

## Test TP - Travail à Domicile

---

Deux travaux. 12 points. Dernier délai : 25/09/2020.

---

### Travail 1 : 7 pts

- 1) Aller sur le site imdb : <https://www.imdb.com/>
- 2) Rechercher ou sélectionner un film de votre choix. Aller sur l'onglet User Reviews (avis utilisateur) du film (en haut) et récupérer 10 avis utilisateur. Stocker chaque avis dans son propre fichier .txt sur votre machine.
- 3) Ecrire un programme python qui permet de :
  - a) Lire un fichier-avis de votre choix et afficher son contenu.
  - b) Calculer et afficher la fréquence de chacun des mots dans chaque fichier.
  - c) Calculer et afficher les 5 mots les plus similaires qu'on retrouve dans tous les fichiers.
  - d) Construire et afficher un index de tous les mots par leurs lemmes (mots dans les 10 fichiers). Vous pouvez trouver la liste des suffixes et préfixes les plus communs en anglais ici :  
<http://www.uefap.com/vocab/build/building.htm#:~:text=The%20most%20common%20prefixes%20used%20to%20form%20new%20nouns%20in,%2C%20%2Dage%2C%20%2Dery.>  
Ou là : <https://examples.yourdictionary.com/list-of-suffixes-and-suffix-examples.html>
  - e) Calculer et afficher pour chaque fichier le pourcentage des stopwords.  
Un stopwords, mot vide en français, est un mot non significatif figurant dans un texte. Un mot qui est tellement commun qu'il est inutile de l'indexer ou de l'utiliser dans une recherche. En anglais, des mots vides évidents pourraient être 'I', 'the', 'a', 'our', 'will', 'an', 'as', etc. La liste des stopwords peut être trouvée ici : <https://gist.github.com/sebleier/554280>

### Travail 2 : 5 pts

Rédiger un compte rendu expliquant le travail 1 ci-haut.

Structure du compte rendu :

- Introduction
- Objectifs du travail : expliquant les problèmes à résoudre et les questions demandées.
- La méthodologie du travail : la technique utilisé (méthodes et algorithmes).
- Les outils : citer les bibliothèques, frameworks, outils pour réaliser le travail.
- Tests et résultats : présenter des exemples des données en entrée et les résultats après exécution.
- Conclusion.
- Code en annexe : structurer le code et doter le avec des commentaires expliquant les choses importantes.

**Notes :**

- Vous pouvez travailler individuellement ou en binôme.
- Dernier délai pour rendre les deux travaux : 25/09/2020.
- Tous les travaux doivent être remis en un seul fichier compressé (rar ou zip). Le compte rendu en PDF et les codes en fichier .py ou .ipynb. A remettre sur la plateforme elearning du département.
- Tout plagiat et copier-coller seront sanctionnés par une note de 0/12.
- Un autre test sur 8 pts sera programmé ultérieurement.