

**UNIVERSITE IBN TOUFAIL**

FACULTE DES SCIENCES

**Master Génie Logiciel pour Le Cloud**

ANALYSE DES SENTIMENTS AVEC LSTM

**Présenté par : Abdellah BIROUK & Ayoub MANÇOUR BILLAH**

**Encadré par : Younes Chihab**

Année Universitaire 2020-2021

Table de matières

[Table de matières 2](#_Toc74761244)

[Liste des figures 4](#_Toc74761245)

[I. INTRODUCTION 5](#_Toc74761246)

[II. Définitions 6](#_Toc74761247)

[1. Le Deep Learning 6](#_Toc74761248)

[2. Analyse des sentiments 6](#_Toc74761249)

[3. La polarité et l’intensité de l’opinion 7](#_Toc74761250)

[4. Domaine d’application d’analyse des sentiments : 7](#_Toc74761251)

[III. RÉALISATION 9](#_Toc74761252)

[5. Python 9](#_Toc74761253)

[6. JupyterLab 10](#_Toc74761254)

[7. Anaconda 10](#_Toc74761255)

[8. Google Colab 11](#_Toc74761256)

[9. Pandas 11](#_Toc74761257)

[10. NumPy 12](#_Toc74761258)

[11. Matplotlib 12](#_Toc74761259)

[12. NLTK 12](#_Toc74761260)

[13. KERAS 13](#_Toc74761261)

[14. LSTM 13](#_Toc74761262)

[15. TensorFlow 14](#_Toc74761263)

[IV. Test et fonctionnement : 15](#_Toc74761264)

[1. Préparation des données (dataset) : 15](#_Toc74761265)

[2. Word Embeddings 15](#_Toc74761266)

[3. Base de données (Dataset): 16](#_Toc74761267)

[4. Réalisation : 17](#_Toc74761268)

[a) Importation des bibliothèques et des plates-formes 17](#_Toc74761269)

[b) Lecture des données 17](#_Toc74761270)

[c) Prétraitement des données 18](#_Toc74761271)

[d) Fractionnement des données 18](#_Toc74761272)

[e) Tokenisation des données 18](#_Toc74761273)

[f) Construction du modèle 19](#_Toc74761274)

[g) Formation du modèle 20](#_Toc74761275)

[h) Score et précision des tests 20](#_Toc74761276)

[i) Précision et perte du modèle pour l'entraînement et les tests 20](#_Toc74761277)

[j) Test et résultat final : 21](#_Toc74761278)

[k) Sauvegarde du programme 22](#_Toc74761279)

[V. CONCLUSION 23](#_Toc74761280)

Liste des figures

[Figure 1. Logo Python 9](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761281)

[Figure 2. Logo JupyterLab 10](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761282)

[Figure 3. Logo Anaconda 10](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761283)

[Figure 4. Logo Google Colab 11](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761284)

[Figure 5. Logo Pandas 11](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761285)

[Figure 6. Logo NumPy 12](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761286)

[Figure 7. Logo Matplotlib 12](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761287)

[Figure 8. Logo NLTK 12](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761288)

[Figure 9. Logo KERAS 13](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761289)

[Figure 10. Schéma d'un réseau LSTM 13](#_Toc74761290)

[Figure 11. Logo TensorFlow 14](file:///C:\Users\hp\Desktop\Master\Semm1\Cours%20Python\Projet%20python\projet_analyse_des_sentimens_LSTM.docx#_Toc74761291)

# INTRODUCTION

Dans ce projet nous allons vous présenter une reconnaissance émotionnelle en utilisant un texte (Sentiment Analysis using text), ce projet est un système de reconnaissance des émotions multimodal qui est construit pour extraire des informations sur les émotions de la saisie de texte. Le système de reconnaissance des émotions classe les émotions selon deux types de base : positif ou négatif. Si la valeur d'intensité d'émotion de l'émotion actuellement reconnue est inférieure à un seuil prédéfini, la sortie d'émotion est déterminée comme étant neutre.

Le système de reconnaissance des émotions proposé peut détecter les émotions à partir du texte par l’algorithme **LSTM** (long short-term memory) pour évaluer l'approche acoustique, un drame diffusé, y compris le contenu textuel. Lors de la sélection des fonctionnalités, un ensemble initial de fonctionnalités acoustiques contenant 33 fonctionnalités est d'abord analysé et la reconnaissance multimodale d'émotion à partir du texte est extraite.

# Définitions

## Le Deep Learning

A présent que nous avons précisé comment fonctionnent les réseaux de neurones de manière générale, nous allons aborder le domaine du Deep Learning.

Cette famille d’algorithmes a permis de faire des progrès importants dans les domaines de la classification des textes et du traitement du langage par exemple.

Les modèles de Deep Learning sont bâtis sur le même modèle que les perceptrons multicouches précédemment décrits. Cependant, il convient de souligner que les différentes couches intermédiaires sont plus nombreuses.

Chacune des couches intermédiaires va être subdivisée en sous partie, traitant un sous problème, plus simple et fournissant le résultat à la couche suivante, et ainsi de suite.

## Analyse des sentiments

L’analyse de sentiment est l’étude computationnelle et sémantique des parties de textes en fonction des opinions, des sentiments et des émotions exprimés dans le texte. Généralement l’expression « analyse des sentiments » est utilisée pour désigner la tâche de classification automatique des unités de texte en fonction de leur polarité (positive, négative, neutre).

## La polarité et l’intensité de l’opinion

La polarité peut être définie par des catégories telles que « positif », « neutre » et « négatif » La polarité d’une opinion exprime la positivité, la négativité ou une information de cette dernière. On dit d’une opinion positive qu’elle possède une polarité positive, et inversement, on dit d’une opinion négative qu’elle possède une polarité négative ou neutre possède une information L’intensité décrit à quel point la polarité d’une opinion est forte. Par exemple, dans une opinion à polarité positive, aimer est plus intense qu’apprécier, ou encore, dans une opinion de polarité négative, haïr est plus intense que de ne pas aimer.

## Domaine d’application d’analyse des sentiments :

* **Politique :** Grâce à l’analyse des sentiments, les décideurs de politique peuvent prendre l'avis des citoyens sur certaines politiques, afin de bénéficier de cette information pour améliorer ou créer une nouvelle politique qui convient avec les citoyens.
* **Prise de décision :** L'opinion et l'expérience des gens sont un élément très utile dans le processus de prise de décision.
* **Domaine de Transport :** Pour assembler et analyser les opinions du public sur le statut de transport.
* **Les systèmes de recommandations :** À travers l’analyse des sentiments on peut classer les opinions des gens de façon positive ou négative, le système définit qui devrait prendre ou pas prendre la recommandation.
* **Domaine médical :** Analyse l’opinion des médecins et des patients sur les médicaments et les services hospitaliers. Ainsi que sur les documents de l’état du patient qui contiennent le diagnostic et la description du résultat d’examen.
* **Domaine d’éducation :** Développer le niveau d'enseignement à travers l'analyse et l'interprétation de l'opinion de l'étudiant par les méthodes d'enseignement pour permettre d’améliorer l'enseignement et l'apprentissage.
* **Marketing :** Du côté d’entreprises, on permet au fournisseur plus de connaissances à propos des besoins des consommateurs. Du côté client, il peut donner son opinion, s’inspirer des opinions des autres et aussi comparer les produits avant de les acheter.

# RÉALISATION

Durant la réalisation de ce projet nous avons été menés à l’utilisation du langage de programmation PYTHON ainsi que plusieurs bibliothèques de ce langage.

* **Langages et Programmes utilisés :**

## Python

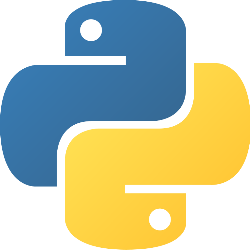
Python est un langage de programmation interprété, multi paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions.

Figure . Logo Python

## JupyterLab

JupyterLab est un environnement de développement interactif basé sur le Web pour les blocs-notes, le code et les données Jupyter. JupyterLab est flexible : configurez et organisez l'interface utilisateur pour prendre en charge un large éventail de flux de travail dans les domaines de la science des données, de l'informatique scientifique et de l'apprentissage automatique. JupyterLab est extensible et modulaire : écrivez des plugins qui ajoutent de nouveaux composants et s'intègrent aux composants existants.

Figure . Logo JupyterLab

## AnacondaAnaconda | The World's Most Popular Data Science Platform

Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique, qui vise à simplifier la gestion des paquets et de déploiement.

Figure . Logo Anaconda

## Google Colab

Google Colab est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l’apprentissage automatique. Cette plateforme permet d’entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d’installer quoi que ce soit sur notre ordinateur à l’exception d’un navigateur.

Figure . Logo Google Colab

* **Outils et Bibliothèques utilisés :**

## Pandas

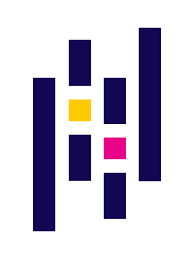
Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. Pandas est un logiciel libre sous licence BSD.

Figure . Logo Pandas

## **NumPy**

NumPy est une bibliothèque pour langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.

Figure . Logo NumPy

## Matplotlib

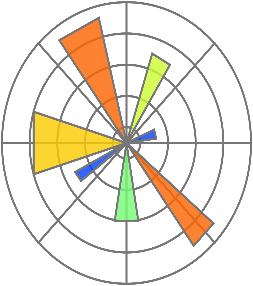
Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy.

Figure . Logo Matplotlib

## NLTK

Natural Language Toolkit (NLTK) est une [bibliothèque logicielle](https://fr.wikipedia.org/wiki/Biblioth%C3%A8que_logicielle) en [Python](https://fr.wikipedia.org/wiki/Python_(langage)) permettant un [traitement automatique des langues](https://fr.wikipedia.org/wiki/Traitement_automatique_des_langues), développée par Steven Bird et Edward Loper du département d'informatique de l'[université de Pennsylvanie](https://fr.wikipedia.org/wiki/Universit%C3%A9_de_Pennsylvanie). En plus de la bibliothèque, NLTK fournit des démonstrations graphiques, des données-échantillon, des tutoriels, ainsi que la documentation de l'[interface de programmation](https://fr.wikipedia.org/wiki/Interface_de_programmation) (API).

Figure . Logo NLTK

## KERAS

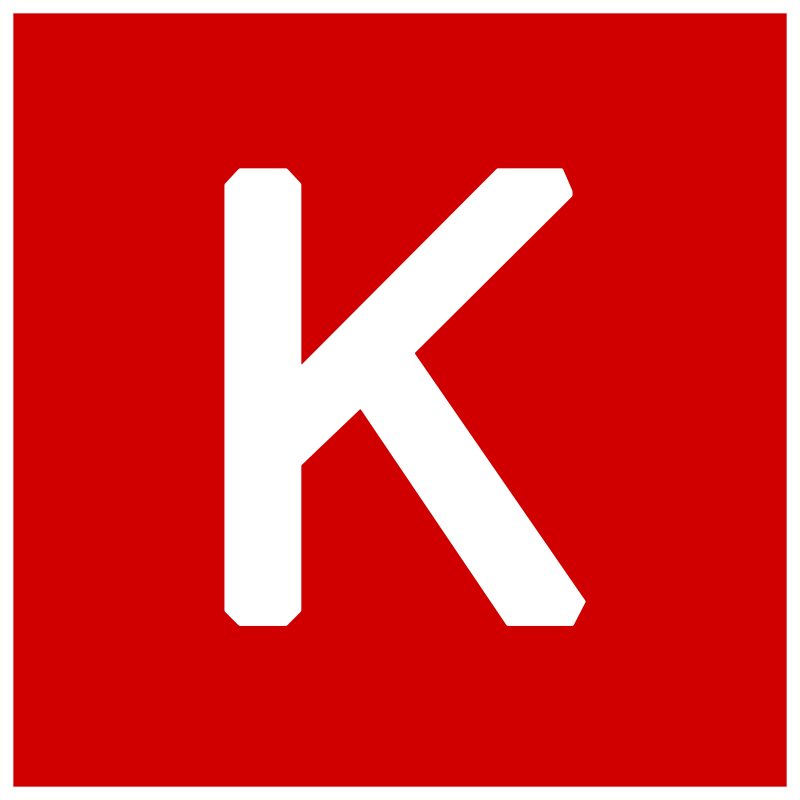
Keras est une bibliothèque open source écrite en python. La bibliothèque Keras permet d'interagir avec les algorithmes de réseaux de neurones profonds et d'apprentissage automatique, notamment Tensorflow, Theano, Microsoft Cognitive Toolkit ou PlaidML. Conçue pour permettre une expérimentation rapide avec les réseaux de neurones profonds, elle se concentre sur son ergonomie, sa modularité et ses capacites d’extension.

Figure . Logo KERAS

## LSTM

Un réseau Long short-term memory (LSTM), en français réseau récurrent à mémoire court et long terme ou plus explicitement réseau de neurones récurrents à mémoire court-terme et long terme, est l'architecture de réseau de neurones récurrents la plus utilisée en pratique qui permet de répondre au problème de disparition de gradient. Le réseau LSTM a été proposé par Sepp Hochreiter et Jürgen Schmidhuber en 1975.

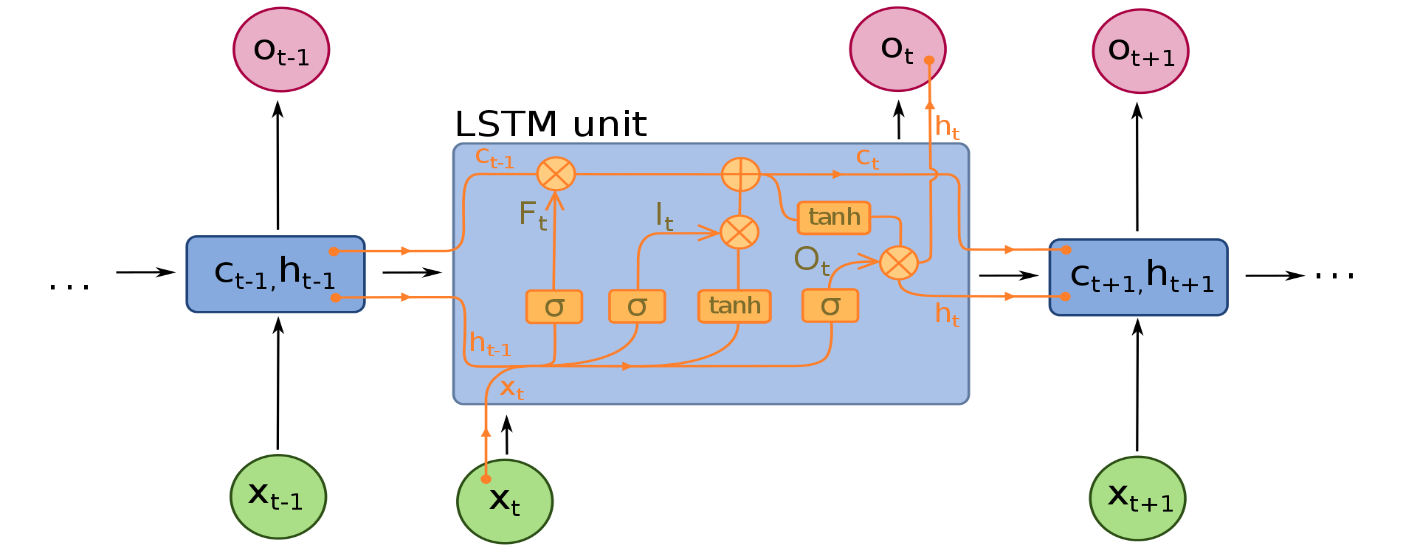


Figure . Schéma d'un réseau LSTM

## TensorFlow

TensorFlow est un framework de programmation pour le calcul numérique qui a été rendu Open Source par Google en Novembre 2015. Depuis son release, TensorFlow n’a cessé de gagner en popularité, pour devenir très rapidement l’un des framework les plus utilisés pour le Deep Learning et donc les réseaux de neurones. Son nom est notamment inspiré du fait que les opérations courantes sur des réseaux de neurones sont principalement faites via des tables de données multidimensionnelles, appelées Tenseurs (Tensor). Un Tensor à deux dimensions est l’équivalent d’une matrice. Aujourd’hui, les principaux produits de Google sont basés sur TensorFlow : Gmail, Google Photos, Reconnaissance de voix

Figure . Logo TensorFlow

# Test et fonctionnement :

Dans ce chapitre, on va traiter les différentes phases de notre méthode d’analyse des sentiments. Durant ce travail, on a fait face à plusieurs phases :

## Préparation des données (dataset) :

Nous considérons d'abord comment diviser les données. Nous avons choisi de diviser les données en trois morceaux : train, développement, test.

1. **Training :** L'échantillon de données utilisé pour l'apprentissage.
2. **Ensemble de développement** (ensemble de validation croisée de blocage) : échantillon de données utilisé pour ajuster les paramètres d'un classificateur et fournir une évaluation impartiale d'un modèle.
3. **Ensemble de test :** échantillon de données utilisé uniquement pour évaluer performance d'un modèle final.

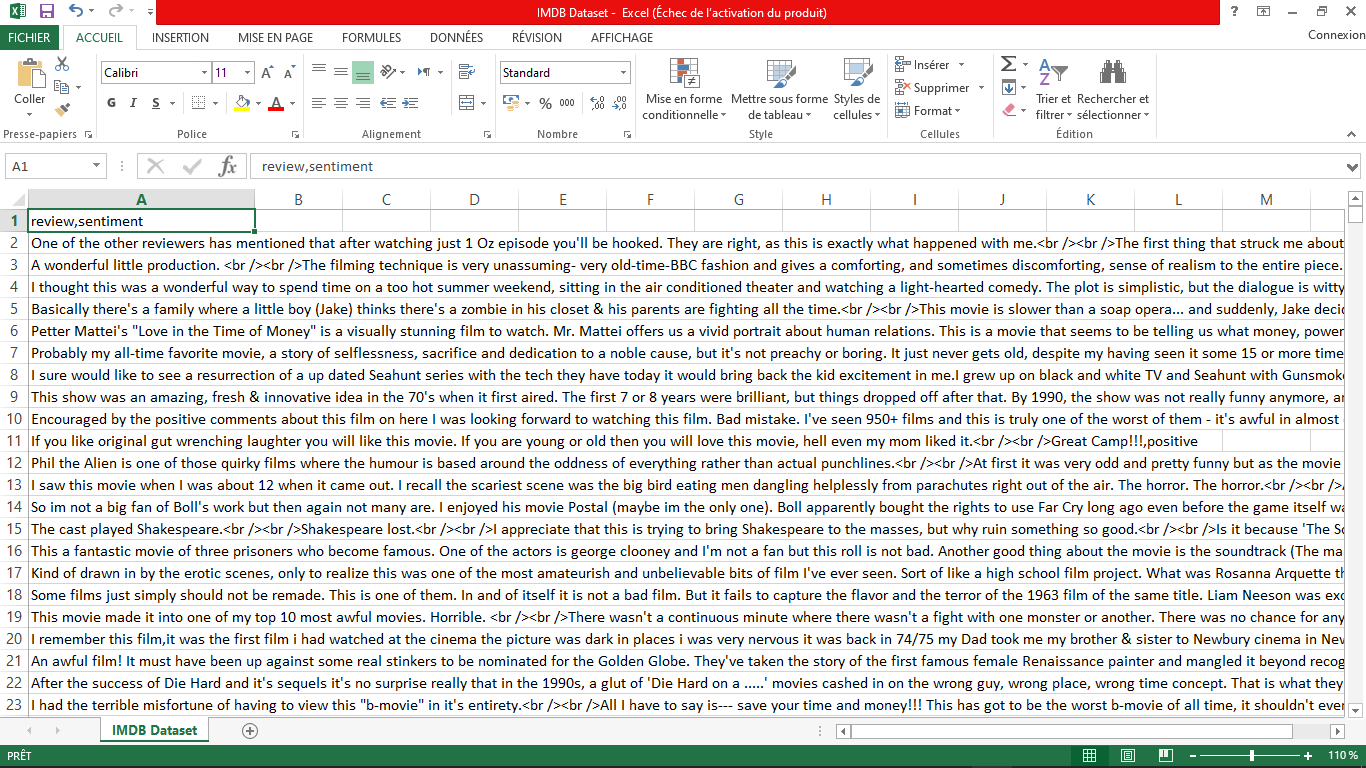
## Word Embeddings

Un mot embedding est une représentation apprise pour le texte où les mots qui ont le même sens ont une représentation similaire. C'est cette approche de représenter des mots et des documents qui peut être considérée comme l'une des principales clés de l'apprentissage en profondeur sur les problèmes de traitement du langage naturel.

## Base de données (Dataset):

On a choisi comme base de données (dataset) une liste d’avis sur plusieurs films. On a +49000 avis (positif / négatif) sur les films.

Voici ci-dessous une partie de notre dataset :



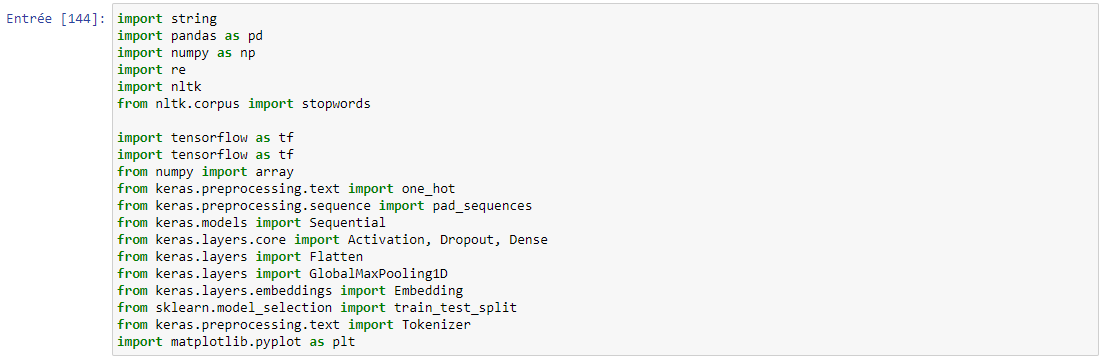
La taille du fichier (‘IMDB Dataset.csv’) est 64mo :



## Réalisation :

### Importation des bibliothèques et des plates-formes

Durant cette phase, on a importé les différentes bibliothèques et plates-formes tels que pandas, tensorflow, numpy…



### Lecture des données

Après, on a fait une lecture de données :



### Prétraitement des données

Dans cette phase, on va traiter nos données (cleaning) en supprimant les ponctuations, les espaces et les différents caractères spéciaux :



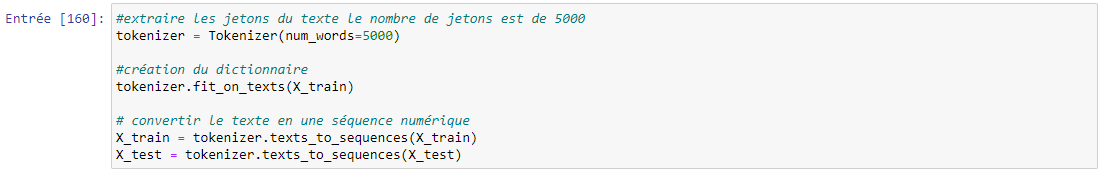
### Fractionnement des données

On a fait un fractionnement ; train de données : 70%, validation : 10% et le test : 20%.



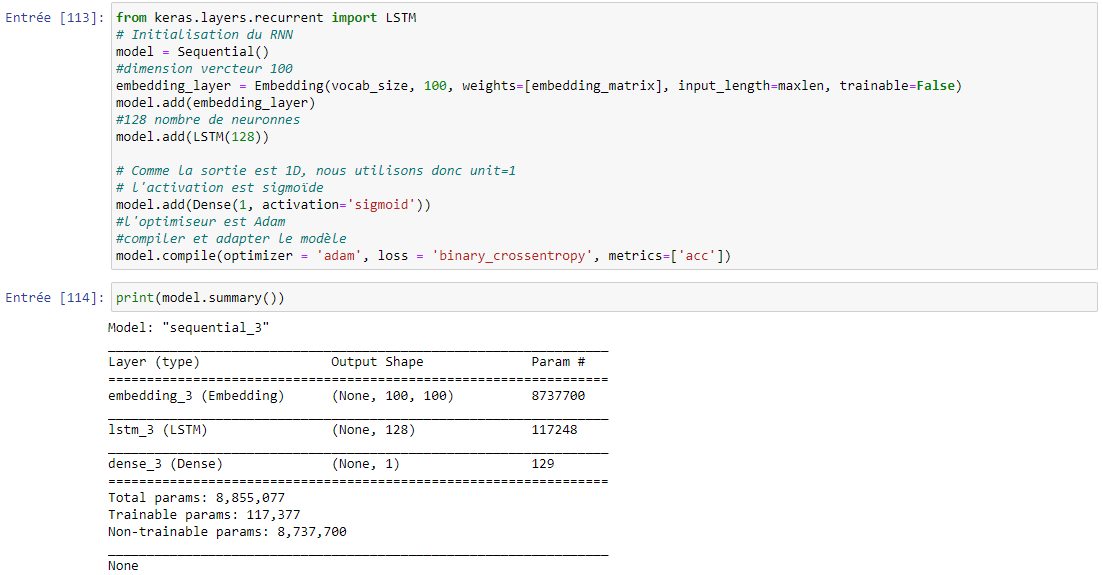
### Tokenisation des données

La tokenisation est le processus de séparation d'un flux de texte en mots, phrases, symboles et d'autres éléments significatifs.

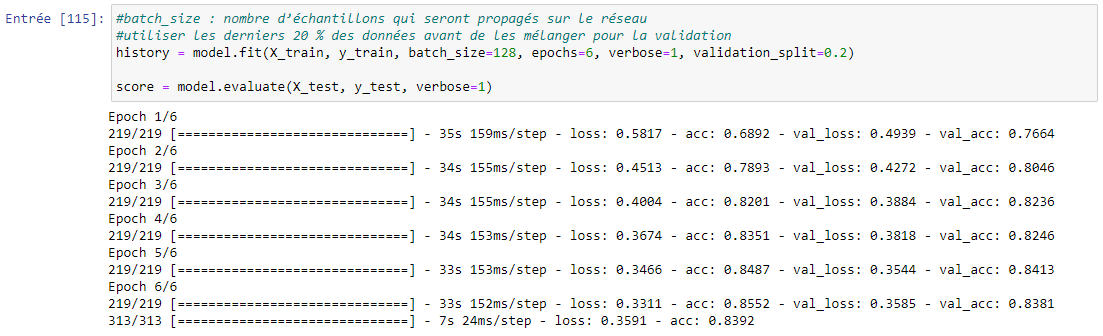




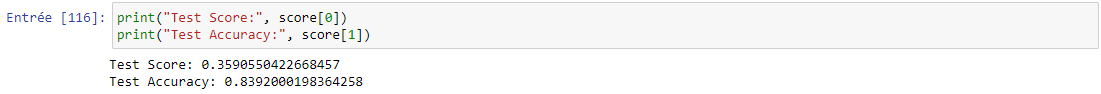
### Construction du modèle



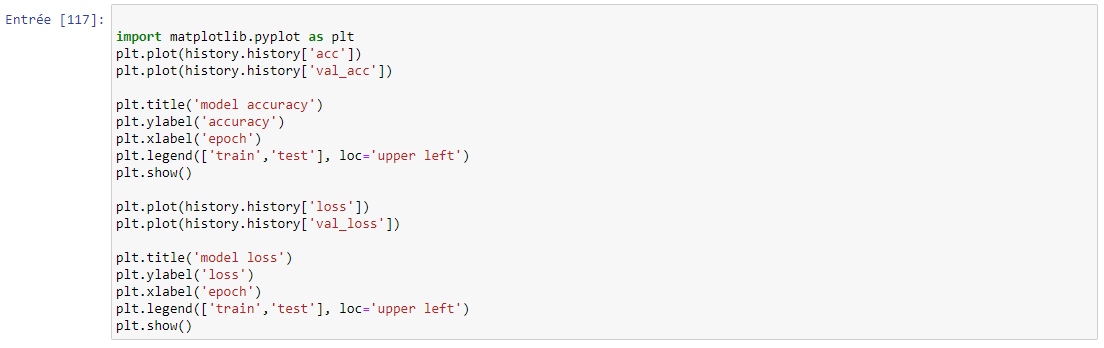
### Formation du modèle

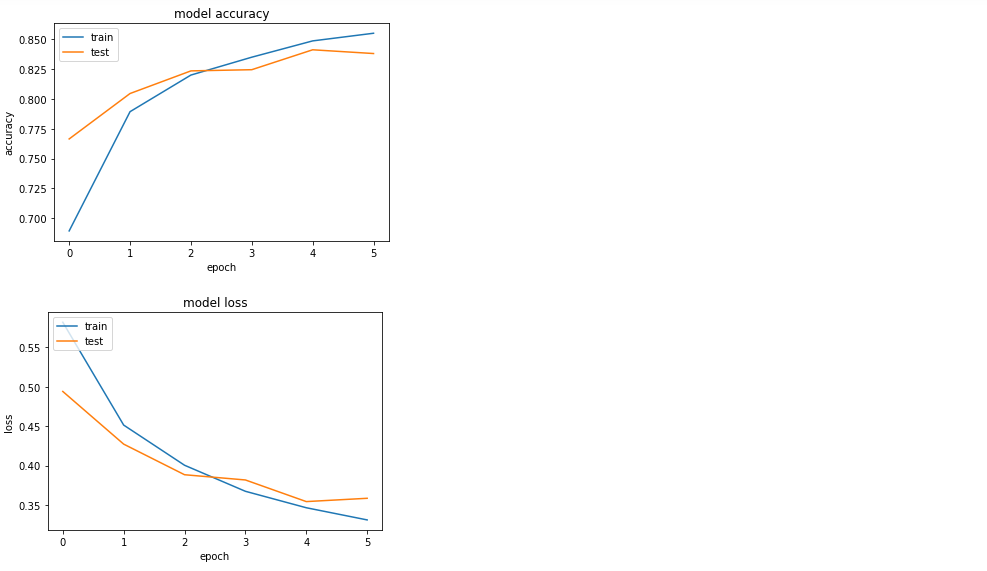


### Score et précision des tests



### Précision et perte du modèle pour l'entraînement et les tests

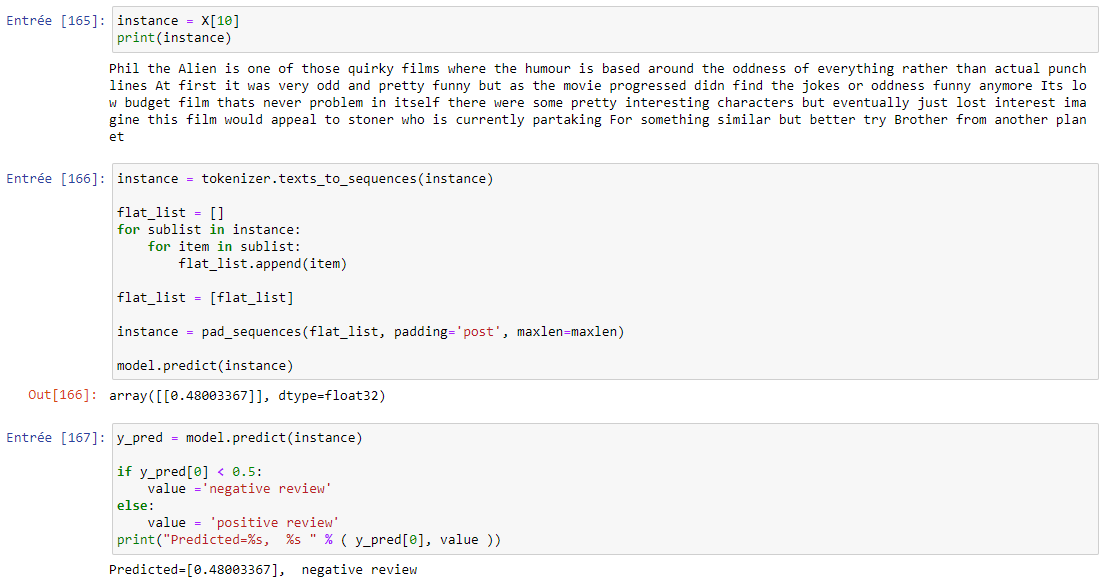




### Test et résultat final :

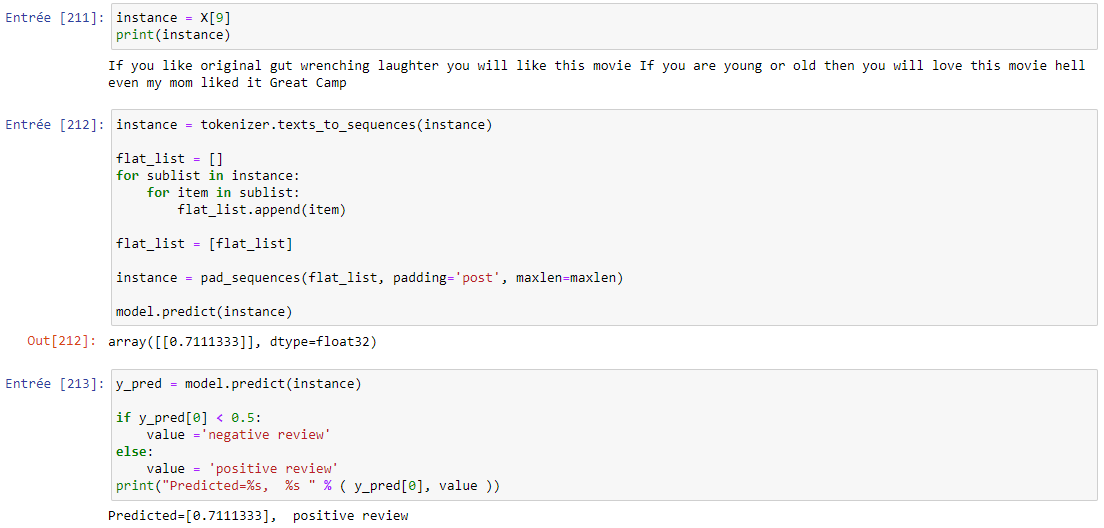
* Résultat Négatif :

On a essayé le 7ème avis de notre nouveau tableau créé X [10] :

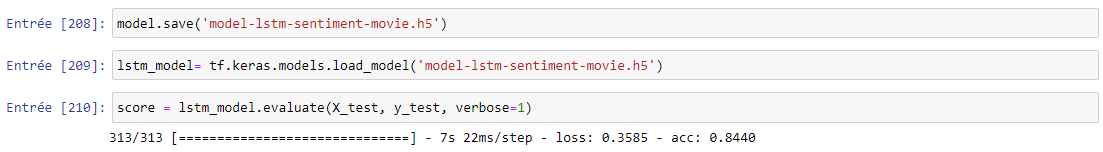


* Résultat Positif :

Maintenant, on a essayé le 9ème avis de notre nouveau tableau créé X [9] :



### Sauvegarde du programme



# CONCLUSION

Durant ce projet nous avons étudié une méthode de reconnaissance textuelle des sentiments. Cette application qui repose sur l’apprentissage profond (Deep Learning), le langage Python et le réseau Long short-term memory (**LSTM**) peut extraire des sentiments depuis des informations textuelles (. Cette reconnaissance textuelle permet d’obtenir une vue d'ensemble sur l'opinion du public au sujet de certains thèmes, cela aide les entreprises à extraire des informations à partir des données du web social.