LOG6308 – TP3 – RAPPORT

Grégoire CHAPEAUX – 2033122 / Hugo CANNEDDU – 2096673

# PARTIE 1.

## Question 2 : Combien y a-t-il d'utilisateurs uniques et de films uniques ?

Nombre de films uniques : 1664

Nombre d’utilisateurs uniques : 943

## Question 3 : Quels sont les avantages et inconvénients d'avoir un nombre de dimension élevé des représentations ?

Plus le nombre de dimensions est élevé, plus le plongement permettra d'extraire des features précises et de dresser un profil de l'utilisateur et de l'objet : un plongement avec un grand nombre de dimensions permettra ainsi d'obtenir un résultat d'autant plus pertinent. Cependant, l'intérêt d'un plongement est de simplifier les entrées du réseau de neurone et de rendre l'apprentissage machine plus facile. Un plongement avec un nombre de dimensions trop important perd tout cet intérêt en ne simplifiant pas l'entrée, voire en la complexifiant.

## Question 4 : Donnez une piste de solution pour complexifier le présent modèle et tirer pleinement profit de l'approche réseau de neurones ? Quelle serait sa principale limite ?

Une piste pour complexifier cette approche serait d'utiliser un réseau plus complexe, constitué de plusieurs couche, avec par exemple une descente de gradients, afin de mettre à jour les poids et de permettre au réseau d'apprendre et de s'actualiser en fonction de ses faiblesses au gré de l'apprentissage.

## Question 5 : Est-ce que le présent réseau de neurones est plus, moins ou également performant par rapport à une approche de factorisation matricielle classique après 1 epoch d'entraînement ?

Après une seule epoch d'entrainement, le réseau de neurones peut être simplement ramené à une approche matricielle, puisque l'on va simplement multiplier les inputs par des coefficients correspondant aux différents neurones. La performance sera donc similaire à une approche de factorisation matricielle classique.

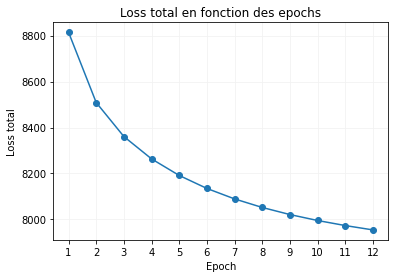
## Question 7 : Entraînez le modèle avec model.fit jusqu’à ce qu’il surpasse de 40% la métrique top\_100\_categorical\_top\_k pour l’entraînement. Après combien d’epochs cela survient-t-il ?

Le modèle dépasse 40% pour la métrique top\_100\_categorical\_top\_k à partir de 10 epochs.

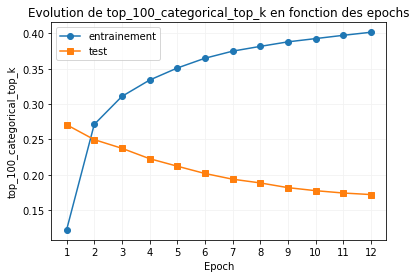
## Question 8 : Si on obtenait des données d'entraînement supplémentaires dans le futur, devrait-on recommencer l'entraînement du modèle sur l'ensemble du corpus augmenté des nouvelles données ? Si oui, pourquoi ? Si non, comment devrait-on s'y prendre ?

Non, il n'y a pas besoin de reprendre l'entraînement à 0. On peut séparer les nouvelles données en données de test et d'entrainement, pour ensuite continuer l'entrainement de notre modèle. Il s'agit finalement de faire une entrainement sur un modèle pré-entrainé. Il faut cependant faire attention à ne pas tomber dans le surapprentissage qui peut survenir peu importe la méthode, et à veiller à ne pas exécuter trop d'epochs d'apprentissage.

## Question 9 : Tracez le graphique du loss total d'entraînement en fonction des epochs.



## Question 10: Tracez, sur un même graphique, l'évolution de la métrique top\_100\_categorical\_top\_k en fonction des epochs pour l'entraînement et le test.



## Question 11: Pourquoi la performance sur les données de test est inférieure à celle sur les données d'entraînement (1 raison)?

La performance sur les données de test est plus faible par rapport aux données d'entrainement en raison d'un surapprentissage : le modèle est "trop entraîné" sur les données de test et donc peu performant sur les données qui n'appartiennent pas à cet ensemble de test.

## Question 12: Est-ce que ces courbes représentent un résultat attendu pour un réseau de neuronne classique ? Pourquoi ? Développez.

Les courbes présentent un résultat inattendu par rapport à un réseau de neurone classique : en effet, bien que l'on observe une diminution du loss lors de l'entrainement, et donc une amélioration de la performance au fur et à mesure que l'on entraine le réseau de neurones, la performance sur les données de test dès la première epoch diminue; on pourrait penser que le réseau est dès la première itération d'entrainement en surapprentissage.

## Question 13: Recommandez les 5 meilleurs films pour l'utilisateur 25.

## Recommandations pour l’utilisateur 25 :

## | Winnie the Pooh and the Blustery Day (1968)

## | Wallace & Gromit: The Best of Aardman Animation (1996)

## | Farewell to Arms, A (1932)

## | Old Man and the Sea, The (1958)

## | Fantasia (1940)

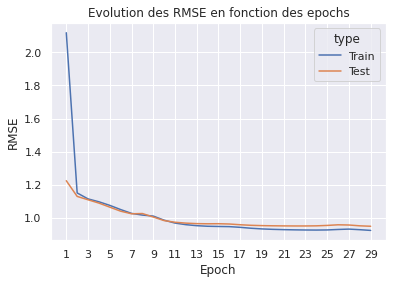
# PARTIE 2.

## Question 3: Entraînez le modèle jusqu'à ce que le RMSE de l'entraînement soit inférieur à 0.925

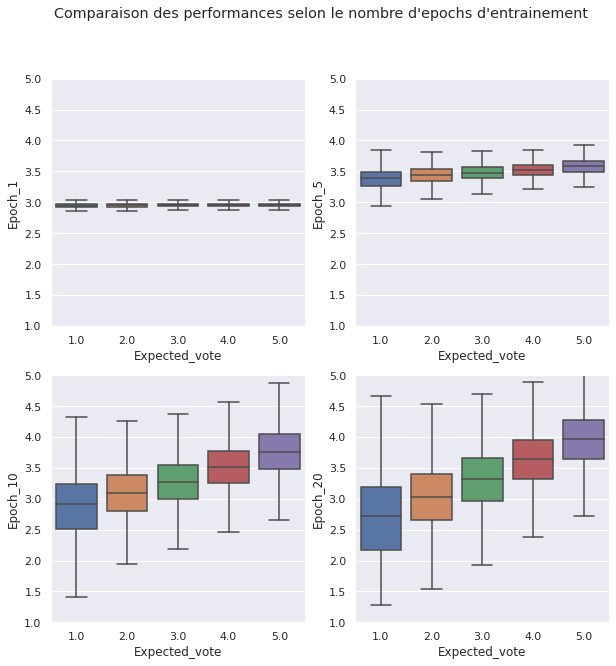
La RMSE devient inférieure à 0.925 au bout de 29 epochs.

## Question 4 : Faites le graphique du RMSE d'entraînement et de test selon le nombre d'epochs. Pour combien d'epochs devraient-on entraîner ce modèle?

## On devrait entraîner ce modèle pour 20 epochs. A partir de 20 epochs, la RMSE ne diminue plus significativement.



## Question 6 : Produisez un boxplot de la prédiction des votes après 1, 5, 10 et 20 epochs.



On observe que pour un petit nombre d'epoch, la prédiction varie peu : en particulier, pour une seule epochs autour de 3 (le vote moyen standard, comme les votes varient entre 1 et 5), et pour 5 epochs autour de 3.5 (la moyenne sur l'ensemble d'entraînement). Cependant, à partir de 5 epochs, on remarque que les prédictions commencent à se différencier selon le vote attendu, et à s'ordonner : plus le vote attendu est grand, plus le vote prédit sera plus grand que le vote moyen, et inversement. En d'autres termes, on constate que les votes prédits se différencient et se précisent pour correspondre aux votes attendus. Cependant, il faut garder en tête qu'un trop grand nombre d'epochs pourrait mener à du surapprentissage.

## Question 7 : Compilez, dans un dataframe Pandas, les moyennes ainsi que les ecart-types des predictions pour chaque modèle et chaque vote

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Moy\_1 | Moy\_2 | Moy\_3 | Moy\_4 | Moy\_5 | Ec\_1 | Ec\_2 | Ec\_3 | Ec\_4 | Ec\_5 |
| 1 ep | 2.946 | 2.951 | 2.954 | 2.956 | 2.958 | 0.033 | 0.031 | 0.031 | 0.031 | 0.031 |
| 5 ep | 3.294 | 3.424 | 3.471 | 3.526 | 3.578 | 0.345 | 0.180 | 0.147 | 0.135 | 0.146 |
| 10 ep | 2.816 | 3.089 | 3.277 | 3.511 | 3.750 | 0.630 | 0.457 | 0.413 | 0.402 | 0.425 |
| 20 ep | 2.681 | 3.024 | 3.305 | 3.624 | 3.936 | 0.713 | 0.570 | 0.514 | 0.480 | 0.486 |

## Question 8 : En vous servant de la distance cosinus, effectuez un calcul de similarité entre les embeddings des films pour retrouver les 5 films les plus semblables à 1) Pulp Fiction (id 3) 2) Silence of the Lambs (id 43) et 3) 2001: A Space Odyssey (id 264). Affichez les résultats.

5 films les plus semblables à Pulp Fiction (id 3)

| Silence of the Lambs, The (1991) - 0.68  
| Killing Fields, The (1984) - 0.62  
| Diva (1981) - 0.61  
| American in Paris, An (1951) - 0.60  
| Raiders of the Lost Ark (1981) - 0.58

= = =

5 films les plus semblables à Silence of the Lambs (id 43)

| Pulp Fiction (1994) - 0.68  
| Casablanca (1942) - 0.63  
| American in Paris, An (1951) - 0.61  
| Garden of Finzi-Contini, The (Giardino dei Finzi-Contini, Il) (1970) - 0.59  
| Once Upon a Time in the West (1969) - 0.58

= = =

5 films les plus semblables à 2001: A Space Odyssey (id 264)

| Paradise Road (1997) - 0.61  
| Down by Law (1986) - 0.59  
| Mad Dog Time (1996) - 0.57  
| Wizard of Oz, The (1939) - 0.55  
| Oscar & Lucinda (1997) - 0.53