ce que veulent vraiment dire les résultats. Mais c'est en y pensant à plusieurs que l'on pouvait trouver la réponse le plus rapidement.

Les apprentissages

Le stage a été ponctué de séquences durant lesquels Mr Chrétiens est venu nous parler de notions en rapport avec les différents sujets de stage. Nous avons ainsi eu une introduction aux signatures, sur ce que sont les tenseurs, la façon de faire des calculs avec ou encore un exposé sur le lien entre la NMF et l'ACP. A cela s'ajoutent toutes les applications et notions mathématiques que j'ai découvert en m'exerçant, en lisant ou à travers des vidéos.

Ce que j'ai apprécié

Ce stage était comme une introduction à la recherche dans un environnement stimulant. En échangeant avec mes collègues nous avons amélioré notre compréhension commune de certaines notions.

De plus, Mr Chrétien a pu nous raconter plein d'anecdotes sur le monde de la recherche et sur l'apprentissage en générale. C'était très enrichissant.

Je me suis rendu compte au cours du stage que le document tableau de bord servait aux autres. Ils le consultaient parfois pour trouver des pistes ou se renseigner sur ce que je faisais. J'ai donc essayé de le rendre plus organisé et compréhensible.

Ce que j'aurai aimé faire

J'aurais aimé approfondir encore plus mes connaissances en python. Même si je pense avoir vu de nouvelles choses au fil du stage, cela m'aurait peut-être permis de corriger les bugs de certains programmes et de mieux les lire.

Je pensais pouvoir maitriser toutes les méthodes de topic modeling à la fin de mon stage mais ce ne fut pas le cas.

Je comptais faire mon rapport de stage en Latex comme dernier défi, mais je n'ai pas pu. Je sais cependant où trouver les ressources pour essayer de refaire un document en latex. Je pense que cela pourrait très bientôt me servir.

Voici le lien d'un tutoriel que j'ai consulté pour commencer à programmer en Latex sur Overleaf.

Lien : https://fr.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes

Annexe

Lien pour accéder aux documents (journal de bord + Programme)

https://github.com/StagiairesMIASHS/Stage-TweetsNMF/tree/main/Abdourahmane