Carnet de notes - Pontoire Julien

Cours 1 : Histoire du Web

A. Historique du développement du Web

Les quelques dates importantes du développement du Web :

- En **1969**, l'Arpanet est créé aux États-Unis par la Defense Advanced Research Projects Agency (DARPA). Son but est de permettre de maintenir une communication rapide et sans interruption dans le contexte de la guerre froide.
- En **1971**, le premier mail est envoyé. Il contenait simplement les premières lettres du clavier ("QWERTYUIOP")
- En **1981**, le protocole TCP/IP est spécifié.
- En **1983**, il est adopté par Arpanet. C'est la "création" d'Internet.
- En **1990**, le premier site web est mis en ligne par le CERN. Il s'agissait du site https://info.cern.ch/.
- En **1993**, les protocoles du web sont déposés sous licence libre.
- En **1998**, Google naît.

B. Technologies et infrastructure du Web

Pour le développement du premier site web, plusieurs technologies ont été développées.

Tout d'abord le protocole **HTTP**, qui est conçu sous le principe de requête-réponse. L'utilisateur envoie une requête GET au serveur qui lui répond avec les données attendues.

Il a également fallu développer un navigateur web et un serveur web pour pouvoir mettre en place le site web.

Enfin, le langage HTML a également été développé, même si on aurait techniquement pu s'en passer.

Aujourd'hui, on utilise de moins en moins HTTP pour plutôt utiliser HTTPS (Hypertext transfer protocol secure) qui offre un chiffrement des données contrairement à HTTP.

À titre informatif, le port du protocole HTTP est le port 80 et pour le protocole HTTPS il s'agit du port 443.

Lorsque les informations sont transportées d'un point à un autre, elles passent par un ensemble de **routeurs**. Le plus gros hub de routeurs de France se trouve à Marseille.

En 1991, les premières universités ouvrent leurs sites web : Toronto (Canada), Stanford (Californie) puis en France en 1992 (à Lyon).

En 1992, on compte entre 10 et 20 sites web dans le monde. Ce chiffre parait assez faible en 2 ans d'existence du web mais à cette époque il s'agissait d'une technologie nouvelle et assez peu connue. On peut comparer son développement à celui de l'IA : il y a quelques années nous étions très peu avancé en matière d'IA et depuis que chatGPT est arrivé on constate une effervescence de nouvelles IA (StableDiffusion, Sora, …)

Avec la dépôt sous licence libre des protocoles du web en 1993, on assiste à une dissémination du web. À la fin de l'année, on compte 623 pages web.

C. Internet Archive

Internet Archive a été créé en 1996 par Brewster Kahle. Son but est de conserver des copies de pages et de médias présents sur le web.

Cette collecte se fait de manière automatisée grâce à des robots appelés des *crawleurs* qui parcourent les pages et les enregistrent.

Cette archive contient 835 milliards de pages web, 44 millions de livres et textes, 15 millions d'enregistrements audio, 10.6 millions de vidéos, 4.8 millions d'images et 1 million de logiciels.

D. Accès au Web

À cette époque on pouvait accéder au web grâce aux équipements mis à disposition dans les universités et les cybercafés, ou bien avec des connexions à domicile.

Le modem monopolise la ligne téléphonique : si quelqu'un est connecté au Web, plus personne ne peut émettre ou recevoir d'appels sur la même ligne.

On ne peut donc pas rester connecté en permanence sans se couper de la communication téléphonique par ailleurs (à moins de disposer de deux lignes, ce qui est coûteux).

E. Publication sur le Web

Pour publier sur le web, il n'y a pas de nécessité de faire appel à une entreprise spécialisée pour avoir un espace sur leurs serveurs, on peut le faire nous sur notre propre serveur.

Cependant il faut avoir une adresse IP fixe (qui peut être délivrée contre un abonnement annuel par exemple).

On fait face à un problème majeur : il n'y a plus d'adresse IPv4.

On a donc développé les IPv6 mais elles ne sont pas globalement adoptées (problèmes de compatibilités).

Cours 2 : Fonctionnement d'Internet

A. Envoi de requête

Une requête part originellement d'un client, c'est-à-dire un ordinateur qui va demander un service à un autre ordinateur. Il y a différents types de requêtes, par exemple les requêtes GET qui permettent de récupérer le contenu d'une page web ou bien les requêtes POST qui permettent d'envoyer du contenu (d'un formulaire par exemple).

Cette requête va passer par des routeurs, c'est-à-dire des ordinateurs dont le but est d'acheminer des paquets d'informations sur Internet.

Pour pouvoir circuler, cette requête va utiliser le protocole TCP/IP. Le protocole IP gère les communications entre plusieurs machines d'un réseau (basé sur l'adressage, le routage et la fragmentation).

Le protocole TCP lui assure que le transfert de données soit fiable en découpant les données en paquets et en s'assurant que ces paquets soient reçus dans le bon ordre. Il peut également transmettre des paquets perdus.

Les requêtes HTTP sont traitées par le serveur web. Il héberge des documents HTML/CSS/JS.

Le navigateur web permet de faire des requêtes HTTP vers le serveur web et d'interpréter les résultats de ces requêtes pourra afficher un résultat compréhensible pour l'utilisateur.

Une requête HTTP va renvoyer un code de retour pour indiquer le résultat de la requête (200 pour OK, 404 pour quand on cherche à accéder à quelque chose qui n'existe pas, 403 quand on cherche à accéder à un fichier non autorisé, …).

Dans le modèle OSI, le navigateur web se trouve dans la dernière couche (application) et permet de créer une interface pour l'utilisateur.

B. Adressage et nom de domaine

L'adresse IP correspond au numéro d'identification d'une machine. Il y en a 2 types : IPv4 et IPv6.

On peut associer à ces adresses IP des noms de domaine pour ne pas avoir à taper l'adresse IP dans la barre de recherche d'un navigateur web. Les noms de domaine peuvent s'acheter auprès de différents services.

Lorsqu'on va faire notre requête pour une page, on va la faire avec le nom de domaine. Mais avant d'être envoyée, une autre requête va être envoyée à un serveur DNS qui a pour but de faire la correspondance entre un nom de domaine et une IP et qui va donc renvoyer l'adresse IP correspondante.

C. Connexion à distance

Il est possible de se connecter à distance à un ordinateur avec certains protocoles tels que Telnet et ssh. Cependant il ne faut pas utiliser Telnet car contrairement à ssh il n’est pas crypté. Pour chaque lettre entrée il va envoyer un paquet non crypté. On peut donc par exemple retrouver le mot de passe de quelqu’un en faisant une analyse de paquets avec des programmes comme tcpdump (SR06).

D. Distinction entre Internet et le Web

Il est important de différencier Internet du Web. Le web utilise le protocole HTTP mais il en existe beaucoup d’autres sur Internet (FTP, IMAP, …).

Internet est un réseau mondial de réseaux informatiques interconnectés et le Web est l'un des nombreux services et applications disponibles sur Internet. C’est un système d'informations hypertexte qui permet aux utilisateurs de consulter et de naviguer à travers des pages web reliées entre elles par des liens hypertextes.

Cours 3 : Lecture et écriture scientifique

A. Principe de réfutabilité

Irréfutable : Cette expression qualifie souvent en langage courant un énoncé toujours vrai, mais en science, c'est plutôt une mauvaise nouvelle que quelque chose soit irréfutable...

Dans les écrits scientifiques, il est intéressant qu'un texte soit réfutable car cela nous donne de la matière pour en parler, que ce soit en bien ou en mal.

Quand on réalise un écrit scientifique il faut appliquer le **principe de réfutabilité**, c'est-à-dire écrire ses textes dans le but qu'ils puissent être contredits empiriquement (avec des données par exemple).

En général pendant nos études on nous a appris une mauvaise façon de réaliser nos écrits. On a tendance à faire des textes longs pour ne dire que très peu de choses et pas très réfutables parce qu'on parle de sujet qu'on ne maîtrise pas trop.

Il faut éviter d'utiliser des énoncés existentiels tels que "X existe" ou encore "dans certaines conditions ceci…" car ce sont des énoncés irréfutables, on ne peut pas décider de la valeur de vérité.

B. Structure d'une publication scientifique

Une publication scientifique est un article de quelques lignes à quelques pages, fait par des spécialistes et pour des spécialistes. Elle est composée d'une problématique à laquelle elle a pour but de répondre. Elle est publiée dans une revue ou un colloque (débat entre plusieurs personnes sur des questions théoriques, scientifiques.), est révisée par ses pairs et est validée par un comité éditorial.

Il est possible d'utiliser un article de blog en tant que référence dans un article ou bien pour l'analyser dans une fiche de lecture. Cependant il est important de vérifier certaines informations avant.

Il faut faire des recherches sur l'auteur, pour savoir si c'est une personne qui parle juste d'un sujet "au hasard" ou si c'est quelqu'un qui connaît vraiment ce sujet.

Il faut vérifier si c'est une source primaire (première personne à apporter cette information) ou si c'est une source secondaire (il ne fait que relayer l'information en changeant la forme). Si c'est une source secondaire, il vaut mieux éviter et retourner à la source car il peut y avoir des risques de pertes d'informations avec une source secondaire.

Quand on fait nos recherches, il faut éviter les recherches superficielles; c'est-à-dire les recherches sur un moteur de recherche classique. Ces moteurs de recherches vont nous renvoyer des résultats avec très souvent très peu de nouvelles informations pour nous. Il est préférable de privilégier des moteurs de recherche comme DuckDuckGo mais ce n'est toujours pas le mieux. On essaye de se servir des moteurs de recherche juste pour trouver des mots clés en rapport avec le sujet.

On peut aussi commencer ses recherches sur Wikipédia qui référence toutes ses sources et donc nous donne énormément de matière à travailler.

On peut faire des recherches plus approfondies sur des moteurs de recherche scientifique (google scholar ou semantic scholar) et ici il faut essayer de trouver un article en rapport avec notre sujet et qui ait un nombre “élevé” de citations (ça va varier en fonction du sujet), ce qui est en général gage de qualité.

On peut également chercher sur des bases de données ou des archives en ligne.

Quand on réalise un document scientifique, il est important de toujours énoncer une problématique (on écrit pour une raison), ensuite il faut faire un état de l’art (dire ce qui est connu sur le sujet (en citant ses sources) et enfin il faut apporter une contribution à ce sujet, pour ne pas juste reformuler ce qu’énoncent les précédents ouvrages.

Attention ! Il faut toujours respecter la problématique. Si on commence à s’en défaire, soit il faut arrêter d’écrire soit il faut changer la problématique. Il est également important de ne pas formuler trop d’opinion.

Cours 4 : Redécentralisation d'Internet

La création d'Arpanet (l'ancêtre d'Internet) en 1969 avait pour but de décentraliser les moyens de communication des États-Unis dans le contexte de la guerre froide pour éviter une panne de communication suite à une attaque de la Russie sur les postes de communication. Cette précaution permet donc de garantir de pouvoir riposter à une attaque.

Cette création a également été influencée par la culture hippie (principalement sur la côte ouest des États-Unis, à San Francisco) dans le but de dénoncer les pouvoirs de l'État.

Cependant nous faisons face depuis les années 2000-2010 à une recentralisation autour des géants du web dont notamment les GAFAM.

Cette recentralisation provoque de nombreux problèmes et ont des impacts massifs, notamment lorsqu'un de ces services n'est plus disponible comme pour le #googledown en 2020 lors duquel Sciences Po avait reporté les délais pour rendre les copies.

Le fait de recentraliser est donc un problème puisqu'une panne d'un de ces services suffit à ébranler totalement notre société. Il est donc nécessaire d'essayer de redécentraliser Internet.

C'est la tâche que s'est lancé Framasoft depuis 2014. Avec une soixantaine de services, Framasoft permet de réduire ce phénomène de recentralisation.

Cependant, vers 2016, "trop" de personnes utilisaient les services de Framasoft et de ce fait on était de nouveau en face d'une recentralisation. L'association a donc décidé de lancer le collectif du chaton, c'est-à-dire de proposer à n'importe qui de copier ce qu'à fait Framasoft dans le but d'augmenter le nombre de services et réduire l'influence des géants du Web.

Cours 5 : Culture libre

A. Introduction au droit d’auteur

Le droit d’auteur s’applique automatiquement dès qu’on produit quelque chose à condition que ce soit quelque chose d’original (pas de plagiat !) et que la création soit mise en forme (les idées ne sont pas protégées).

Les lois sur ce qu’on peut faire et ce qu’on ne peut pas faire en matière de droit d’auteur sont assez floues.

Un professeur peut diffuser un court extrait d’un film dans un amphi (exception du contexte pédagogique et citation courte) mais il n’y a pas de temps défini pour un “court extrait” (par exemple 3 min c’est trop long).

Quand une œuvre tombe dans le domaine public, elle appartient toujours aux ayants droits mais on peut la diffuser librement. Par exemple, pour un livre une fois qu’il est tombé dans le domaine public, n'importe quel éditeur peut l’imprimer et le vendre.

En Europe, une œuvre tombe dans le domaine public au bout de 70 ans.

B. Évolution et et aspect technique

Le droit d’auteur est ratifié en 1791 par l’Assemblée Constituante.

En 1886, elle est harmonisée au niveau international avec la convention de Berne.

En 1948, la durée pour qu’une œuvre tombe dans le domaine public passe à 50 ans et en 1993 à 70 ans.

À noter, le droit d’auteur n’a rien à voir avec les brevets. Le droit d’auteur protège le texte alors que le brevet protège l’idée du texte.

C. Droits des créateurs et licences

Les élèves et les étudiants possèdent les droits sur les œuvres qu'ils créent au cours de leurs études.

La mention “copyright” n’a pas d’intérêt en France (elle permet juste d’indiquer qu’on en est l’auteur).

Le droit moral est inaliénable et perpétuel. Il s’applique sans limite de temps et comprend le droit à la paternité (on est obligé de respecter l’intention de l’auteur). On ne peut pas par exemple utiliser un extrait d’un discours en le coupant pour faire dire l’inverse.

Le droit d'auteur empêche la libre réutilisation des œuvres mais il est possible d’attribuer une licence à son œuvre qui va apposer certaines règles à la réutilisation de cette dernière.

Les licences les plus utilisées sont les licences CC (Creative Commons).

Important à noter sur les licences :

On peut faire une citation d'un article payant même si on ne l'a pas acheté, cependant on ne peut pas le reproduire.

La licence CC BY-ND est à éviter, elle interdit par exemple de recadrer une image car cela est considéré comme une modification.

Les licences les plus conseillées à utiliser sont CC BY et CC BY-SA.

Il est également possible de "créer" sa propre licence, comme avec le cas de la *do what the fuck you want to public license*.

Disney avait réussi à rallonger la durée avant que le personnage de Mickey ne tombe dans le domaine public. Cependant ça a fini par arriver le 1er janvier 2024 après 95 ans et dès le lendemain un trailer pour un film d'horreur centré sur le personnage de Mickey était diffusé sur Internet (Mickey's Mouse Trap).

Point sur le TD :

J'ai trouvé ce TD assez intéressant parce qu'il nous a permis de découvrir le site OpenStreetMap, qu'on pourrait qualifier de "Google Maps open source".

D'un point de vue personnel, OpenStreetMap va sûrement m'être utile pour réaliser un projet personnel. Je comptais utiliser l'API de Google Maps pour pouvoir gérer des adresses sur une application mais je vais plutôt essayer de me tourner vers OpenStreetMap ou une autre alternative open source.

Cours 6 : Le modèle économique des grandes plateformes (Capitalisme de surveillance I)

A. Le pognon de dingue des GAFAM

Entre 2010 et 2021, la capitalisation boursière des GAFAM a explosé et leur chiffre d'affaires également. La plupart de leurs investissements se concentrent maintenant autour de l'IA et des produits dont on est sûrs qu'ils sont rentables.

Les GAFAM ont deux types de clients : les entreprises spécialisées dans l'achat/revente de données (data brokers) et des organisations souhaitant cibler des personnes selon leur profil (par exemple, un vendeur de casque VR va vouloir cibler un certain type de population).

Mais il ne faut pas sous-estimer les autres sources de revenus des GAFAM (par exemple pour Apple c'est la vente de ses produits qui lui rapporte la majeure partie de son CA).

B. Mais d'où vient le pognon ?

Les data brokers sont des entreprises qui vont stocker le maximum de données sur le maximum de personnes. Ce ne sont pas forcément des données numériques mais tout type de donnée (adresse IP, postale, identifiant des différents appareils, données de cartes de fidélité, souscriptions à des magazines, …).

Ensuite une entreprise qui souhaite faire de la publicité ciblée peut contacter un data broker pour savoir qui cibler et ensuite démarrer les démarches de publicités auprès d'entreprises comme Meta par exemple.

Le profilage et le ciblage peuvent être utilisés dans différents cas :

- Ils peuvent être utilisés dans un but commercial, où les GAFAM vont se placer en intermédiaire entre les entreprises qui cherchent à faire de la publicité et les consommateurs dans le but de cibler des profils spécifiques.
- Ils peuvent également être utilisés pour faire de la communication ciblée, par exemple pour recruter un certain type de personnes pour participer à une enquête ou encore pour faire des campagnes politiques lors des élections.

Point sur le TD :

En TD nous avons parlé du réseau social Mastodon. Personnellement je ne suis pas un grand fan des réseaux sociaux dans la manière traditionnelle de les utiliser. J'utilisais pas mal Twitter anciennement mais j'ai totalement arrêté il y a un an car je trouve que les réseaux sociaux créent une atmosphère anxiogène et un besoin constant d'être informé sur tout ce qui se passe. Cependant j'ai tout de même compris comment il pouvait être utile pour partager des contenus scientifiques.

Cours 7 : Anatomie de l'économie de l'attention (Capitalisme de surveillance II)

A. Générer des données brutes

Pour générer des données brutes, il faut réussir à capter l'attention du consommateur. Pour cela il faut réussir à créer une habitude chez ce dernier. Cette habitude va se créer en 4 étapes.

La première est de créer un **Trigger**, un déclencheur externe (une notification par exemple qui va attirer le consommateur).

La seconde est de rendre **l'action** de l'utilisateur la plus simple possible (cliquer sur une notification, scroller sur un fil, …).

La troisième est de lui offrir une **récompense variable** avec du contenu intéressant ou inintéressant. On peut rapprocher cette étape à la boîte de Skinner, qui avait montré que les oiseaux étaient beaucoup plus intéressés pour avoir de la nourriture quand cette dernière est variable plutôt que lorsque c'est toujours le même type de nourriture.

Enfin la dernière étape est celle de créer un **investissement chez l'utilisateur**, que ces actions améliore le service lors de sa prochaine visite (par exemple s'abonner à des comptes qu'il apprécie ou bloquer ceux qui ne l'intéressent pas).

Une fois que cette habitude est créée, il faut réussir à l'exploiter : on crée de l'inertie avec le scroll infini ou encore les recommandations, on renforce l'habitude en créant de la peur pour les utilisateur de rater quelque chose d'important (FOMO) et on augmente les interactions entre les utilisateurs (possibilité de liker / commenter des publications par exemple).

B. Capter les données libres

Pour collecter les informations des utilisateurs, les entreprises disposent de deux principaux moyens.

Le premier est l'utilisation de **cookies**. Ce sont des petits fichiers stockés en local sur l'appareil de l'utilisateur. Ces fichiers sont utilisés pour stocker des informations de l'utilisateur dans un premier temps dans le but de faciliter l'expérience de l'utilisateur sur le site. Cependant les entreprises se servent aussi de ces cookies pour récupérer des informations sur les utilisateurs. C'est notamment grâce aux cookies que Meta arrivait à récolter des informations sur des personnes qui n'avaient pas de compte Facebook.

L'autre moyen est l'utilisation de **Pixels**. Les pixels sont des petites images invisibles sur une page web qui lorsqu'ils sont chargés sur une page envoient une information au serveur indiquant que la page a été consultée.

Les pixels sont par exemple utilisés par le site suicide france. Dans l'absolu ce n'est pas un problème que ce site utilise un pixel pour recueillir des informations sur ses utilisateurs mais le problème est que ces informations sont ensuite renvoyées sur les serveurs de la

compagnie ayant créé le pixel, et c'est personnes n'ont aucune raison valable d'avoir accès à une telle donnée.

Pour obtenir des statistiques sur la consultation d'un site, il est ensuite possible d'utiliser Google Analytics. Cet outil utilise les cookies et les pixels pour collecter des infos sur les personnes consultant le site web.

En 2021, sur les 20000 applications médicales et de santé analysées, la plupart utilisaient ces techniques de collecte de données, ce qui signifie que des données personnelles / confidentielles tombent entre les mains des géants du numérique.

C. Extraire l'information

Aujourd'hui avec toutes ces techniques une quantité inquantifiable est récoltée et de ce fait il est impossible de la traiter à la main. C'est pourquoi on utilise le Machine Learning pour y parvenir.

Le machine learning est une branche de l'IA dans laquelle l'IA est programmée pour apprendre par elle-même en lui fournissant des jeux de données. Elle va ensuite être capable de traiter des données qu'on va lui fournir en utilisant ce qu'elle a appris des jeux de données d'entraînement.

Aujourd'hui le Machine Learning est très développé. À partir des likes Facebook il est par exemple possible de déterminer l'ethnie de la personne, son genre, ou encore son orientation sexuelle qui sont des données personnelles qui ne devraient pas tomber entre les mains des géants du net.

Une étude a également montré en 2011 qu'il était possible de déterminer à 80% l'émotion / l'humeur d'une personne à sa façon de taper sur son clavier.

Point sur le TD:

En TD nous avons fait un atelier **« Questions éthiques autour du Web »** au cours duquel nous devions nous informer sur un sujet pour ensuite le présenter aux autres.

J'ai trouvé le sujet assez intéressant et le format était très pédagogue.

Le sujet sur lequel je devais travailler était les logiciels libres. C'est un sujet sur lequel j'avais déjà quelques connaissances de base mais que j'ai pu étoffer.

Cours 8 : Surveillance étatique et privée (Capitalisme de surveillance III)

A. Les états face aux plateformes

Les États profitent également de cette possibilité de récolter des données massivement.

La NSA (National Security Agency) avait fait des partenariats avec les géants du numérique pour pouvoir collecter différentes données de personnes vivant en dehors des États-Unis à l'aide du logiciel PRISM.

À l'opposée en Europe, l'accent est mis sur la protection des données avec le Règlement Général sur la Protection des Données (RGPD). Mis en application en 2018, il permet de garantir la protection des données personnelles numériques des européens et des personnes se trouvant en Europe.

Cependant cette mise en avant de la protection est nuancée. Europol par exemple s'oppose au chiffrement de bout en bout qui permet de garantir que seuls la personne qui envoie le message et celle qui le reçoit puisse accéder au contenu du message.

En Australie, une loi a été votée qui force les entreprises à casser le chiffrement de bout en bout.

B. La surveillance dans l'espace public et privé

La vidéo surveillance dans les lieux publics est généralisée, l'utilisation de drones est augmentée, la reconnaissance faciale est également développée avec la détection automatisée de comportements "suspects", de bruits "suspects".

Les communes peuvent aujourd'hui faire financer leurs installations de surveillance à hauteur de 50%, ce qui les encourage grandement à adopter ces technologies.

En 2015, les forces de l'ordre ont acquis en secret un logiciel d'analyse d'images de vidéosurveillance de la société israélienne Briefcam. Elles ont obligé le maire de Brest à mettre des caméras équipées de cette technologie dans sa ville.

On note également une effervescence des technologies de surveillance autour des événements importants. Pour les Jeux Olympiques de Paris 2024 par exemple, des caméras ont été installées dans plusieurs endroits stratégiques de la ville en faisant passer cette installation pour une "expérimentation vidéoprotection augmentée".

Pour finir, ces technologies sont également utilisées par certaines enseignes comme la Fnac pour surveiller leurs clients et éviter les vols, par exemple avec la détection d'un bras qui se rapproche du sac à dos qui pourrait s'apparenter à une tentative de cacher un objet à l'intérieur pour le voler.

C. La surveillance dans les administrations françaises

Les administrations sociales nous surveillent également. La CAF par exemple possède pour chaque personne un “score de suspicion" qui va être impacté par plusieurs facteurs tels que le fait de disposer de revenus faibles ou encore le fait d’être au chômage.

Cours 9 : Le capitalisme de surveillance (Capitalisme de surveillance IV)

A. Capitalisme et Surveillance

Le capitalisme de surveillance consiste à “capter l'expérience humaine dont on se sert comme d'une matière première pour la transformer en prévisions comportementales monnayables sur un nouveau marché.” (Shoshana Zuboff)

Il est important de noter que la notion de surveillance est souvent vue d’une manière péjorative mais ce n’est pas toujours le cas. Par exemple, la surveillance nous permet d’avoir un suivi continu de la fonte des glaces, ou encore d’avoir des données statistiques sur l’ensoleillement d’une exploitation agricole.

Il est également important de ne pas limiter la surveillance aux GAFAM mais bien de l’étendre à tous les acteurs de surveillance (États, Plateformes et Industriels).

B. Quid de l'utilité et de l'efficience ?

La seule données métrique qui permet de l’évaluer est l’efficience économique.

Il n’est par exemple pas possible de faire une corrélation entre l’existence de dispositifs de vidéosurveillance et une évolution du niveau de délinquance.

Une analyse récente a indiqué que la criminalité a diminué de 24 à 28% dans les rues et les métros mais nous n’avons aucun moyen de savoir si c’est vraiment lié à la vidéosurveillance.

Mais le système socio-économique de la surveillance répond juste à un logique capitaliste et libérale et n’a donc pas besoin de savoir si c’est efficace ou non.

Cours 10 : Surveillance et pouvoir de nuisance (Capitalisme de Surveillance V)

A. Centralisation : structurellement néfaste ?

La centralisation du web autour des GAFAM a permis de faciliter la surveillance.

Cette centralisation crée une absence d'alternative qui nous oblige à subir les conditions générales d'utilisation.

Par exemple Facebook et Reddit peuvent lire nos messages privés, Amazon peut nous traquer lorsqu'on visite d'autres sites.

Les GAFAM multiplient également les partenariats avec les think tanks, centres de recherches, médias ou fondations sociales. Ils ont ainsi une influence sur les décisions des lois.

B. Des effets bien réels sur les populations

Cette centralisation autour des GAFAM leur permet d'avoir accès à des données auxquelles il ne devrait pas avoir accès. Par exemple Meta a accès à des données de santé grâce à son pixel qui est utilisé sur un équivalent de Doctolib.

En 2018 sur youtube, 70% des vidéos visionnées étaient recommandées par l'algorithme. Les utilisateurs ne font plus l'effort de chercher ce qu'ils veulent regarder et préfèrent se laisser guider par les recommandations. Les contenus qui désinforment le plus et qui sont les plus extrêmes sont ceux qui sont le plus recommandés car ceux avec lesquels on interagit le plus. L'algorithme considère alors que ces vidéos sont plus intéressantes.

En 2021 il y a eu une grosse mise à jour pour réduire cette mise en avant du contenu "borderline" mais ces problèmes étaient connus depuis 2011. Youtube n'a agit que lorsque les annonceurs commençaient à ne plus vouloir être associés avec Youtube, ce qui aurait entraîné une grosse perte de revenus.

Les données collectées peuvent également avoir un impact direct sur notre vie. Par exemple, les banques et assurances ou encore les chefs d'entreprise peuvent acheter des données aux data brokers et s'en servir pour discriminer les gens en fonction de leurs problèmes de santé, leurs pauvreté, etc.

C. Données et autoritarisme : danger

Aux États-Unis, certains data brokers revendent les informations des personnes qui sont allés dans des plannings familiaux aux États depuis que l'avortement a été rendu illégal.

Nokia aidait la Russie à interconnecter ses réseaux avec ceux de surveillance de la Russie.

Google a cherché à développer un prototype “Dragonfly” qui pourrait être déployé en Chine (Google interdit là-bas) mais a arrêté le développement après une lettre ouverte de 1400 employés de Google et une autre de Amnesty International.

Jusqu’en février 2024, Amazon avait un partenariat avec la police au sujet de leur visiophone Ring. La police mettait en avant cet outil et en échange Amazon leur fournissait des données. Ce projet a été arrêté sous les pressions.

Ces exemples montrent que les organisations cherchent un équilibre entre l’impact sur l’image publique et l’intérêt au niveau économique.

En 1940, IBM a vendu des machines aux nazis pour leur permettre de recenser les populations juives. Cette collaboration a grandement accéléré le génocide juif, en augmentant le nombre de recensés de 400000 à plus de 2 millions.

D. Capitalisme, surveillance et éthique : l'équation impossible ?

Cependant ces organisations ont également des bonnes facettes. Google par exemple est impliquée dans plusieurs projets humanitaires qui leurs sont déficitaires. Il avait travaillé avec le ministère de la santé américain sur la transmission de maladies par piqûre de moustiques.

Le plus important pour le capitalisme de surveillance n’est pas qu’il fonctionne mais qu’il permette d’en tirer des bénéfices. À l’échelle Macro, les GAFAM sont guidés par les prévisions économiques de leurs actions mais cela n’empêche pas à des gens qui œuvrent pour des bonnes choses d’exister à l’échelle micro.

Cours 11 : Société et surveillance : une question d'échelle (Capitalisme de surveillance VI)

A. Surveillance : pas le choix ?

Le besoin de surveillance est nécessaire dans certains cas. Par exemple, la surveillance nous permet de contrôler l'évolution de la fonte des glaciers, le taux d'humidité dans une plantation, etc.

Le problème viendrait donc de l'aspect capitalisme. Est-ce que ce serait possible d'imaginer une smart-city bienveillante à ce sujet ?

- Pas possible car cela nécessiterait qu'on soit constamment surveillé, ce qui n'est pas envisageable d'un point de vue moral
- Pose également problème au niveau des matières premières et des droits humains : les matières nécessaires sont des matières rares et les personnes chargées de l'extraction (dont parfois des enfants) sont souvent très peu payés.
- Notre vision du monde est alors réduite à ce qui est capté par la surveillance.

B. IA et innovation : un changement de nature

Le déploiement massif des IA a infusé toutes les sphères où le numérique était présent et a totalement changé notre rapport au monde. Par exemple, le déploiement d'IA comme ChatGPT, Dall-E ou encore Sora change radicalement notre rapport au monde. Lorsque les étudiants ont une recherche à faire sur une définition ils vont plus utiliser ChatGPT qui leur permet d'avoir une réponse claire et en une seule recherche plutôt que de faire des recherches sur plusieurs sites et d'ensuite rassembler toutes les informations récoltées.

Microsoft a annoncé que ses nouveaux ordinateurs Microsoft Copilot+ seraient équipés de la fonctionnalité recall qui va enregistrer toutes les quelques secondes une capture de l'écran du pc qui va ensuite être envoyée sur un serveur où elle sera traitée par une IA qui va lui assigner un tag.

Cette fonctionnalité stockerait énormément de nos informations personnelles dans un même endroit ce qui serait assez dangereux d'un point de vue sécurité, si un hacker parvient à avoir accès à cette base de données il aurait accès à beaucoup trop d'informations sensibles. Microsoft a d'ailleurs décidé de repousser la sortie de cette fonction qui aurait normalement dû être disponible à l'heure actuelle pour pouvoir retravailler sur la sécurité de nos données (https://www.lemondeinformatique.fr/actualites/lire-la-cnil-au-rendez-vous-des-amendes-rgpd-94077.html).

Ces nouvelles technologies se font accepter petit à petit par la société et elles deviennent banales. Par exemple, la fonction Recall a fait l'objet de controverse mais elle a tout de même passé les étapes de conception chez Microsoft. À une époque pas si lointaine ce genre de technologie aurait été impensable.

Aujourd’hui on ne cherche plus à progresser mais seulement à “innover”, même si cette innovation n’a pas de réelle utilité.

Enfin, l'IA est imposée, sans demande et sans consentement explicite, comme un mode d'accès « naturel » à l'existence.

Cours 12 : Contre-mesures et autodéfense (Capitalisme de Surveillance VII)

A. Modèle de menace

Il y a différents acteurs dont on peut se protéger :

- de ses proches (segmentation de l'identité numérique, pas qu'ils puissent accéder à certaines infos) (on le traite pas ici)
- de la surveillance publicitaire
- de la surveillance de masse
- de la surveillance ciblée

Comment se protéger ?

B. Surveillance publicitaire

Cas de Google Chrome :

On sait aujourd'hui que Google enregistre ce qu'on tape au clavier, notre historique de navigation et les cookies et pixels sur les sites web.

Pour contrer cela on peut utiliser un navigateur libre comme Firefox et utiliser des extensions comme uBlock Origin, Privacy Badger ou AdAway sur Android.

Pour contrer la revente de données, les prédictions comportementales et les algorithmes d'influence, on peut utiliser des services libres comme ceux du collectif Chatons, Mastodon pour les réseaux sociaux ou encore Invidious et NewPipe en tant que client youtube sans publicité.

Pour ce qui est des CGU abusives, le RGPD permet de réduire leur impact sur nos données en apportant :

- une obligation de préciser les données collectées et les traitements
- une obligation de recueillir de consentement avant la collecte des données
- des droit d'accès, de rectification, d'opposition, de suppression
- des amendes jusqu'à 4% du CA mondial

Cependant ces directives ne sont pas toujours respectées, c'est pourquoi les entreprises ont très souvent une partie de leur budget qui est réservé pour les amendes du RGPD (vu en cours de SR06).

C. Surveillance de masse

Cas de l'envoi de données (mail, sms, etc.) :

Il y a deux types d'informations qui transitent :

- les données : le contenu du message
- les métadonnées : les données sur les données (émetteur, destinataire, date, localisation, etc.)

Les données sont beaucoup plus faciles à sécuriser que les métadonnées. Les métadonnées sont d'ailleurs souvent plus intéressantes à récupérer pour les hackers que les données elles-mêmes car elles permettent de savoir où on se situe, avec qui on discute, etc.

Pour rendre le message inintelligible, on utilise un chiffrement. Il en existe deux dans ce contexte : le chiffrement symétrique et le chiffrement asymétrique (dans d'autres contextes on a aussi le hashage).

Le chiffrement symétrique utilise la même clé secrète pour chiffrer et déchiffrer le contenu du message et utilise un algorithme de chiffrement. Le plus utilisé est AES.

Le chiffrement de César par exemple est un chiffrement symétrique simple. On substitue juste chaque lettre par une autre lettre de l'alphabet avec un décalage constant. Si on fait un décalage de 13 lettres, on va se retrouver avec la même clé pour chiffrer et déchiffrer (décaler de 13 lettres pour chiffrer et pour déchiffrer, on appelle cet algorithme le ROT13).

Le chiffrement asymétrique se base lui sur deux clés, une privé et une publique. Alice va avoir ces deux clés. Elle va envoyer sa clé publique à Bob (pas besoin de l'envoyer de manière sécurisé, c'est une clé publique). Bob va ensuite utiliser la clé publique d'Alice pour chiffrer le message qu'il souhaite lui envoyer. Le seul moyen de déchiffrer le message est alors d'utiliser la clé privée de Alice, qui est la seule à la posséder. Il est cependant techniquement toujours possible pour un hacker de récupérer le contenu du message s'il se place au milieu de la conversation, intercepte l'envoi de la clé publique d'Alice et la remplace par sa clé publique. Ainsi lorsque Bob renverra le message chiffré, le hacker pourra le déchiffrer puisque le message a été chiffré avec sa clé. (via le cours de SR06 sur la cryptanalyse).

Pour le transit des données, il faut du chiffrement de bout en bout où les intermédiaires n'ont pas de moyen de connaître les clés de chiffrement (ex : Signal).

D. Surveillance ciblée

Cas de Windows :

On retrouve énormément de problèmes de sécurité sur Windows :

- le code est fermé est troué de failles
- il y a des leaks de masse de données
- le chiffrement des disques sur Windows est volontairement très faible
- longue histoire de collaboration

La solution est d'utiliser un OS libre comme Linux. Cependant j'ai vu en SR06 que Linux est plus sécurisé parce que moins de gens cherchent à le cracker contrairement à Windows

étant donné que Windows est beaucoup plus utilisé que Linux, et est donc une meilleure source d'informations à pirater.

Cas d'un Linux chiffré :

Il n'y a pas d'anonymat natif sur Linux, on peut utiliser un VPN mais certains d'entre eux collectent et stockent des données même s'ils affirment ne pas le faire.

Une vraie solution est d'utiliser un navigateur comme Tor.

Pour ce qui est de l'attaque Man in the Middle, la solution est d'utiliser des certificats. Durant un projet de SR06 ce semestre, j'ai eu à mettre en place une chaîne de confiance d'autorités de certifications pour sécuriser un serveur web apache2 que nous avions mis en ligne précédemment. Nous avons donc dû créer les différentes autorités de certifications, les certificats et ensuite modifier la configuration du serveur apache. Nous avons également créé une CRL (Certificate Revocation List) pour tester de se connecter avec un certificat révoqué.

E. Bilan

Cependant, malgré toutes ces mesures il reste toujours des risques. Nous ne sommes jamais à l'abri d'une mise à jour d'un logiciel que l'on utilise avec du code malveillant.

Au final, la meilleure façon de se protéger c'est de ne pas avoir de données numériques

Cours 13 : Au temps des IA

A. L’essor des IA

Au XXe siècle, les IA ne marchaient pas du tout et la majorité des gens pensaient que ça ne marcherait jamais.

À cette époque, “The spirit is willing, but the flesh is weak” traduit en russe puis en anglais donnait “The vodka is good, but the meat is rotten”.

Aujourd’hui sur Internet, 50% du trafic global est réalisé par des robots.

Les IA envahissent également les écoles. En 2021 ChatGPT était capable d’obtenir 18.5/20 au final de WE01 et 11/20 au baccalauréat de philosophie. Son utilisation est d’ailleurs encouragée par certains professeurs d’informatique, notamment en ce qui concerne le développement web.

Mais les IA ont un impact assez important sur l’environnement. AWS a par exemple acquis un datacenter en 2024 avec la capacité de production de plusieurs centrales nucléaires.

Aux USA , il était prévu de fermer les usines à charbons mais obligé de les laisser tellement les IA demandent de l'énergie.

B. L’histoire des IA

En 1936, la machine de Turing a été créée. Sa création correspond à la naissance des ordinateurs et la mécanisation du traitement de l’information.

Elle représentait toute l’information sous forme binaire sur une bande et on peut faire des déplacements des éléments de ces bandes pour traiter les infos.

On pourrait tout réduire à ce simple ruban avec des opérations unitaires (mais ça nécessiterait un ruban infini et une vitesse de calcul très puissante). Ce test n’a pas pour but de savoir si une machine est intelligente, mais si elle est capable d’imiter un humain.

En 1950, on retrouve le test de Turing (The imitation game). Son but est de déterminer si une intelligence artificielle est capable de réfléchir comme un être humain. Les captcha sont en quelque sorte un test de Turing inversé (on cherche à s’assurer que ce n’est pas un robot qui agit comme un humain).

La même année, Isaac Asimov a formulé les 3 lois de la robotique qui sont aujourd’hui restaient ancrées dans le fonctionnement des robots :

1. Un robot ne peut porter atteinte à un être humain, ni, en restant passif, permettre qu'un être humain soit exposé au danger.
2. Un robot doit obéir aux ordres qui lui sont donnés par un être humain, sauf si de tels ordres entrent en conflit avec la première loi.
3. Un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.

Le terme IA a été utilisé à partir de 1956 par John McCarthy (l'inventeur du LISP) lors de la Conférence de Dartmouth. À cette époque ce terme faisait référence à des algorithmes du type minmax qui utilisent des arbres de raisonnement. (voir Annexe 1)

Les systèmes experts sont une autre application de l'IA. Ils consistent en une base de faits et une base de règles. Le moteur d'inférence va ensuite faire des comparaisons et des opérations suivant l'ordre du SE (0, 0+, 1). (voir Annexe 2)

On a ensuite vu apparaître les réseaux de neurones. Ces modèles sont inspirés par le fonctionnement du cerveau humain et sont énormément utilisés dans le domaine du Machine Learning.

C. Les IA aujourd'hui

Aujourd'hui il y a deux visions radicalement opposées quant aux IA :

- Les machines ne pensent pas, elles ne font que copier les humains (anthropocentrisme)
- Les machines ne pensent pas, elles ne font pas comme les humains (anthropomorphisme)

Une Intelligence Artificielle possède plus de paramètres qu'un humain ne possède de neurones connectés (autant pour GPT3 et le cerveau humain).

De nos jours, on retrouve des IA pour tout faire : génération de texte, d'images, de vidéos, d'audio, etc.

On retrouve également certaines IA spécialisées pour coder comme Github Copilot qui est énormément utilisé par les développeurs pour réaliser des morceaux de codes génériques (par exemple dans le domaine du développement web lors du développement de tests unitaires).

Les IA peuvent également être utilisées pour tromper d'autres IA. Par exemple, au niveau des entreprises, le tri des CV peut se faire par IA. Il est alors possible d'utiliser une IA pour injecter du code dans le fichier PDF qui va mettre en avant son CV.

Cours 14 : Lowtechisation

A. Introduction

Aujourd'hui on ne sait pas réellement si le télétravail permet de réduire les émissions de gaz à effet de serre. Il permet en effet de réduire l'utilisation des véhicules qui dégagent énormément de ces gaz mais la voiture lors de son utilisation sert par exemple de chauffage à son utilisateur, alors que chez soi on va allumer le chauffage spécialement pour se chauffer.

B. Technosolutionnisme

Le technosolutionnisme correspond à l'essai de l'utilisation de l'ingénierie ou de la technologie pour résoudre un problème (qui a souvent été créé par une ancienne intervention technologique).

Plutôt que de simplement chercher la source de ce problème et essayer de la supprimer, on préfère créer des surcouches pour réduire ces problèmes, mais ces surcouches finiront par avoir des problèmes auxquels on appliquera d'autres surcouches.

On peut émettre différentes critiques au solutionnisme :

- Il empêche de penser le problème différemment (ex : voiture individuelle) (pourquoi il y a ça plutôt que comment le résoudre)
- Non prise en compte de la complexité (ex : monoculture)
- Minoration des effets indirects, ex : effet rebond → gain d'optimisation d'un carburant donc on diminue le coût donc on va plus l'utiliser donc on va autant voir plus consommer (paradoxe de Jevons)
- Mauvaise échelle de temps, si on reconnaît par ailleurs l'urgence d'agir pour inverser les courbes : mettre en place des solutions (neutre en carbone par exemple alors que ce qu'on fait c'est juste replanter des arbres du coup il y a un décalage temporelle de 20-30 ans et cela nécessite également qu'il n'y ait pas de soucis au niveau des arbres plantés)
- Non prise en compte des effets de généralisation, à l'échelle mondiale (ex : nucléaire)
- Logique de pari (si on ne trouve pas de solution, c'est grave) : on va tenter mais on ne peut jamais être sûr, d'autant plus qu'il peut y avoir des conséquences assez importantes
- Lien à la croissance, on peut continuer à "croître" à peu près de la même façon si on fait "attention", puisqu'on trouvera toujours des solutions techniques

Beaucoup des phénomènes qu'on essaie de traiter son à croissance exponentielle. Le problème avec ces phénomènes est qu'on a du mal à comprendre cette fonction exponentielle et souvent le problème dépasse le seuil où on ne peut plus le gérer avant qu'on ait pu le régler.

C. Impact environnemental

On nous vend beaucoup la voiture électrique comme un outil qui pollue très peu, ce qui est vrai **dans sa phase d'utilisation**. Pour un objet numérique, 75% de son impact environnemental se situe au moment de sa fabrication à cause de l'extraction des terres rares nécessaires.

Pour un Iphone 12 par exemple, sa fabrication représente 83% de son impact environnemental.

Cependant il est totalement possible de faire des objets numériques ayant des performances décentes tout en ayant un impact environnemental réduit. L'eXtreme Defi de Ademe consiste à concevoir un objet roulant véhiculant 1 à 3 personnes et étant 10x moins coûteux, 10x plus durable, etc.

Il est important que les outils qu'on développe aujourd'hui ne nuisent pas à l'environnement.

Installer un bloqueur de publicités permet par exemple de réduire l'impact direct du numérique sur l'environnement. En effet, cela permet de limiter les données transmises et collectées et permet également de limiter la fonction de génération de consommation du produit de la publicité. On ne visionne pas la publicité donc on n'est pas poussé à acheter le produit.

D. Lowtechisation

La lowtechisation est un processus qui désigne la promotion de technologies plus simples, durables et accessibles que les technologies actuelles. Ces alternatives sont plus respectueuses de l'environnement, faciles à réparer et requiert moins de ressources que les technologies classiques.

La marque Fairphone est par exemple un producteur de produits lowtech.

Conclusion et avis sur l'UV

Pour conclure, j'ai trouvé cette UV passionnante. Ça faisait un an que j'hésitais à faire l'UV à cause des retours que j'avais eu qui disaient que c'était une UV qui demandait beaucoup de travail mais j'ai trouvé la nouvelle formule de l'UV vraiment bien équilibrée. Elle demande une charge de travail raisonnable et est très intéressante.

J'ai également trouvé que c'était une UV très intéressante du point de vue de "l'initiation au numérique". Étant en deuxième semestre de Génie Informatique, j'ai fait beaucoup d'UV qui traitent de concepts communs à WE01 et j'ai trouvé qu'ils étaient très bien amenés et expliqués pour un public plus général.

Les informations que j'ai acquises au cours du semestre en matière de fonctionnement du Web m'ont déjà été très utiles car elles m'ont en partie permis d'obtenir mon stage TN09 dans la fouille de données sur le Web que je vais réaliser le semestre prochain.

J'ai cependant trouvé dommage qu'il y ait beaucoup de cours que nous n'ayons pas eu le temps de terminer.

Annexe :

1. Projet d'IA jouant à Gopher et Dodo

Ce semestre en IA02 nous avons beaucoup travaillé sur ces arbres de raisonnement. Nous avons réalisé en tant que projet de fin de semestre des IA basées sur ce type d'arbre qui devaient jouer aux jeux Gopher et Dodo de Mark Steere. Nous avons donc dû coder le jeu en python et ensuite créer des algorithmes qui devaient gagner le plus possible aux jeux. Pour ce projet j'ai exploré plusieurs algorithmes tels que le negamax ou encore l'algorithme de Monte Carlo.

Lien du projet : https://github.com/tomoriolu/IA02_project

2. Système Expert sur les associations de l'UTC

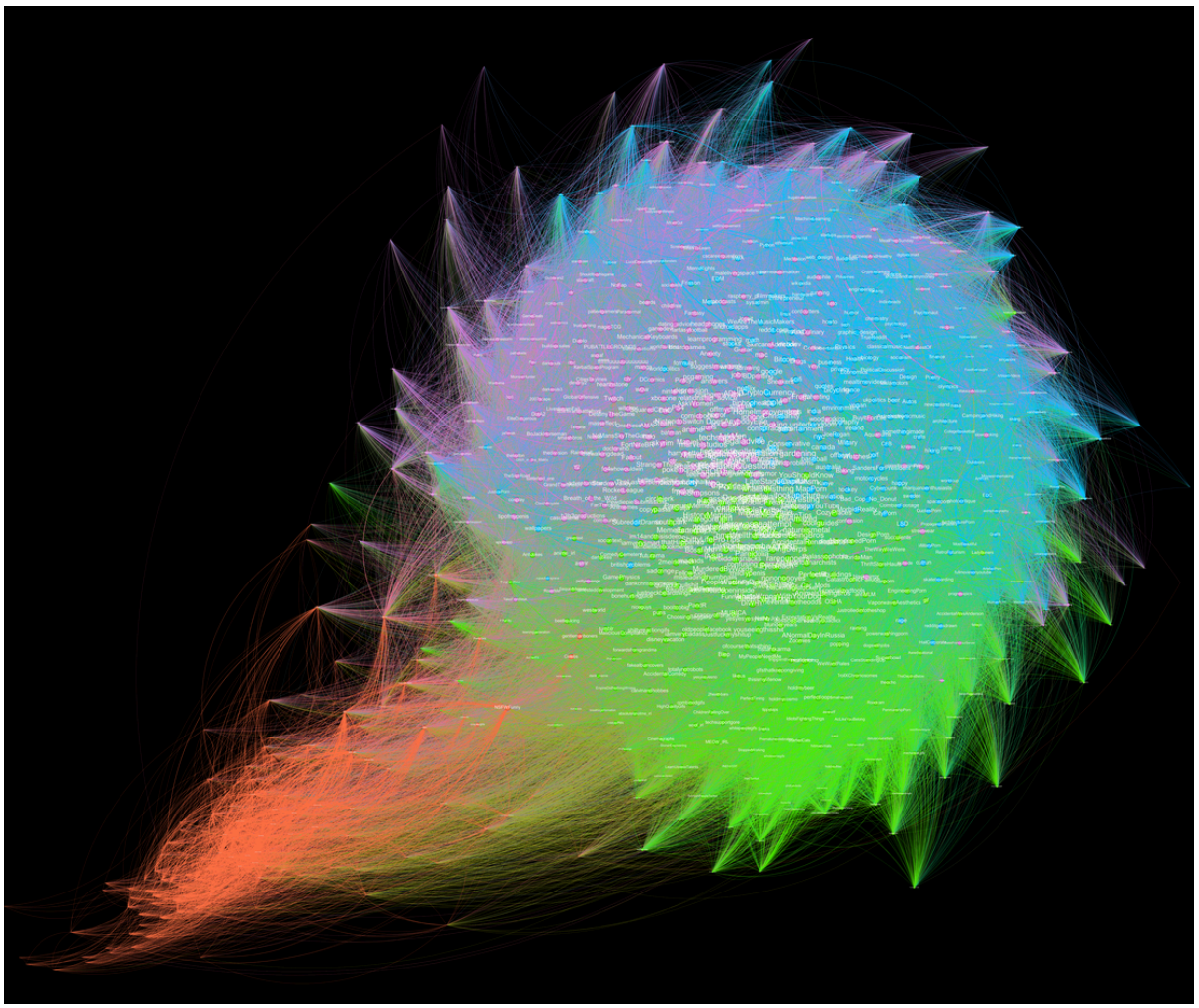
Nous avons aussi étudié les systèmes experts en IA02, notamment lorsqu'on codait en Prolog. J'ai également réalisé un système expert d'ordre 0+ en Lisp le semestre dernier en IA01 sur les associations de l'UTC. Le moteur d'inférence permettait grâce à la base de faits et la base de règles de savoir quelle association de l'UTC nous correspondait le plus (bien que cela reste très subjectif).

Lien du projet : https://github.com/jpontoire/IA01/tree/main/TP/TP3

3. Cartographie de Reddit

Je rajoute aussi un autre projet dont je suis assez fier et que je trouve en rapport avec le sujet de l'UV. Le semestre dernier en IC05 j'ai réalisé avec mes collègues une cartographie de Reddit. Nous avons donc travaillé sur comment récolter les données nécessaires, comment automatiser cette collecte et comment ensuite traiter ces données. Ce projet n'est cependant pas parfait, il y a plusieurs biais que nous n'avons pas réussi à supprimer lors de la collecte des données. J'ai aussi trouvé dommage que nous n'ayons eu aucune mention du cadre légal de la collecte des données que nous effectuions au moment de réaliser ce projet.

Lien du projet : https://github.com/jpontoire/IC05_reddit



Cartographie de Reddit