Sciences
Technologies
Santé

2D

Travaux dirigés

Réseaux informatiques : analyse du chargement d'une page web

Auteur :

GROUPE INSA

Réalisation :

fondation UVED

Fiche d'exercices
Réseaux informatiques : analyse du chargement d'une page web

Ce TD, qui se déroule sur deux séances de 2h chacune, propose d'étudier les échanges réseaux qui prennent place pendant le chargement d'une page Web. À partir de cette analyse, il s'agit d'extraire des enjeux économiques/écologiques/sociaux associés à la technologie Web en général, et à de la presse en ligne en particulier. Pour ce faire, un corpus de textes sera étudié et discuté. Pour conclure, nous utiliserons un outil d'analyse systématique pour articuler l'ensemble des conclusions obtenues.

Séance 1/2 : analyse du chargement d'une page Web

Objectifs :

- Connaître les principes de fonctionnement du protocole HTTP
- Connaître les principes du chargement d'une page Web
- Savoir analyser et interpréter des données issues des réseaux

Consignes générales :

- La partie technique (collecte HAR et programmation) doit être effectuée sur une machine Linux (bureau virtuel 2D Linux)

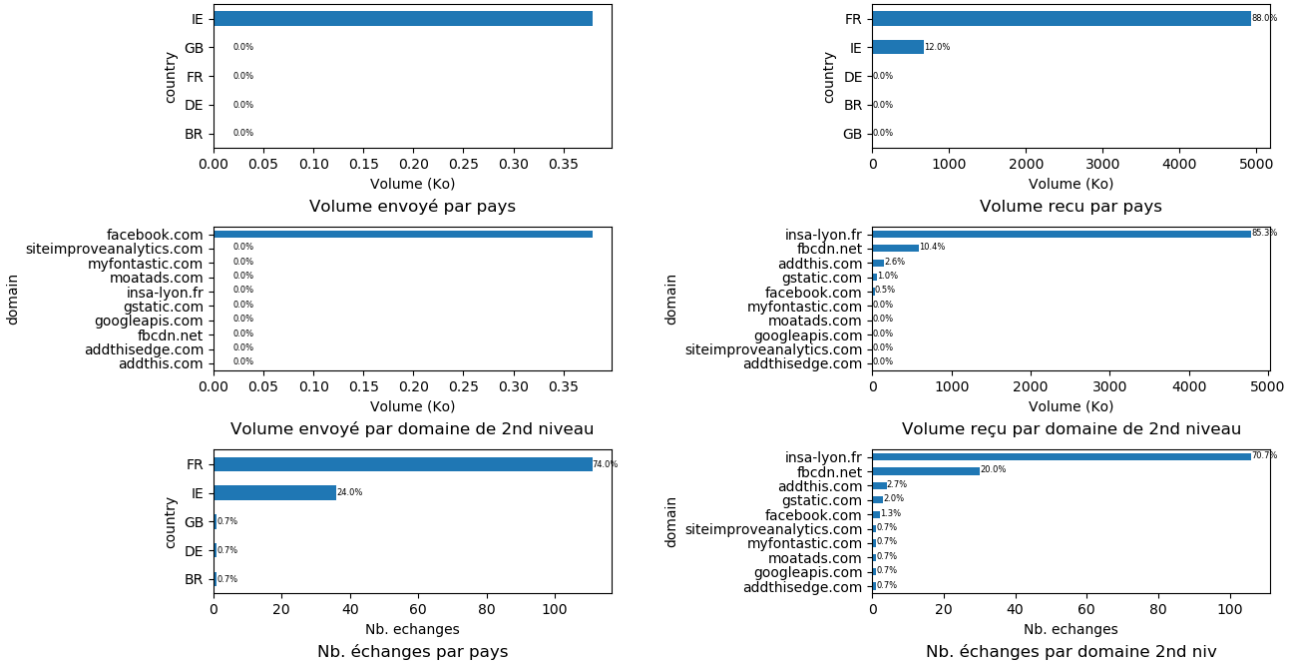


Figure 1 – Exemple de résultats issus de l'analyse d'un fichier HAR issue du chargement du site Web de l'INSA-Lyon.

Exercice 1 : Introduction

Prenez connaissance du fonctionnement des outils qui seront utilisé dans le cadre de ce TP en lisant la section suivante :

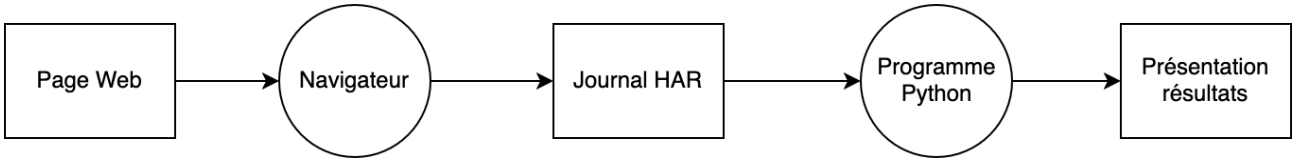


FIGURE 2 – Analyse du chargement d'une page Web à travers différentes étapes.

L'objectif de cette première séance est de produire des graphes qui décrivent les caractéristiques des échanges effectués lors du chargement d'une page Web : nombre de serveurs contactés, localisation et entreprise propriétaire de ces serveurs, ou encore volumes données transférées. La Figure 1 présente un exemple des résultats que nous souhaitons générer.

Pour produire ces résultats, nous allons procéder en plusieurs étapes consécutives :

1. chargement d'une page Web dans un navigateur
2. récupération d'un journal d'évènement réseau HAR
3. analyse du journal par un script Python afin de produire des résultats synthétiques

1.1 Journal HAR

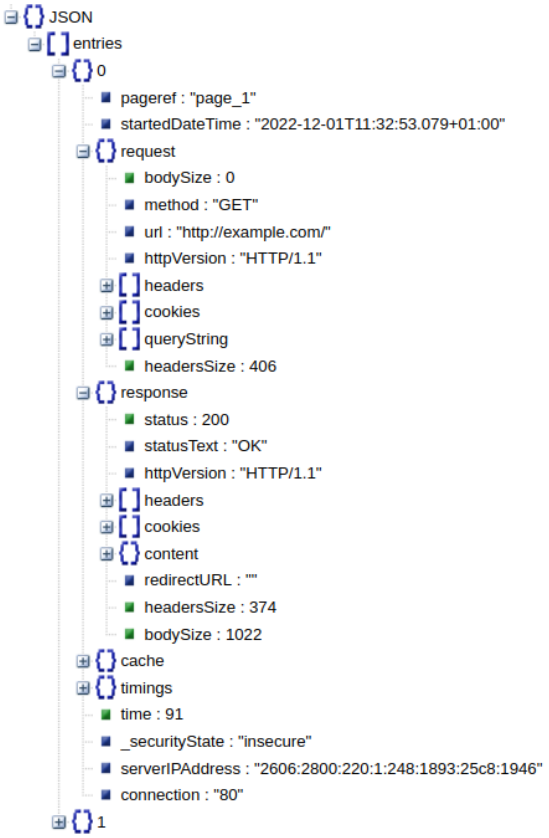


FIGURE 3 – Exemple de contenu de fichier HAR avec une organisation en arbre via le format JSON.

Un journal HAR (pour HTTP Archive) est habituellement stocké dans un fichier avec .har et contient des données sur le chargement d'une page Web dans un navigateur. En particulier, il contient des données sur chacun des échanges réseaux (ex. : requête HTTP) effectués par le navigateur. Un fichier .har contient des données en format JSON, qui permet une représentation en arbre des informations (voir Figure 3). Cette structure est composée d'*entries* (entrées), qui correspondent

à des évènements. Sur l'exemple de la Figure 3, l'entrie 0 a un attribut *serverIPAddress* (adresse IP du serveur) ainsi que des sous-structures *request* et *response* qui correspondent à la requête et la réponse de cet échange. Dans *request* on peut par exemple trouver des attributs comme l'*url* demandé et son *bodySize* (taille du contenu de la requête).

Note : Sauf mention contraire, vous devez désactiver le bloqueur de publicité (uBlock origin, Adblock, ...) de votre navigateur Web (si il en a un).

Exercice 2 : Prise en main des outils

Dans cet exercice, nous allons découvrir et apprendre à manipuler les outils permettant l'analyse des données réseaux et en particulier celles issues d'un fichier HAR. Nous découvrirons comment effectuer des tâches comme décomposer un nom d'hôte, ou obtenir la localisation d'une machine à partir de son adresse IP. Dans cet exercice nous utiliserons le fichier `example.har` qui correspond au chargement de la page `http://example.com`

Cet exercice est à faire dans le fichier `etapes_preliminaires.py`.

2.1 Prérequis

Téléchargez l'ensemble des fichiers fournis et travaillez dans le zip extrait. Ce dossier contient notamment :

1. Le fichier `etapes_preliminaires.py`
2. Le fichier `analyse_fichier_HAR.py`
3. Le fichier `example.har`, contenant le journal HAR du chargement de la page `http://example.com`.
4. Les fichiers `IP2LOCATION-LITE-DB5.BIN` et `IP2LOCATION-LITE-DB5.IPV6.BIN`, contenant des bases de géolocalisation d'IP (à récupérer sur le site `https://lite.ip2location.com/`).

Le code fourni utilise plusieurs autres dépendances à installer **via le terminal de Spyder** :

1. Le paquet `Haralyzer`. Commande d'installation : `pip install haralyzer`
2. Le paquet `IP2Location`. Commande d'installation : `pip install IP2Location`

2.2 Analyse d'un nom réseau (hostname)

Un nom hostname permet d'identifier une machine (ou serveur) et est composé de plusieurs parties : Top Level Domain (TLD) et domaine de second niveau. Par exemple dans le hostname `moodle.insa-lyon.fr` désignant le serveur moodle de l'INSA-Lyon, `.fr` est le TLD et `insa-lyon.fr` est le domaine de second niveau (et non pas `moodle.insa-lyon.fr`).

(2.1) Complétez les fonctions `get_tld` et `get_2nd_lvl_domain` qui prennent toutes les deux en paramètre un hostname (sous la forme d'une chaîne de caractère) et qui retournent respectivement le TLD et le domaine de second niveau.

Note : Pour répondre à la question précédente, vous pourrez vous aider de la fonction split qui permet de décomposer une chaîne de caractère étant donné un séparateur (ici '.').

2.3 Localisation à partir de l'adresse IP

Il est possible d'obtenir la position géographique d'une machine en consultant une base donnée dédiée comme la base IP2Location. La librairie IP2Location stocke les information de géolocalisation sous forme d'enregistrement. La documentation de cette librairie est disponible ici.

Le code fournis contient déjà le chargement des bases de localisation pour les adresses IPV4 et IPV6 :

```
baseIPV4 = IP2Location.IP2Location("IP2LOCATION-LITE-DB5.BIN")
baseIPV6 = IP2Location.IP2Location("IP2LOCATION-LITE-DB5.IPV6.BIN")
```

ainsi que l'initialisation de l'outil `ipTools` :

```
ipTools = IP2Location.IP2LocationIPTools()
```

L'outils `ipTools` (voir sa documentation) intègre une fonction qui permet de tester si une adresse IP est de type IPV4 ou non (donc IPV6) et dont la docstring est la suivante :

```
def is_ipv4(ip):
'''
Vérifie si une adresse IP est de type IPV4
Parameters :
    ip : une adresse IP
Return :
    boolean : True si l'IP est V4, False sinon
'''
```

(2.2) À l'aide de la fonction `is_ipv4` d'`ipTools` écrivez la fonction `affiche_type_adresse` qui affiche le type de l'adresse : IPV4 ou IPV6.

Il est possible d'interroger une base IP2Location avec la fonction `get_all` afin d'obtenir un enregistrement regroupant les information de géolocalisation d'une adresse IP donnée en paramètre. Celle-ci prend en paramètre une adresse IP et retourne un enregistrement :

```
rec = base.get_all(ip)
```

(2.3) Complétez le code de la fonction `get_IP2Loc_record` afin qu'elle renvoie l'enregistrement IP2Location d'une adresse IP passée en paramètre (IPV4 ou IPV6).Pensez à choisir la base adaptée à l'adresse en fonction de son type (IPV4 ou IPV6).

Un enregistrement `rec` d'IP2Location est composé de plusieurs attributs (`latitude`, `longitude`, `country`, `country_short` ...) accessibles via la notation pointée : `rec.latitude`, `rec.country`, ...

(2.4) Complétez le code de la fonction `get_country_code` afin qu'elle retourne le code `country_short` de l'adresse IP passée en paramètre. La signification des codes est disponible ici.

2.4 Ouverture et lecture d'un fichier HAR

Un fichier HAR contient une information structurée dans un format JSON. Il peut être ouvert et lu grâce à la librairie haralyzer. L'information du fichier HAR est organisée en entrées (entries) qui représentent chacune un échange réseau. L'ensemble des entrées peut être parcouru comme une liste à l'aide d'une boucle for. Chaque entrée est composée de plusieurs attributs qui représentent les caractéristiques de l'échange : serveur contacté, volume de données envoyées/reçues, date, ...

Pour cette question, on utilisera le fichier fournis `example.har` correspondant au chargement de la page example.com.

(2.5) Complétez le code pour afficher, pour chaque entrée, l'adresse IP du serveur, son port et son `hostname`.

Note : la liste des attributs d'une entrée et leur noms sont disponible dans la documentation de haralyzer :

https://pypi.org/project/haralyzer/1.9.0/

Note : l'attribut `request` permet d'obtenir des informations complémentaires (dont le nom d'hôte), également décrites dans la documentation.

Exercice 3 : Observation du chargement d'une page Web

Dans cet exercice, nous allons observer le chargement d'une page Web et récupérer des informations sur les échanges réseaux qui ont lieux au cours de ce chargement et qui permettent la récupération des différents éléments composant la page.

Dans un premier temps, nous travaillerons avec la page d'accueil du site 20Minutes : `https://www.20minutes.fr/`.

3.1 Console développeur et journal HAR

Firefox, comme la plupart des navigateurs Web disposent d'une console développeur qui permet d'afficher des informations avancées sur la page Web et son chargement. En particulier, l'onglet réseau présente une liste des ressources qui ont été chargées par le navigateur pour afficher la page (voir Figure 4). Chaque élément de cette liste correspond donc à un élément (image, vidéo, script, feuille de style, police de caractère, icône ...), et apportent des informations sur l'élément : son type, sa taille, sa provenance, etc ... Ce journal peut être exporté dans un format nommé HAR (HTTP Archive format) pour être analysé par un programme externe.

(3.1) Récupérez le journal HAR correspondant au chargement de la page `https://www.20minutes.fr/` en suivant la procédure suivante.

Procédure pour récupérer le journal HAR :

1. Charger la page `https://www.20minutes.fr/`
2. Acceptez les cookies
3. Activer la console développeur : "Ctrl-Shift-I" ou menu en haut à gauche "Outils supplémentaires > Outils de développement web" ("More Tools > Developper Tools" si votre interface est en anglais).
4. Sélectionner l'onglet Network/Réseau
5. Recharger la page avec la touche F5
6. Faire défiler la page tout en bas et attendre que la liste se stabilise (environ 1 minute).
7. Sauvegarder le journal HAR : cliquer sur la roue en haut à droite de la console puis sélectionner "Save all as HAR"

Le journal HAR sera analysé dans le cadre de l'exercice 4.

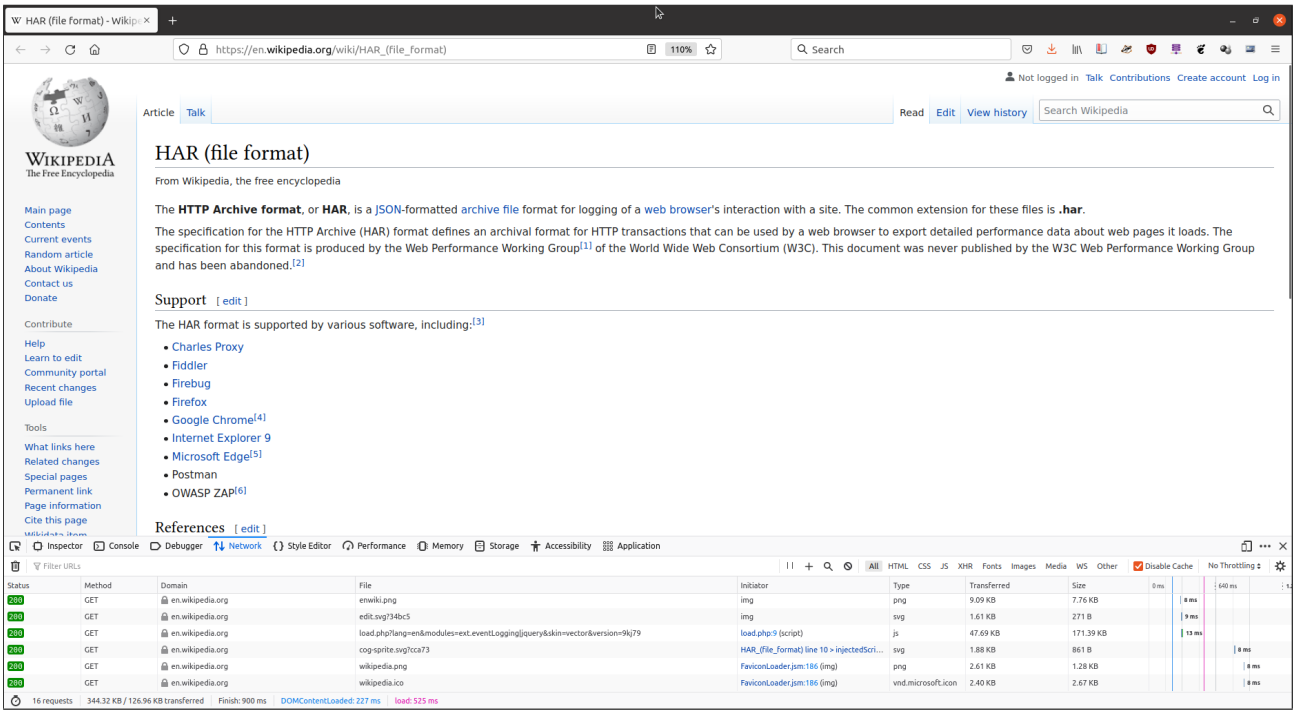


FIGURE 4 – Navigateur Web Firefox avec console développeur activée et affichant l'onglet Réseau/-Network.

Exercice 4 : Analyse des traces de chargement d'une page Web

Nous allons maintenant travailler sur l'analyse du journal HAR obtenu à l'exercice Exercice 3 : afin de générer des résultats comme illustrés sur la Figure 1.

Dans cet exercice, vous devrez travailler dans le fichier `analyse_fichier_HAR.py` fourni. C'est un squelette de code effectuant l'ouverture et l'analyse d'un fichier HAR. Ce programme opère de la manière suivante :

- Ouverture du fichier HAR
- Parcours du fichier HAR (entrée par entrée)
  - Pour chaque échange réseau (entrée) : calcul des informations suivantes : hostname, TLD, domaine de second niveau, taille de la requête et de la réponse, code du pays
- Agrégation des données de tous les échanges et affichage des résultats sous forme de graphes.

Le squelette contient des éléments que vous n'avez pas à modifier :

- Une fonction `plot_data` effectuant la génération des graphes.
- Une boucle parcourant le journal HAR, entrée par entrée, et stockant les données dans une liste.

La fonction `analyse_entry(entry)` est en charge d'analyser chaque échange représenté par la variable `entry`. Votre travail consiste à compléter cette fonction afin de récupérer et/ou calculer les valeurs suivantes :

- hostname
- TLD (Top Level Domain)
- domaine de second niveau
- tailles de la requête et de la réponse
- code du pays où est situé le serveur

Il est fortement recommandé de vous inspirer de l'exercice de prise en main pour répondre aux

questions de cet exercice.

4.1 Travail à faire (dans analyse_entry)

(4.1) A l'aide d'un copier-coller, intégré dans ce fichier les fonctions que vous venez d'écrire dans le cadre de l'exercice 3.

(4.2) Écrivez le code qui récupère la valeur de hostname

(4.3) À partir du hostname, calculez le TLD et le domaine de second niveau

(4.4) Écrivez le code qui permet d'obtenir la taille de la requête et la taille de la réponse (attribut `bodySize`)

(4.5) Écrivez le code qui permet d'obtenir le code du pays dans lequel est situé le serveur (identifié par l'adresse IP).

(4.6) Exécutez votre code et vérifiez qu'il génère bien les graphes.

Exercice 5 : Interprétation des résultats

Dans cet exercice, nous allons analyser le graphique qui a été généré lors de l'exercice précédent.

(5.1) *Lorsque la page a été chargée, combien de domaines tiers (appartenant à des domaines différents de celui de l'hôte principal) ont envoyé des informations ?*

(5.2) *Quelle fraction du volume total de données reçues fait intervenir ces domaines tiers ?*

(5.3) *À quel pays est destinée la page Web chargée ?*

(5.4) *Avec quel pays y a-t-il le plus d'échange ? Quel pays reçoit le volume de données le plus important ?*

(5.5) *Mediasquare.fr fait partie de la liste des domaines tiers. D'après vous, pourquoi est-ce le cas ? Quelles relations économiques ce domaine entretient-il avec 20min.fr ? (faire des recherches sur internet pour répondre à ces questions)*

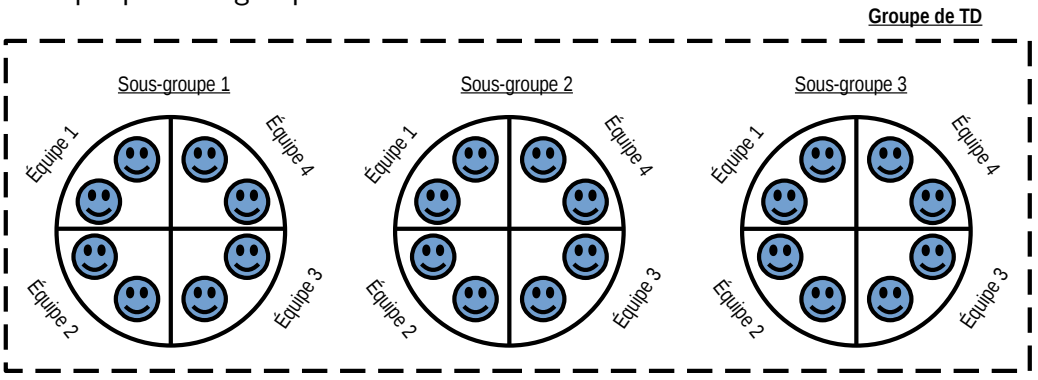
(5.6) *Lorsque vous avez chargé cette page, avez-vous donné votre consentement à quelque chose ? Avez-vous lu le document de « Politique de protection des Données Personnelles » ? Pourquoi ?*

(5.7) *Sachant que l'adresse 20min.fr va être intégrée dans chacune des requêtes effectuées dans le cadre de son chargement, quels serveurs sont au courant que vous avez consulté cette page ?*

Consigne du travail à faire pour la séance prochaine :

1. finir l'Exercice 5 : ; ;
2. former 3 sous-groupes (par groupe de TD) ;
3. au sein de chaque sous-groupe, former 4 équipes numérotées de 1 à 4, de taille aussi homogène que possible. Les membres d'une même équipe liront le(s) même(s) texte(s).

Exemple pour un groupe de TD de 24 étudiants :



Exercice 6 : BONUS : une autre page

Les résultats obtenus concernent une page Web spécifique. Vous avez maintenant les outils pour analyser n'importe quelle page.

(6.1) *Choisissez une autre page Web que vous avez l'habitude de consulter, et générez les résultats pour cette page.*

(6.2) *Constatez-vous des différences ? Si oui, lesquelles ? Comment expliquez-vous ces différences ?*

(6.3) *Refaites la même expérience après avoir activé et installé (si nécessaire) un bloqueur de publicité (ex. : Privacy Badger et/ou Ublock Origin)*

Séance 2/2 : sensibilisation au modèle économique du Web et à ses implications sociales, le cas de la presse en ligne

Pour cette séance, regroupez-vous en sous-groupes.

Objectifs :

- À partir des graphes et analyses produites lors de la séance précédente, identifier une variété d'implications et d'acteurs qui se « cachent » derrière le téléchargement d'une page web.
- Autour de l'exemple de la presse en ligne, appréhender un outil d'analyse systémique pour étudier un phénomène socio-technique numérique et travailler ainsi à l'articulation entre sciences pour l'ingénieur (SPI) et sciences humaines et sociales (SHS)
- Se construire une culture générale dans le domaine du numérique par l'analyse d'extrait de littérature SHS dédiée
- Se rendre compte des enjeux liés à une activité quotidienne, banale

Exercice 7 : Correction de l'Exercice 5 : (15 min)

Exercice 8 : Lecture de textes et réponses aux questions (1h15)

(8.1) *Lire individuellement et attentivement le(s) texte(s) de votre équipe (30min). Chaque équipe lit le(s) même(s) textes(s).*

(8.2) *Au sein d'une même équipe, répondre collectivement aux questions figurant dans l'annexe A relative à votre (vos) texte(s) (15 min).*

(8.3) *Toujours au sein d'une même équipe, identifier l'élément en lien avec votre texte que vous estimez le plus marquant ou le plus important (5min).*

(8.4) *Par sous-groupe, présentez votre texte et l'élément important/marquant que vous avez identifié. Choisissez un seul de ces éléments par sous-groupe. Vous le présenterez à tout le groupe en 2 min (au cours de l'Exercice 10 :), en expliquant votre choix (25 min).*

Exercice 9 : BONUS : textes supplémentaires

(9.1) *Choisir un des textes supplémentaires de l'annexe A, et répondre aux questions associés.*
Attention à ne pas choisir deux fois le même texte au sein d'un même sous-groupe.

Exercice 10 : Restitution orale (30 min)

A Questions relatives aux textes à lire

1.1 Texte de l'équipe 1

Texte [1] : Høstaker, Roar. « L'immatérialité de l'information », Esprit, vol. -a, no. 7-8, 2021, pp. 147-159.

1. De quoi est constitué « le récit de libération » qui s'est développé dans les années 1990 ?
2. Sur quoi repose le récit d'immatérialité de l'information actuel ?
3. Quels problèmes ce récit d'immatérialité engendre-t-il ?
4. Quel lien faites-vous avec les histogrammes produits lors du téléchargement de la page web ?

1.2 Texte de l'équipe 2

Texte [2] : Le Floch, P. & Sonnac, N. (2013). II. Les caractéristiques économiques d'une publication de presse. Dans : Patrick Le Floch éd., Économie de la presse à l'ère numérique (pp. 13-40). Paris : La Découverte. Url.

1. Définir ce qu'est « la désintermédiation de la chaîne de valeur ». En donner un exemple dans le cas de la numérisation.
2. Dans le cas de la numérisation de l'information, peut-on dire que l'étape de diffusion se fait sans intermédiaire ?
3. Expliquer la phrase « On constate in fine qu'il existe moins de barrières à l'entrée ».
4. Comment un contenu éditorial peut-il être « vendu deux fois » ? Est-ce le cas de 20min.fr ?
5. Dans le cas de 20min.fr, qui sont les deux groupes désignés dans la phrase suivante, et quels sont les bénéfices de chacun d'eux « le bénéfice d'un agent d'un des deux groupes dépend positivement du nombre d'agents présents dans l'autre groupe, et inversement. »
6. Ce mécanisme de marché à deux versants est-il spécifique aux plateformes numériques ?
7. L'article parle d'une caractéristique essentielle de la presse diffusée sur papier, structurant le rapport entre employeur et employés. Quelle est cette caractéristique ?
8. La presse en ligne présente-t-elle cette même caractéristique ? Par rapport à la presse diffusée sur papier, comment est-ce que cela affecte le rapport entre employeur et employés ?

1.3 Textes de l'équipe 3

Texte [3a] : Mathilde Aubry et al. « Bonnes feuilles : La nouvelle géographie de la transformation digitale. » The conversation. Publié le 13 mai 2021. Url
Texte [3b] : Dominique Cardon et al. "Mines de rien" Atlas du numérique, Paris : Presses de ScPo, p. 46-47.

1. Pourquoi la géopolitique et le digital sont-ils fortement liés ?
2. Qu'est-ce qui fait la puissance des USA ?
3. En quoi le digital reste-t-il « physique » ?
4. En quoi le numérique perturbe la souveraineté des Etats ?
5. Sur quoi repose la stratégie chinoise en matière de géopolitique numérique et que vise-t-elle ?
6. Pourquoi l'UE est en situation de faiblesse et comment tente-t-elle d'y remédier ?
7. Quel lien faites-vous avec les histogrammes produits lors du téléchargement de la page web ?

1.4 Textes de l'équipe 4

Texte [4] : « Le réchauffement médiatique est publicitaire », in Boullier, D. (2020), Comment sortir de l'emprise des réseaux sociaux, Paris : Le Passeur, p. 39-46.

1. Un marché économique particulier est décrit dans le texte. Quel est le produit ou service vendu ? Qui sont les offreurs ? Qui sont les demandeurs ?
2. Dans quel mesure ce marché est-il nouveau ?
3. À quelle configuration de la typologie de Stackelberg (cf. tableau ci-dessous) le marché décrit dans le texte correspond-il ?

Demandeurs \Offreurs	Un	Quelques	Nombreux
Un	Monopole bilatéral	Monopsone contrarié	Monopsone
Quelques	Monopole contrarié	Oligopole bilatéral	Oligopsone
Nombreux	Monopole	Oligopole	Concurrence

4. Que cherche-t-on à capter chez l'internaute et comment ?
5. Pour quoi faire ?
6. Qu'est-ce que font les machines des traces/données ?
7. Sur quoi les GAFAM fondent-elles leur puissance ?
8. Quel lien faites-vous avec les histogrammes produits lors du téléchargement de la page web ?

1.5 Textes supplémentaires

Texte [5] : France Culture, « L'Irlande, paradis européen des Big Tech ». Diffusée le lundi 3 janvier 2022. Url.

1. Quel est le pays qui accueille 8 des 10 entreprises leader dans le numérique ?
2. Pour quelle raison L'Irlande et le Luxembourg accueillent les entreprises leader dans le numérique ?
3. Quelles conséquences cela a-t-il ?

Texte [6] : « Les revenus du "Monde", des sources diversifiées ». Lemonde.fr. Publié le 26 janvier 2021. Url.

1. Expliquez ce qu'est le modèle d'accès aux contenus « freemium » ?
2. Commenter la phrase suivante « Les dirigeants de la SEM ont toujours considéré que la multiplication des sources de revenus était une garantie d'indépendance. Un socle important d'abonnés protège ainsi le journal contre les risques de pressions de la part d'annonceurs, alors que les revenus publicitaires lui permettent d'investir dans des projets sans mettre davantage ses lecteurs à contribution. »
3. Pourquoi le monde met-il en avant sa charte publicitaire ? Pourquoi y a-t-il des « règles assurant une séparation « sans ambiguïté » entre les contenus rédactionnels et les contenus publicitaires » ? Connaissez-vous des exemples où ces règles ne sont pas respectées ?
4. Qui commercialise les espaces publicitaires sur Lemonde.fr ?
5. Pourquoi Facebook rémunère-t-il Le Monde ?

Texte [7] : CNIL, « Publicité ciblée en ligne : quels enjeux pour la protection des données personnelles ? ». Publié le 14 janvier 2020. Url.

1. Sur quoi est fondée l'économie des données personnelles ?
2. Qu'est-ce que le Real-Time Bidding et à quoi sert-il dans cette économie des données personnelles ?
3. Quels sont les grands enjeux que la CNIL identifie à propos des publicités ciblées ?
4. Le consentement tel qu'il est recueilli aujourd'hui est-il suffisant ? Pourquoi ?
5. Quel lien faites-vous avec les histogrammes produits lors du téléchargement de la page web ?

Ce TD a été mutualisé dans le cadre d'un partenariat entre le Groupe INSA et la Fondation UVED pour la déclinaison disciplinaire de l'enseignement des enjeux de transition écologique et sociétale (TEDS).

Il est mis à disposition selon les termes de la Licence Creative Commons- 4.0 International : Attribution- Pas d'Utilisation Commerciale- Partage dans les Mêmes Conditions

Pour la formation continue ou professionnelle, les modalités d'usage sont à déterminer avec UVED et doivent faire l'objet d'un contrat définissant les conditions d'usage et de commercialisation. Contact : contact@fondation-uved.fr

Première édition : 7 Mars 2025