

Master of Science HES-SO in Engineering
Av. de Provence 6
CH-1007 Lausanne

Master of Science HES-SO in Engineering

Orientation : Technologies de l'information et de la communication

Web digital footprints and data privacy

Fait par

Kewin Dousse

Sous la direction de
Prof. Fatemi Nastaran
à la HEIG-VD

Félicien Fleury (NGSENS)

Lausanne, HES-SO//Master, 2017

Résumé

Avec l'augmentation incessante de la présence d'Internet dans la vie de tous les jours depuis des années, posséder et entretenir une identité digitale est un point de plus en plus important dans la vie de nombreuses personnes. Aujourd'hui, des décisions et des carrières peuvent se jouer sur l'apparence qu'une personne montre en ligne, volontairement ou involontairement par des tiers.

Dans un but de sensibilisation, l'objectif de ce projet est de concevoir et d'implémenter un outil d'analyse de comportements d'utilisateurs du Web, dans le but de révéler la possibilité de détecter des aspects du profil d'une personne comme ses préférences, ses centres d'intérêts, ses orientations ou ses opinions.

Ce document présente la poursuite de cet objectif à travers le développement d'une extension de navigateur récoltant des informations sur les pages visitées par l'utilisateur. Les informations amassées sont anonymisées puis centralisées sur un serveur, où des algorithmes évaluent les données et tentent de reconstituer un profil de l'utilisateur. Celui-ci a accès à tout moment à des statistiques et visualisations sur ses informations.

Nous terminons le projet par une analyse et une évaluation des méthodes utilisées, révélant qu'il est en effet possible de déterminer certains aspects du profil d'un utilisateur avec un nombre suffisant d'un set de données assez restreint.

Keywords. Web, Big Data, Privacy, Profiling

Table des matières

| | | |
|----------|------------------------------------|-----------|
| 1 | Introduction | 9 |
| 1.1 | Contexte | 9 |
| 1.2 | Objectifs | 9 |
| 1.3 | Méthodologie | 10 |
| 2 | Etat de l'art | 11 |
| 2.1 | Projet similaire | 11 |
| 2.1.1 | Introduction | 11 |
| 2.1.2 | Résumé | 11 |
| 2.1.3 | Données | 13 |
| 2.1.4 | Acquisition | 14 |
| 2.1.5 | Conclusion | 14 |
| 2.2 | Outils de tracking | 15 |
| 2.2.1 | Trackers | 15 |
| 2.2.2 | Marché | 15 |
| 2.2.3 | Google Analytics | 17 |
| 2.3 | Extensions de navigateur | 17 |
| 2.3.1 | Introduction | 17 |
| 2.3.2 | Etat de l'art | 17 |
| 2.3.3 | Conclusion | 21 |
| 2.4 | Analyse de texte | 22 |
| 2.4.1 | Keyword extraction | 22 |
| 2.4.2 | TF-IDF | 24 |
| 2.4.3 | Topic Modeling | 25 |
| 2.4.4 | LDA | 26 |
| 3 | Design du système | 27 |
| 3.1 | Introduction | 27 |
| 3.1.1 | Idée | 27 |
| 3.1.2 | Architecture | 28 |
| 3.1.3 | Données | 29 |

| | | |
|----------|-------------------------------|-----------|
| 3.2 | Architecture | 31 |
| 3.2.1 | Stack technologique | 31 |
| 3.2.2 | Extension | 31 |
| 3.2.3 | Serveur | 32 |
| 3.2.4 | Base de données | 36 |
| 3.2.5 | Interface | 42 |
| 4 | Développement des vues | 48 |
| 4.1 | Technologies | 48 |
| 4.1.1 | Introduction | 48 |
| 4.1.2 | Offline | 48 |
| 4.1.3 | Serveur | 49 |
| 4.1.4 | Client | 49 |
| 4.2 | Wordcloud | 50 |
| 4.2.1 | Concept | 50 |
| 4.2.2 | Données | 51 |
| 4.2.3 | Traitement | 51 |
| 4.2.4 | Visualisation | 52 |
| 4.3 | Topics List | 54 |
| 4.3.1 | Concept | 54 |
| 4.3.2 | Données | 54 |
| 4.3.3 | Traitement | 55 |
| 4.3.4 | Visualisation | 56 |
| 4.4 | Most Watched | 58 |
| 4.4.1 | Concept | 58 |
| 4.4.2 | Données | 59 |
| 4.4.3 | Traitement | 59 |
| 4.4.4 | Visualisation | 60 |
| 4.5 | History | 62 |
| 4.5.1 | Concept | 62 |
| 4.5.2 | Données | 62 |
| 4.5.3 | Traitement | 63 |
| 4.5.4 | Visualisation | 64 |
| 4.6 | Trackers | 66 |
| 4.6.1 | Concept | 66 |
| 4.6.2 | Données | 68 |
| 4.6.3 | Traitement | 68 |
| 4.6.4 | Visualisation | 69 |
| 4.7 | Stats | 70 |
| 4.7.1 | Concept | 70 |
| 4.7.2 | Données | 70 |

| | | |
|----------|----------------------------------|------------|
| 4.7.3 | Traitement | 70 |
| 4.7.4 | Visualisation | 71 |
| 4.8 | Settings | 72 |
| 4.8.1 | Concept | 72 |
| 4.8.2 | Données | 73 |
| 4.8.3 | Traitement | 73 |
| 4.8.4 | Visualisation | 73 |
| 5 | Résultats | 75 |
| 5.1 | Processus de test | 75 |
| 5.1.1 | Inputs | 75 |
| 5.2 | Evaluation | 76 |
| 5.2.1 | Modèles | 76 |
| 5.2.2 | Vues et résultats | 84 |
| 5.3 | Statistiques | 90 |
| 5.3.1 | Profiling | 91 |
| 5.3.2 | Trackers | 99 |
| 6 | SDIPI | 105 |
| 7 | Conclusion | 107 |
| 7.1 | Conclusion technique | 107 |
| 7.1.1 | Réalisations | 108 |
| 7.2 | Travaux futurs | 109 |
| 7.3 | Conclusion personnelle | 111 |
| A | Historique des versions | 119 |
| B | Cahier des charges | 120 |
| B.1 | Activités | 120 |
| B.2 | Planification | 121 |
| B.3 | Diagramme de Gantt | 122 |
| C | Documentation | 123 |
| C.1 | Localisation | 123 |
| C.2 | Contenu | 123 |
| C.2.1 | GitHub | 123 |
| D | Procès-verbaux | 124 |

Table des figures

| | | |
|------|--|----|
| 2.1 | Précision moyenne du modèle prédisant la personnalité d'un utilisateur en fonction du nombre de likes analysés[5]. | 12 |
| 2.2 | Déviation de la personnalité moyenne estimée d'un visiteur régulier du site ““deviantart.com”” selon les cinq axes psychologiques employés[5]. | 13 |
| 2.3 | Nombre moyen de domaines contactés au chargement d'une page web[20]. | 15 |
| 2.4 | Marché occupé par Google Analytics dans les domaines d'analyse, de tracking et de mesure d'audience sur le Web. | 16 |
| 2.5 | Logo de la solution Google Analytics[8]. | 16 |
| 2.6 | Image de présentation de timeStats[12]. | 18 |
| 2.7 | Page d'accueil de Ghostery[13]. | 18 |
| 2.8 | Interface de base de Privacy manager[14]. | 19 |
| 2.9 | Interface de TheGoodData[15]. | 20 |
| 2.10 | Premier paragraphe de la page web de Noiszy[16]. | 21 |
| 2.11 | Contrôle des actions face aux trackers de Privacy Badger[17]. | 22 |
| 2.12 | Flux de données de Kraken.me[18]. | 22 |
| 3.1 | Flux de données de l'extension. | 28 |
| 3.2 | Exemple d'URL et traitement | 29 |
| 3.3 | Technologies utilisées pour chaque partie | 31 |
| 3.4 | Téléchargement et enregistrement du contenu d'une page | 33 |
| 3.5 | Traitements du contenu des pages | 35 |
| 3.6 | Schéma des tables de la base de données | 37 |
| 3.7 | Maquette de la page de Profil | 44 |
| 3.8 | Maquette de la page de Trackers | 45 |
| 3.9 | Suite de la maquette de la page de Trackers | 46 |
| 4.1 | Code de la méthode de vérification de connexion | 49 |
| 4.2 | Chargement des données | 49 |
| 4.3 | Formulaire du filtre de dates | 50 |

| | | |
|------|--|-----|
| 4.4 | Maquette initiale et résultat final de la vue Wordcloud | 51 |
| 4.5 | Algorithme utilisé pour le Wordcloud | 53 |
| 4.6 | Maquette initiale et résultat final de la vue Wordcloud | 54 |
| 4.7 | Algorithme utilisé pour le Topics List | 57 |
| 4.8 | Maquette initiale et résultat final de la vue Most Watched | 58 |
| 4.9 | Algorithme utilisé pour les pages "Most Watched" et "Most Visited" | 61 |
| 4.10 | Maquette initiale et résultat final de la vue History | 62 |
| 4.11 | Algorithme utilisé pour les graphiques de la page "History" | 65 |
| 4.12 | Domaine activé, puis désactivé | 66 |
| 4.13 | Maquette initiale et résultat final d'une des vues Trackers | 67 |
| 4.14 | Maquette initiale et résultat final de la vue détaillée lors d'un clic sur un Tracker | 67 |
| 4.15 | Algorithme utilisé pour les données des pages "Trackers" | 69 |
| 4.16 | Maquette initiale et résultat final d'une vue Stats | 70 |
| 4.17 | Algorithme utilisé sur les données de la page "Stats" | 71 |
| 4.18 | Champ d'entrée des centres d'intérêts | 72 |
| 4.19 | Sélection d'intérêt sur un topic | 72 |
| 4.20 | Algorithme utilisé pour les données de la page "Settings" | 74 |
| 5.1 | 20 URLs les plus regardées et leurs meilleurs mots selon TF-IDF . . | 78 |
| 5.2 | Vue de l'interface Wordcloud | 85 |
| 5.3 | Vue de l'interface Topics List | 86 |
| 5.4 | Vue de l'interface Most Visited | 87 |
| 5.5 | Vue de l'interface Most Watched | 87 |
| 5.6 | Vue du premier graphique de l'interface History | 88 |
| 5.7 | Vue du deuxième graphique de l'interface History | 89 |
| 5.8 | Vue de l'interface Wordcloud | 89 |
| 5.9 | Temps total de visionnage par domaine | 91 |
| 5.10 | Taux d'association de topics par rapport au temps de visualisation avec courbe de tendance | 93 |
| 5.11 | Taux d'association de topics par rapport au nombre de domaines différents visités avec courbe de tendance | 94 |
| 5.12 | Taux d'association de topics par rapport au nombre de pages différentes avec courbe de tendance | 95 |
| 5.13 | Taux d'association de topics par rapport au temps de visualisation avec courbe de tendance | 96 |
| 5.14 | Taux d'association de topics, brut et pondéré, par rapport au temps de visualisation avec courbes de tendances | 97 |
| 5.15 | Visualisation de la fréquence et de la taille de l'envoie des données des domaines | 102 |

| | | |
|-----|---|-----|
| 6.1 | Logo officiel de l'association Swiss Digital Identity and Privacy Institute | 105 |
| 6.2 | Page d'accueil du site web https://sdipi.ch | 106 |

Chapitre 1

Introduction

1.1 Contexte

En janvier 2014, l'ONG Internet Society a publié le document Digital footprints[3] qui aborde la question de la capacité que les web trackers ont de définir le profil personnel des utilisateurs d'Internet.

En 2016, Michal Kosinski[1], chercheur à Stanford, révèle les possibilités de définir un profil précis simplement en analysant les préférences (likes) enregistrées dans un profil Facebook[2]. L'étude révèle que ce type d'analyse permet de mieux connaître une personne que ses proches et même de prévoir de probables comportements avec une grande précision. De plus, lors d'événements politiques majeurs ces techniques de profiling auraient été utilisées, comme dans le cadre des campagnes pour le Brexit ou pour l'élection du président américain Trump.[4]

1.2 Objectifs

Le but de ce projet est de concevoir et d'implémenter un outil d'analyse de comportements d'utilisateurs d'applications Web pour révéler les potentiels de détection de profils des personnes (préférences, centre d'intérêt, orientations et opinions) en analysant les interactions et les informations échangées avec les applications Web. L'application développée dans ce projet a pour le but principal de sensibiliser le public et les médias à la question du profiling sur internet.

L'objectif technique du projet est de développer un plugin pour les navigateurs Mozilla Firefox et Google Chrome qui permettraient de :

1. Définir un profil utilisateur selon des critères de préférence, d'intérêt, d'habitude, d'opinion, etc.
2. De construire le profil d'un utilisateur en se basant sur sa navigation sur Internet ainsi que sur les métadonnées (durée de consultation des pages,

heure de consultation, etc.). Des algorithmes de machine learning seront utilisés pour apprendre les profils en se basant sur des collections de profils annotées telle que la collection kaggle[6].

3. Identifier des trackers qui ont la possibilité de construire des profiles utilisateurs en intégrant des données de plusieurs sources.

1.3 Méthodologie

Le développement du code sera open-source. Le déroulement du projet sera divisé en deux phases distinctes :

1. La première phase du projet consistera en une analyse des études et résultats actuels afin de proposer des concepts innovants à travers l'outil développé, tout en collectant les données des utilisateurs pour la deuxième phase.
2. La seconde phase mettra l'accent sur les données récoltées par le plug-in développé durant la première phase : Le but sera d'analyser les données et d'en tirer des conclusions intéressantes.

Chapitre 2

Etat de l'art

2.1 Projet similaire

2.1.1 Introduction

Afin de placer notre recherche dans les connaissances actuelles, nous nous intéressons d'abord aux recherches récentes partageant un objectif semblable au nôtre.

Michal Kosinski se présente sur son site web[1] comme un "psychologist and data scientist". L'étude qu'il a co-rédigée à l'Université de Stanford en 2016 a eu un impact important sur le monde académique et même industriel, en montrant les possibilités techniques ouvertes par la récolte de données simples d'utilisateurs : les "likes" Facebook.

Il est montré qu'avec un peu plus de 300 "likes" tirés une personne, il est possible de définir avec une précision remarquable (mieux que son époux/épouse) des traits psychologiques, ainsi que d'autres caractéristiques personnelles.

2.1.2 Résumé

Une enquête a été menée auprès d'une population variée de personnes possédant un compte Facebook. Les données concernant leurs "likes" ont été récoltées, ainsi que des données personnelles pouvant être disponible (ou non) selon le souhait de l'utilisateur sur Facebook, comme ses informations démographiques. Des tests psychologiques ont été également réalisés par une certaine partie des utilisateurs afin de pouvoir trouver des corrélations entre les pages likées et certains traits psychologiques.

Cette enquête a rencontré un succès très large, et le nombre de personne ayant répondu à l'enquête, au moins en partie, se compte en millions.

Les résultats présentés à la fin de l'étude sont inattendus : Michal annonce qu'il est possible de prédire certains comportements d'une personne mieux que

son entourage le plus proche.

Un des modèles créés avec les données récoltées, permet d'estimer le profil psychologique d'un participant selon cinq axes différents, en se basant sur ses likes Facebook. La figure 2.1 montre la précision obtenue par le modèle en fonction du nombre de likes utilisé en entrée.

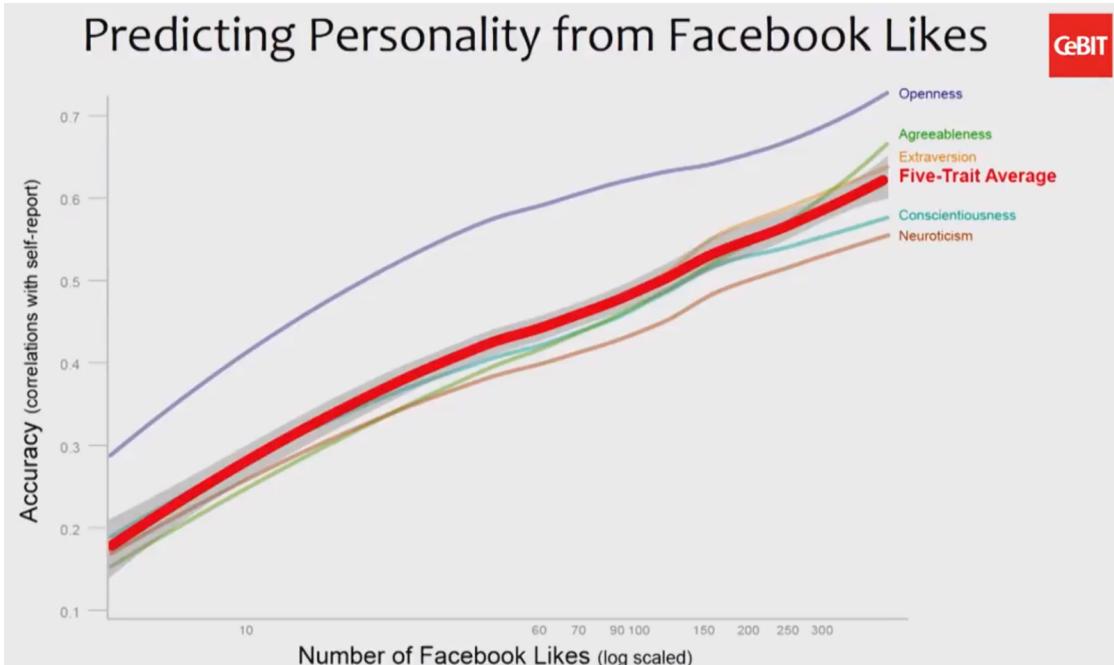


FIGURE 2.1 – Précision moyenne du modèle prédisant la personnalité d'un utilisateur en fonction du nombre de likes analysés[5].

On remarque que la précision de la prédiction de tous les critères augmente avec le nombre de likes utilisés, ce qui n'est pas surprenant. En revanche, le tableau 2.1 montre le lien entre le nombre de likes utilisés et la précision moyenne atteinte par l'algorithme, et compare ces valeurs à la précision atteinte par d'autres êtres humains.

On peut voir que la précision de la prédiction de l'algorithme surpassé celle même l'époux/se d'une personne avec 250 likes, ce qui se trouve être légèrement au-dessus du nombre de likes moyen par personne, qui est de 227.

Les possibilités de prédiction du modèle ne se limitent pas à une simple personne, et les possibilités sont nombreuses. Par exemple, Michal montre qu'il est possible de montrer une corrélation entre les visiteurs d'un certain site web, et une tendance vers certains traits psychologiques. La figure 2.2 montre la personnalité moyenne estimée des visiteurs du site web ““deviantart.com”” par rapport à la moyenne de tous les utilisateurs.

| | Précision | Nombre de likes |
|----------|-----------|-----------------|
| Collègue | 0.27 | 10 |
| Ami | 0.44 | 80 |
| Famille | 0.5 | 100 |
| Epoux/se | 0.58 | 250 |

TABLE 2.1 – Précision atteinte par type de relation avec une personne, et nombre de likes nécessaires au modèle pour égaler sa précision

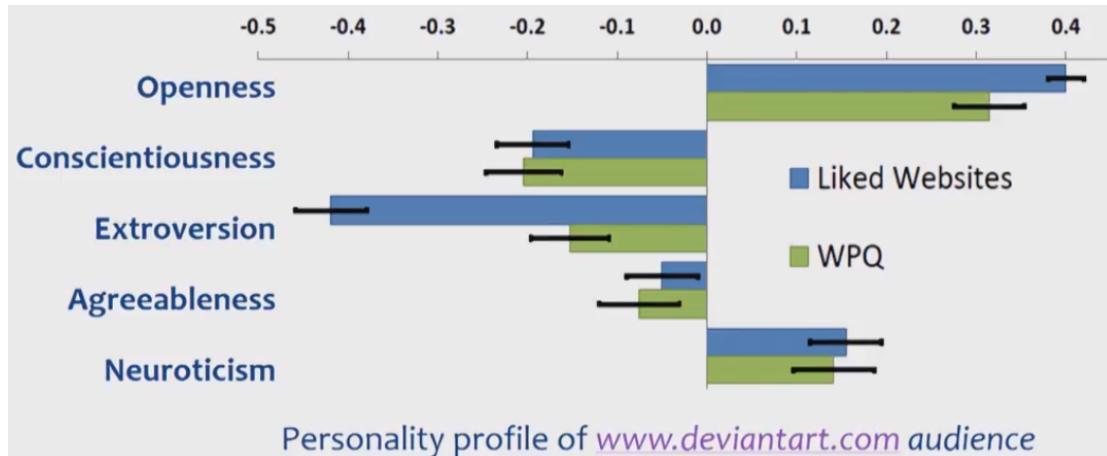


FIGURE 2.2 – Déviation de la personnalité moyenne estimée d'un visiteur régulier du site ““deviantart.com”” selon les cinq axes psychologiques employés[5].

Ces corrélations ne sont que quelques exemples parmi un très large éventail de possibles corrélations que le modèle est capable de mettre en lumière. Les implications de telles découvertes sont massives : Il serait par exemple possible de déterminer si un utilisateur sera réceptif ou non à un certain type de publicité, par exemple. Ce genre de problématique touche à plusieurs domaines et n'est pas exactement de notre ressort ici : Des principes éthiques sont en jeu, et le sujet devient de plus en plus délicat. Mais une chose est certaine : Des likes Facebook peuvent révéler énormément d'informations.

2.1.3 Données

La quantité de données amassée par l'étude est massive. Non seulement en quantité d'utilisateurs, mais également en diversité de données. Michal Kosinski a mis en place le site web "myPersonnality Project"[7] permettant de partager cette source de données avec d'autres chercheurs. Les données comprennent, entre autres :

- Scores de personnalité selon la méthode BIG5 de >3 millions de personnes
- Données démographiques de >4 millions de personnes
- Localisation géographique de >1.5 million de personnes
- Vues politiques de >500'000 personnes
- Likes Facebook de >19 millions de personnes

Le type de données présenté ici n'est qu'un sous-ensemble restreint de l'ensemble des tables présentées, bien qu'il s'agisse ici des données comprenant le plus d'entrées au total.

2.1.4 Acquisition

Bien que l'objectif du site web soit de partager l'accès à cette énorme base de données, l'accès à celle-ci est loin d'être aisé. Tout d'abord, Kosinski ne met ces données à disposition que de milieux académiques, il interdit l'utilisation de ces données à des fins commerciales.

Cependant l'accès n'est pas donné pour autant : Une demande d'accès est à lui envoyer, comprenant une présentation du projet et de ses buts par le biais d'un mail ainsi que le remplissage et l'enregistrement du projet de recherche sur des sites spécialisés.

Cette étape ne semblait constituer qu'une marche nécessitant un temps restreint, mais un prérequis à l'envoi d'une demande d'accès à la base de données est l'approbation de l'"IRB" (Institutional Review Board), ce qui correspond à un comité d'éthique.

2.1.5 Conclusion

Etant donné les délais estimés de l'envoi de la demande à un comité d'éthique responsable puis de la demande d'accès aux données à Kosinski, nous avons écarté cette source de données de la liste principale du projet car nous n'avions pas l'assurance de disposer des données à temps pour la suite de l'étude. Bien qu'il s'agisse certainement d'un ajout conséquent aux données amassées par le projet, nous ne pouvons pas nous permettre de mettre en péril tout l'agenda du projet sur cette source de données.

Bien que cette base de connaissance ait pu être utile, notre étude va changer de direction. Nous décidons de baser la recherche sur des données que nous récupérerons nous-même.

2.2 Outils de tracking

2.2.1 Trackers

Un tracker est un serveur contacté lors du chargement d'une page web par un utilisateur. De nos jours, les pages web sont souvent constituées de contenu provenant de plusieurs serveurs ou domaines différents. Il n'est pas rare qu'une seule page web fasse appel à plus d'une dizaine de domaines différents pour charger une seule page. La figure 2.3 montre l'évolution du nombre moyen de domaines contactés pour le chargement d'une seule page web, sur les 1'000 sites web les plus visités mondialement.

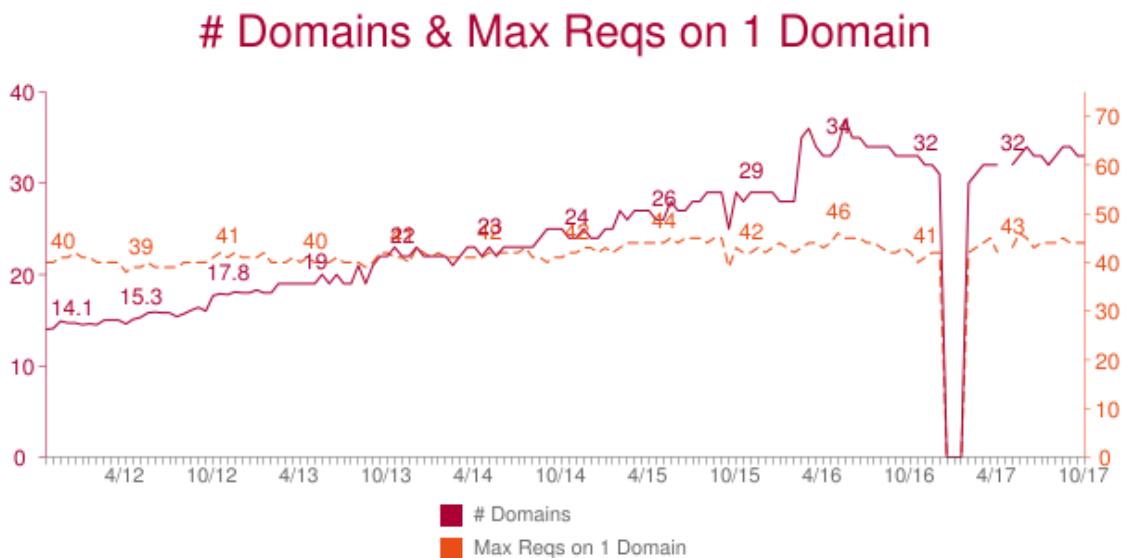


FIGURE 2.3 – Nombre moyen de domaines contactés au chargement d'une page web[20].

Bien qu'une partie des domaines soient nécessaires à contacter afin de charger du contenu indispensable à la page, une partie d'entre eux ne sert également qu'à des fins statistiques ou publicitaires. Par exemple, ceux-ci peuvent récupérer des informations sur l'utilisateur et son navigateur afin de lui proposer des publicités ciblées sur ses intérêts. Cette pratique est aujourd'hui courante, comme le montre la prochaine sous-section.

2.2.2 Marché

Etant donné que nous nous intéressons aux données des utilisateurs récupérées lors de la navigation Web, nous cherchons à connaître quels sont les plus grands trackers sur le web.

La figure 2.4 montre la part de marché qu'occupe Google Analytics ainsi que ses compétiteurs sur les sites web Suisses.

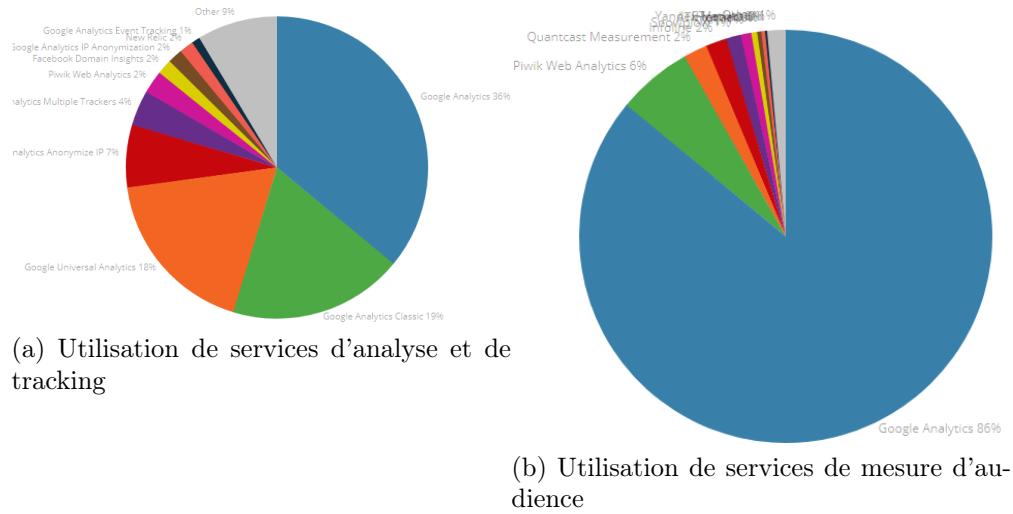


FIGURE 2.4 – Marché occupé par Google Analytics dans les domaines d'analyse, de tracking et de mesure d'audience sur le Web.

Nous pouvons calculer grâce au premier graphique que l'ensemble des produits de Google, y compris Google Analytics et ses versions proches, représentent plus de 83% des installations de solutions dans le domaine de l'analyse et du tracking. De plus pour la sous-catégorie du marché de la mesure d'audience uniquement, Google Analytics a lui seul représente 86% d'installations sur le Web.



FIGURE 2.5 – Logo de la solution Google Analytics[8].

Il est donc de plus en plus évident que s'intéresser aux fonctionnalités de Google Analytics est intéressant pour les buts du projet. Nous souhaitons nous poser la question du risque encouru par les utilisateurs en se connectant sur un site web utilisant Google Analytics. Quelles informations sont prélevées ? Lesquelles sont envoyées ? Les données sont-elles anonymisées ?

2.2.3 Google Analytics

Google Analytics se présente comme une solution d'analyse de statistiques d'utilisateurs dans le but d'améliorer les résultats des sites web sur lesquels il est installé. Ce produit étant totalement gratuit pour les PME, il est aujourd'hui très répandu sur le net et particulièrement en Suisse[9].

2.3 Extensions de navigateur

2.3.1 Introduction

Au vu de l'objectif du projet qui est à la fois de récolter des données tout en montrant un feedback à l'utilisateur, l'extension pour navigateurs est le moyen le plus facile à la fois pour nous de distribuer notre code, et pour les utilisateurs de l'installer. Cependant, de nombreuses extensions dont le but est de montrer des statistiques sur la navigation de l'utilisateur existent déjà. L'objectif n'est donc pas seulement d'implémenter les mesures adéquates pour notre étude, mais également de fournir des fonctionnalités à l'utilisateur novatrices afin que l'extension se démarque des concurrents.

Une analyse des extensions existantes est donc requises afin de prendre des décisions sur la direction que vont prendre les fonctionnalités implémentées.

2.3.2 Etat de l'art

Nous nous intéressons aux extensions disponibles pour deux des navigateurs les plus utilisés : Google Chrome, et Mozilla Firefox. Chaque navigateur possède son propre éventail d'extensions, bien que parfois certaines se retrouvent disponibles dans les deux catalogues. Chrome Web Store[10] est le catalogue officiel d'extensions pour Google Chrome, et Modules Firefox[11] est celui correspondant à Mozilla Firefox. Quelques recherches avec des mots-clé adaptés sur chaque catalogue vont nous fournir les extensions les plus populaires pour un thème semblable aux nôtre.

timeStats

timeStats[12] est une extension disponible pour Google Chrome. La figure 2.6 montre comment l'extension se présente via une image montrée sur le Google chrome Store.

Cette extension se focalise sur la visualisation du temps passé sur les différents sites web, parfois regroupés en domaines. La plupart des informations représentées sont le temps passé, et l'extension s'organise en plusieurs pages permettant de

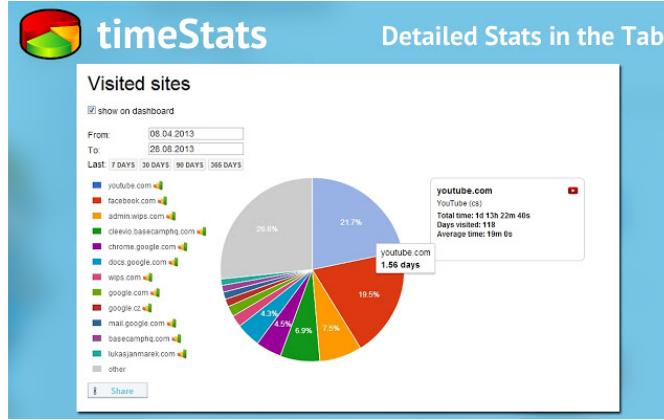


FIGURE 2.6 – Image de présentation de timeStats[12].

voir des visualisations différentes. On remarque la présence de plusieurs types de graphiques (en ligne, en secteurs) adaptés à la mesure affichée. timeStats est disponible pour Google Chrome uniquement.

Ghostery

Ghostery est une extension Google Chrome qui possède également sa propre page web en dehors du catalogue. La figure 2.7 montre la page d'accueil du site “ghostery.com”, qui est le domaine officiel de l'extension listée sur Google Chrome.

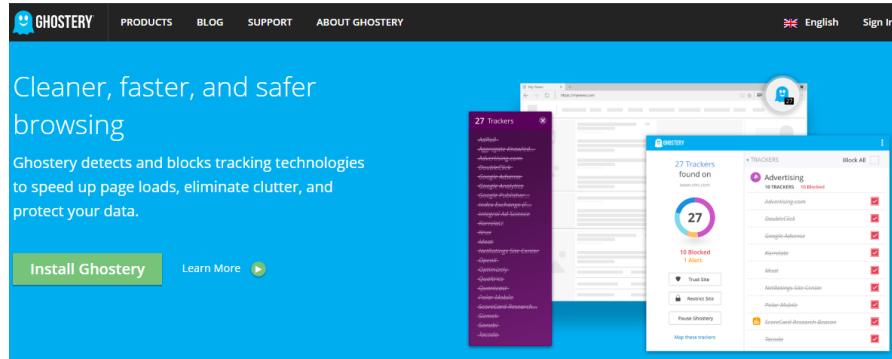


FIGURE 2.7 – Page d'accueil de Ghostery[13].

Ghostery semble donc se concentrer sur la détection et le blocage des informations envoyées aux trackers tiers lors de la navigation. Quelques options de personnalisation y sont présenter, comme la possibilité d'autoriser des trackers particuliers, ou des domaines choisis.

Privacy manager

Privacy manager se montre comme une extension permettant la gestion de mécaniques liées à la préservation de la vie privée. La figure 2.8 montre l'interface principale utilisée par l'extension. Bien que certaines options existent pour la protection de la vie privée, presque la moitié les options activables n'ont pas directement à faire avec la vie privée, et sont plutôt des désactivation ou activations de fonctionnalités de productivité.

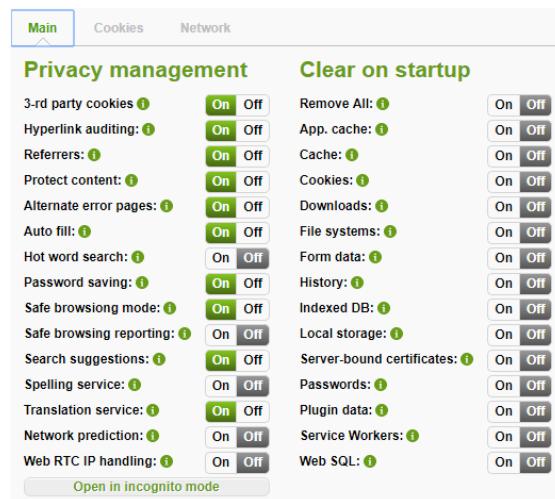


FIGURE 2.8 – Interface de base de Privacy manager[14].

TheGoodData

TheGoodData remplit à priori la même mission que Ghostery, mais propose des outils légèrement différents, et son thème est centré sur l'utilisation de la valeur des données de navigation pour une bonne cause. Un tableau de bord montré à la figure 2.9 permet de se renseigner sur l'état actuel de sa navigation avec des analyses basiques sur les dangers trouvés.

Noiszy

Noiszy cherche quand à lui à brouiller les pistes des trackers existants, sans les bloquer. Son hypothèse de base est qu'il est presque impossible de dissimuler complètement ses "Digital Footprints", et que la meilleure solution est de tenter de les brouiller en les "falsifiant", par exemple en envoyant des données erronées aux trackers, ou en quantité trop élevées. La figure 2.10 montre le premier paragraphe de présentation de Noiszy, présent sur leur site web.

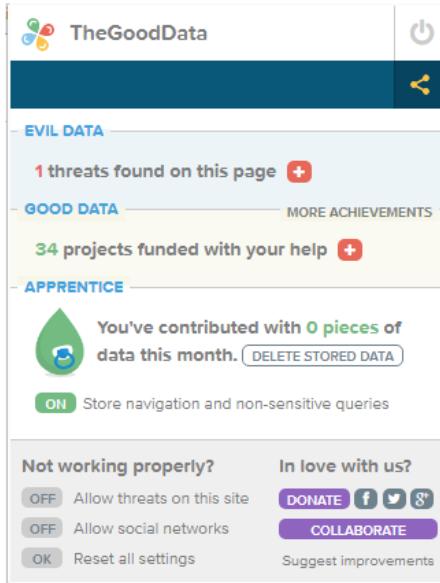


FIGURE 2.9 – Interface de TheGoodData[15].

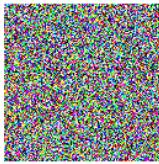
Privacy Badger

Privacy Badger est une extension développée par l'EFF[19]. Disponible à la fois sur Google Chrome et Mozilla Firefox, cette extension a également comme objectif de contrôler l'envoi de données à des trackers. Plutôt que de strictement bloquer toute requête, cette extension laisse à l'utilisateur décider quel niveau de danger représenter chaque tracker, et adapte son comportement entre un blocage total, la retenue de certaines informations ou aucune action entreprise pour chaque tracker détecté. La figure 2.11 montre l'interface de l'application une fois celle-ci installée. On peut y voir les lignes de présentant chacune un tracker, et la possibilité de définir son niveau de danger, et par conséquent l'action appropriée associée.

Kraken.me

Kraken.me est une extension de navigateur, mais également une application pouvant s'installer sur smartphone. Cette application analyse le flux de données de certains services comme Facebook, Twister, LinkedIn et encore d'autres. L'objectif est ici de donner à l'utilisateur une vue sur ses propres données, et la manière que celles-ci sont utilisées par les applications. La figure 2.12 montre le modèle présenté par le site web.

Cette application est probablement une des plus semblable à l'objectif général de notre projet, il serait donc intéressant de voir quels ont été les débouchés de cette étude. Notons que la plupart de l'activité de celle-ci ainsi que de l'outil semblent



You are being tracked.

Whatever you do online, you leave digital tracks behind.

These digital footprints are used to market to you - and to influence your thinking and behavior.

On April 3, President Donald Trump signed a repeal of online privacy rules that would have limited the ability of ISPs to share or sell customers' browsing history for advertising purposes. Erasing these footprints - or not leaving them in the first place - is becoming more difficult, and less effective.

Hiding from data collection isn't working.

Instead, we can make our collected data less actionable by leaving misleading tracks, camouflaging our true behavior.

We can resist being manipulated by making ourselves harder to analyze - both individually, and collectively.

We can take back the power of our data.

FIGURE 2.10 – Premier paragraphe de la page web de Noiszy[16].

avoir cessé en 2014.

2.3.3 Conclusion

Après avoir dressé une liste des extensions de navigateur les plus populaires et utilisés, nous pouvons prendre position sur les fonctionnalités que notre extension va posséder afin de se démarquer et de répondre à la problématique de l'étude. Nous allons choisir les fonctionnalités que nous estimons avoir un impact pour la sensibilisation du public aux traces que les internautes laissent, et des informations que nous pouvons en retirer. Ainsi, le plug-in se concentrera sur les deux aspects suivants :

- Détection et mise en lumière des différents trackers présents sur les pages visitées par l'utilisateur.
- Tentative de reconstitution du profil de l'utilisateur à partir de la fréquence de la visite des pages web et de leur contenu.



FIGURE 2.11 – Contrôle des actions face aux trackers de Privacy Badger[17].



FIGURE 2.12 – Flux de données de Kraken.me[18].

2.4 Analyse de texte

Nous allons utiliser le contenu des pages pour trouver des informations sur les centres d'intérêt de l'utilisateur. Pour ceci, nous avons besoin d'analyser et de tenter de comprendre le contenu des pages.

2.4.1 Keyword extraction

Le keyword extraction est le nom donné à un algorithme dont le but est de ressortir, parmi les mots d'un document, quels en sont les mots les plus représentatifs, ou les plus expressifs. Cette opération ne se fait généralement pas sur un seul document, mais sur un ensemble de documents, appelé corpus. Il s'agit d'une sous-catégorie des algorithmes de Natural Language Processing.

Dans notre cas, les techniques de keyword extraction que nous prenons en compte sont les techniques dites "non supervisées". Cela signifie que nous n'avons

pas d'information à priori sur quels sont les mots qui vont être potentiellement importants dans le texte, et que nous laissons une complète liberté à l'algorithme.

Parmi les techniques de keyword extraction, on peut citer les plus connus : Term Frequency-Inverse Document frequency, TextRank et Rapid Automatic Keyword Extraction. Etant donné que ces algorithmes fonctionnent de manière semblable, il a fallu se pencher sur des études montrant leur efficacité afin de décider lequel de ceux-ci nous allons utiliser.

Recherche

Etant donné qu'il s'agit ici d'entraîner des modèles à l'aides d'algorithmes non-supervisés, il n'existe pas de méthode efficace cartésienne pour prouver que nous obtenons des bons résultats à l'aide des modèles générés. Nous ne cherchons ici une vérité absolue, mais nous voulons trouver des modèles dans les pages parcourues par l'utilisateur. Les publications étudiées font donc également souvent appel à un jugement humain afin de définir si une méthode est adaptée ou non.

Un total de trois ressources a été étudié pour baser notre choix de la technique de keyword extraction à utiliser.

Première étude La première source[28] présente un cas concret où on cherche à classifier une série de documents connus à l'avance, présentant des sujets communs. Le but est de créer un modèle capable de reconnaître le thème de nouveaux documents en créant un modèle à partir d'une liste de documents connus. Bien que la comparaison soit faite entre TF-IDF, RAKE et TextRank, le cas ici assez différent du nôtre. En effet, le cas décrit ici cherche à entraîner un modèle de manière supervisée.

L'ensemble des documents en entrée est une liste de documents lus par l'auteur de l'article durant les 6 dernières années. Après avoir testé les trois algorithmes et comparé les différents résultats, la conclusion est que pour ce cas, TextRank et RAKE sont les algorithmes préférables. Le reproche est fait à TF-IDF de classer trop bas certains mots qui apparaissent dans de nombreux documents. Par exemple, nous savons que l'auteur a lu beaucoup de document sur JavaScript, et donc TF-IDF met un poids très faible à ce mot.

Cette conclusion est sensée, mais ce problème ne s'applique pas à notre cas. En effet, nous cherchons spécifiquement à avoir un éventail très large de pages afin d'éviter ce genre de cas. Également, notre entraînement ne sera pas supervisé, nous ne pourrons donc pas évaluer les résultats finaux de cette manière.

Deuxième étude La deuxième étude[29] se penche sur l'extraction automatisée et non supervisée de keywords, ce qui est cette fois très proche de notre cas. Cette fois-ci, la publication se base sur des méthodes statistiques pour tenter de prédire

quel est l'algorithme le plus efficace pour le keywords extraction en comparant les résultats de ceux-ci avec l'avis d'humains auxquels ont demande d'exécuter la même tâche.

Sont comparées ici la méthode TF-IDF ainsi que certaines de ses variantes, et plusieurs méthodes basées sur l'utilisation de graphes (référencées simplement comme "Graph-based methods" à plusieurs endroits du texte).

Après une comparaison des résultats de chaque méthode avec les résultats trouvés par des humains, il est conclu que "Our results on the human transcripts show that the simple TFIDF based method is very competitive"[29]. Certaines variantes de TF-IDF où la position des mots dans une phrase ajoute un poids aux mots offre des scores très comparables à la méthode de base, mais les autres techniques ont montré des résultats plus faibles.

Troisième étude Le troisième document[30] est en réalité une question posée sur le site Quora, où des experts ont proposé une réponse à la question "What are the best keyword extraction algorithms for natural language processing and how can they be implemented in Python?".

Les trois méthodes populaires Term Frequency-Inverse Document frequency, TextRank et Rapid Automatic Keyword Extraction sont à nouveau citées, mais il ressort que RAKE et TF-IDF semble être consensuellement les méthodes les plus efficaces.

À la fin de l'analyse de ces sources, nous choisissons d'utiliser l'algorithme TF-IDF, qui semble être à la fois efficace et adapté à notre cas.

2.4.2 TF-IDF

TF-IDF est une méthode attribuant un poids à chaque mot de chaque document d'un corpus. Ce poids mesure l'importance relative du mot dans ce document. Le nom "TF-IDF" signifie "Term Frequency - Inverse Document Frequency". Le poids final d'un mot dans un document se calcule en prenant en compte uniquement deux mesures :

TF Term Frequency : La quantité d'apparition de ce mot dans ce document

DF Document Frequency : Le nombre de documents dans lesquels ce mot apparaît

Le score final d'un mot multiplie le TF d'un mot à l'inverse de son DF. Cela signifie que pour avoir une grande importance dans un document, un mot sera typiquement :

- Présent de nombreuses fois dans ce document
- Présent dans très peu d'autres documents

Le calcul des poids se fait sur l'ensemble du corpus en une fois, car il nécessite que l'on connaisse le nombre d'occurrences de chaque mot dans l'entièreté du corpus de documents. Il n'est donc pas possible de mettre à jour les poids de manière "online" en utilisant ce modèle.

2.4.3 Topic Modeling

Le topic modeling est un exercice différent de celui du keyword extraction. Nous ne cherchons plus ici à trouver les meilleurs mots parmi des textes, mais nous cherchons à les regrouper afin d'en former des thèmes, ou "topic"s.

Le but est cette fois de déterminer par des méthodes statistiques quels sont les thèmes communs à plusieurs documents, nous permettant ainsi de les regrouper. Ceci part du principe que nous appliquons l'apprentissage de l'algorithme sur un corpus de documents qui contient des textes qui ont effectivement des thèmes en commun. Ce qui est sans doute notre cas lorsque notre corpus est composé de pages web visitées par des utilisateurs.

Recherche

Etant donné que nous sommes dans le même cas que TF-IDF, à savoir en recherche d'un algorithme pour le cas d'un apprentissage non supervisé, la plupart des méthodes tentant de comparer des modèles se basent sur des mesures empiriques. Le jugement humain est à nouveau indispensable dans ce cas pour déterminer de l'adéquation d'un algorithme ou non avec les buts recherchés.

Nous recherchons ici une méthode relativement populaire car nous allons nous baser sur une librairie l'implémentant déjà. Il aurait été possible de développer nous-même ce module, mais nous avons estimé que les efforts à déployer pour ce faire étaient trop élevés pour justifier le temps passé.

Après quelques recherches sur le web, les méthodes les plus populaires semblent être LSA, pLSA et LDA[33]. Cependant, les études comparatives entre ces méthodes semblent très rares, et les seules sources trouvées furent des déclarations de chercheurs. Voici les ressources ayant contribué au choix de la méthode :

Première source La première source comparant ces algorithmes est une réponse à une question posée sur le site web Quora[34]. À la question "What's the difference between Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA)?", la réponse votée le plus positivement par un consultant en NLP conclut "In practice, LSI is much faster to train than LDA, but has lower accuracy.".

Seconde source La deuxième source est une question posée sur le site Reddit, intitulée "LSA vs pLSA vs LDA"[35]. L'auteur demande la différence et les avan-

tages de chaque méthode, et une réponse résume chaque algorithme en une simple ligne :

LSA -> uses SVD, and as a result the topics are assumed to be orthogonal.

pLSA -> Treats topics as word distributions, uses probabilistic methods, and topics are allowed to be non-orthogonal.

LDA -> similar to pLSA, but with dirichlet priors for the document-topic and topic-word distributions. This prevents over-fitting, and gives better results.

Semblablement, le modèle LDA semble primer dans l'avis sur la qualité des résultats. Après une recherche de librairies, il se trouve que LDA possède une implémentation en Python, dans une librairie nommée gensim. Nous nous arrêtons donc sur ce choix, comme le modèle semble à la fois précis et adapté à notre cas d'utilisation avec Python.

2.4.4 LDA

LDA (de l'anglais Latent Dirichelet Allocation) est un modèle de topic modeling, et va nous permettre de révéler des topics relatifs aux pages visitées par les utilisateurs. Ici, un topic est défini par une liste de mots, et un poids associé à chaque mot pour le topic.

La génération d'un modèle LDA prend plusieurs paramètres en entrée, mais le plus important pour nous est de définir un nombre de topics que nous souhaitons voir en sortie. On fixe ce nombre de topics, puis on lance l'apprentissage du modèle sur l'ensemble du corpus de documents, opération que peut durer plusieurs heures.

À la fin de l'apprentissage, nous sommes en possession d'un modèle que nous pouvons questionner de plusieurs manières, par exemple :

- Quels sont les mots les plus contribuant à un topic ?
- Quels sont les topics les plus probables pour un document ?

Nous allons donc par exemple utiliser le modèle afin d'assigner des topics au contenu d'URLs, et ainsi tenter de trouver quels sont les thèmes communs aux pages visitées par un utilisateur.

Chapitre 3

Design du système

3.1 Introduction

3.1.1 Idée

Le but recherché de l'outil est de sensibiliser les utilisateurs aux informations que ceux-ci dévoilent potentiellement en naviguant sur le web. Pour ce faire, nous avons besoin d'amasser des données sur leurs habitudes de navigation afin de les analyser.

Ces données seront centralisées sur un serveur afin que nous puissions lancer des traitements sur l'ensemble des données plus tard dans le but de tenter de révéler des tendances, habitudes ou corrélations entre les données.

De plus, nous souhaitons également offrir un service direct à l'utilisateur afin que celui-ci ait un bénéfice à installer l'extension et nous autoriser à accéder à ces données. Nous allons lui montrer via une interface web les données que nous avons pu amasser sur sa navigation depuis l'installation du plug-in, au travers de plusieurs pages et visualisations.

Nous souhaitons également que les données récupérées ne puissent pas être utilisées pour reconnaître une personne particulière. C'est pourquoi le plug-in ne nécessite aucune connexion avec un compte externe, et ne demande pas d'information directement divulgateuse d'une identité.

Nous pouvons ainsi résumer les caractéristiques principales du plug-in en quelques points.

Le plug-in :

- Récupère les informations de navigation de l'utilisateur
- Envoie ces informations de manière anonyme à un serveur centralisé
- Propose une visualisation des données récoltées et calculées sur l'utilisateur

3.1.2 Architecture

Le projet dans son ensemble requiert le développement d'un minimum de deux parties différentes :

- Une extension pour navigateur afin de récupérer et d'envoyer les données
- Un serveur recevant les données des extensions installées

Une troisième partie s'occupant de l'interface utilisateur est également à prévoir, celle-ci pouvant se situer autant dans l'extension que sur le serveur. La décision est finalement prise d'héberger l'interface utilisateur sur un différent serveur, auquel se connecte l'interface lorsque l'utilisateur souhaite accéder à sa page.

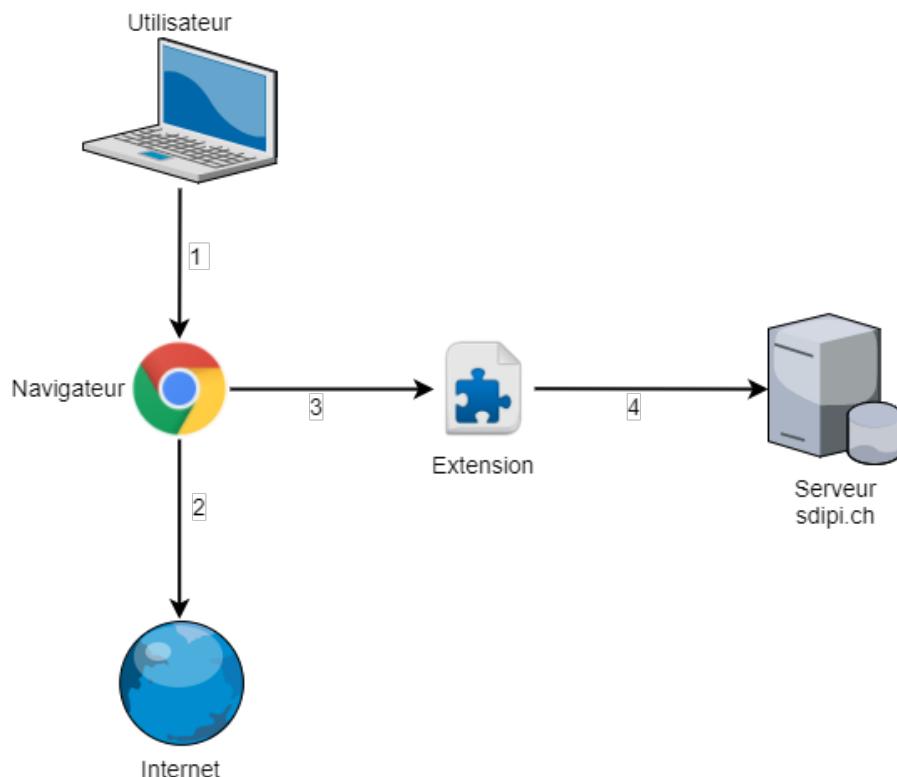


FIGURE 3.1 – Flux de données de l'extension.

La figure 3.1 schématisé la récolte de données effectuée par l'extension.

1. L'utilisateur entre une URL dans son navigateur
2. Le navigateur accède à la ressource concernée
3. Le navigateur transmet au plug-in les informations concernant la navigation
4. Le plug-in contacte le serveur SDIPI pour lui transmettre les informations

3.1.3 Données

Les possibilités de récolte de données depuis une extension de navigateur sont extrêmement nombreuses. Nous allons cependant nous concentrer sur l'amassage de données utiles à l'étude, et qui ne représentent pas une menace à l'intimité de l'utilisateur. Nous devons donc nous limiter à un set de données adéquat.

Voici les différents types d'informations que nous récoltons, et à quelles fins chaque type d'information est utilisé :

Visite d'une URL

Lorsque l'utilisateur accède à une nouvelle URL dans son navigateur, qu'il s'agisse d'un clic sur un lien ou d'une entrée dans la barre d'adresse, l'extension enregistre une partie de l'URL accédée ainsi que la date d'accès. Pour des raisons de protection de la vie privée, seule une partie de l'URL est conservée et envoyée au serveur.

https://www.google.ch/search?q=Recherche+test
Partie conservée

FIGURE 3.2 – Exemple d'URL et traitement

La figure 3.2 montre que tous les paramètres de la requête ne sont pas conservés. Seuls le protocole (**http** ou **https**), le nom de domaine, l'éventuel numéro de port ainsi que le chemin d'accès à la ressource sont conservés. Nous évitons ainsi la possibilité de stocker des informations sensibles comme le nom d'utilisateur, qui peut parfois se trouver dans cette partie de l'URL de certains sites web.

De plus, nous savons déjà que certaines URLs ne seront pas utiles à notre étude. Par exemple, nous pouvons d'avance dire que les URLs menant à des pages dont le contenu est constitué de l'une de ces manières est peu adéquat à analyser :

- Du texte généré spécialement lorsque l'utilisateur est connecté à un compte, par exemple une page d'accueil d'un service
- De contenu principalement autre que du texte, par exemple une page mettant en avant une vidéo
- De contenu régénéré très fréquemment, par exemple une première page d'un site de news

Nous décidons donc de définir une liste d'URLs que nous ne prendrons volontairement pas en compte lors de l'analyse. Le but de cette liste n'est bien sûr pas d'exhaustivement ignorer les sites dont le contenu textuel est très peu fiable ou présente très peu d'intérêt, mais d'améliorer la qualité possible des résultats en écartant certains cas dont nous pouvons raisonnablement dire qui apportent des biais par rapport à la réalité.

Ainsi, nous avons dressé une simple liste d'une dizaine de patterns d'URL que nous allons volontairement ignorer. Dans cette liste nous retrouvons des pages d'accueil de sites comme `^https://www.facebook.com/?$` ou `^https://www.twitter.com/?$`, mais également des sites de lecture de vidéo comme `^https://www.youtube.com` et d'autres cas plus spécifiques, comme les URLs ne pointant pas vers des ressources accessibles en HTTP ou HTTPS : `^(?!http)`.

Activité sur une page

À tout moment, l'utilisateur a probablement plusieurs onglets ou plusieurs fenêtres de navigateur ouvertes. Nous souhaitons nous intéresser à quelle page est actuellement en train d'être parcourue par l'utilisateur. À cette fin, nous détectons les événements sur la page web : Appui sur une touche, ou clic de souris par exemple. Dès lors qu'il se passe plus de 30 secondes sans aucun événement de la part de l'utilisateur, nous estimons qu'il ne regarde plus activement la page. Ce temps passé à s'intéresser à chaque page est également envoyé au serveur central toutes les 30 secondes.

Requêtes du navigateur

Lorsque le navigateur accède à une page web ou à d'autres moments, le navigateur doit charger des ressources qui se trouvent sur un serveur distant. Ce chargement peut prendre place pour afficher par exemple une image, un morceau de la page web elle-même, ou être demandé par un script chargé.

Pour chaque requête que le navigateur envoie, l'extension mémorise certaines informations :

Origine L'extension mémorise l'URL de la page qui demande la ressource.

Cette information est traitée de la même manière que décrit à la figure 3.2.

Hôte Toujours d'une manière identique à la figure 3.2, l'extension mémorise également le serveur contacté.

Taille L'extension mémorise également la taille de la requête en question, qui correspond à l'addition du contenu envoyé dans le contenu de celle-ci, ainsi que la taille des paramètres (ceux qui ne sont pas retenus par l'extension).

Identificateur

Lors de l'installation de l'extension, un nombre aléatoire est généré pour l'installation. Cet identificateur est envoyé envoyé au serveur central en plus de chaque autre information : Elle nous est utile pour assigner chaque donnée de navigation avec un navigateur particulier.

3.2 Architecture

3.2.1 Stack technologique

L'ensemble des éléments constituant le projet peuvent être regroupés en 3 parties différentes où s'exécute le code.

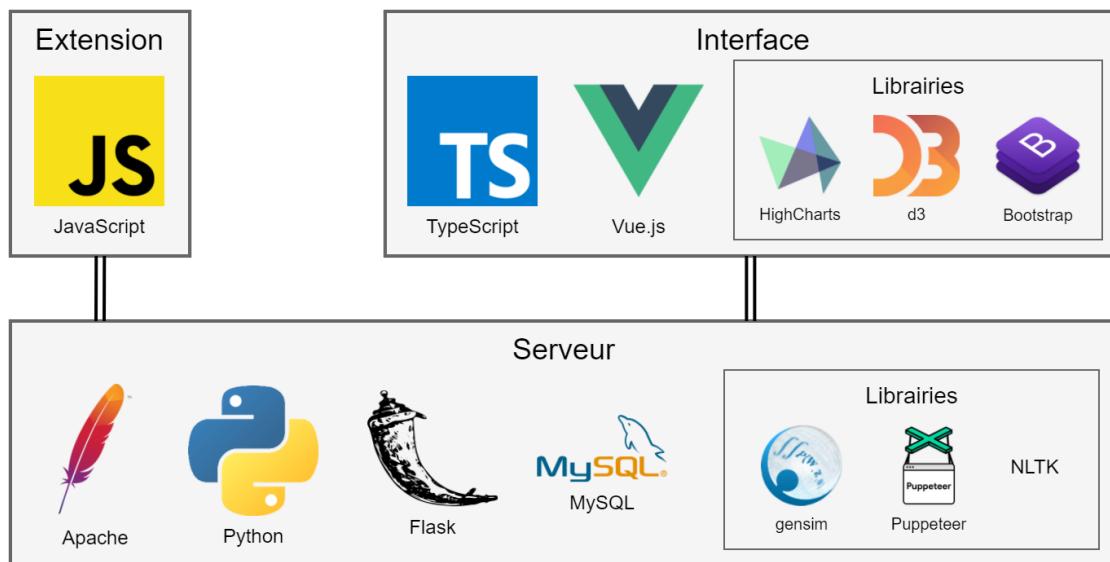


FIGURE 3.3 – Technologies utilisées pour chaque partie

La figure 3.3 montre les trois parties ainsi que les technologies utilisées dans chacune.

- L'extension comprend le code exécuté dans le navigateur du client, et qui communique avec l'API de Google Chrome afin de pouvoir récolter et envoyer les informations.
- L'interface comprend le code des diverses pages de visualisations montrées à l'utilisateur. On peut accéder à cette série de pages via un lien montré dans l'extension, ou par leur URL directement.
- Le serveur comprend le code exécuté notre machine.

3.2.2 Extension

L'extension de navigateur est sans aucun doute la partie la plus simple du projet. Etant donné que nous avons décidé d'héberger l'interface utilisateur sur un serveur différent, l'extension ne va principalement s'occuper que de récupérer les données de l'utilisateur et les transmettre à notre serveur.

Etant donné qu'une extension de navigateur n'est disponible que pour un type de navigateur à la fois, la question du support de plusieurs navigateurs s'est posée. D'après le site populaire w3schools.com[21], Google Chrome représente plus de 75% des visites au moins de décembre 2017. Nous décidons donc de ne pas adapter le code de l'extension pour plusieurs types de navigateurs, car nous estimons que le gain en utilisateurs serait insuffisant pour justifier le développement supplémentaire.

L'extension sera donc développée pour le navigateur Google Chrome, en utilisant l'API JavaScript que celui-ci met à disposition. Les fonctionnalités implémentées sont la récolte et l'envoi des types de données décrites à la section 3.1.3.

L'extension proposera également à l'utilisateur d'accéder à l'interface grâce à un lien, ainsi que la possibilité de se "lier" ce navigateur au profil d'un autre navigateur existant, en entrant son ancien identificateur. Ceci permet à un utilisateur de profiter d'un seul profil au travers de plusieurs machines possédant l'extension, par exemple.

3.2.3 Serveur

Rôles

La partie du serveur est probablement la plus complexe du projet. Le serveur va devoir assurer le fonctionnement de plusieurs tâches clés :

- Récolte et enregistrement des données de l'extension
- Traitement des données utilisateurs
- API au service de l'interface

Récolte

Avant tout traitement, le serveur doit être capable de recevoir et d'enregistrer les données des clients. Etant donné que l'extension est développée en JavaScript, les données seront transmises par HTTP au format JSON pour des raisons de simplicité.

Installation du plug-in Au moment où le plug-in est installé, une requête est envoyée au serveur afin de l'avertir qu'un nouvel utilisateur a installé l'extension. Le serveur génère un identifiant, l'envoie à l'extension en réponse et est désormais prêt à recevoir des informations de ce nouvel identifiant.

Récolte continue Afin que le serveur soit capable de supporter une certaine charge d'utilisateurs, il est nécessaire que celui-ci reçoive un nombre réduit de requêtes de la part des clients. Pour cette raison, l'extension ne contacte le serveur

qu'une seule fois toutes les 30 secondes afin de le tenir informé des événements ayant eu lieu.

Le serveur va donc pouvoir exposer une API simple : L'extension contactera toujours le même endpoint, et chaque requête contiendra la liste des informations concernant les événements qui se sont passés chez le client.

Lorsque nous détectons qu'un utilisateur visite une URL pour la première fois, le serveur va télécharger le contenu de cette page. Notre serveur ouvre un navigateur virtuel afin de simuler le chargement complet de la page - y compris l'exécution de scripts - et enregistre le contenu final de la page dans la base de données.

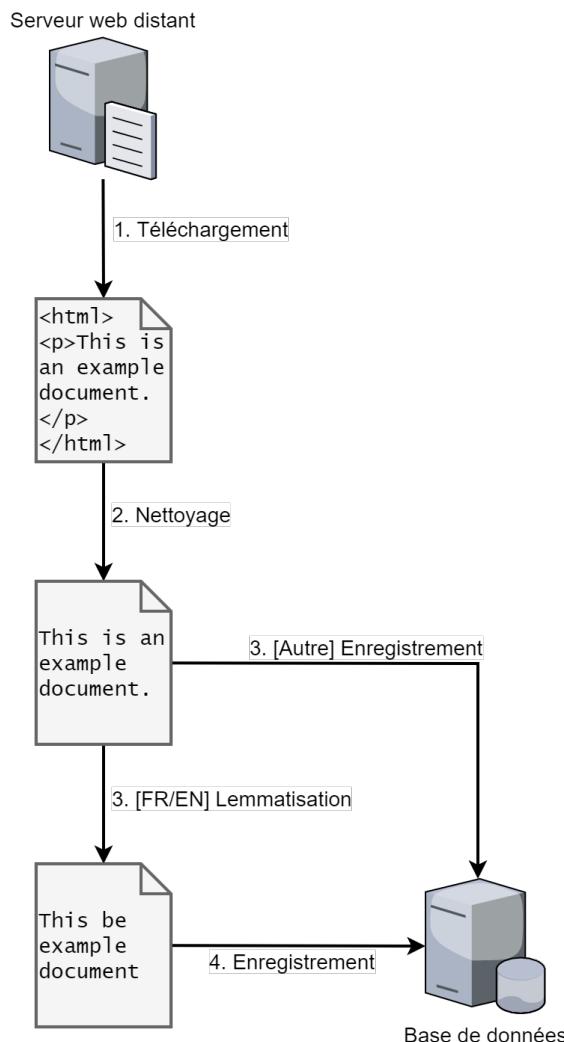


FIGURE 3.4 – Téléchargement et enregistrement du contenu d'une page

Une série d'opérations est ensuite effectuée sur le contenu de la page, afin de

le rendre utilisable par les prochains algorithmes. La figure 3.4 montre les étapes qui entrent en compte dans le pré-traitement du contenu :

1. **Téléchargement** Le serveur lance, dans un navigateur virtuel, le téléchargement de la page ainsi que l'exécution des scripts présents sur celle-ci. Une fois la page complètement chargée, on conserve le DOM de la page chargée.
2. **Nettoyage** À partir du HTML de la page, on ne cherche à garder que le texte de celle-ci, sans balise. Un parseur enlève toutes les balises `<script>` et `<style>`, puis ne garde que le contenu des éléments restants.
3. **[FR/EN Lemmatisation]** Un détecteur de langue nous renseigne sur la langue du texte. Si celui-ci est en anglais ou en français, nous lemmatisons chaque mot du texte. Ce processus analyse lexicalement les mots présents, et tente de ramener chaque mot à une forme plus simple pour le représenter. Par exemple, le temps des verbes est changé en infinitif, et les noms communs perdent leur pluriel. La liste complète des traitements effectuée est en réalité bien plus longue, et propre à la langue du texte.
- 3/4. **Enregistrement** Si le texte n'est pas dans une langue supportée par la lemmatisation, ou après la lemmatisation du texte anglais ou français, celui-ci est enregistré dans la base de données. Tous les traitements futurs sur le contenu de la page se feront sur cette version-ci.

Enregistrement

Le serveur se charge également de la gestion du stockage des données reçues (et calculées). Une base de données MySQL sera continuellement alimentée par les nouvelles données reçues. La base de données comprendra généralement une table par type de données à enregistrer, ainsi que des tables temporaires dans lesquelles seront placées des informations pré-calculées afin de répondre plus rapidement aux requêtes de l'interface.

Traitement des données

Une fois des données enregistrées, celles-ci sont traitées par différentes méthodes en fonction des besoins de l'interface. Voici le traitement que subit chaque type de données. Les traitements décrits ici ne sont pas effectués directement à la réception de données d'un client : Ils sont effectués régulièrement lorsque nécessaire.

Quantité de visites Deux mesures sont récoltées sur l'intérêt que peut avoir un utilisateur par rapport à une page web : Le nombre de fois que cette URL a été ouverte, et le temps passé à être actif sur la page en question. Chacune de ces informations est également datée.

Contenu des pages Les traitements les plus lourds que nous effectuons prennent en entrée le contenu des pages visitées.

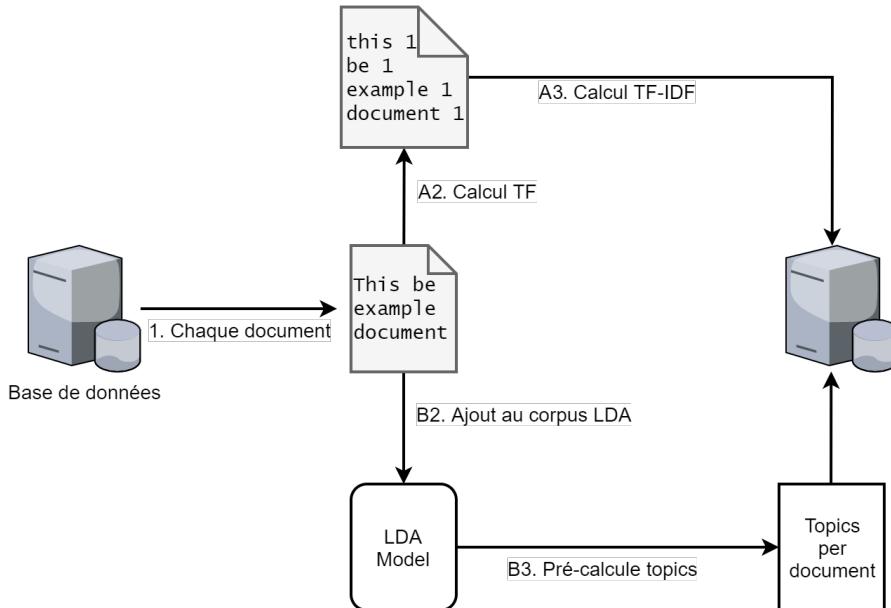


FIGURE 3.5 – Traitements du contenu des pages

Le contenu de la page est analysé indépendamment par deux algorithmes différents, chacun permettant de révéler un type d'information différent.

La figure 3.5 montre les deux traitements effectués aux documents de la base de données.

1. **Chaque document** le serveur va charger la liste entière des contenus enregistrés des pages web, obtenues comme décrites à la section 3.2.3.
- A2. **Calcul TF** La fréquence de chaque mot du document est calculée, puis enregistrée
- A3. **Calcul TF-IDF** Une fois en possession de la fréquence de chaque mot dans chaque document, le poids final TF-IDF normalisé est calculé et enregistré dans la base de données, pour chaque mot à l'intérieur de chaque document.
- B2. **Ajout au corpus LDA** Avant de générer un modèle, un traitement semblable à la branche A2 est effectué pour chaque document. Une fois tous les documents chargés, on lance l'exécution de la génération du modèle LDA.
- B3. **Pré-calcule topics** Une fois le modèle entraîné, processus qui peut facilement durer plusieurs heures, il est enregistré sur le disque. Après quoi, une multitude de requêtes sont effectuées sur le modèle afin de connaître déjà

quels sont les topics les plus probables pour chaque page de la base de données. On enregistre les résultats à nouveau dans la base de données. Ce pré-calcul va accélérer considérablement les traitements futurs sur la reconnaissance des topics significatifs pour un utilisateur en fonction des pages web qu'il a visité.

TF-IDF TF-IDF est une méthode permettant de détecter quels sont les mots les plus importants dans un document parmi l'ensemble d'un corpus. La méthode consiste purement en l'analyse de la fréquence de chaque mot dans chaque document, et ne s'occupe absolument pas de la signification des mots.

LDA LDA est un modèle qui permet de générer un nombre de sujets, thèmes ou topics, en fonction du contenu textuel d'un corpus de documents. Le modèle suppose que chaque document parle de un ou plusieurs topics, et tente de les retrouver en se basant sur la fréquence d'utilisation de ses mots en comparaison avec le reste du corpus de documents.

Requêtes du navigateur Comme mentionné précédemment, quelques informations de chaque requête du navigateur du client sont enregistrées. Ces données n'ont pas besoin d'un traitement particulier.

Etant donné que nous nous intéressons particulièrement à leur quantité, nous n'allons principalement que les compter. Cependant dû au fait de leur énorme quantité, il nous est nécessaire de pré-calculer certaines sommes avant de les servir à l'interface.

API

Le serveur a également le rôle de répondre aux demandes de l'interface, et de lui fournir les informations nécessaires pour afficher les données du client. Ces communications se font au travers d'une série de requêtes initiées par le client.

Une partie des données servies au client sont pré-calculées, comme la liste des topics tirés de LDA ou le poids TF-IDF des mots, et ne sont donc rafraîchies que périodiquement lorsque demandé.

Le reste des données, comme le nombre d'ouvertures d'une page ou le temps actif passé sur chaque page, est continuellement rafraîchi. Ces données sont donc toujours à jour.

3.2.4 Base de données

Toutes les informations sont centralisées dans une base de données MySQL. La figure 3.6 montre le schéma global des tables de la base de données. L'utilisation

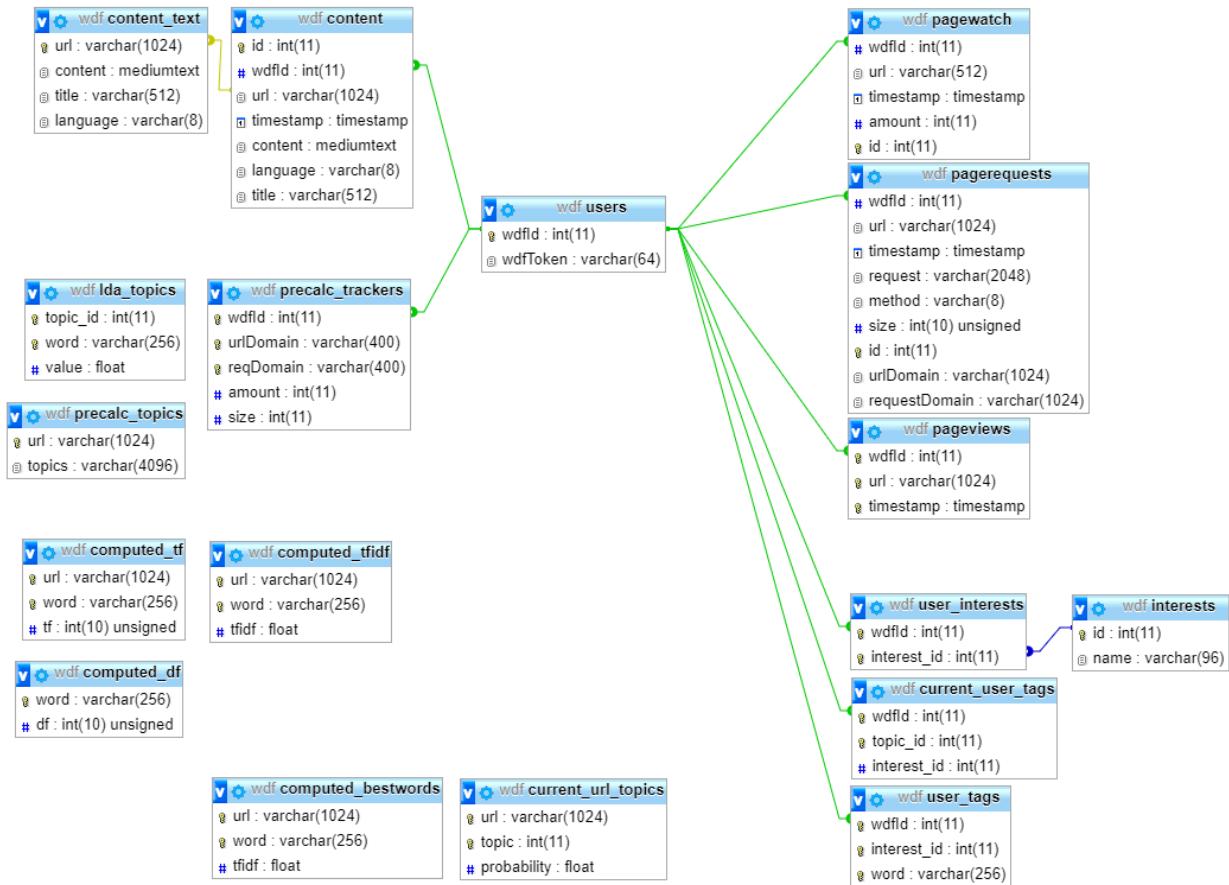


FIGURE 3.6 – Schéma des tables de la base de données

de chaque table est décrite dans la section suivante, lors de leur accès. La section suivante décrit chacun des champs des tables de la base de données.

Table **content**

La table **content** sert à stocker le contenu HTML initial des pages téléchargées. Il s'agit du contenu DOM de la page après la fin du chargement et de l'exécution du JavaScript présent sur celle-ci.

| Colonne | Description |
|------------------|---|
| id | Clé primaire artificielle |
| wdfId | Identifiant de l'utilisateur ayant accédé à la page |
| url | URL de la page |
| timestamp | Date du téléchargement |
| content | Contenu DOM entier |
| language | Langue détectée du texte |
| title | Contenu de la balise title |

Table content_text

La table **content_text** sert à stocker le contenu textuel des pages après le traitement décrit au paragraphe 3.2.3.

| Colonne | Description |
|-----------------|-----------------------------------|
| url | URL de la page |
| content | Contenu textuel lemmatisé |
| title | Contenu de la balise title |
| language | Langue détectée du texte |

Table users

La table **users** sert à stocker les utilisateurs enregistrés sur l'extension.

| Colonne | Description |
|-----------------|----------------------|
| wdfId | Numéro d'identifiant |
| wdfToken | Token du client |

Table pagewatch

La table **pagewatch** sert à enregistrer le temps passé par les utilisateurs sur les différentes pages web qu'ils visitent.

| Colonne | Description |
|------------------|---------------------------|
| wdfId | Numéro d'identifiant |
| url | URL de la page |
| timestamp | Date de visualisation |
| amount | Temps regardé [sec] |
| id | Clé primaire artificielle |

Table pagerequests

La table **pagerequests** sert à enregistrer les requêtes envoyées par le navigateur des utilisateurs.

| Colonne | Description |
|----------------------------|---------------------------|
| <code>wdfId</code> | Numéro d'identifiant |
| <code>url</code> | URL de la page |
| <code>timestamp</code> | Date de visualisation |
| <code>request</code> | URL requêtée |
| <code>method</code> | Méthode HTTP |
| <code>size</code> | Taille de la requête |
| <code>id</code> | Clé primaire artificielle |
| <code>urlDomain</code> | Domaine de l'URL actuelle |
| <code>requestDomain</code> | Domaine de l'URL requêtée |

Table `pageviews`

La table `pageviews` sert à enregistrer les ouvertures d'URL de la part des utilisateurs.

| Colonne | Description |
|------------------------|-----------------------|
| <code>wdfId</code> | Numéro d'identifiant |
| <code>url</code> | URL de la page |
| <code>timestamp</code> | Date de visualisation |

Table `user_interests`

La table `user_interests` sert à enregistrer les centres d'intérêt que les utilisateurs déclarent.

| Colonne | Description |
|--------------------------|----------------------|
| <code>wdfId</code> | Numéro d'identifiant |
| <code>interest_id</code> | Numéro d'intérêt |

Table `interests`

La table `interests` sert à enregistrer la liste de centres d'intérêt que les utilisateurs peuvent choisir.

| Colonne | Description |
|-------------------|------------------|
| <code>id</code> | Numéro d'intérêt |
| <code>name</code> | Nom de l'intérêt |

Table `current_user_tags`

La table `current_user_tags` sert à enregistrer les associations que les utilisateurs ont créée entre un topic LDA actuel et un centre d'intérêt renseigné.

| Colonne | Description |
|--------------------|----------------------|
| wdfId | Numéro d'identifiant |
| topic_id | Numéro du topic |
| interest_id | Numéro de l'intérêt |

Table user_tags

La table **user_tags** sert à enregistrer les associations que les utilisateurs ont créée entre un précédent topic LDA représenté par quelques mots et un centre d'intérêt.

| Colonne | Description |
|-----------------|----------------------|
| wdfId | Numéro d'identifiant |
| topic_id | Numéro du topic |
| words | Trois mots du topic |

Table precalc_trackers

La table **precalc_trackers** sert à enregistrer les informations pré-calculées concernant le nombre de requêtes entre les domaines visités par l'utilisateur.

| Colonne | Description |
|----------------------|----------------------------|
| wdfId | Numéro d'identifiant |
| urlDomain | Domaine de l'URL actuelle |
| requestDomain | Domaine de l'URL requêtée |
| amount | Nombre de requêtes |
| size | Taille totale des requêtes |

Table precalc_topics

La table **precalc_topics** sert à enregistrer les informations pré-calculées sur les topics relatifs à chaque page web.

| Colonne | Description |
|---------------|--------------------------|
| url | URL de la page |
| topics | JSON des topics associés |

Table lda_topics

La table **lda_topics** sert à enregistrer les informations des topics LDA du modèle actuel.

| Colonne | Description |
|-----------------|--------------------|
| topic_id | Numéro du topic |
| topics | Mot associé |
| value | Probabilité du mot |

Table computed_tf

La table **computed_tf** sert à enregistrer la valeur TF de chaque mot dans chaque page web. Cette table est uniquement là dans un but d'archivage, et n'est pas lue directement.

| Colonne | Description |
|-------------|-----------------------------|
| url | URL de la page |
| word | Mot |
| tf | Term Frequency selon TF-IDF |

Table computed_df

La table **computed_df** sert à enregistrer la valeur DF de chaque mot dans chaque page web. Cette table est uniquement là dans un but d'archivage, et n'est pas lue directement.

| Colonne | Description |
|-------------|---------------------------------|
| word | Mot |
| df | Document Frequency selon TF-IDF |

Table computed_tfidf

La table **computed_tfidf** sert à enregistrer la valeur TF-IDF de chaque mot dans chaque page web.

| Colonne | Description |
|-------------|--------------------|
| url | URL de la page |
| word | Mot |
| df | Score final TF-IDF |

Table computed_bestwords

La table **computed_bestwords** sert à enregistrer les meilleurs mots selon TF-IDF de chaque page web.

| Colonne | Description |
|--------------|--------------------|
| url | URL de la page |
| word | Mot |
| tfidf | Score final TF-IDF |

Table current_url_topics

La table **current_url_topics** sert à enregistrer les meilleurs topics du modèle LDA actuel pour chaque page web.

| Colonne | Description |
|--------------------------|-----------------------------------|
| <code>url</code> | URL de la page |
| <code>topic</code> | Numéro du topic |
| <code>probability</code> | Probabilité du topic pour la page |

3.2.5 Interface

L'interface a connu de nombreuses versions au fur et à mesure du projet. Cependant, le thème et le but commun de ces pages n'a pas changé : Montrer à l'utilisateur les informations qu'il révèle, ainsi que des possibles utilisations de celles-ci. Le design initial des visualisation était très visuel et varié et a progressé vers des pages plus utilitaires.

L'interface se divise en trois onglets distincts, chacun tentant de représenter une partie des informations : Settings, Profil, et Trackers.

Settings

La page Settings laisse la possibilité à l'utilisateur de renseigner ces centres d'intérêt en les sélectionnant parmi une liste d'une centaine d'entre-eux. Cette centaine d'intérêts sont ceux que Google utilise pour "classifier" les visiteurs de sites web utilisant Google Analytics, nous estimons donc que ces intérêts font sens.

Profil

La page Profil cherche à montrer le résultat de l'analyse des pages visitées par l'utilisateur, tentant de retrouver et de lui montrer quels sont ses centres d'intérêts. La figure 3.7 montre les différentes vues prévues initialement :

- 1 Word Cloud** Cette vue cherche à mettre rapidement en valeur les mots les plus consultés par l'utilisateur en affichant un nuage de mots, où les plus grands seraient les plus vus.
- 2 Interests graph** Cette visualisation cherche à rassembler les mots fréquemment lus par l'utilisateur en topics, eux-mêmes liés entre eux. Le but est de montrer une synthèse du Word Cloud.
- 3 Website themes** Cette partie cherchait à montrer les mots redondants ainsi que les thèmes trouvés sur certains sites web.
- 4 Most visited sites** Ce graphique en barres cherche à montrer à l'utilisateur quels sont les sites qu'il a le plus souvent visité, c'est-à-dire ouverts l'URL, peu importe le temps passé sur chaque site.
- 5 Most watched sites** Ce graphique, contrairement au 4, cherche à montrer à l'utilisateur le temps total passé à regarder chaque page.

6 History of websites Ce graphique cherche à montrer à l'utilisateur la fluctuation de sa visite de sites web sur la durée.

7 History of interests Ce graphique cherche à mettre en lumière les sujets les plus visités par l'utilisateur sur une période de temps afin de potentiellement détecter des tendances ou des changements dans son comportement.

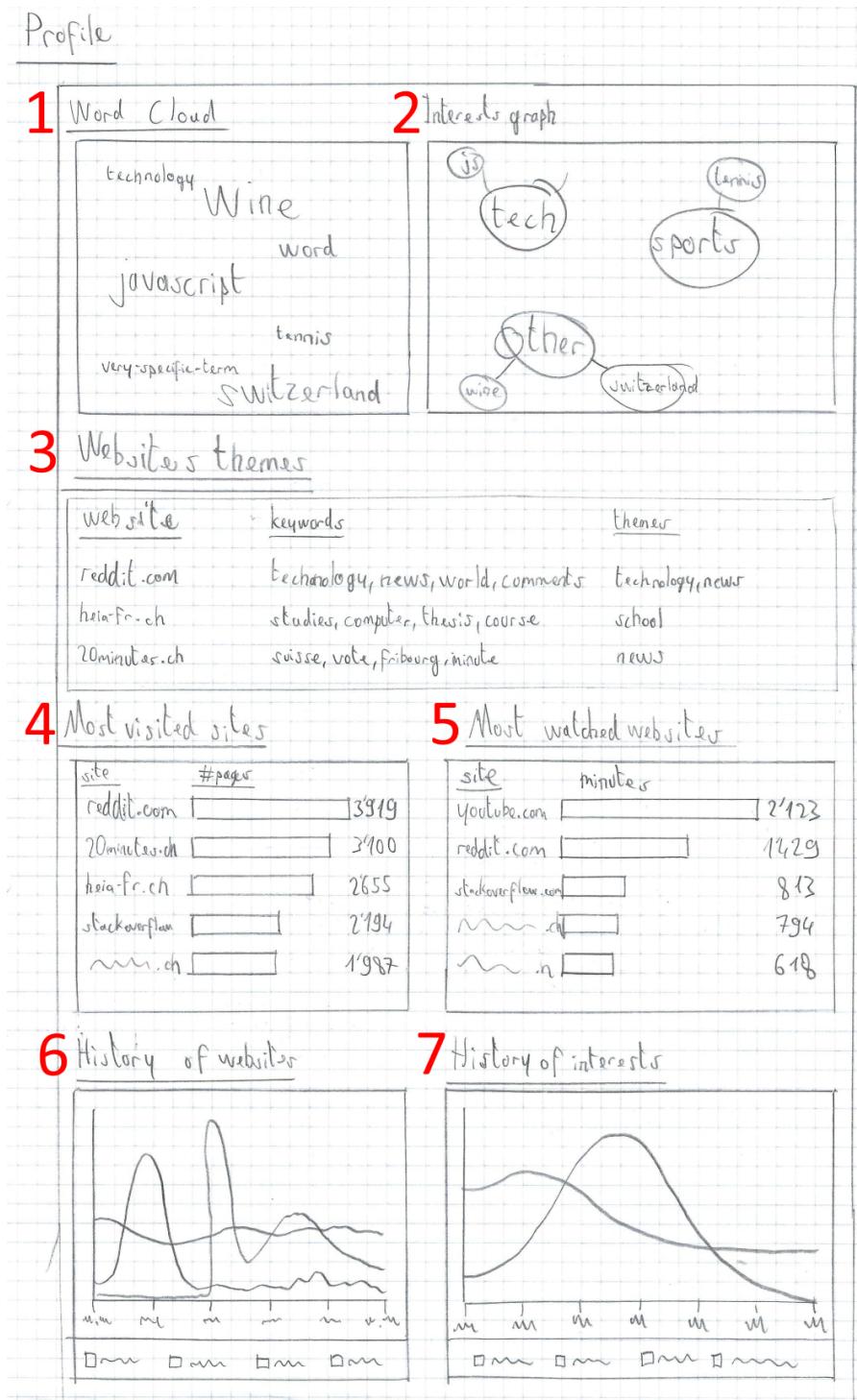


FIGURE 3.7 – Maquette de la page de Profil

Trackers

La page Trackers montre la liste des différents trackers contactés au cours de la navigation, ainsi que les domaines les ayant contactés.

Les requêtes effectuées depuis une page vers le même domaine ne sont pas comptées car on estime qu'il s'agit de trafic que l'on sait qui va prendre place, peu importe la page contactée : Il est évident qu'elle va chercher à charger du contenu provenant du même domaine.

La figure 3.8 montre les 4 premières visualisations conceptualisées de la page Trackers, et la figure 3.9 montre les 4 dernières.

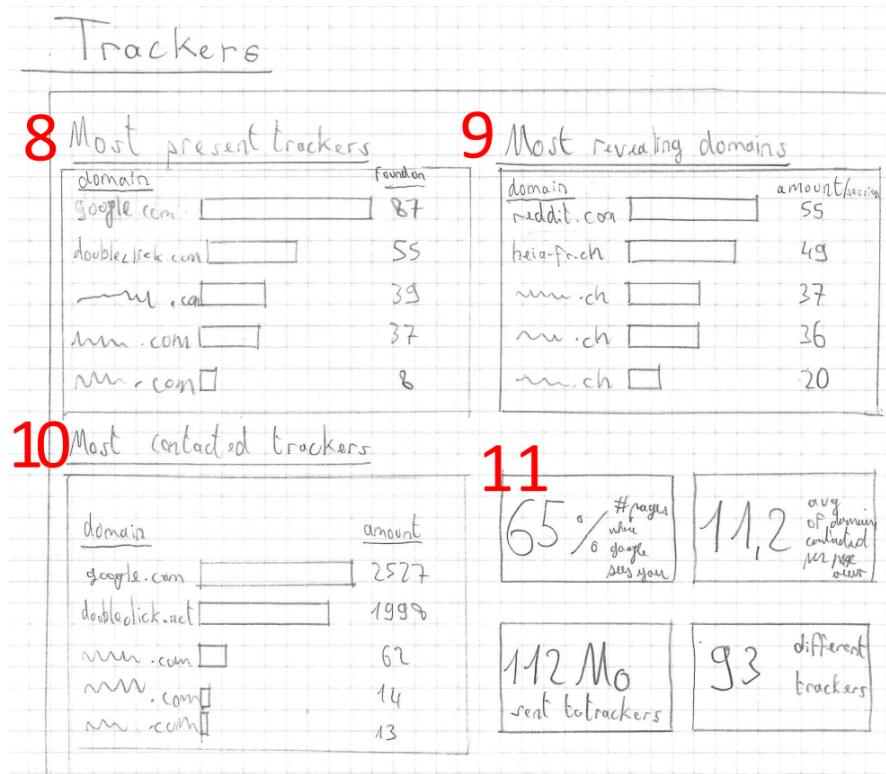


FIGURE 3.8 – Maquette de la page de Trackers

8 Most present trackers Ce graphique en barres montrait les trackers potentiels contactés depuis le plus grand nombre de pages différentes.

9 Most revealing trackers Ce graphique en barres montrait une moyenne par domaine du nombre de requêtes effectuées vers des trackers potentiels.

10 Most contacted trackers Ce graphique en barres montrait les potentiels trackers les plus contactés au total, depuis n'importe quelle page.

11 Stats Quelques nombres montrent des statistiques générales de l'utilisateur afin de lui faire prendre compte de certaines mesures. Par exemple, la taille totale d'informations envoyées aux trackers potentiels, ou le nombre de ceux-ci contactés.

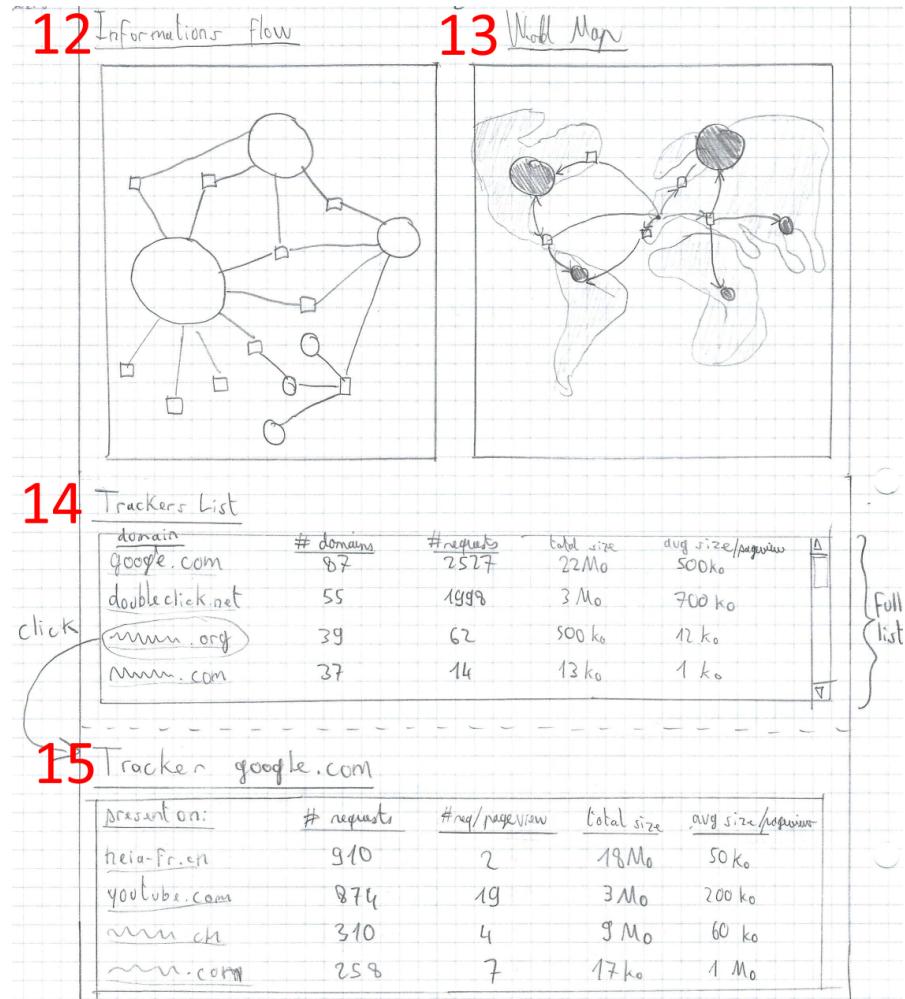


FIGURE 3.9 – Suite de la maquette de la page de Trackers

12 Informations flow Cette visualisation sous forme de graphe cherche à montrer quels sont les domaines les plus connectés entre eux. Le but est de rassembler des domaines et sous-domaines afin de montrer lesquels de ceux-ci communiquent le plus.

13 World map Cette visualisation est très semblable à la précédente. Les domaines sont à présent placés sur leur emplacement géographique afin de se rendre compte du trafic réel physique engendré par ces requêtes.

14 Trackers list Cette partie montre de manière exhaustive l'ensemble des requêtes effectuées vers un domaine. Le but est de permettre aux utilisateurs curieux de parcourir l'ensemble des données de manière plus fine. Un clic sur un tracker ouvre la vue 15.

15 Selected tracker Ce tableau s'ouvre en sélectionnant un tracker potentiel de la vue 14. Il montre l'ensemble des domaines ayant contacté le potentiel tracker sélectionné.

Topics graph Le "topics graph" cherche à rassembler les mots en thèmes, et monte d'une manière plus synthétique les thèmes estimés que l'utilisateur parcourt fréquemment. À chaque thème est lié un ou plusieurs mots, qui représentent le thème d'une manière générale. Chaque cercle du graphe représente soit un thème, soit un mot.

Le but de cette visualisation est de montrer que nous pouvons déduire des thèmes et ainsi montrer un traitement plus fin des intérêts de l'utilisateur, que simplement additionner une liste de mots. Dans le marketing, les thèmes découverts pourraient être utilisés pour labelliser les utilisateurs à qui faire apparaître une publicité.

Chapitre 4

Développement des vues

4.1 Technologies

4.1.1 Introduction

Dans cette section se trouveront des schémas représentant des algorithmes. Les étapes de ceux-ci peuvent être exécutés à trois moments différents, indiqué sur leurs schémas :

Offline Les étapes effectuées "offline" requièrent une intervention de l'administrateur. C'est à lui de décider quand ces traitements doivent intervenir car ils ne peuvent pas être effectuées pendant que le serveur fonctionne.

Serveur Les étapes effectuées sur le "serveur" sont calculées pendant que le serveur est en ligne, typiquement lorsque celui-ci reçoit une requête.

Client Les étapes s'effectuant sur le "client" sont calculées dans le navigateur de l'utilisateur actuel.

4.1.2 Offline

Certains scripts doivent être lancées ponctuellement par l'administrateur afin d'exécuter des opérations coûteuses de calcul. Le lancement de ces opérations peut se faire manuellement, ou leur lancement peut être programmé à l'aide d'un script d'automatisation.

Chaque opération est lancée à l'aide d'un script Python3.6 indépendant, se trouvant dans un dossier nommé `/script`. Ceux-ci nécessitent généralement d'accéder à la base de données pour effectuer des opérations, il est donc nécessaire de leur fournir les identifiants de connexion à la base de données, ou de les laisser puiser dans le script de configuration à la racine du projet.

4.1.3 Serveur

Le serveur est développé en Python3.6, et utilise le framework Flask. Flask permet de définir rapidement le comportement d'un serveur web basique en mappant des fonctions à des endpoints de l'API.

Par exemple, à l'aide d'une combinaison de décorateurs fournis par Flask et créés par nous-même, nous pouvons très facilement spécifier le comportement d'une méthode. La figure 4.1 montre le peu de code nécessaire pour définir l'endpoint de vérification de la connexion de l'utilisateur.

```

1 @app.route("/api/connectionState", methods=['GET'])
2 @userConnected
3 @apiMethod
4 def connectionState(wdfId):
5     return jsonify({'success': "Connected", "wdfId": wdfId})

```

FIGURE 4.1 – Code de la méthode de vérification de connexion

Naturellement, la plupart des autres méthodes nécessitent un accès à la base de données MySQL et donc du code plus conséquent.

Ici, Flask n'écoute pas directement sur les ports concernés. Le programme entier est accessible derrière un serveur Apache pour des questions de sécurité : la communication entre l'extension, l'interface et le serveur se fait en HTTPS, grâce à la gestion de SSL par Apache, et un certificat obtenu à l'aide de Let's Encrypt.

4.1.4 Client

L'interface est développée en JavaScript et TypeScript en utilisant le framework Vue.js. Comme plusieurs de ses autres équivalents, Vue.js permet d'organiser le code en composants réutilisables. Le but étant de développer une one-page app pour des raisons de réactivité de l'interface. Un des avantages de ce type de framework est donc la possibilité de configurer un router.

Ceci permet à l'utilisateur d'avoir l'impression de naviguer entre plusieurs pages distinctes, alors que tout le processus prend place dans le framework lui-même, qui se charge de changer le contenu affiché sans nécessiter de requête supplémentaire au serveur.

Au lancement de la page, l'interface entière est donc chargée sur le navigateur, et l'ensemble des données de l'utilisateur est demandée au serveur. La figure 4.2 montre le spinner affiché à l'utilisateur pendant le chargement.



Loading data from server

FIGURE 4.2 – Chargement des données

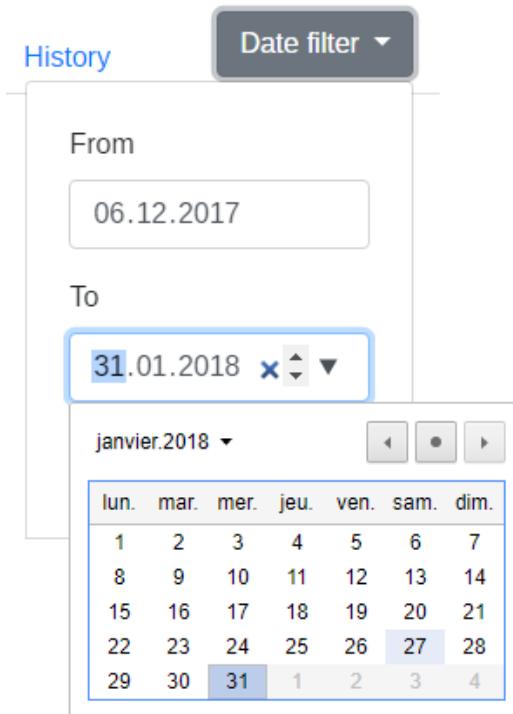


FIGURE 4.3 – Formulaire du filtre de dates

Sur les pages de Profil, l'utilisateur a ensuite la possibilité de filtrer les données affichées par date. La figure 4.3 montre le formulaire du filtre de données que l'utilisateur peut décider d'activer. Une fois activé après la sélection d'une date de début et d'une date de fin, la page entière se recharge en demandant au serveur le nouveau jeu de données correspondant aux dates entrées. Il s'agit d'un paramètre de recherche supplémentaire passé aux endpoints.

4.2 Wordcloud

4.2.1 Concept

Le wordcloud montre à l'utilisateur la liste des mots qu'il lit le plus fréquemment. Le visualisation est un amassage de mots de différentes tailles, placés d'une manière aléatoire sur un rectangle. Les mots les plus lus ont une taille plus grande afin d'attirer l'attention de l'utilisateur.

Cette visualisation cherche à donner très rapidement une impression générale des thèmes que l'utilisateur parcourt lors de sa navigation.

La figure 4.4 montre la différence entre la vue imaginée initialement et le ré-

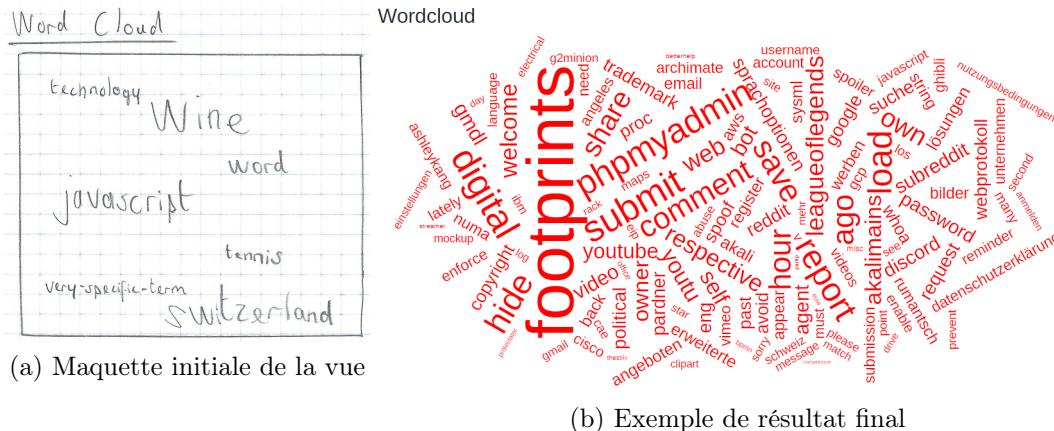


FIGURE 4.4 – Maquette initiale et résultat final de la vue Wordcloud

sultat final.

4.2.2 Données

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table `pagewatch` (figure 3.2.4).

Poids TF-IDF : Poids final selon l'algorithme TF-IDF de chaque mot. Ces données sont stockées dans la table `computed_tfidf` (figure 3.2.4).

4.2.3 Traitement

Afin de déterminer quels sont les mots affichés ainsi que leur taille sur la visualisation, on assigne un "poids" à chaque mot.

La figure 4.5 illustre le fonctionnement de l'algorithme utilisé :

- A On calcule le poids de chaque mot dans chaque document en effectuant la méthode de TF-IDF. Le poids TF-IDF de chaque mot est stocké dans la base de données, mais n'est pas constamment rafraîchi. L'opération de calcul des poids TF-IDF est une opération ponctuelle qui doit être lancée sur l'entièreté de la base de données par l'administrateur. Cette opération ne nécessite cependant pas de redémarrage du serveur.
 - B On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. Cette opération est effectuée sur le serveur et est calculée en direct par une commande MySQL, elle est donc constamment à jour. L'interface obtient ce résultat en appelant l'endpoint `/api/mostWatchedSites`

du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web visitées, comprenant entre autres pour chaque page :

- Son URL
- Le temps total de visite, en secondes
- Une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1)

On initialise un dictionnaire qui va contenir le poids de chaque mot.

- C** Pour chaque page web, on multiplie l'indice TF-IDF de chaque mot avec le temps de visualisation de la page. On additionne ce résultat au poids actuel du mot.

Une fois tous les mots de toutes les pages web traités, nous sommes en possession d'un dictionnaire nous indiquant le poids final de chaque mot. Ce poids est donc égal à la somme de l'indice TF-IDF du mot sur chaque page multiplié par le temps de visite sur cette page.

- D** On trie les mots par leur poids final, et on ne conserve que les 200 premiers. Il s'agira des 200 mots présents sur le wordcloud.

Pour chacun des 200 mots, leur taille sur le Wordcloud est égale à leur poids final.

4.2.4 Visualisation

La page va s'occuper d'agréger les résultats reçus du serveur. Ensuite, elle utilise les librairies **d3-cloud** ainsi que **d3** pour générer la visualisation du Wordcloud.

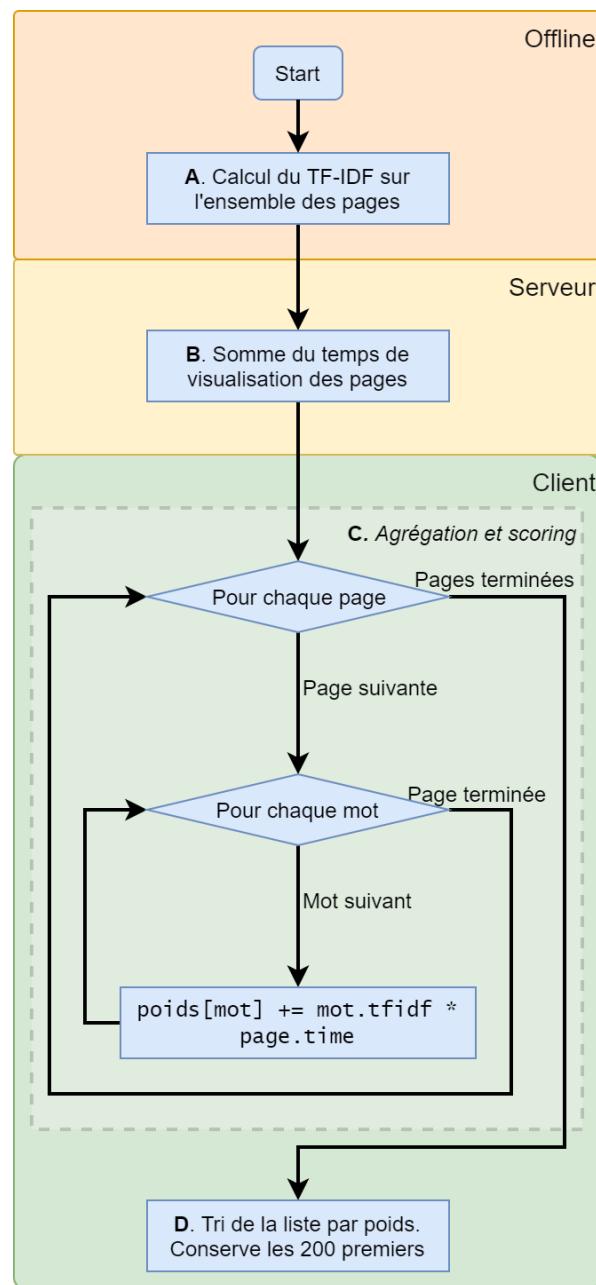


FIGURE 4.5 – Algorithme utilisé pour le Wordcloud

4.3 Topics List

4.3.1 Concept

Le "Topics List" cherche à rassembler les mots en thèmes, et montre d'une manière plus synthétique les thèmes estimés que l'utilisateur parcourt fréquemment. À chaque thème est lié un ou plusieurs mots, qui représentent le thème d'une manière générale. Chaque cercle du graphe représente soit un thème, soit un mot.

Le but de cette visualisation est de montrer que nous pouvons déduire des thèmes et ainsi montrer un traitement plus fin des intérêts de l'utilisateur, que simplement additionner une liste de mots. Dans le marketing, les thèmes découverts pourraient être utilisés pour labelliser les utilisateurs à qui faire apparaître une publicité.

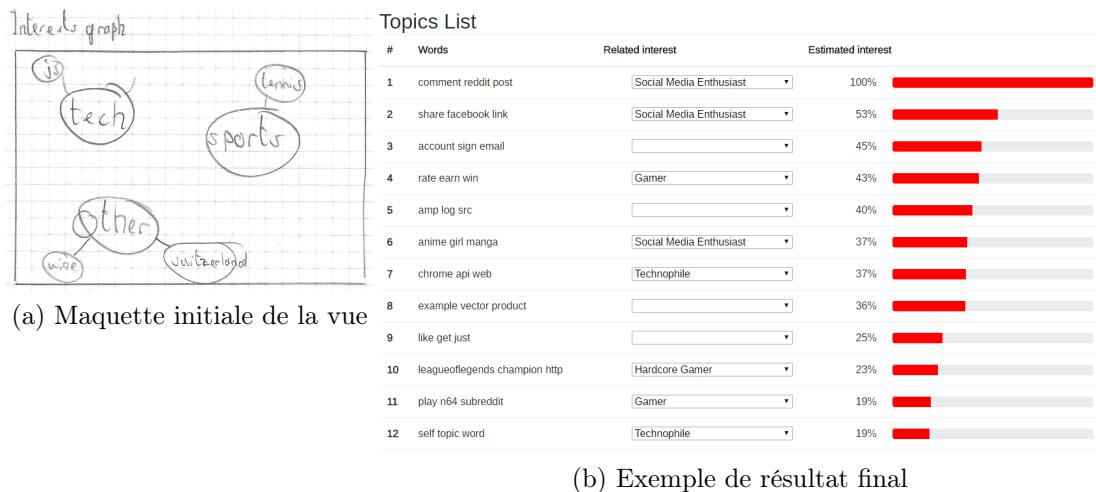


FIGURE 4.6 – Maquette initiale et résultat final de la vue Wordcloud

La figure 4.6 montre la différence entre la vue imaginée et le résultat final. On note ici que le principe même de la vue ainsi que son nom sont différents qu'initialement.

4.3.2 Données

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table `pagewatch` (figure 3.2.4).

Topics LDA : Liste des topics générés par le modèle LDA. Ces données sont stockées dans la table `lda_topics` (figure 3.2.4).

Topics par page : Liste pré-calculée des topics trouvés pour chaque page. Ces données sont stockées dans la table `current_url_topics` (figure 3.2.4).

Intérêts utilisateur : Liste des intérêts renseignés par l'utilisateur. Ces données sont stockées dans la table `user_interests` (figure 3.2.4).

Correspondances topic-intérêt : Liste des correspondances entre topic et intérêt renseignés par l'utilisateur. Ces données sont stockées dans la table `current_user_tags` (figure 3.2.4).

4.3.3 Traitement

Afin de déterminer quels sont les topics affichés ainsi que leur intérêt estimé, on assigne un "score" à chaque topic pour l'utilisateur.

L'algorithme suivant, illustré par la figure 4.7, est appliqué aux données sources :

- A** On entraîne un modèle LDA avec un nombre défini de topics (typiquement 100) sur le contenu de l'ensemble des pages web, une page web représentant un document.

Le modèle LDA est enregistré sur le disque local, et les résultats tirés de celui-ci, comme une représentation en 5 mots de chaque topic, est maintenant stockés dans la base de données. Tout ceci n'est donc pas constamment rafraîchi. L'opération d'entraînement du modèle LDA est une opération ponctuelle qui doit être lancée sur l'entièreté de la base de données par l'administrateur. Cette opération nécessite le redémarrage du serveur, car de nombreuses mesures temporaires sont touchées.

- B** Pour chaque page enregistrée, on demande au modèle LDA quels sont les 5 topics les plus probables avec leur score de probabilité. Ces informations sont également enregistrées dans la base de données. Jusqu'ici, toutes ces opérations sont donc déjà calculées et se font avant le lancement du serveur. Elles ne sont pas mises à jour en temps réel.

- C** On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. Cette opération est effectuée sur le serveur et est calculée en direct par une commande MySQL, elle est donc constamment à jour. L'interface obtient ce résultat en appelant l'endpoint `/api/mostWatchedSites` du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web visitées, comprenant entre autres pour chaque page :

- Son URL
- Le temps total de visite, en secondes
- Une liste des topics les plus significatifs selon le modèle LDA ainsi que leur probabilité

- D** L'endpoint `/api/allTopics` renvoie la liste des topics générés par LDA ainsi que leur numéro d'identifiant.

- E** L'endpoint `/api/getCurrentTags` renvoie la liste des associations que l'utilisateur a créée pour le modèle LDA courant. Il s'agit d'une liste de couples `topicId ↔ interestId`.
- F** L'endpoint `/api/interestsList` renvoie la liste des 101 intérêts globaux à tous les utilisateurs.
- G** On initialise un dictionnaire qui va contenir le score de chaque topic.
Pour chaque page web, on multiplie la probabilité de chaque topic LDA avec le temps de visualisation de la page. On additionne ce résultat au score actuel du topic.
Une fois tous les topics de toutes les pages web traités, nous sommes en possession d'un dictionnaire nous indiquant le score final de chaque topics. Ce score est donc égal à la somme de la probabilité du topic sur chaque page multiplié par le temps de visite sur cette page.
- H** On trie les topics par leur score final, et on ne conserve que les 20 premiers. Il s'agira des 20 topics présents sur la page.
- I** On sélectionne les centres d'intérêt de l'utilisateur, ainsi que les associations qu'il a déjà créée pour le modèle LDA actuel. On ajoute les associations aux topics de l'interface.

4.3.4 Visualisation

La page va s'occuper d'agréger les résultats reçus du serveur. La liste est ensuite générée sous forme d'un tableau HTML en passant par un composant Vue personnalisé. Les barres d'intérêt sont des éléments `progressbar` venant de Bootstrap.

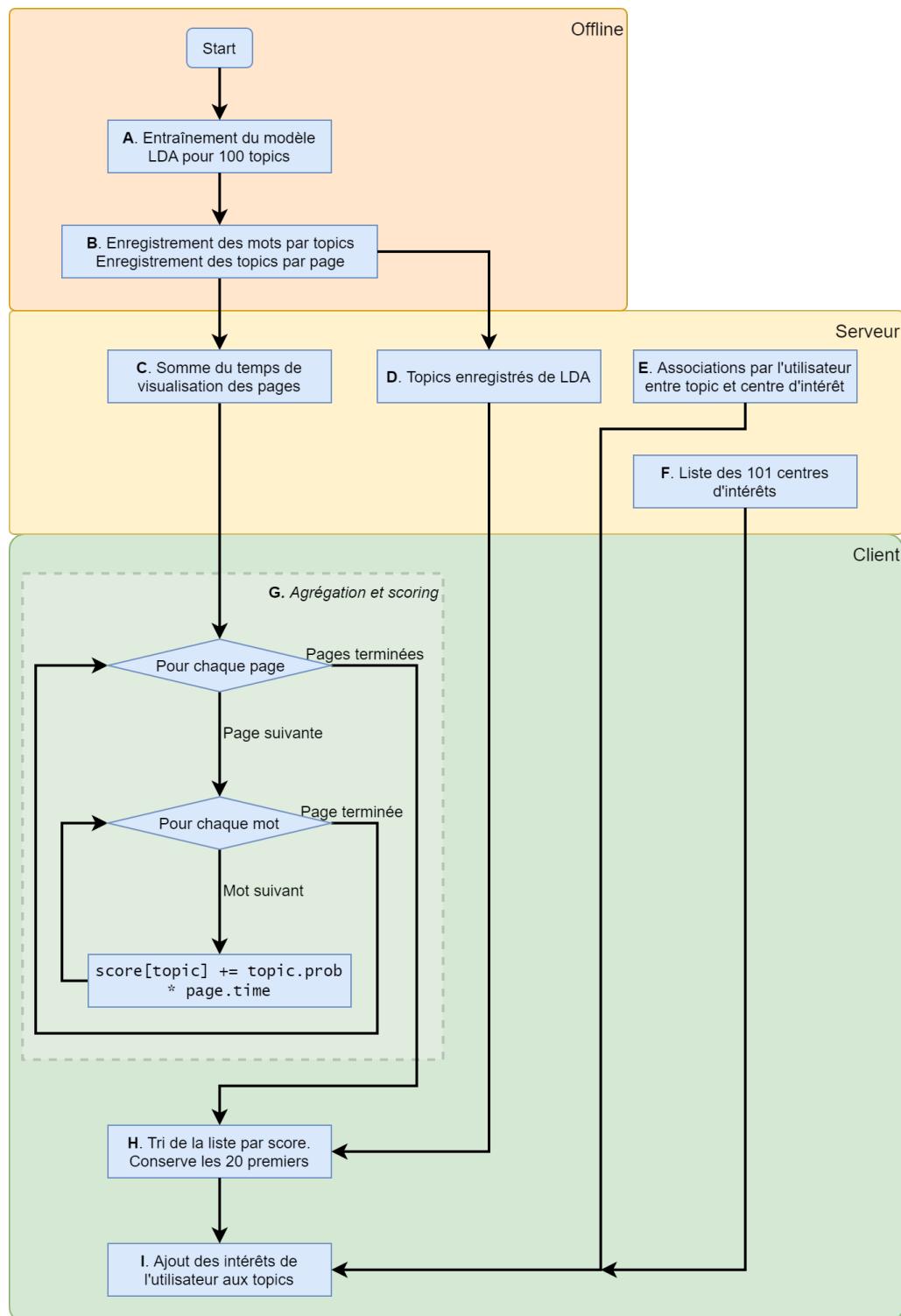


FIGURE 4.7 – Algorithme utilisé pour le Topics List

4.4 Most Watched

4.4.1 Concept

Les pages "Most Watched" et "Most Visited" montrent deux informations, mais sous une forme semblable. Ces pages affichent quelles pages web et les domaines que l'utilisateur a visité le plus. Plus précisément, "Most Watched" s'intéresse au temps réel passé à lire chaque page, et "Most Visited" s'intéresse au nombre d'ouvertures de l'URL.

Le but est ici de faire prendre conscience à l'utilisateur qu'il est possible de se rendre compte de son activité une page web, et un grand nombre d'ouvertures d'un lien ne veut pas forcément dire un grand intérêt pour cette page.

On profite également de cet espace pour afficher les mots relatifs aux pages web qui ont le plus d'intérêt, afin de montrer qu'il est possible de déterminer quels sont les mots importants d'une page web simplement en la comparant au contenu des autres pages.

La figure 4.8 montre la différence entre la vue imaginée et le résultat final.



FIGURE 4.8 – Maquette initiale et résultat final de la vue Most Watched

Comme certaines pages web demandent une connexion pour être visualisées, par exemple la page d'accueil de <https://www.facebook.com>, nous omettons volontairement une liste de pages web dans ce classement, car nous n'avons pas d'informations intéressante sur leur contenu à montrer. En effet, nous ne téléchargeons volontairement pas de copie de la page vue par l'utilisateur pour des questions de protection de données privées. Notre serveur télécharge la version publique de l'URL visitée par l'utilisateur pour en déterminer son contenu. Ainsi, il ne fait pas sens d'analyser le contenu des pages générées dynamiquement par l'utilisateur.

4.4.2 Données

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table `pagewatch` (figure 3.2.4).

Nombre de visites : Nombre total d'ouvertures de chaque URL. Ces données sont stockées dans la table `pageviews` (figure 3.2.4).

Poids TF-IDF : Poids final selon l'algorithme TF-IDF de chaque mot. Ces données sont stockées dans la table `computed_tfidf` (figure 3.2.4).

4.4.3 Traitement

Afin de déterminer quels sont les pages et les domaines affichés, on demande au serveur la liste triée des URLs les plus regardées et ouvertes.

L'algorithme suivant, illustré par la figure 4.9, est appliqué aux données sources :

- A On calcule le poids de chaque mot dans chaque document en effectuant la méthode de TF-IDF. Le poids TF-IDF de chaque mot est stocké dans la base de données, mais n'est pas constamment rafraîchi. L'opération de calcul des poids TF-IDF est une opération ponctuelle qui doit être lancée sur l'entièreté de la base de données par l'administrateur. Cette opération ne nécessite cependant pas de redémarrage du serveur.
- B On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. Cette opération est effectuée sur le serveur et est calculée en direct par une commande MySQL, elle est donc constamment à jour. L'interface obtient ce résultat en appelant l'endpoint `/api/mostWatchedSites` du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web regardées, comprenant entre autres pour chaque page :
 - Son URL
 - Le temps total de visite, en secondes
 - Une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1)
- C On effectue la somme du nombre d'ouvertures de chaque page visitée. Cette opération est effectuée sur le serveur, et l'interface obtient ce résultat en appelant l'endpoint `/api/mostVisitedSites` du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web visitées, comprenant entre autres pour chaque page :
 - Son URL
 - Le nombre total d'ouvertures
 - Une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1)

- D Une fois la liste des pages les plus regardées obtenue, on ne conserve que les 10 premières d'entre-elles pour des raisons visuelles. Ces 10 premières pages sont alors affichées.
- E Ensuite, on cherche à agréger le temps de visualisation par domaine plutôt que par page, afin d'avoir une vue d'ensemble. On additionne donc le temps passé à regarder les pages d'un même domaine.
- F On trie la nouvelle liste de domaines créée par leur temps total de visualisation, et on ne garde également que les 10 premiers d'entre-eux pour les afficher.
- G On s'occupe ensuite du traitement du nombre d'ouvertures de chaque page. On ne conserve également que les 10 plus ouvertes d'entre-elles pour des raisons visuelles, elles sont alors affichées dans la liste.
- H Ensuite, on agréger le nombre d'ouvertures par domaine plutôt que par page, afin d'avoir une vue d'ensemble. On additionne donc le temps passé à regarder les pages d'un même domaine.
- I On trie la nouvelle liste de domaines créée par leur temps total de visualisation, et on ne garde également que les 10 premiers d'entre-eux pour les afficher.

4.4.4 Visualisation

La page va s'occuper d'agréger les résultats reçus du serveur. Chaque liste est ensuite générée sous forme d'un tableau HTML en passant par un composant Vue personnalisé. Les barres relatives à la quantité exprimée par chaque tableau ajoute un élément de comparaison visuel, et sont des éléments `progressbar` venant de Bootstrap.

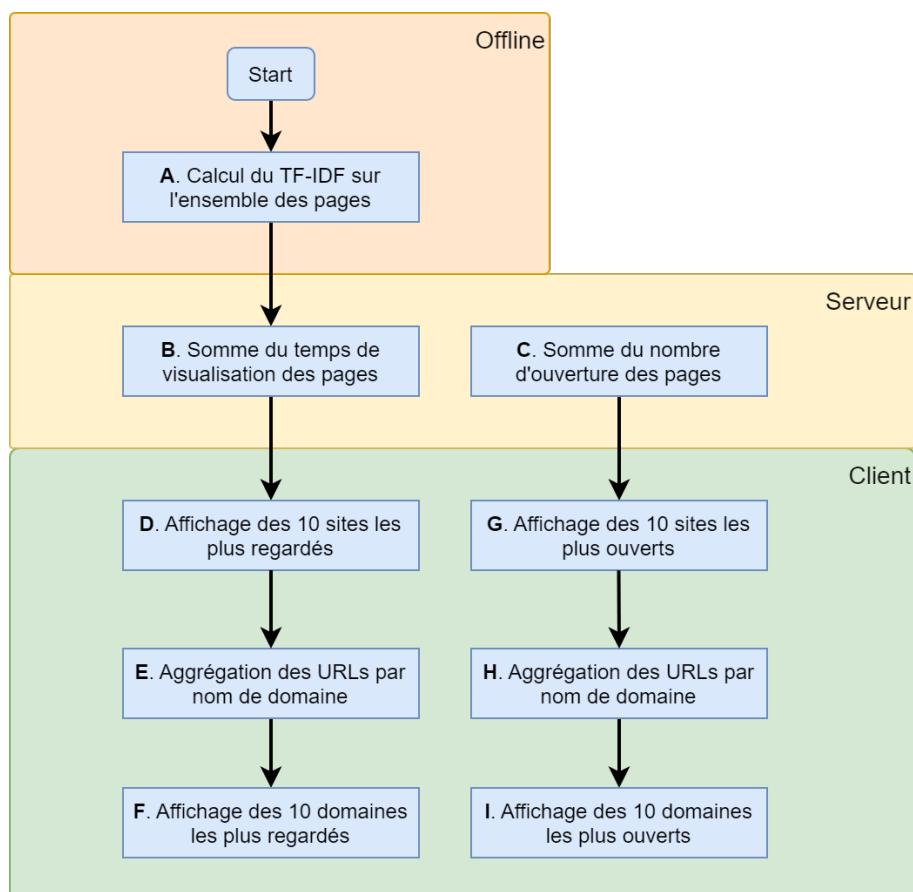


FIGURE 4.9 – Algorithme utilisé pour les pages "Most Watched" et "Most Visited"

4.5 History

4.5.1 Concept

La page "History" permet de montrer à l'utilisateur la variation de ses habitudes au cours du temps durant lequel il a utilisé l'extension. Deux graphiques sont présents sur cette page : Le premier montre la tendance à visiter des pages relatées à certains topics, et l'autre montre la tendance dans la visite de pages contenant certains mots particuliers. Le but est ici de détecter d'éventuels intérêts passagers dans le temps.

La figure 4.10 montre la différence entre la vue imaginée et le résultat final.

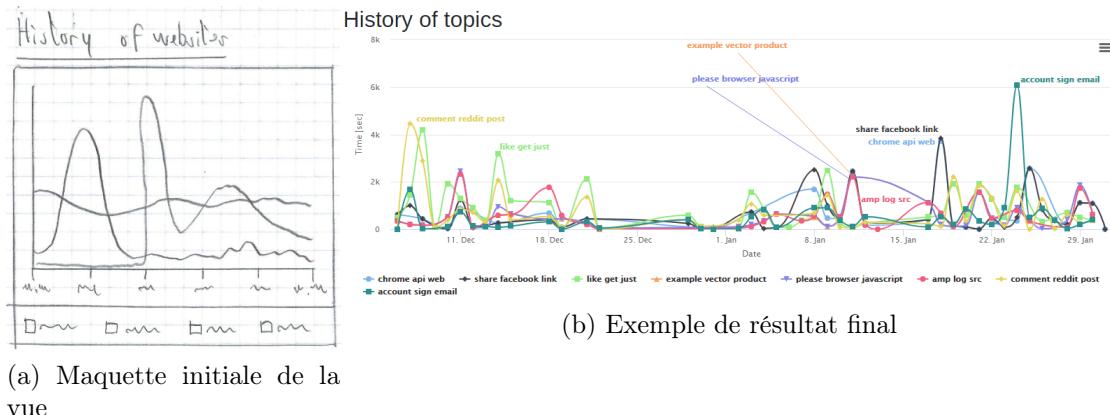


FIGURE 4.10 – Maquette initiale et résultat final de la vue History

4.5.2 Données

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table `pagewatch` (figure 3.2.4).

Historique de visualisation : Temps passé par jour sur chaque URL. Les données non agrégées proviennent de la table `pageviews` (figure 3.2.4).

Poids TF-IDF : Poids final selon l'algorithme TF-IDF de chaque mot. Ces données sont stockées dans la table `computed_tfidf` (figure 3.2.4).

Topics LDA : Liste des topics générés par le modèle LDA. Ces données sont stockées dans la table `lda_topics` (figure 3.2.4).

4.5.3 Traitement

Afin de déterminer quels sont les topics et les mots cumulant le plus d'intérêt, il est nécessaire de disposer de plusieurs sources de données et de les assembler afin d'arriver au résultat voulu. Ceci se fait en plusieurs étapes, distribuées entre le serveur et client.

L'algorithme suivant, illustré par la figure 4.11, est appliqué aux données sources :

- A** On calcule le poids de chaque mot dans chaque document en effectuant la méthode de TF-IDF.
- B** On entraîne un modèle LDA avec un nombre défini de topics (typiquement 100) sur le contenu de l'ensemble des pages web, une page web représentant un document.
- Une fois le modèle LDA entraîné, on lui demande la liste des 100 topics générés par leur représentation en 5 mots. Cette liste de topics est enregistrée dans la base de données.
- C** Pour chaque page enregistrée, on demande au modèle LDA quels sont les 5 topics les plus probables avec leur score de probabilité. Ces informations sont également enregistrées dans la base de données. Jusqu'ici, toutes ces opérations sont donc déjà calculées et se font avant le lancement du serveur. Elles ne sont pas mises à jour en temps réel.
- D** On demande à la base de données de grouper le temps de visionnage (en secondes) en une somme par jour et par URL différente. Ceci se fait au travers d'une commande MySQL et est calculé en temps réel, et s'occupe également d'"arrondir" chaque date de visualisation d'une page à la journée (au lieu de la seconde près, qui est la granularité utilisée dans la base de données).
- E** Le résultat de l'étape précédente est disponible via l'endpoint **/api/historySites**.
- F** On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. L'interface obtient ce résultat en appelant l'endpoint **/api/mostWatchedSites** du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web regardées, comprenant entre autres une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1) pour chaque page.
- G** L'endpoint **/api/allTopics** renvoie la liste des topics générés par LDA ainsi que leur numéro d'identifiant.
- H** On cherche à savoir quels sont les mots où lesquels l'utilisateur a montré le plus d'intérêt afin de les afficher sur le graphe. Pour ceci, on multiplie la valeur TF-IDF de chaque mot par le temps passé à visualiser la page. La

somme de ce calcul sur toutes les pages va nous donner l'"intérêt" final de l'utilisateur pour un mot particulier. On ne gardera ici que les 8 mots avec le plus grand intérêt estimé.

I Finalement, pour chacun des 8 mots retenus, on affiche leur intérêt journalier sur le deuxième graphique, "Words history".

J On cherche à savoir quels sont les topics où lesquels l'utilisateur a montré le plus d'intérêt afin de les afficher sur le graphe.

Voici ce que l'on effectue sur chaque page : Pour chaque topic où sa valeur selon le modèle LDA sur cette page est au-dessus de 0.1, on estime que la page parle de ce topic et on compte le temps passé à visualiser la page dans la valeur de ce topic pour la journée.

Finalement, on somme le temps passé sur chaque "topic". Le résultat de ce calcul sur toutes les pages va nous donner l'"intérêt" final de l'utilisateur pour un topic particulier. On ne gardera ici que les 8 topics avec le plus grand intérêt estimé.

K Finalement, pour chacun des 8 topics retenus, on affiche leur intérêt journalier (qui est la même somme que précédemment, mais agrégée par jour au lieu de toute la période) sur le premier graphique, "Topics history".

4.5.4 Visualisation

Les données sont finalement transformées dans un format compatible et passées à une instance configurée de la librairie HighCharts, qui génère la visualisation du graphique sur la page.

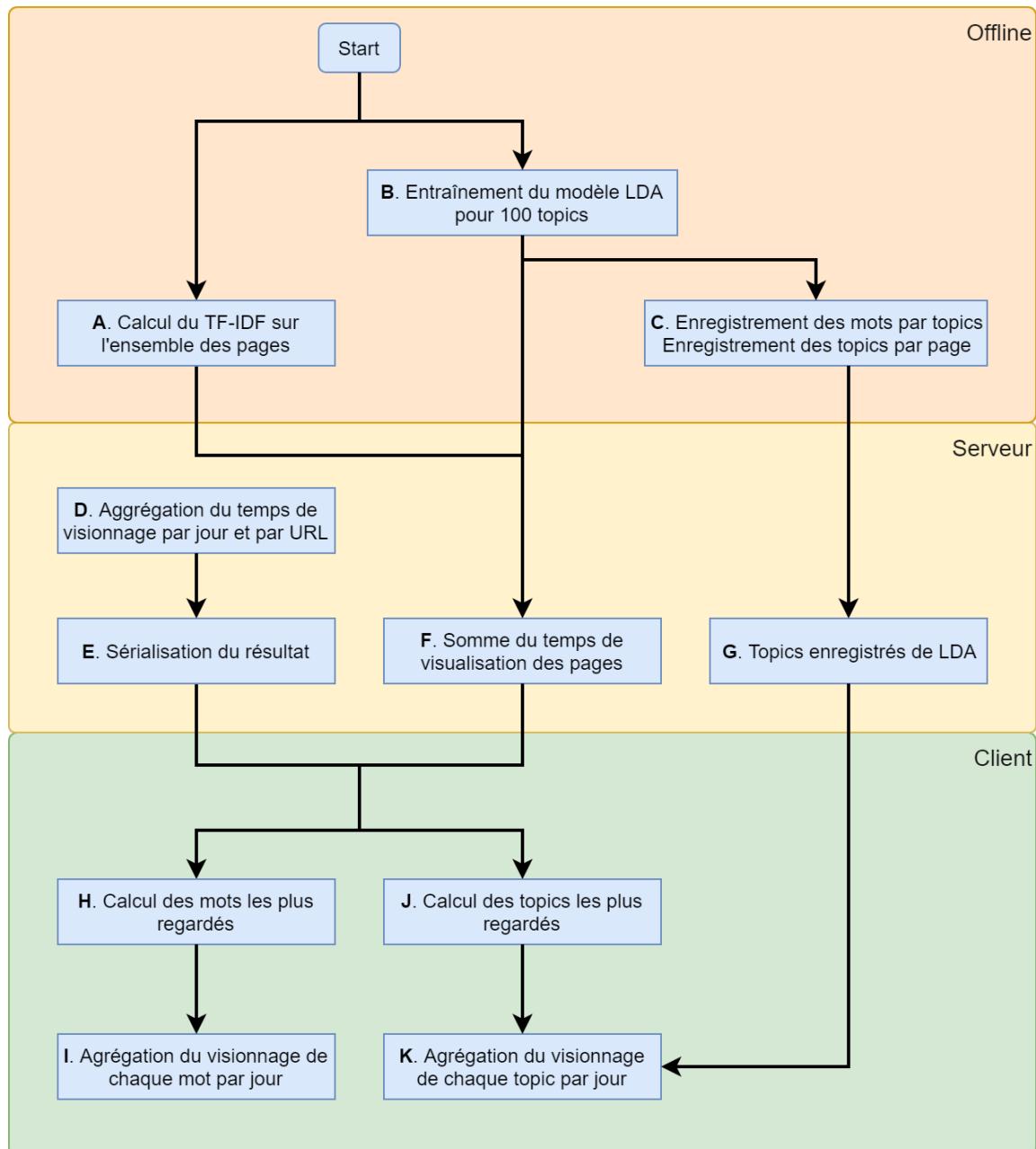


FIGURE 4.11 – Algorithme utilisé pour les graphiques de la page "History"

4.6 Trackers

4.6.1 Concept

La vue "Trackers" comporte deux pages liées à une seule source de données. Le but est de montrer à l'utilisateur que lorsqu'une page web est visitée, des informations peuvent tout de même être transmises à d'autres domaines.

La première vue de l'interface, "Most Sending", montre quels sont les domaines qui contactent beaucoup des domaines différents. Il peut ainsi voir sur cette vue quels sont les serveurs qui sont contactés lorsqu'il accède à une page web.

L'inverse est possible également. Grâce à la deuxième vue "Most receiving" il est possible de découvrir quels sont les domaines - trackers potentiels - qui sont fréquemment contactés par d'autres pages web. Chaque vue donne donc un point de vue différent sur le flux des données lorsque l'utilisateur parcourt le web.

De plus, ces deux vues peuvent interagir : Il est possible de décider de cacher certains domaines de l'une ou de l'autre vue, car par exemple l'utilisateur souhaiterait ne pas prendre en compte les données d'un certain site web, ou à l'inverse ignorer les données envoyées vers un potentiel tracker particulier.

Un bouton permettant d'activer ou de désactiver les données du domaine est présent à chaque ligne, et la désactivation de celui-ci impacte la vue des données de l'ensemble des deux pages "Trackers". La figure 4.12 montre un bouton de domaine activé et désactivé.

Ainsi il est par exemple possible de désactiver les données émises par un domaine, et de regarder quelle est la répercussion sur les données reçues par les autres.

La figure 4.13 montre la différence entre la vue imaginée et le résultat final d'une des deux pages Trackers. La figure 4.14 montre la différence entre la vue imaginée et le résultat final de la page détaillée d'un tracker.

De plus, il est possible sur chacune des deux pages d'accéder aux statistiques détaillées sur le trafic de données d'un domaine particulier. En cliquant sur un domaine, l'interface ouvre une troisième vue qui montre en détail le nombre de requêtes liées à un domaine particulier.



FIGURE 4.12 – Domaine activé, puis désactivé

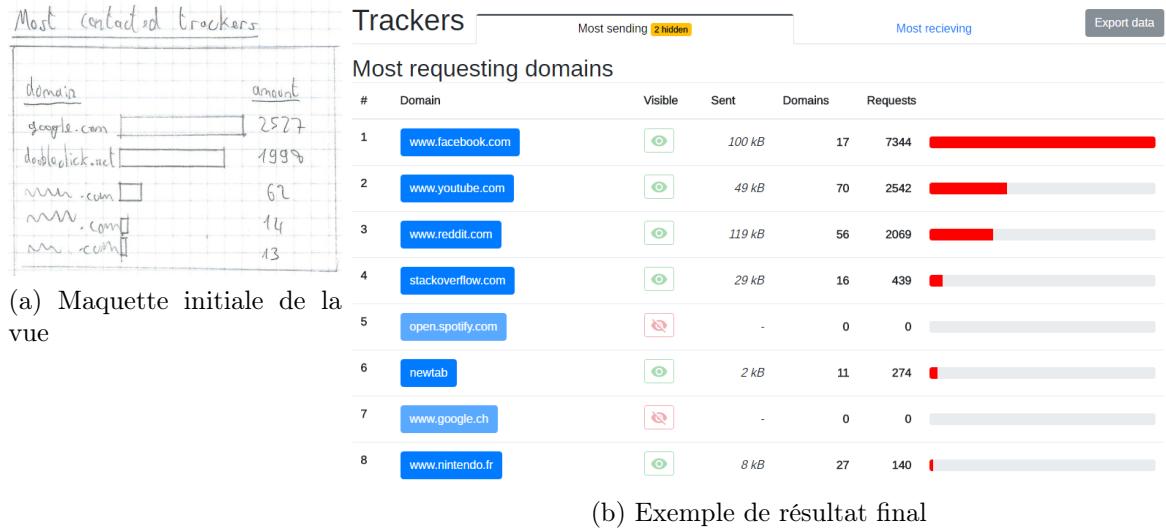
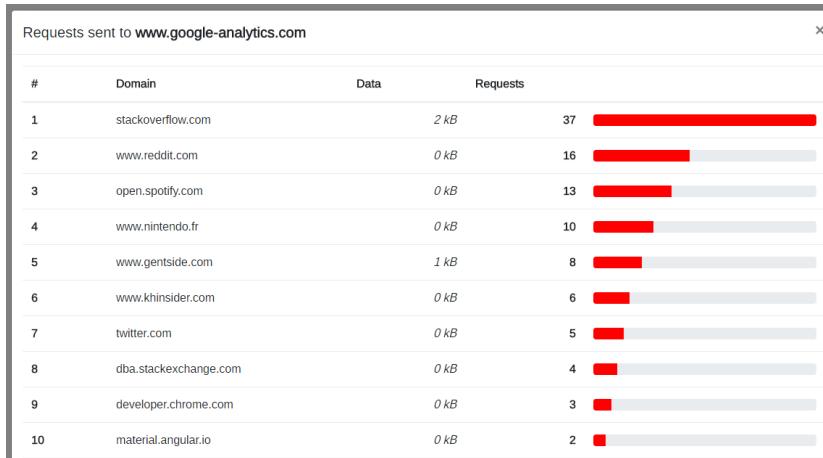


FIGURE 4.13 – Maquette initiale et résultat final d'une des vues Trackers

Tracker google.com

| présentation: | # requests | #req/pageview | Total size | Avg size/request |
|---------------|------------|---------------|------------|------------------|
| heia-fr.cn | 910 | 2 | 18 Mo | 50 ko |
| youtube.com | 874 | 19 | 3 Mo | 200 ko |
| www.ch | 310 | 4 | 9 Mo | 60 ko |
| www.com | 259 | 7 | 17 ko | 1 Mo |

(a) Maquette initiale de la vue



(b) Exemple de résultat final

FIGURE 4.14 – Maquette initiale et résultat final de la vue détaillée lors d'un clic sur un Tracker

4.6.2 Données

Une seule source de données est nécessaire à constituer cette visualisation :

Requêtes pré-calculées : Nombre de requêtes entre chaque domaine. Ces données sont stockées dans la table `precalc_trackers` (figure 3.2.4).

4.6.3 Traitement

Peu de traitements entrent en jeu dans la génération de la page Trackers. Il s'agit principalement de calculer la somme d'une liste de domaines. Ceci est illustré par la figure 4.15 :

- A On additionne le nombre de requêtes faite pour un domaine vers un autre domaine, et on enregistre le total pour chaque paire par utilisateur. Toutes ces données sont pré-calculées avant le lancement du serveur. Il s'agit de calculer le nombre de requêtes de chaque domaine vers chaque autre domaine pour chaque utilisateur. La nécessité de pré-calculer ces données vient de leur quantité brute. Les compter sur le moment pour chaque requête demande trop de temps.
- B L'endpoint `/api/getTrackers` sert l'ensemble des résultats enregistrés pour l'utilisateur.
- C On cherche ici les domaines ayant reçu le plus de requêtes. Nous allons donc effectuer regrouper les requêtes faites par leur nom de domaine de destination, et effectuer la somme des autres mesures.
- D Inversement à l'étape précédente, on cherche cette fois les domaines ayant envoyé le plus de requêtes. Nous allons regrouper les requêtes par leur domaine d'envoi, et effectuer la somme des autres mesures.
- E Les deux listes obtenues aux étapes précédentes sont alors affichées dans leur page correspondante.
- F L'utilisateur peut choisir de cacher ou d'afficher un certain domaine d'une des vues. Ceci lance alors un nouveau calcul à partir de l'étape C. Aucune communication avec le serveur n'est nécessaire : Les données reçues initialement sont préservées.
- G L'utilisateur peut également choisir d'afficher les requêtes d'un domaine particulier.
- H Dans le cas de la sélection d'un domaine à afficher en détail, la liste des requêtes est filtrée est l'interface n'affiche que les requêtes concernant le domaine souhaité.

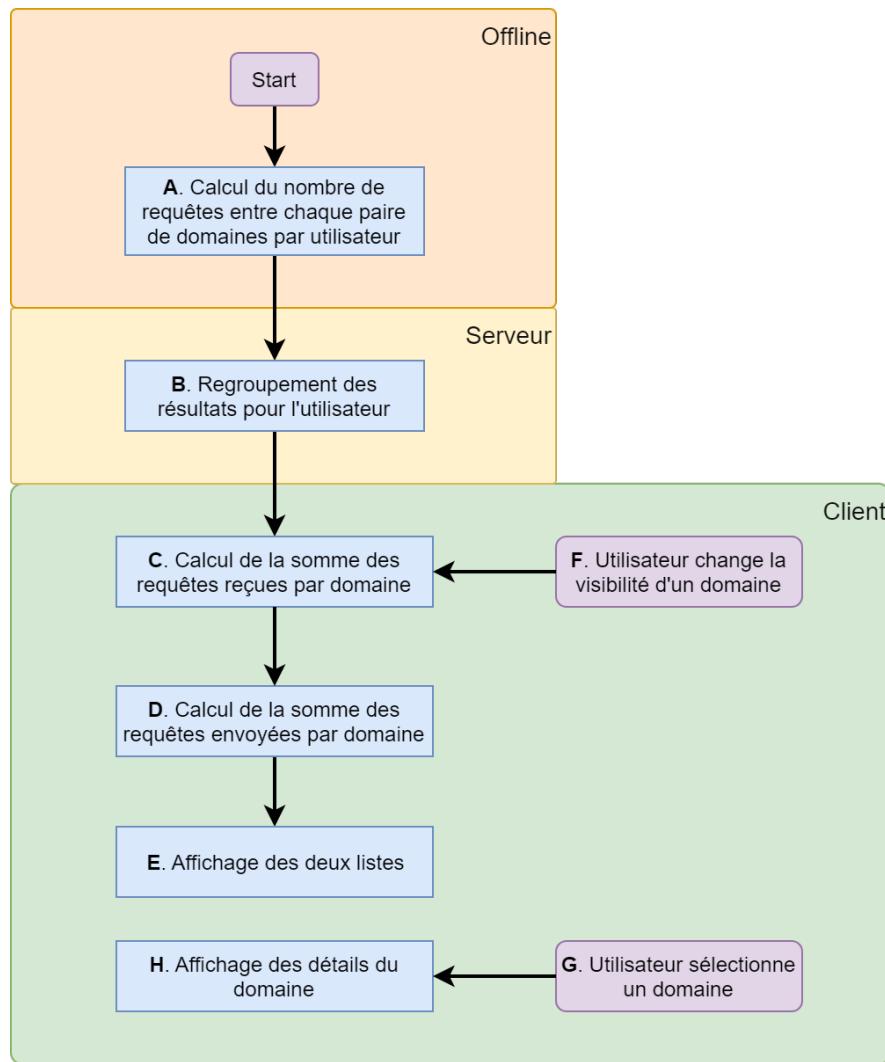


FIGURE 4.15 – Algorithme utilisé pour les données des pages "Trackers"

4.6.4 Visualisation

Les deux listes sont des tableaux HTML stylisés par Bootstrap. La gestion de leur interaction et de leur affichage est gérée par plusieurs composants Vue imbriqués.

4.7 Stats

4.7.1 Concept

Le but de la vue Stats est de résumer très rapidement en quelques nombres la quantité de données échangées entre les différents domaines visités par l'utilisateur.

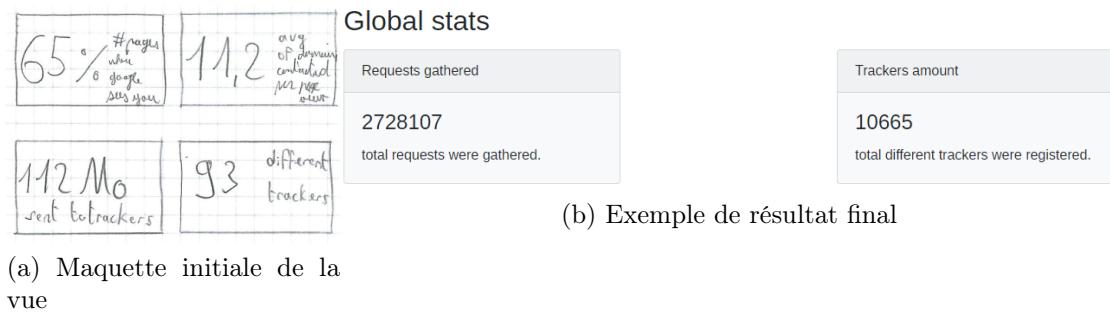


FIGURE 4.16 – Maquette initiale et résultat final d'une vue Stats

4.7.2 Données

Une seule source de données est nécessaire à constituer cette visualisation :

Requêtes pré-calculées : Nombre de requêtes entre chaque domaine. Ces données sont stockées dans la table `precalc_trackers` (figure 3.2.4).

4.7.3 Traitement

Très peu de traitements sont nécessaires pour cette vue. La figure 4.17 montre le traitement effectué aux données avant de les afficher.

Les données nécessaires à cette vue sont calculées en temps réel : Il s'agit simplement de compter le nombre de requêtes enregistrées dans la table pré-calculée, ainsi que le nombre unique de nom de domaines.

- A On additionne le nombre de requêtes faite pour un domaine vers un autre domaine, et on enregistre le total pour chaque paire par utilisateur. Toutes ces données sont pré-calculées avant le lancement du serveur. Il s'agit de calculer le nombre de requêtes de chaque domaine vers chaque autre domaine pour chaque utilisateur. La nécessité de pré-calculer ces données vient de leur quantité brute. Les compter sur le moment pour chaque requête demande trop de temps.

- B** On calcule le total de certaines valeurs pour les requêtes de l'utilisateur :
Par exemple, le nombre total effectuées, et le nombre total de domaines différents.
- C** On calcule ici certaines valeurs globales pour l'ensemble des utilisateurs.
Par exemple le nombre total effectuées, et le nombre total de domaines différents.
- D et E** On affiche les nombres résultats des opérations précédentes dans leur page respective, à savoir "Your stats" présentant les statistiques de l'utilisateur uniquement, ou "Global stats" affichant les statistiques globales sur la base de données.

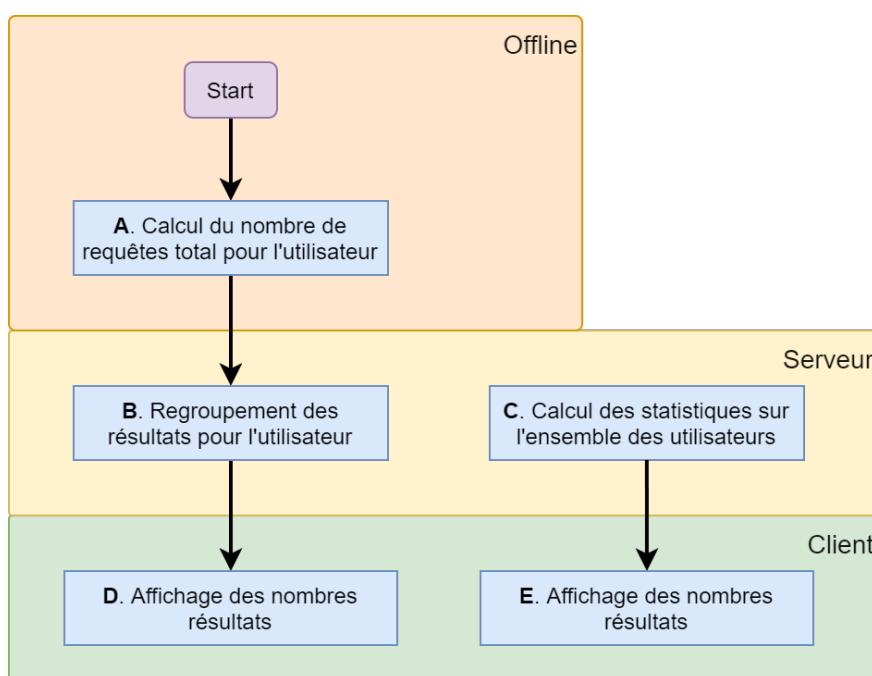


FIGURE 4.17 – Algorithme utilisé sur les données de la page "Stats"

4.7.4 Visualisation

Les nombres renvoyés par le serveur sont simplement affichés dans des cards de Bootstrap.

4.8 Settings

4.8.1 Concept

Le but de la page Settings est de montrer à l'utilisateur la possibilité de renseigner ces centres d'intérêt. Ces informations nous seront utiles par la suite pour tenter d'associer les centres d'intérêt entrés ici avec les topics que nous estimons être importants.

La figure 4.18 montre un formulaire à remplir par l'utilisateur. Celui-ci cherche ces centres d'intérêts parmi une liste de 101, organisés hiérarchiquement. Un maximum de 10 centres d'intérêts peuvent être choisis.

Settings

Interest fields

FIGURE 4.18 – Champ d'entrée des centres d'intérêts

Une fois que l'utilisateur a défini des centres d'intérêt, il peut donner des informations sur la page Topics (voir section 4.3). Les centres d'intérêts choisis sur la page Settings sont donc "simplement" utilisés pour définir quels seront les choix de centres d'intérêts possibles sur la page Topics List. La figure 4.19 montre la fenêtre déroulante de sélection d'un centre d'intérêt pour un topic donné.

| # | Words | Related interest |
|---|---------------------|--|
| 1 | comment reddit post | <div style="border: 1px solid #ccc; padding: 5px;"> Social Media Enthusiast <ul style="list-style-type: none"> Foodie Nightlife Enthusiast Gamer Hardcore Gamer Roleplaying Game Fan Electronica & Dance Music Fan Social Media Enthusiast Technophile </div> |

FIGURE 4.19 – Sélection d'intérêt sur un topic

Il est demandé aux utilisateurs d'ajouter une association entre un topic et un intérêt lorsque cela semble lui faire sens. Ainsi, nous pouvons savoir quels topics de l'utilisateur nous avons réussi à identifier. En effet, lorsqu'un utilisateur associe un centre d'intérêt à un topic, cela signifie non seulement que le topic trouvé a du sens en soi, mais en plus qu'il est intéressant pour l'utilisateur.

4.8.2 Données

Deux sources de données sont nécessaires à constituer la vue Settings :

Liste des 101 centres d'intérêts : Ces données sont stockées dans la table `interests` (figure 3.2.4).

Centres d'intérêt de l'utilisateur : Ces données sont stockées dans la table `user_interests` (figure 3.2.4).

4.8.3 Traitement

Très peu de traitements sont nécessaires pour cette vue. La plupart des données nécessaires à cette vue sont calculées en temps réel : Il s'agit simplement de compter le nombre de requêtes enregistrées dans la table pré-calculée, ainsi que le nombre unique de nom de domaines.

La figure 4.17 montre le traitement effectué aux données avant de les afficher.

- A La liste pré-définie de centres d'intérêts est ajoutée à la base de données. Cette opération n'est à effectuer qu'une seule fois, le but de cette liste n'est pas de changer. Dans notre cas, la liste des centres d'intérêts provient de ceux utilisés par la classification des utilisateurs par Google.
- B On envoie la liste des centres d'intérêts et de leur nom au client
- C On envoie également la liste des centres d'intérêts que l'utilisateur a déjà renseignés, afin de les afficher directement comme étant entrés dans le formulaire.
- E L'utilisateur peut donc changer ses centres d'intérêts, puis envoyer sa nouvelle liste au serveur.
- F Les nouveaux centres d'intérêts sont donc enregistrés, et la nouvelle liste est envoyée une fois supplémentaire au client, reprenant au point C de l'algorithme.

4.8.4 Visualisation

Le champ du formulaire permettant d'entrer des centres d'intérêts en affichant leur hiérarchisation est une instance configurée de la librairie JavaScript nommée

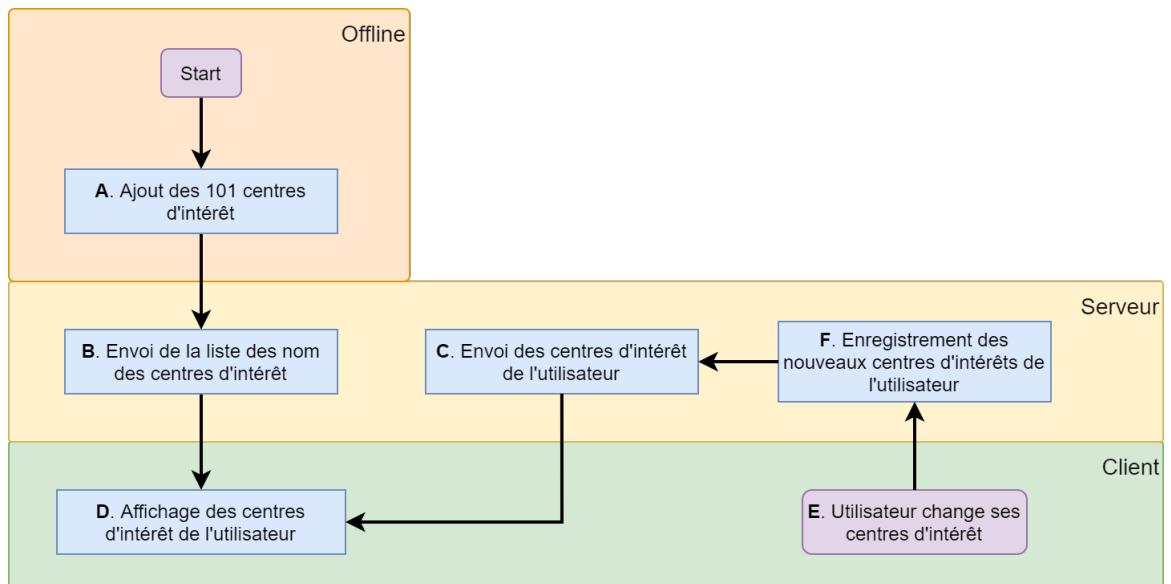


FIGURE 4.20 – Algorithme utilisé pour les données de la page "Settings"

Selectize[32], dont le but est de proposer des champs de formulaire personnalisés tels que celui-ci.

Chapitre 5

Résultats

5.1 Processus de test

Une fois toutes les parties du projet réalisées, à savoir l'extension navigateur, le serveur et l'interface, nous avons cherché des utilisateurs volontaires pour installer l'extension et l'utiliser pendant une période de 4 semaines. Bien qu'initialement prévue pour un grand panel d'utilisateurs, les restrictions temporelles ont limité la quantité d'utilisateurs que nous avons pu atteindre.

Un total de 10 utilisateurs volontaires ont installé l'extension. Parmi eux, 8 ont été identifiés comme ayant une activité de navigation sur Chrome assez grande pour contribuer à l'étude. (Ceux étant jugés inactifs totalisent moins de 10 minutes d'activité).

5.1.1 Inputs

En plus de récolter les données des utilisateurs et de les afficher, il leur a également été demandé de remplir quelques informations sur eux-même afin de pouvoir valider certains points de notre étude. Deux formulaires ont été mis en place à cette fin.

Centres d'intérêts

La figure 4.18 montre un premier formulaire à remplir par l'utilisateur. Accessible via le lien vers la page "Settings", on lui demande ici de trouver et renseigner quelques uns de ces centres d'intérêt parmi une centaine.

Un maximum de 10 centres d'intérêts peuvent être définis. Le but de laisser à l'utilisateur entrer des centres d'intérêts est de pouvoir nous rendre compte si les topics que nous lui proposons sont proches de ces centres d'intérêt.

Association topic et intérêt

Une fois que l'utilisateur a défini des centres d'intérêt, il peut donner des informations sur les topics que nous lui suggérons sur la page Topics List (voir section 4.3). La figure 4.19 montre la fenêtre déroulante de sélection d'un centre d'intérêt pour un topic donné.

Il est demandé aux utilisateurs d'ajouter une association entre un topic et un intérêt lorsque cela semble lui faire sens. Ainsi, nous pouvons savoir quels topics de l'utilisateur nous avons réussi à identifier. En effet, lorsqu'un utilisateur associe un centre d'intérêt à un topic, cela signifie non seulement que le topic trouvé a du sens en soi, mais en plus qu'il est intéressant pour l'utilisateur.

Sur l'échantillon de 8 personnes actives, 6 personnes ont ajouté des associations aux 20 topics proposés sur leur page.

5.2 Evaluation

Après une récolte des données sur environ 4 semaines, nous pouvons nous intéresser aux résultats que nous avons récoltés.

5.2.1 Modèles

Une première étape est de se pencher sur les modèles que nous avons générés sur la base des données elles-mêmes. Nous utilisons principalement deux algorithmes qui se basent sur le contenu des pages pour en déterminer leur thème : TF-IDF, et LDA.

Ces modèles se basent sur le contenu des pages visitées des utilisateurs. Etant donné que nous ne récupérons pas directement le contenu de la page visitée par un utilisateur mais uniquement une partie de son URL, nous ne pouvons pas garantir que le contenu que nous récupérons d'une page soit effectivement le contenu que l'utilisateur voit sur son écran. En effet, beaucoup de pages web aujourd'hui sont liées à une application qui demande une authentification de l'utilisateur pour être affichée.

Par exemple, une grande partie des pages d'un réseau social peuvent nécessiter la connexion de l'utilisateur pour être affichée.

De plus, il n'existe pas de méthode autre que empiriques pour juger de la qualité des résultats de ce type d'algorithmes. Nous nous attendons donc à des résultats imparfaits, mais allons tenter de juger au mieux leur efficacité malgré ceci.

TF-IDF

Résultats Le fonctionnement de TF-IDF est décrit à la section 2.4.2. Juger les résultats du calcul de TF-IDF sur des documents est une tâche non triviale. Cela revient principalement à vérifier manuellement que les mots ayant le poids le plus élevé pour certaines pages soit significatif de leur sujet.

Intéressons-nous donc aux mots ayant le plus de poids trouvés pour les 20 pages les plus regardées, par exemple. Le tableau 5.1 illustre les 20 pages les plus regardées avec leurs mots associés, ainsi que plusieurs étapes amenant à une estimation finale de l'adéquation des mots trouvés avec le contenu de la page. Voici comment se lit le tableau :

URL URL de la page concernée. Certaines URLs trop longues ont été raccourcies ici

Mot 1, 2, 3 3 meilleurs mots dans l'ordre décroissant décrivant la page selon TF-IDF.

Pub(lique) Est-ce que la page nécessite une connexion afin d'accéder à son contenu principal.

Con(tenu) Est-ce que le principal contenu de la page est textuel ?

Mot Est-ce que chacun des mots trouvés sur la page fait sens dans une langue connue ?

Adé(quat) Est-ce que l'ensemble des mots trouvés forme un potentiel résumé adéquat du contenu de la page ?

Les 4 premières colonnes (URL et 3 mots) proviennent de la base de données, tandis que les 4 dernières colonnes sont le résultat d'une évaluation manuelle des critères décrits. Un "OUI" dans une colonne indique que la page a passé le critère défini, contrairement à un "NON". Un "NON" dans une colonne entraîne automatiquement un "NON" dans les colonnes situées les plus à droite.

Réflexion Le fonctionnement de TF-IDF est décrit à la section 2.4.2. Le tableau 5.1 montre un résumé des résultats que l'on peut récupérer précédent tableau. On remarque que sur les 20 URLs entrées, seules 7 valent vraiment la peine d'être parcourues par notre algorithme, par élimination à causes des deux premières raisons énoncées. Cependant, sur les 7 URLs contenant du texte intéressant, TF-IDF a été capable d'en résumer adéquatement 5 d'entre-elles.

Ce résultat est loin d'être parfait, mais il montre tout de même qu'il est possible d'automatiser la recherche de mots importants sur des pages lorsque les conditions sont favorables à notre approche.

FIGURE 5.1 – 20 URLs les plus regardées et leurs meilleurs mots selon TF-IDF

| URL | Mot 1 | Mot 2 | Mot 3 | Pub | Con | Mot | Adé |
|---|-----------------|-------------|---------|-----|-----|-----|-----|
| http://wdf.sdiipi.ch/ | footprints | digital | web | OUI | OUI | OUI | OUI |
| https://www.draw.io/ | gndl | eng | proc | OUI | NON | NON | NON |
| https://www.reddit.com/r/videos/ | submit | load | report | OUI | OUI | OUI | NON |
| https://www.google.co.uk/search | eingabetaste | suche | drücke | OUI | NON | NON | NON |
| http://game110.idlekiller.com/ | explorer | chrome | browser | OUI | OUI | OUI | OUI |
| http://df.sdiipi.ch/phpmyadmin/sql.php | phpmyadmin | past | welcome | NON | NON | NON | NON |
| https://web.whatsapp.com/ | whatapp | macos | mozilla | NON | NON | NON | NON |
| http://hexaclicker.github.io/ | hexa | dp | level | OUI | OUI | OUI | OUI |
| http://blankmediagames.com/TownofSalem/ | salem | adobe | town | OUI | NON | NON | NON |
| http://www.jeuvideo.com/ | jeu | annonce | bande | OUI | OUI | OUI | OUI |
| https://discordapp.com/channels/217...408/217...408 | own | respective | owner | NON | NON | NON | NON |
| https://www.reddit.com/r/leagueoflegends/ | leagueoflegends | submit | self | OUI | OUI | OUI | OUI |
| https://www.reddit.com/ | bot | agent | pardner | OUI | OUI | OUI | NON |
| https://s3-fr.gladiatus.gameforge.com/game/index.php | eingabetaste | suche | drücke | OUI | NON | NON | NON |
| https://docs.google.com/presentation/d/1IB...15w/edit | de | gameforge | vous | NON | NON | NON | NON |
| https://twitter.com/ | row5w | gecb...el5w | slide | NON | NON | NON | NON |
| http://df.sdiipi.ch/phpmyadmin/db_structure.php | tweet | foto | hast | OUI | NON | NON | NON |
| | phpmyadmin | past | welcome | NON | NON | NON | NON |

TABLE 5.1 – Résumé des résultats de TF-IDF

| | |
|----------------------------------|----|
| URLs initiales | 20 |
| Pages publiques | 14 |
| Contenu textuel principal | 7 |
| Mots sensés | 7 |
| Mots adéquats | 5 |

LDA

Résultats Dans la première semaine de récolte des résultats, une recherche empirique sur les paramètres à fournir au modèle a été effectuée. Le paramètre du nombre de topics a été fixé à 100 ; cela semblait un bon compromis, car 50 générait un nombre de trop restreint pour définir de manière assez précise quels étaient les thèmes d'un utilisateur, et 200 générait beaucoup de topics qui n'avaient pas de sens en eux-mêmes.

Le tableau 5.2 montre l'ensemble des 100 topics générés, ainsi que plusieurs étapes amenant à une estimation finale de la pertinence du thème trouvé. Voici comment se lit le tableau :

Mots 5 meilleurs mots pour décrire le topic

Mot Est-ce que chaque mot, pris séparément, a un sens ?

Thè(me) Est-ce que les mots ont un thème en commun discernable ?

Thème commun Quel est le thème en commun des mots ?

S(e)ns Est-ce que le thème trouvé peut avoir un sens en tant que centre d'intérêt ?

Les mots de la première proviennent de la base de données, tandis que les 4 dernières colonnes sont le résultat d'une évaluation manuelle des critères décrits. Un "OUI" dans une colonne indique que la page a passé le critère défini, contrairement à un "NON". Un "NON" dans une colonne entraîne automatiquement un "NON" dans les colonnes situées les plus à droite.

TABLE 5.2 – Topics LDA

| Mots | Mot | Thè | Thème commun | Sns |
|--|-----|-----|--------------------|-----|
| play, n64, subreddit, game, vox | OUI | OUI | Jeu vidéo | OUI |
| add, ad, free, div, class | OUI | NON | - | NON |
| game, studio, software, entertainment, ... | OUI | OUI | Dévelop. jeu vidéo | OUI |
| dog, sign, dogbuddy, amp, style | OUI | NON | - | NON |

| Mots | Mot | Thè | Thème commun | Sns |
|--|-----|-----|----------------------|-----|
| vue, component, data, will, render | OUI | OUI | Vue.js framework | OUI |
| example, vector, product, displaystyle, cross | OUI | OUI | Librairie de graphes | OUI |
| please, browser, javascript, intel, vi | OUI | NON | - | NON |
| self, topic, word, model, corpus | OUI | OUI | Topic modelling | OUI |
| antwort, google, bleiben, div, thema | OUI | NON | - | NON |
| chf, prix, km, ch, gris | OUI | NON | - | NON |
| chrome, api, web, devtools, headless | OUI | OUI | Chrome API | OUI |
| github, sign, data, issue, tab | OUI | OUI | Github | OUI |
| also, system, love, note, may | OUI | NON | - | NON |
| view, u, duration, minute, add | OUI | NON | - | NON |
| jan, oct, dec, nov, feb | OUI | OUI | Mois de l'année | NON |
| v0, td, selectize, class, nowrap | OUI | OUI | Librairie JS | OUI |
| laptop, lenovo, driver, thinkpad, nvidia | OUI | OUI | Produits Lenovo | OUI |
| vue, composant, composants, propriétés, ... | OUI | OUI | Vue.js framework | OUI |
| point, reply, child, give, permalink | OUI | OUI | Reddit | OUI |
| support, english, product, security, steam | OUI | NON | - | NON |
| school, child, social, research, study | OUI | OUI | Ecole | OUI |
| facebook, retweets, registrieren, seite, konto | OUI | OUI | Réseaux sociaux | OUI |
| kid, baby, girl, child, parent | OUI | OUI | Enfants | OUI |
| kindle, commentaire, the, produit, amazon | OUI | OUI | Produits Amazon | OUI |
| root, size, review, id, common | OUI | NON | - | NON |
| string, type, return, class, json | OUI | OUI | Types JSON | NON |
| twitter, gefällt, account, antworten, tweets | OUI | OUI | Twitter | OUI |
| plus, bien, autres, voir, jour | OUI | NON | - | NON |
| ago, year, month, day, switch | OUI | OUI | Durée | NON |
| leagueoflegends, champion, http, img, png | OUI | OUI | League of Legends | OUI |
| webpack, j, loader, file, module | OUI | OUI | Webpack | OUI |
| replay, del, normal, lunatic, extra | OUI | OUI | Niveaux de jeu vidéo | OUI |
| the, ch, hast, to, and | NON | NON | - | NON |
| amazon, game, badge, book, home | OUI | NON | - | NON |
| window, microsoft, windows, store, phantomjs | OUI | OUI | Windows | OUI |
| property, method, data, server, application | OUI | OUI | Dev. backend | OUI |
| yes, flag, prototype, array, strict | OUI | NON | - | NON |
| post, publish, revision, code, microspot | OUI | NON | - | NON |
| zelda, wild, breath, link, legend | OUI | OUI | The Legend of Zelda | OUI |
| head, coin, room, stone, ll | OUI | NON | - | NON |
| icon, copy, font, arrow, fill | OUI | OUI | Graphisme | OUI |
| modifier, code, états, unis, article | OUI | OUI | Article Etats-Unis | OUI |
| class, div, html, cs, button | OUI | OUI | Elements HTML | OUI |
| child, food, see, year, book | OUI | NON | - | NON |

| Mots | Mot | Thè | Thème commun | Sns |
|--|-----|-----|---------------------------|-----|
| cuisinières, min, laver, service, cuisson | OUI | OUI | Cuisine | OUI |
| time, season, episode, adventure, part | OUI | OUI | Série télévisée | OUI |
| image, png, data, width, item | OUI | OUI | Métadonnées image | OUI |
| mysql, syntax, table, statement, create | OUI | OUI | SQL | OUI |
| account, sign, email, please, click | OUI | OUI | Compte mail | OUI |
| url, request, flask, app, response | OUI | OUI | Flask Framework | OUI |
| chart, plot, data, column, babilsh | OUI | OUI | Graphique | OUI |
| video, youtube, watch, vimeo, makeup | OUI | OUI | Vidéo | OUI |
| file, npm, package, run, version | OUI | OUI | Packaging application | OUI |
| date, prototype, support, object, document | OUI | NON | - | NON |
| function, var, return, array, option | OUI | OUI | Keywords (prog) | OUI |
| light, book, lamp, night, éclairage | OUI | OUI | Lumière | OUI |
| answer, stack, vote, overflow, question | OUI | OUI | StackOverflow | OUI |
| like, get, just, one, make | OUI | NON | - | NON |
| flex, div, align, item, content | OUI | OUI | CSS | OUI |
| typescript, code, test, type, file | OUI | OUI | TypeScript | OUI |
| thread, method, call, will, module | OUI | OUI | Concurrence (prog) | OUI |
| character, draw, art, design, forward | OUI | OUI | Personnage fiction | OUI |
| amazon, plus, prime, jeux, nav | OUI | OUI | Produits Amazon | OUI |
| option, value, select, input, label | OUI | OUI | Formulaire HTML | OUI |
| table, td, th, silhouette, tr | OUI | OUI | Tableau HTML | OUI |
| hsbc, business, banking, website, http | OUI | OUI | E-Banking | OUI |
| november, history, world, december, retrieve | OUI | NON | - | NON |
| anime, girl, manga, meme, irl | OUI | OUI | Anime | OUI |
| top, middle, support, jungle, stockage | OUI | OUI | Rôles (League of Legends) | OUI |
| net, number, random, insert, openx | OUI | OUI | Nombres | NON |
| advertising, design, ad, forward, marketing | OUI | OUI | Marketing | OUI |
| channel, message, princess, margaret, jump | OUI | NON | - | NON |
| accessoires, cm, autres, iphone, jeux | OUI | NON | - | NON |
| canada, québec, france, ainsi, guerre | OUI | OUI | Pays | NON |
| merci, janvier, classe, enfants, répondre | OUI | NON | - | NON |
| de, la, le, nutzt, kommentare | OUI | NON | - | NON |
| vaiselle, poche, tuner, longueur, gcn | OUI | NON | - | NON |
| page, edit, fire, hero, list | OUI | NON | - | NON |
| stock, temp, come, soon, sell | OUI | OUI | Bourse | OUI |
| jeu, amp, ps4, pc, jeux | OUI | OUI | Jeux vidéo | OUI |
| pm, january, follow, http, www | OUI | NON | - | NON |
| img, src, alt, http, jpg | OUI | OUI | Métadonnées image | OUI |
| net, post, internet, neutrality, trump | OUI | OUI | Net neutrality | OUI |
| comment, reddit, post, save, report | OUI | OUI | Reddit | OUI |

| Mots | Mot | Thè | Thème commun | Sns |
|--|-----|-----|-------------------|-----|
| game, card, switch, update, now | OUI | OUI | Jeu vidéo | OUI |
| game, league, team, legend, comment | OUI | OUI | League of Legends | OUI |
| nintendo, switch, mario, demande, iwata | OUI | OUI | Nintendo | OUI |
| pack, ce, café, news, bbc | OUI | NON | - | NON |
| reimu, interest, manipulation, shrine, touhou | OUI | OUI | Touhou Project | OUI |
| moon, blood, tattoo, artwork, madj0hn | OUI | OUI | Tatouages | OUI |
| select, mysql, set, value, sql | OUI | OUI | SQL | OUI |
| amp, google, web, android, app | OUI | OUI | Android | OUI |
| amp, log, src, hide, style | OUI | NON | - | NON |
| eps3, robot, mr, season, wiki | OUI | OUI | Mr.Robot | OUI |
| suisse, protecteur, vapeur, services, sécurité | OUI | OUI | Sécurité | OUI |
| rate, earn, win, th, panda | OUI | NON | - | NON |
| share, facebook, link, pinterest, twitter | OUI | OUI | Réseaux sociaux | OUI |
| function, class, name, object, decorator | OUI | OUI | Fonction (prog) | NON |
| unit, attack, build, hero, hp | OUI | OUI | Theorycraft | OUI |
| the, saison, série, big, film | OUI | OUI | Télévision | OUI |

Réflexion Le fonctionnement de LDA est décrit à la section 2.4.4. Le tableau 5.3 montre un résumé des résultats que l'on peut récupérer précédent tableau. Sur les 100 topics générés par le modèle, on a pu estimer qu'au final 65 de ceux-ci sont des topics potentiellement désirables, et décrivant un thème commun et sensé.

TABLE 5.3 – Résumé des résultats de LDA

| | |
|-----------------------------|-----|
| Topics générés | 100 |
| avec mots sensés | 99 |
| avec un topic commun | 71 |
| avec un topic sensé | 65 |

Bien que ceci laisse un tiers des topics comme étant potentiellement inutiles ou indésirables, ce résultat montre tout de même qu'une partie conséquente des utilisateurs peut être retrouvée. Notons qu'un nombre final de topic sensé de 100% n'est pas le but recherché, et serait même totalement utopique : Il est impossible de définir précisément qu'est-ce qu'un topic, ou même définir quelle page web parle de quel topic, ou même encore combien de topics différents une page web touche.

De plus, il semble fort probable que certains topics soient définis autour de concepts qui se présentent dans des proportions semblables à des centres d'intérêts, mais qui forment des topics qui ont peu de sens. Comme le concept de temps par exemple, ou de géographie, qui constituent tous deux des topics dans notre modèle.

| URL | Probabilité |
|---|-------------|
| http://www.transformice.com/tutorials/index.php | 0.735 |
| https://play.google.com/books/reader | 0.643 |
| https://www.facebook.com/mieletfleur/ | 0.632 |
| https://www.transformice.com/ | 0.620 |
| https://www.facebook.com/herominutebuzz/ | 0.618 |
| http://www.20min.ch/schweiz/ | 0.563 |
| https://www.buecher.de/shop/bilderbuecher/mats-und... | 0.525 |

TABLE 5.4 – URLs possédant le topic "De, la, le, nutzt, kommentare" avec >0.5 de probabilité

Exemple Certains topics non retenus générés sont tout de même surprenants, et nous pouvons nous poser la question comment ceux-ci ont été généré. Par exemple, regardons le topic "de, la, le, nutzt, kommentare" qui semble plutôt étrange car il contient des déterminants. Intéressons-nous aux URLs qui présentent ce topic avec une probabilité supérieure à 0.5. Le tableau 5.4 montre la liste des URLs concernées.

En regardant les données enregistrées dans la base ainsi que le contenu des pages de ces URLs, des caractéristiques semblables émergent. Tout d'abord, il s'agit de pages affichant du texte en plusieurs langues, ou empruntant des mots à une langue différente. Etant donné que notre algorithme part du principe que le texte de chaque page est exprimé en une langue donnée, la présence de mots d'autre langue que celle détectée influence la détection des mots importants dans les documents concernés. Ainsi, ici nous pouvons voir que la langue principale de ces pages est autre que français (généralement allemand, ici), et c'est la raison pour laquelle des mots comme "de" ou "la" prennent une importance trop élevée. D'habitude, ces mots sont filtrés à l'aide d'une liste de stopwords. Mais ici, ceux-ci sont pris en compte et influencent les résultats dans une direction non souhaitée.

Il s'agit là de la mise en évidence d'une des imperfections du système, mais d'autres cas se cumulent à celui-ci. On peut citer par exemple les cas dans lesquels :

- L'URL accédée nécessite une connexion de la part de l'utilisateur. Dans ce cas, notre serveur ne verra probablement qu'un contenu très restreint semblable à "Merci de vous connecter", biaissant l'importance de ces mots.
- L'URL accédée a un contenu dynamique qui varie fortement au fur et à mesure du temps. Dans ce cas, notre serveur n'accordera d'importance qu'à la dernière version en date, et ne traitera les mots que de la dernière version de la page.
- L'URL accédée nécessite des paramètres d'accès, comme ?search=Test. Dans ce cas, l'extension ignore volontairement ces paramètres et il se peut

que la page renvoie une erreur à notre serveur si ces paramètres ne sont pas présents. Dans ce cas, tout accès à ces URLs résultera en un biais vers les mots d'erreur de la page.

- L'URL accédée présente un contenu qui n'est principalement pas du texte. Par exemple, la page pourrait afficher une vidéo, et ne montrer que quelques boutons d'actions comme "Play", "Pause" ou "S'abonner". Dans ce cas, notre algorithme se concentre sur ces quelques mots, qui auront une importance biaisée dans la génération de topics.

Notons qu'ils ne s'agit que d'exemples et que la cette liste n'est pas exhaustive. Nous avons conscience que des solutions existent pour pallier à certains de ces différents cas, mais les solutions à déployer demandaient trop d'efforts pour être raisonnablement implémentées dans le cadre de ce projet.

Des solutions ont déjà été implémentées pour parer à des problèmes survenus initialement. Par exemple, notre serveur, en plus de simplement télécharger le contenu de la page web, simule son comportement au chargement initial avec l'exécution de JavaScript. Cette technique nous a déjà permis de nous rapprocher du résultat visuel final auxquels accèdent les utilisateurs.

Cependant le modèle lui-même serait tout à fait améliorable. En effet, le jeu de données que nous disposons est fortement influencé en quantité par les intérêts des quelques utilisateurs, et LDA possède encore plusieurs paramètres modifiables, qui pourraient amener à des résultats encore meilleurs.

5.2.2 Vues et résultats

Afin de se rendre compte d'un profil type généré par l'extension ainsi que ses représentations visuelles, nous nous intéressons ici au profil généré par le projet de moi-même. Nous allons passer en revue les différentes visualisations et les commenter. Le but est ici d'avoir un avis humain reflétant la précision et l'adéquation des données générées avec avec un profil ; Il aurait été possible de mener ce questionnement sur n'importe-quel participant à l'étude, mais je prends exemple sur moi-même par commodité.

Wordcloud

La figure 5.2 montre la vue de l'interface Wordcloud.

Tout d'abord visuellement, cette vue remplit son objectif qui est de donner très rapidement une information sur les mots les plus représentatifs de l'utilisateur. En effet, les mots qui sont jugés les plus représentatifs sont affichés en plus grand, et ce sont également ceux vers lesquels l'œil se pose en premier.

Ensuite, les données représentées semblent correspondre avec la réalité de mon profil. Les mots que j'ai le plus consultés pendant la période durant laquelle j'ai

Wordcloud



FIGURE 5.2 – Vue de l’interface Wordcloud

conservé l'extension installée sont effectivement ceux qui sont le plus mis en valeur ici.

En conclusion, cette vue assez simple me semble convenir dans à peu près tous les aspects car elle remplit tout à fait son rôle. La seule amélioration possible reste le choix d'une librairie graphique différente, pouvant éventuellement éviter le chevauchement de certains mots, cas qui apparaît ici plusieurs fois de manière non dérangeante, mais tout de même remarquable.

Topics List

La figure 5.3 montre la vue de l'interface Topics List.

On remarque ici que j'ai associé 5 des 8 meilleurs topics à certains de mes centres d'intérêt. Sur les 20 topics affichés sur la page, j'ai un total de 11 assignations. Ceci représente donc un taux de topics "corrects" légèrement supérieur à la moitié. Pour les 3 topics que j'ai laissé sans liens font ici :

- Représente un thème que je n'estime pas être important (account sign email)
 - Fait peu de sens en soi (amp log src)
 - Ne représente pas un thème auquel je m'identifie (anime girl manga)

En ce qui concerne le score d'intérêt estimé, seul le premier topic me parle vraiment, en le sens que je suis tout à fait d'accord et conscient que j'ai plus d'intérêt pour un topic défini par "comment reddit post" que les autres, probablement dû

Topics List

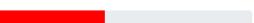
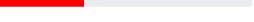
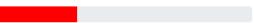
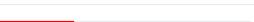
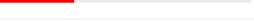
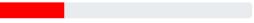
| # | Words | Related interest | Estimated interest |
|---|------------------------|-------------------------|--|
| 1 | comment reddit post | Social Media Enthusiast | 100%  |
| 2 | share facebook link | Social Media Enthusiast | 52%  |
| 3 | example vector product | Technophile | 45%  |
| 4 | account sign email | | 43%  |
| 5 | amp log src | | 42%  |
| 6 | anime girl manga | | 39%  |
| 7 | rate earn win | Gamer | 37%  |
| 8 | chrome api web | Technophile | 37%  |

FIGURE 5.3 – Vue de l’interface Topics List

au fait de ma grande fréquentation du site reddit par rapport aux autres.

J’arrive donc à expliquer l’énorme différence d’intérêt entre ce premier topic et les autres, et je le trouve correct. En revanche, je trouve que les différences minimales de pourcentage entre les autres topics fait peu de sens ; Je n’arriverais pas moi-même à les classer par ordre d’importance. Leur donner une valeur aussi précise me semble donc inutile, selon moi.

Je trouve donc que les données représentées ici sont partiellement adéquates. Certaines d’entre elles sont pertinentes comme les mots des topics trouvés et une estimation de l’intérêt, même si les topics en eux-mêmes ne sont pas toujours corrects. En ce qui concerne la forme dont ces données sont présentées, je pense qu’une échelle avec une granularité diminuée serait moins confuse, par exemple un score de 1 à 5 pour chaque topic.

Au final, je pense qu’il serait possible d’améliorer cette vue par deux ces moyens :

- Meilleure détection des topics d’intérêt
- Visualisation simplifiée du degré d’intérêt

Most watched and visited

La figure 5.4 montre la vue de l’interface Most Viewed, et la figure 5.5 montre celle de l’interface Most Watched.

À l’inverse des autres, ces deux vues-ci présentent des données qui n’ont pas été interprétées, ou en tout cas très peu. Le nombre de visites des pages web ne m’étonnent pas ni leur le temps passé sur celles-ci, je peux donc dire que je pense que ces données sur moi sont correctes.

Most Visited Pages

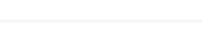
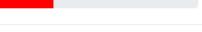
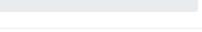
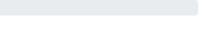
| # | Page | Keywords | Views |
|---|---|------------------------------|---|
| 1 | https://www.google.ch/search | angeboten anmelden bilder | 677  |
| 2 | https://www.google.ch/url | error find know | 614  |
| 3 | http://wdf.sdipi.ch/ | digital footprints web | 578  |
| 4 | https://s3-fr.gladiatus.gameforge.com/game/index.php | adversaires battleknight cgu | 371  |
| 5 | http://df.sdipi.ch/phpmyadmin/sql.php | enable javascript language | 358  |
| 6 | https://www.youtube.com/watch | ago coldplay day | 232  |
| 7 | http://df.sdipi.ch/phpmyadmin/db_structure.php | enable javascript language | 160  |
| 8 | https://www.reddit.com/ | abuse agent appear | 117  |

FIGURE 5.4 – Vue de l'interface Most Visited

Most Watched Pages

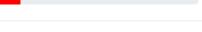
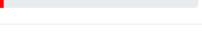
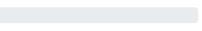
| # | Page | Keywords | Time |
|---|---|------------------------------|--|
| 1 | http://wdf.sdipi.ch/ | digital footprints web | 5h 56min  |
| 2 | https://www.draw.io/ | archimate aws bpmn | 3h 48min  |
| 3 | https://www.reddit.com/r/videos/ | ago comment ghibli | 3h 31min  |
| 4 | http://df.sdipi.ch/phpmyadmin/sql.php | enable javascript language | 2h 29min  |
| 5 | https://www.google.ch/search | angeboten anmelden bilder | 2h 10min  |
| 6 | https://s3-fr.gladiatus.gameforge.com/game/index.php | adversaires battleknight cgu | 1h 57min  |
| 7 | https://www.reddit.com/r/leagueoflegends/ | ago angeles ashleykang | 1h 53min  |

FIGURE 5.5 – Vue de l'interface Most Watched

En ce qui concerne les "Keywords" trouvés, je trouve que plus de la moitié d'entre eux sont confus. Je trouve ceci dommage car je pense qu'il existe des meilleurs mots pour décrire les sites qui arrivent en tête des deux classements. Cependant, comme nous avons déjà expliqué dans un chapitre précédent les problèmes qui peuvent survenir pour détecter les mots idéaux sur certaines pages web, je remarque ici que les pages web que j'ai le plus visité sont des cas typiques de pages où il est difficile de déduire des mots clés idéaux.

En effet, il s'agit principalement de pages web qui demandent une connexion (comme wdf.sdipi.ch et draw.io), de pages web qui ne présentent pas ou peu de contenu texte (comme google.ch/search ou youtube.com), ou encore de pages web

qui nécessiteraient l'URL non tronquée pour être correctement affichées (comme google.ch/url).

Malheureusement, la plupart de ces cas ne semble pas montrer une résolution simple sans une acquisition de données supplémentaires. Il faudrait en effet que l'extension soit capable de voir le contenu effectivement affiché au client, et pas le contenu affiché de manière publique. Cette solution demanderait donc plus de données de la part de l'utilisateur, ce que nous avons volontairement évité.

Au final, cette vue montre des informations quantitatives correctes comme le temps ou le nombre de visites, mais l'évaluation des mots des pages est peu précis. Le potentiel d'amélioration de cette vue se trouve probablement dans l'acquisition de données utilisateur supplémentaires.

History

La figure 5.6 montre la vue de l'interface Wordcloud, et la figure 5.7 montre celle de l'interface Most Watched.

History of topics

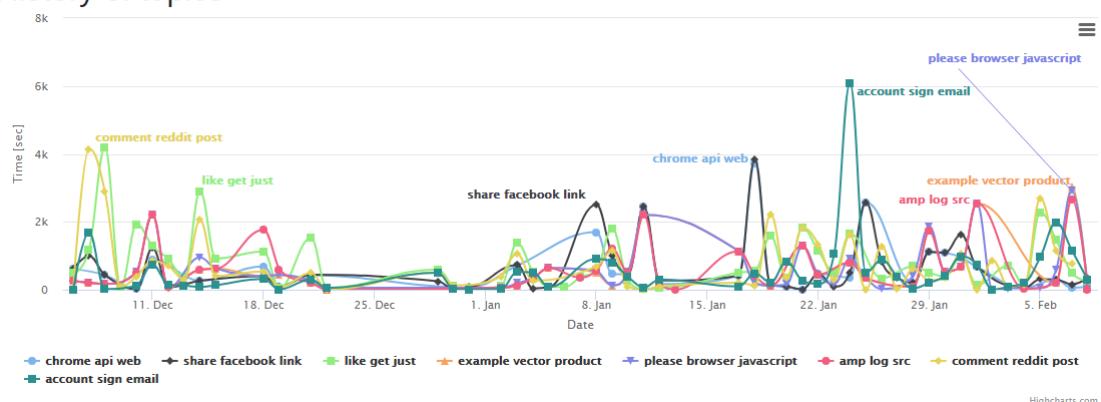


FIGURE 5.6 – Vue du premier graphique de l'interface History

Tout d'abord visuellement, je pense que les vues montrent des tendances intéressantes, mais la quantité d'informations présentes entrave la lecture.

En effet, on peut voir rapidement qu'il y a eu une période de pause entre le 23 décembre et le 3 janvier dans les deux graphiques. Cependant, il reste difficile d'isoler une tendance particulière pour un topic ou un mot particulier à première vue. La version web permet toutefois de cacher ou d'afficher des séries de données en cliquant dessus, ce qui permet d'alléger les graphiques.

Je pense que l'information représentée est intéressante, mais qu'un travail supplémentaire sur la forme de la représentation pourrait valoir la peine. Par exemple, la possibilité de diminuer la granularité des points dans le temps, et de par exemple

History of words

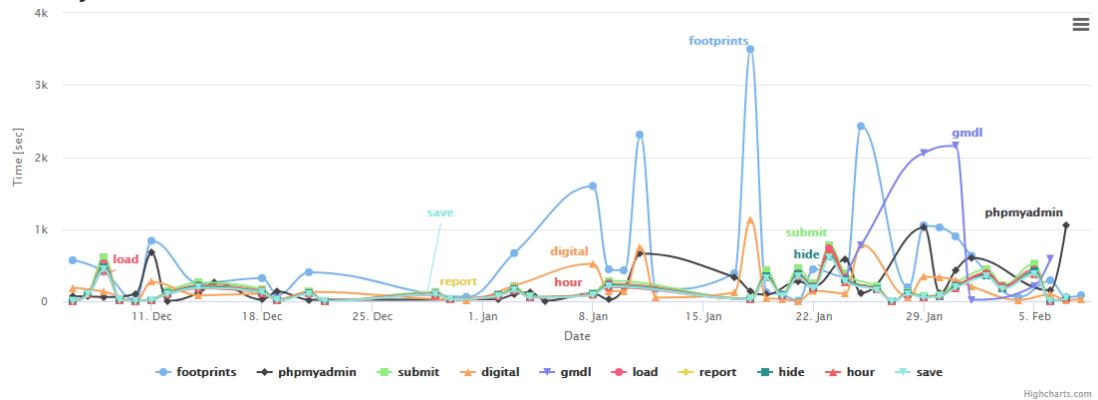


FIGURE 5.7 – Vue du deuxième graphique de l’interface History

regrouper le temps passé à regarder certains mots ou topics peut être intéressant afin de voir plus clairement une tendance dans les deux visualisations.

Trackers

La figure 5.8 montre la vue de la page "Most recieving" de l’interface Trackers.

Most contacted domains

| # | Domain | Visible | Received | Domains | Requests | |
|---|--------------------------------------|-------------|----------|---------|----------|---|
| 1 | video.fgva1-1.fna.fbcdn.net | eye icon | 33 kB | 1 | 3626 | <div style="width: 100%; background-color: red;"></div> |
| 2 | static.xx.fbcdn.net | no eye icon | - | 0 | 0 | <div style="width: 100%; background-color: lightgray;"></div> |
| 3 | scontent.fgva1-1.fna.fbcdn.net | eye icon | 9 kB | 1 | 1044 | <div style="width: 100%; background-color: red;"></div> |
| 4 | e.reddit.com | eye icon | 65 kB | 1 | 668 | <div style="width: 100%; background-color: red;"></div> |
| 5 | 4-edge-chat.facebook.com | eye icon | 23 kB | 1 | 543 | <div style="width: 100%; background-color: red;"></div> |
| 6 | i.ytimg.com | eye icon | 7 kB | 2 | 405 | <div style="width: 100%; background-color: red;"></div> |
| 7 | r2---sn-oxu2a0n-b85l.googlevideo.com | eye icon | 6 kB | 1 | 358 | <div style="width: 100%; background-color: red;"></div> |
| 8 | www.redditstatic.com | no eye icon | - | 0 | 0 | <div style="width: 100%; background-color: lightgray;"></div> |

FIGURE 5.8 – Vue de l’interface Wordcloud

La vue Trackers est celle qui présente le plus d’interactivité, et je me suis en effet pris au jeu d’essayer d’explorer les données en cachant certains domaines

envoyant des données, puis certains autres domaines recevant les données, tout en vérifiant de temps en temps les détails de certains domaines.

Il est difficile de juger des données présentes ici, car nous n'avons pas d'élément de comparaison. Le but de la vue Trackers est plutôt de faire découvrir à l'utilisateur des informations qu'il ignorait lui-même sur sa navigation. Il est donc nécessaire ici de faire confiance à l'extension sur le traitement correct des informations.

Visuellement et interactivement, je trouve que cette vue est particulièrement réussie. J'ai pu naviguer à travers elle et découvrir des nouvelles informations sur ma navigation de manière assez aisée et tout semble clair.

Les possibilités d'amélioration de cette vue sont principalement l'ajout d'une ou deux fonctionnalités supplémentaires d'ergonomie : Par exemple la possibilité de trier les domaines par colonne faciliterait un peu la synthèse des données présentées.

Conclusion

D'une manière générale, j'ai trouvé que les données représentées correspondent assez bien à mon profil, ou du moins à ce que j'imagine que ma navigation dévoile de mon profil. Des modifications et des améliorations dans la manière dont sont traitées les données peuvent sans aucun doute augmenter la fidélité apparente avec laquelle le profil est généré, car certaines informations déduites sont tout de même très imprécises, comme les keywords de mes pages les plus vues.

Cependant malgré les imperfections, je peux dire avec confiance que je me reconnais dans la plupart des informations qui me sont présentées via l'interface. L'objectif de la création d'un profil correspondant à l'utilisateur est donc atteint, avec un potentiel d'amélioration certain.

5.3 Statistiques

Un total de 10 utilisateurs volontaires ont installé l'extension, dont 8 d'entre eux ont été identifiés comme ayant une activité de navigation sur Chrome assez grande pour contribuer à l'étude. Bien que la taille de l'échantillon soit peu impressionnante, le fait que ceux-ci aient utilisé l'extension pendant environ 1 mois nous permet tout de même de montrer des statistiques.

En effet, la taille de la base de données finale est d'environ 4.5 Gio. La moitié de ces données sont le stockage du contenu des pages, l'autre moitié est remplie par des données sur les utilisateurs. Voici ce que nous pouvons dire sur ces données après les avoir analysées.

Tout d'abord, quelques statistiques générales. La figure 5.9 montre le temps passé total à visualiser des pages par domaine. On remarque déjà à cette étape certaines tendances de notre audience.

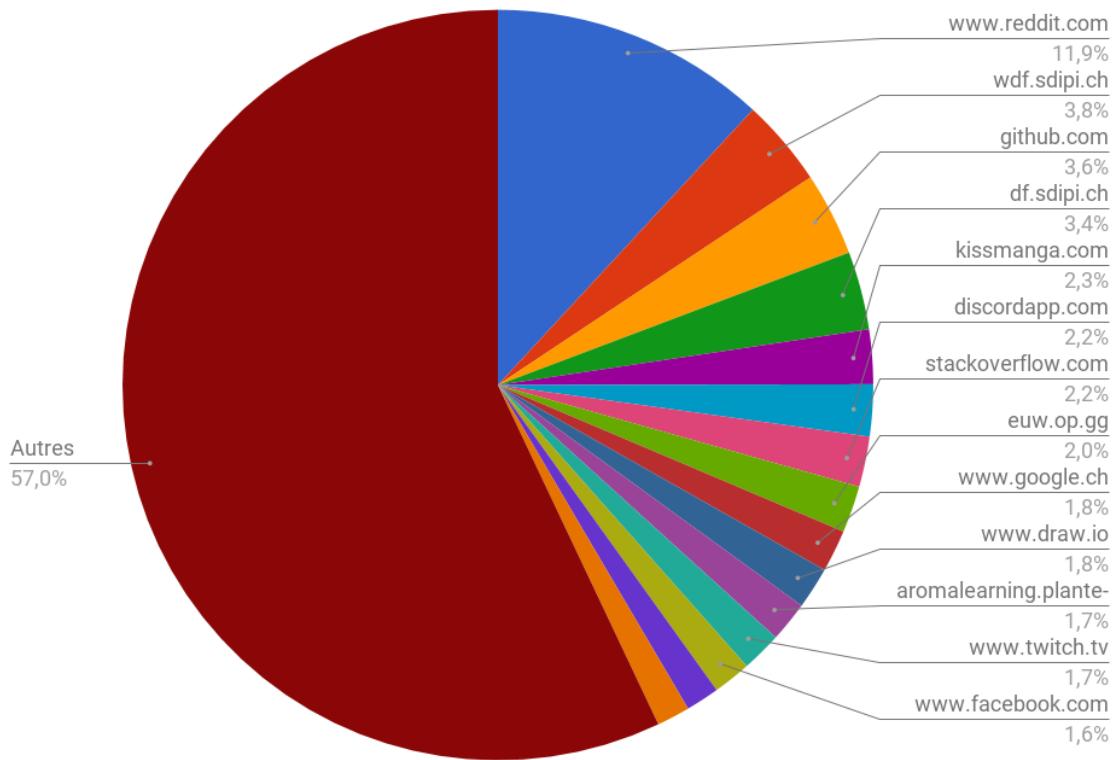


FIGURE 5.9 – Temps total de visionnage par domaine

5.3.1 Profiling

Pour chacun des utilisateurs qui a renseigné ses centres d'intérêt dans son profil, nous nous sommes intéressé à la correspondance entre les profils que les utilisateurs ont renseigné, et les topics que nos algorithmes ont estimé être intéressants pour eux.

Une critique de l'ensemble des vues relatives à un profil se trouve à la section précédente 5.2.2. Nous allons nous intéresser ici aux statistiques générales et aux performances relevées des algorithmes que nous avons utilisé relatifs au profiling.

| Personne | Tags | Int. | Temps | Domaines | Pages | Taux 1 | Taux 2 |
|----------|------|------|-------|----------|-------|--------|--------|
| A | 1 | 8 | 11 | 161 | 551 | 5% | 9,9% |
| B | 11 | 8 | 120 | 998 | 5778 | 55% | 64,2% |
| C | 8 | 4 | 21 | 194 | 732 | 40% | 47,4% |
| D | 7 | 5 | 25 | 193 | 1063 | 35% | 54,0% |
| E | 3 | 6 | 8 | 84 | 462 | 15% | 15,4% |
| F | 4 | 7 | 17 | 145 | 527 | 20% | 16,1% |
| Moyenne | 5,7 | 6,3 | 34 | 296 | 1519 | 28,3% | 34,51% |

TABLE 5.5 – Tableau récapitulatif des statistiques sur les 6 personnes volontaires

Résultats généraux

Parmi les 8 utilisateurs ayant une activité enregistrée suffisante sur le serveur, 6 ont défini des centres d'intérêts et ont participé à la tâche d'assigner des centres d'intérêts aux topics proposés. Le tableau 5.5 montre une série de mesures sur les données amassées des volontaires.

Voici comment se lit chaque colonne :

Personne : Identifiant du volontaire.

Tags : Nombre de topics parmi son top 20 associés à un centre d'intérêt

Int(érêts) : Nombre de centres d'intérêts renseignés sur le profil

Temps : Temps total de visualisation sur des pages, exprimé en heures

Domaines : Nombre de domaines différents visités

Pages : Nombre de pages différentes visitées

Taux 1 : Taux d'association, pourcentage de topics associés à des centres d'intérêts

Taux 2 : Taux pondéré, pourcentage pondéré de topics associés à des centres d'intérêts

On remarque déjà ici que les six personnes présentent des données très différents, autant en activité générale (de 8 à 120 heures passées) ainsi qu'en satisfaction au niveau des topics générés (de 5% à 64% suivant la mesure).

Critique

Tentons à présent de donner du sens aux observations tirées de ces données. Nous voulons vérifier que notre modèle fonctionne, et savoir dans quelle mesure celui-ci est précis. Tout d'abord, nous pouvons nous intéresser à une statistique relativement explicite : Sachant que la page "Topics List" donne le top 20 des topics estimés intéresser l'utilisateur, regardons combien de topics est-ce que les utilisateurs ont associé avec leur centre d'intérêt.

Taux d'association La figure 5.10 montre une nuage de points où chaque point est un utilisateur. En abscisse se trouve le temps total en minutes que l'utilisateur a passé à visiter des pages web, et en ordonnée se trouve le pourcentage de topics parmi les 20 qu'il a associé à des centres d'intérêts. En clair se trouve une courbe de tendance pour les points du graphe.

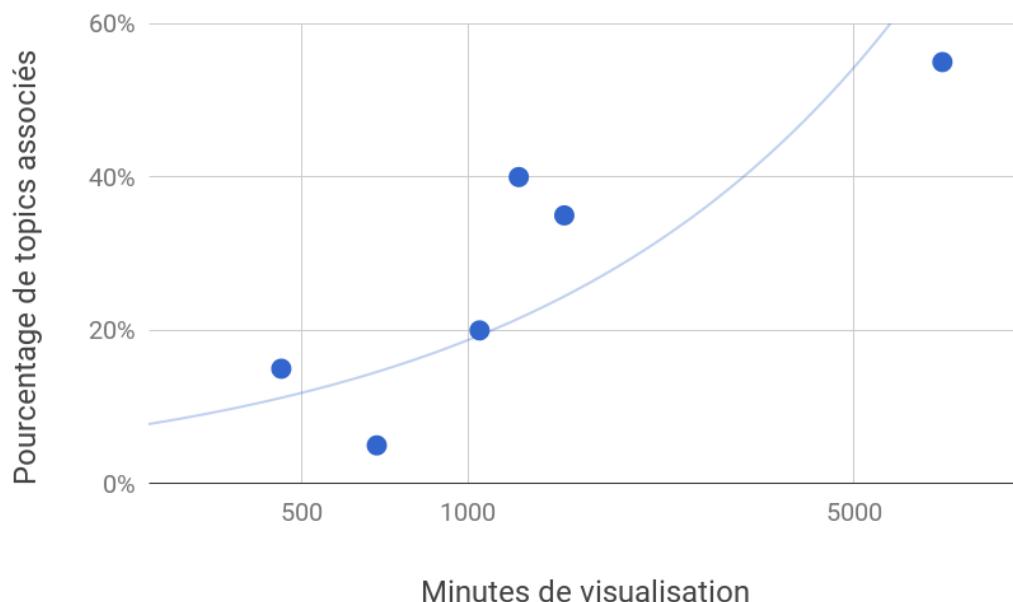


FIGURE 5.10 – Taux d'association de topics par rapport au temps de visualisation avec courbe de tendance

Malgré le fait qu'aucun des utilisateurs soit proche d'associer l'entièreté des topics (le meilleur est à 11/20 topics, soit 55%), on remarque une tendance certaine à afficher un meilleur taux d'association plus l'utilisateur a passé de temps à visiter des sites web. Ceci est plutôt rassurant, car il s'agit en effet du résultat que l'on attend de manière instinctive.

L'algorithme a besoin d'un nombre conséquent de données pour être capable de discerner les préférences et les sujets pouvant intéresser un utilisateur. Plus le nombre de données est grand, plus l'algorithme sera capable de différencier un utilisateur des autres, et donc plus on s'attend à ce que les topics montrés à cet utilisateur soit pertinents à ses propres intérêts.

La variance entre les moins bons et les meilleurs résultats est ici assez grande. L'utilisateur ayant trouvé le moins d'associations est à 1/20 topics, soit 5%. Etant donné que les résultats sont beaucoup plus satisfaisants pour le meilleur utilisateur, il est raisonnable de penser qu'un minimum de données en entrée soit nécessaire

afin d'atteindre un résultat satisfaisant. Dans des expériences futures, on pourrait par exemple situer ce minimum de temps de visualisation quelque part entre le plus faible (8 heures) et le plus élevé (120 heures).

Cependant, comme le montrent les figures 5.11 et 5.12, le temps de visualisation n'est pas forcément la seule méthode d'évaluer la performance des algorithmes. Nous n'allons ici pas calculer un nouveau top 20 pour les utilisateurs d'une manière différente, car cela aurait demandé que les utilisateurs participent une nouvelle fois à l'association des nouveaux topics sortis. Ce n'est malheureusement pas possible car non prévu initialement, et que la phase de récolte des données est à présent terminée.

Nous allons donc nous contenter de projeter ces données sur des abscisses différentes afin de voir si la tendance est similaire. La figure 5.11 montre donc à nouveau le pourcentage de topics associés pour chaque utilisateur en fonction cette fois-ci du nombre de domaines différents qu'ils ont visité au total. La figure 5.12 montre elle en abscisse le nombre de pages web différentes visitées par l'utilisateur.

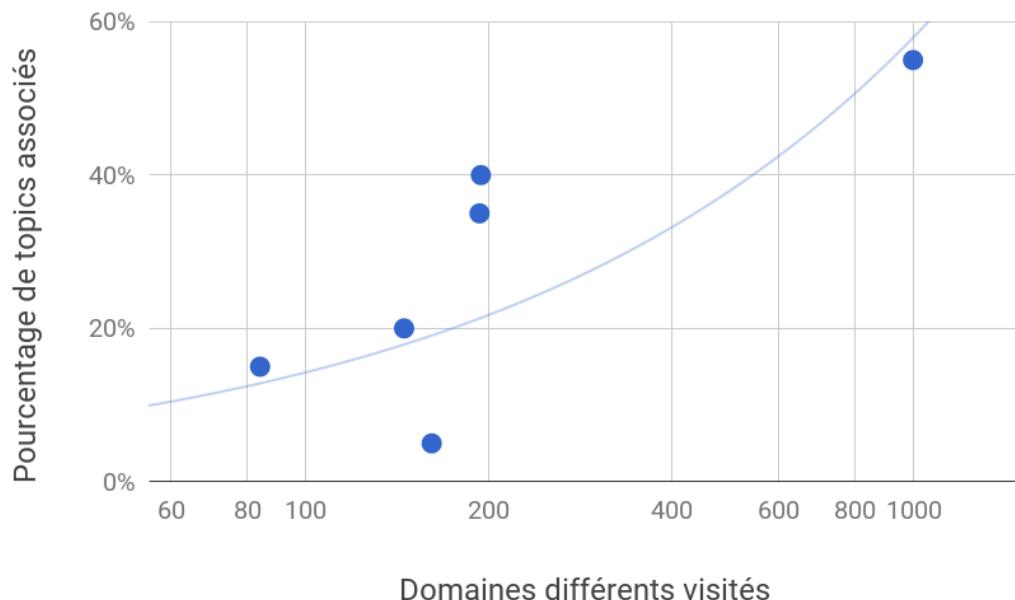


FIGURE 5.11 – Taux d'association de topics par rapport au nombre de domaines différents visités avec courbe de tendance

Sur les deux graphes, la tendance est similaire. Certains des points sont légèrement déplacés car les utilisateurs ont certainement des manières différentes de naviguer sur le web, mais la tendance générale reste que plus notre algorithme dispose de données, plus celui-ci est capable d'afficher des résultats correspondants

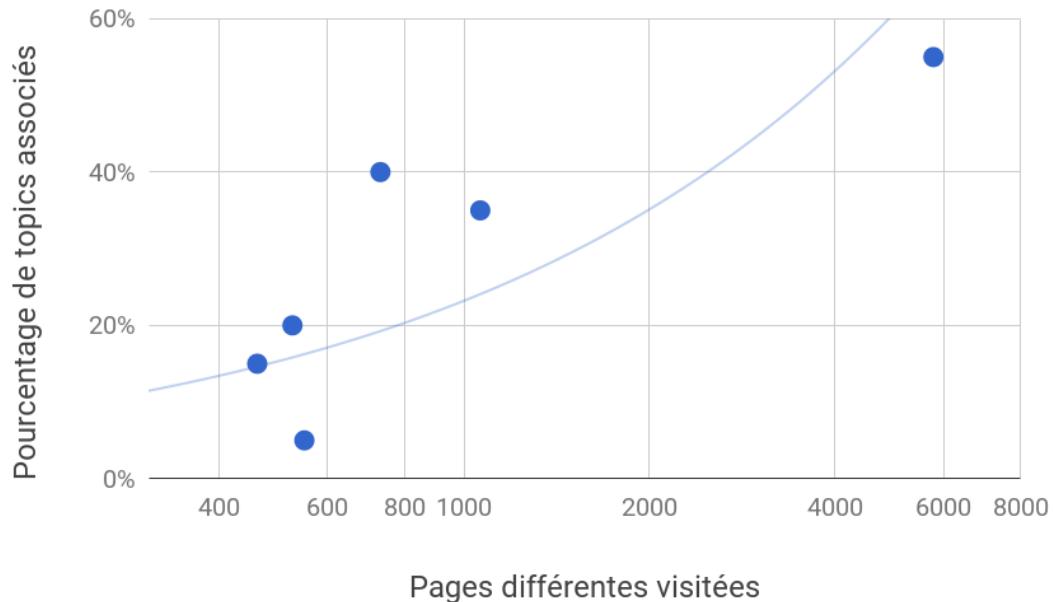


FIGURE 5.12 – Taux d’association de topics par rapport au nombre de pages différentes avec courbe de tendance

aux utilisateurs.

Malgré ceci, nous avons jusqu’ici compté chacun des 20 topics comme également important au résultat final de l’utilisateur. Alors qu’en pratique, nous affichons une barre ainsi qu’un indice d’intérêt estimé, entre 0% et 100% par intérêt. Le top 20 est d’ailleurs constitué d’après cette mesure. Il serait bien à présent d’en prendre compte lors de la mesure de la correspondance entre les topics trouvés et les associations de l’utilisateur.

Taux pondéré C’est pourquoi nous allons nous intéresser à nouvelle mesure intéressante : le taux d’associations, pondéré par l’estimation d’intérêt que l’algorithme associe à chaque topic. Nous allons donc prendre chaque topic en compte non plus équitablement, mais de manière pondérée par l’estimation d’intérêt montrée par l’interface. À titre de rappel, cette estimation d’intérêt par topic par utilisateur est calculée en multipliant la probabilité des topics détectés sur les pages que l’utilisateur a visitées, multiplié par le temps que l’utilisateur a passé sur ces pages.

Ainsi avec cette nouvelle valeur, nous pouvons une nouvelle fois nous intéresser à la tendance de l’évolution de la satisfaction de l’algorithme par rapport à l’utilisateur en fonction de l’augmentation de la quantité de données.

La figure 5.13 montre une mesure du taux pondéré d'association des topics de chaque utilisateur en fonction du temps total qu'il a passé à visualiser des pages web.

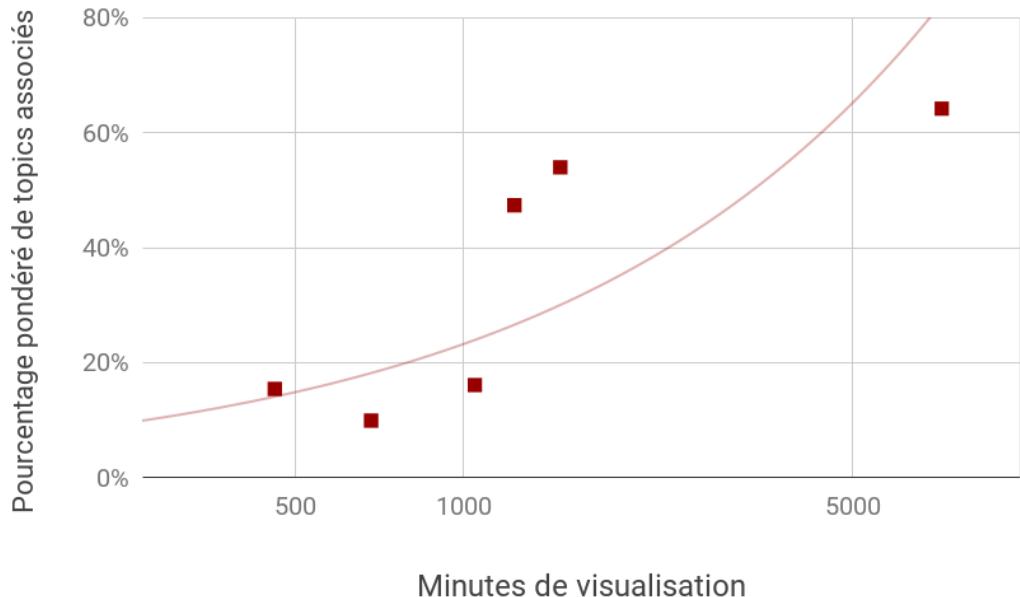


FIGURE 5.13 – Taux d'association de topics par rapport au temps de visualisation avec courbe de tendance

On remarque une nouvelle fois que quelques données ont changé, mais que la tendance reste toujours positive.

Sur les 6 utilisateurs, un seul cas présente une correspondance moins bonne en utilisant le taux pondéré, passant de 20% à 16.4%, les 5 autres se sont améliorés. Ceci résulte en une nette amélioration du taux moyen de correspondance : Celui-ci est passé de 28.3% avec la mesure non pondérée, à 34.51% avec la mesure pondérée. Nous avons donc gagné 6% de taux moyen de "correspondance" entre les centres d'intérêts d'un utilisateur et les topics qu'il a associés. Cette pondération est donc bénéfique à la mesure des topics d'intérêts aux utilisateurs. Egalement, la fourchette des mesures s'est améliorée :

- Les taux non pondérés sont situés entre 5% et 55% de correspondance.
- Les taux pondérés sont situés entre 9.9% et 64.2%.

Conclusion

Pour résumer la manière dont notre algorithme évolue au fur et à mesure de l'augmentation des données disponibles, ainsi que pour justifier l'utilisation d'une pondération, nous pouvons représenter les données de génération de topics des profils par la figure 5.14, qui assemble en une vue les figures précédentes 5.10 et 5.13.

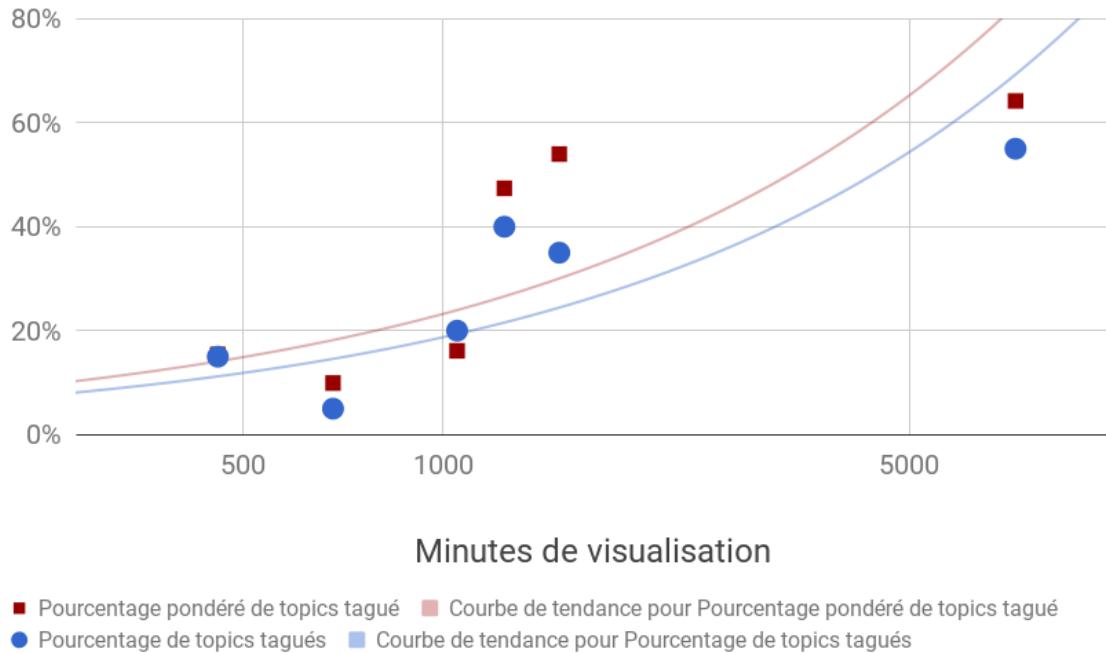


FIGURE 5.14 – Taux d'association de topics, brut et pondéré, par rapport au temps de visualisation avec courbes de tendances

En jetant un coup d’œil à l’échelle des ordonnées, nous pouvons voir ici que nous sommes loin d’avoir des résultats absolument corrects, avec une moyenne finale d’une correspondance d’environ 34.5% entre les topics générés pour l’utilisateur et ses centres d’intérêt, comme le montre la table 5.5.

Cependant, il est bien de se rendre compte en prenant du recul qu’une correspondance de 100% entre les topics générés et les centres d’intérêt de l’utilisateur est utopique. Le fait qu’un utilisateur navigue sur une page peut simplement signifier de la curiosité passagère par exemple, sans que cette page présente effectivement des topics qu’il a signalé comme centre d’intérêts. De plus, tous les modèles entraînés pour atteindre ces résultats sont non supervisés, signifiant qu’il est pratiquement impossible d’estimer leur performance sans y ajouter une composante

humaine, subjective.

Un taux de 100% dans cette mesure ne serait donc pas simplement utopique, mais elle serait également très certainement fausse, car elle aurait d'une manière négligé ou surpassé toutes les décisions humaines imprévisibles. Nous n'avons ici essayé de mesurer les performances de notre méthodologie d'une manière cartésienne, même si celle-ci s'y prête peu du fait la nature des données elles-mêmes.

Au final, nous voyons tout de même que notre algorithme est à la fois capable de générer des topics qui correspondent à certains centres d'intérêts des utilisateurs plus celui-ci possède de données de navigation, et que la méthodologie de pondération utilisée pour générer et estimer l'intérêt de ces topics fait sens car les résultats surpassent ceux lorsque nous ne pondérons pas les résultats.

Nous pouvons donc dire que les résultats que nous observons ici sont satisfaisants car ils sont en ligne avec les objectifs de la partie de génération de profil du projet : Nous sommes capables de générer un profil approximatif d'une personne en se basant uniquement sur ses données de navigation, en les combinant avec des données publiquement accessibles (le contenu des pages) et en y appliquant soigneusement une suite d'algorithmes.

5.3.2 Trackers

Un total de 3'227'000 requêtes ont été enregistrées dans la table `pagerequests`, représentant environ 1 Gio de données dans la base.

Requêtes envoyées

Deux tableaux ont été compilés pour l'ensemble des requêtes analysées par le serveur. Un total de 1058 domaines différents ont été visités par les utilisateurs. Parmi ceux-ci, nous nous intéressons ici à ceux qui ont reçu un minimum de 4 visites de pages. En effet, nous cherchons ici à montrer des statistiques sur des domaines connus des utilisateurs, et donc nous nous restreignons aux domaines qui ont suscité un minimum d'intérêt des utilisateurs tout de même, et ne souhaitons pas voir de domaines de publicité ouvert aléatoirement par une pop-up, par exemple.

| Domaine | Visites | Req. | R/V | Taille [kB] | T/R | T/V |
|---------------------------|---------|--------|------|-------------|------|------|
| www.diablofans.com | 14 | 28776 | 2055 | 12472 | 444 | 891 |
| www.netflix.com | 4 | 8212 | 2053 | 15551 | 1939 | 3888 |
| euw.op.gg | 146 | 147797 | 1012 | 40639 | 282 | 278 |
| www.youtube.com | 378 | 339450 | 898 | 135113 | 408 | 357 |
| www.messenger.com | 34 | 19304 | 568 | 14656 | 777 | 431 |
| inbox.google.com | 27 | 12431 | 460 | 5096 | 420 | 189 |
| www.twitch.tv | 108 | 49048 | 454 | 82062 | 1713 | 760 |
| fr-mg42.mail.yahoo.com | 4 | 1354 | 339 | 562 | 425 | 141 |
| www.lolking.net | 24 | 8016 | 334 | 3695 | 472 | 154 |
| www.facebook.com | 782 | 242727 | 310 | 166256 | 701 | 213 |
| www.draw.io | 35 | 9861 | 282 | 50311 | 5224 | 1437 |
| leagueoflegends.wikia.com | 56 | 15423 | 275 | 2307 | 153 | 41 |
| mlp.wikia.com | 6 | 1393 | 232 | 168 | 124 | 28 |
| www.sabishiidesu.com | 5 | 1146 | 229 | 178 | 159 | 36 |
| www.eclypsia.com | 9 | 2041 | 227 | 1444 | 724 | 160 |
| www.france24.com | 10 | 2161 | 216 | 2102 | 996 | 210 |
| web.whatsapp.com | 44 | 9240 | 210 | 2557 | 283 | 58 |
| clips.twitch.tv | 45 | 9248 | 206 | 6235 | 690 | 139 |
| www.pond5.com | 4 | 784 | 196 | 279 | 364 | 70 |

TABLE 5.6 – Requêtes envoyées par visite sur une page d'un domaine

La table 5.6 montre la liste des 20 domaines ayant envoyé le plus grand nombre de requêtes par visite de page. Voici à quoi chaque colonne correspond :

Domaine : Nom de domaine des pages

Visites : Nombre d'ouvertures d'URL appartenant à ce domaine

Req(uêtes) : Nombre total de requêtes émises par les pages de ce domaine

R/V : Nombre de requêtes émises par visite d'une page

Taille : Taille totale en kB de données utiles envoyées par le navigateur

T/R : Taille moyenne d'une requête

T/V : Taille moyenne de données envoyées par visite d'une page

On peut remarquer que les domaines envoient des quantités de données très différentes, à la fois en nombre et en taille. Les pages qui donnent accès à des applications web tel que le visionnage de vidéo ou le chat semblent aussi envoyer un nombre conséquent d'informations, ce qui n'est pas totalement étonnant étant donné qu'il est nécessaire que l'application soit synchronisée à tout moment avec le serveur pour rester à jour.

Cependant, ce tableau révèle également certains autres domaines qui envoient un nombre suspect d'informations. Rappelons-nous ici qu'il ne s'agit toujours que de données potentiellement personnelles envoyées. Par exemple si un grand nombre d'images doit être chargé pour une page, il est normal que celle-ci émette un nombre élevé de requêtes.

La table 5.7 montre la liste des 20 domaines ayant envoyé le plus grand nombre de requêtes par temps de visualisation de page. Reprenant les valeurs de cette table, la figure 5.15 illustre que pendant que certains domaines lancent un nombre très élevé de requêtes, certains autres compensent en envoyant peu de requêtes, mais chargées d'informations. Voici à quoi chaque colonne de la table correspond :

Domaine : Nom de domaine des pages

Vu [sec] : Temps en secondes passé à regarder la page

Req(uêtes) : Nombre total de requêtes émises par les pages de ce domaine

R/sec : Nombre de requêtes émises par seconde

Taille : Taille totale en kB de données utiles envoyées par le navigateur

T/R : Taille moyenne d'une requête

T/sec : Taille moyenne de données envoyées par seconde

On remarque que certains domaines se retrouvent à la fois dans le haut du classement des deux mesures des tableaux précédents : Nombre de visites, et temps de visualisation, comme par exemple www.netflix.com. Cependant, il est possible que l'estimation du temps passé sur ce domaine soit faussé.

En effet, notre algorithme compte l'utilisateur comme étant inactif après 30 secondes passées durant lesquelles il n'y a eu aucune interaction avec la fenêtre (mouvement de souris, clic, clavier etc). Ceci explique sans doute pourquoi un site de visionnage de vidéo se retrouve en haut de cette liste, avec de plus un temps

| Id | Domaine | Vu[sec] | Req. | R/sec | Taille[kB] | T/R | T/sec |
|-----------|----------------------------|----------------|-------------|--------------|-------------------|------------|--------------|
| A | www.netflix.com | 58 | 8212 | 142 | 15551 | 1939 | 268 |
| B | www.pythonforbeginners.com | 31 | 2623 | 85 | 1235 | 482 | 40 |
| C | newtab | 335 | 25741 | 77 | 1296 | 52 | 4 |
| D | www.lolking.net | 146 | 8016 | 55 | 3695 | 472 | 25 |
| E | mtg.gamepedia.com | 31 | 1181 | 38 | 540 | 468 | 17 |
| F | www.facebook.com | 12413 | 242727 | 20 | 166256 | 701 | 13 |
| G | www.sabishiidesu.com | 75 | 1146 | 15 | 178 | 159 | 2 |
| H | www.cabaneaidees.com | 46 | 576 | 13 | 93 | 165 | 2 |
| I | www.androidcentral.com | 39 | 515 | 13 | 200 | 397 | 5 |
| J | www.rightmove.co.uk | 143 | 1861 | 13 | 786 | 432 | 5 |
| K | zattoo.com | 196 | 2432 | 12 | 305 | 128 | 2 |
| L | www.les-coccinelles.fr | 60 | 658 | 11 | 52 | 81 | 1 |
| M | www.magicmadhouse.co.uk | 227 | 2508 | 11 | 1132 | 462 | 5 |
| N | www.bestbuy.com | 37 | 379 | 10 | 268 | 725 | 7 |
| O | euw.op.gg | 15358 | 147797 | 10 | 40639 | 282 | 3 |
| P | mlp.wikia.com | 135 | 1393 | 10 | 168 | 124 | 1 |
| Q | www.techrepublic.com | 75 | 725 | 10 | 295 | 417 | 4 |
| R | fireemblem.wikia.com | 273 | 2567 | 9 | 466 | 186 | 2 |
| S | www.upc.ch | 70 | 656 | 9 | 244 | 381 | 3 |
| T | www.usatoday.com | 63 | 540 | 9 | 594 | 1127 | 9 |

TABLE 5.7 – Requêtes envoyées par visite sur une page d'un domaine

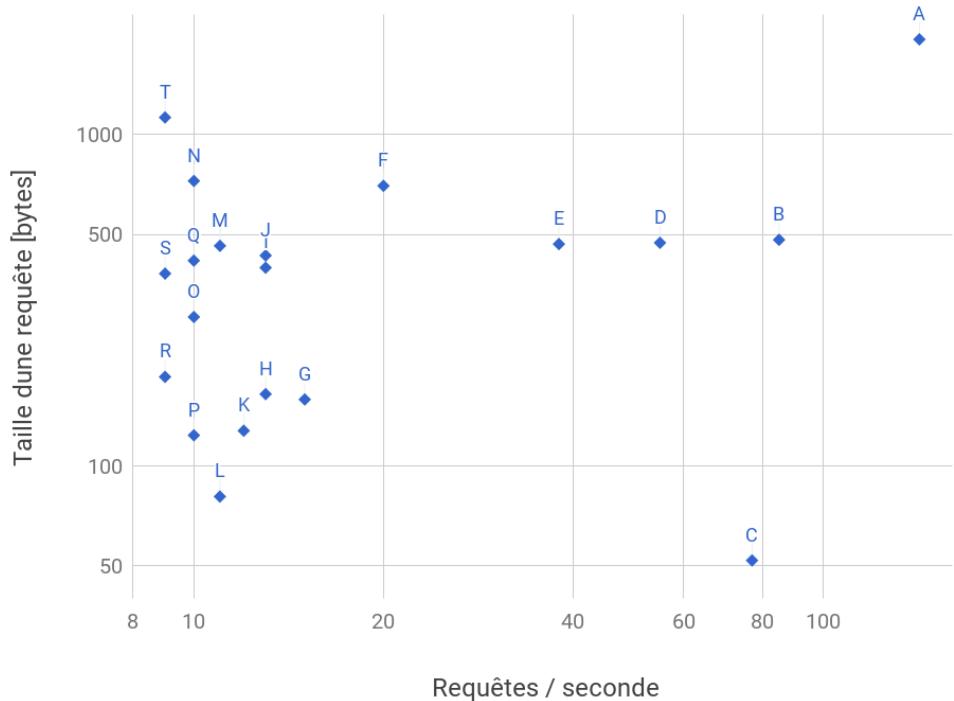


FIGURE 5.15 – Visualisation de la fréquence et de la taille de l'envoie des données des domaines

de visualisation qui semble extrêmement bas pour celui-ci. Notre algorithme a certainement compté le temps passé à regarder la vidéo comme étant inactif. Il aurait fallu faire une distinction, ou déterminer lorsque la page joue une vidéo et considérer ce temps comme étant de l'activité utilisateur.

Intéressant tout de même : Le domaine **newtab** correspond à la page ouverte lorsqu'un nouvel onglet est ajouté. On remarque que cette page visiblement sans contenu envoie tout de même des requêtes.

Requêtes reçues

Nombre de requêtes Le tableau 5.8 montre les domaines ayant reçu le plus grand nombre de requêtes de la part de domaines différents, et donc pouvant être potentiellement considérés comme des trackers. Voici à quoi chaque colonne correspond :

Domaine : Nom de domaine des pages

Requêtes : Nombre de requêtes ayant contacté ce domaine

Taille[MB] : Taille totale des requêtes envoyées vers ce domaine, en MB.

Nous avons donc avec ce tableau la liste des sites ayant reçu beaucoup de données. Il s'agit d'une information à interpréter à nouveau, car nous ne sommes pas en possession des données qui ont été transmises, nous savons uniquement leur quantité. Bien qu'il ne s'agisse pas forcément d'informations personnelles, ces domaines ont tout de même reçu une importante quantité de données avec lesquelles il est potentiellement possible de déterminer des habitudes et autres traits d'un utilisateur.

Domaines différents Le tableau 5.9 montre les domaines ayant été contacté par le plus grand nombre de domaines différents. Les colonnes de ce tableau sont assez explicites. Nous remarquons ici qu'en tête du classement, nous avons un domaine dont nous sommes certain qu'il s'agisse d'un tracker : www.google-analytics.com. La plupart des places suivantes sont également occupées par d'autres sous-domaines de Google.

En dehors de quelques exceptions comme des domaines appartenant à Facebook, il est évident que Google est le possesseur de presque tous les domaines les plus contactés. Nous estimons donc que la mesure "Nombre de domaines différents ayant contacté ce domaine" est un très bon indicateur d'un potentiel tracker. Cette conclusion confirme sans surprise cette intuition initiale.

| Domaine | Requêtes | Taille [MB] |
|-------------------------------|----------|-------------|
| www.youtube.com | 100724 | 98 |
| static.xx.fcdn.net | 97245 | 22 |
| e.reddit.com | 61425 | 59 |
| i.ytimg.com | 45215 | 7 |
| www.redditstatic.com | 40093 | 18 |
| www.google.ch | 33824 | 27 |
| www.facebook.com | 32833 | 114 |
| video.fgval-1.fna.fcdn.net | 31928 | 2 |
| d1u5p3l4wpay3k.cloudfront.net | 28037 | 2 |
| scontent.fgval-1.fna.fcdn.net | 26768 | 5 |
| yt3.ggpht.com | 22661 | 3 |
| www.google-analytics.com | 19519 | 7 |
| b.thumbs.redditmedia.com | 17650 | 6 |
| cdn.discordapp.com | 17021 | 4 |
| s3-fr.gladiatus.gameforge.com | 16854 | 11 |

TABLE 5.8 – 15 domaines ayant reçu le plus de requêtes provenant de domaines différents

| Domaine contacté | Domaines d'origine |
|-----------------------------|--------------------|
| www.google-analytics.com | 713 |
| fonts.gstatic.com | 661 |
| fonts.googleapis.com | 578 |
| www.google.com | 444 |
| stats.g.doubleclick.net | 343 |
| connect.facebook.net | 316 |
| ajax.googleapis.com | 300 |
| www.facebook.com | 297 |
| googleads.g.doubleclick.net | 252 |
| www.gstatic.com | 247 |
| www.googletagmanager.com | 238 |
| apis.google.com | 228 |
| maxcdn.bootstrapcdncdn.com | 214 |
| ssl.gstatic.com | 203 |
| staticxx.facebook.com | 199 |

TABLE 5.9 – 15 domaines ayant été contacté par le plus grand nombre de domaines différents

Chapitre 6

SDIPI

SDIPI signifie "Swiss Digital Identity and Privacy Institute", pouvant se traduire par "Institut Suisse de l'Identité Digitale et de la Vie Privée". Il s'agit d'une association créée pendant ce projet, dans le but initial de soutenir cette étude dans sa visibilité et dans sa légitimité. L'association aspire maintenant objectifs plus larges : Le but est de sensibiliser le public Suisse à la manière dont ses informations privées sont enregistrées, traitées, croisées et utilisées.

Ce projet marque donc la fondation de cette association.

À l'occasion, un site web pour l'association a été créé. Celui-ci présente l'association y compris sa mission, ses membres, ses études et ses statuts.

« We want to raise awareness about how private data are handled online, what kinds of footprints people leave, and how they can control how they appear to the web. », présente le paragraphe "Our work" de la page d'accueil.

La figure 6.1 représente le logo de l'association. La figure 6.2 montre un aperçu de la page d'accueil, accessible à l'adresse <https://www.sdiipi.ch>.



FIGURE 6.1 – Logo officiel de l'association Swiss Digital Identity and Privacy Institute



What is the SDIPI?

SDIPI is the shorthand for "Swiss Digital Identity and Privacy Institute".

As the Internet grew over the years, Digital Identity is an even more important part of a lot of people's lives. As of today, decisions and careers can be decided by how you appear to people online. It's no secret that it's become increasingly more important to know how to control this aspect of one's appearance.

This association was created around the start of a Master's thesis project : [Web Digital Footprints and Data Privacy](#). The goal of this project was to raise awareness about how private informations are handled online, and the study to appeal to the general Swiss population. The three people involved with this project wanted to make this possible and visible through an independant and non-profit medium : That's how the idea of creating this association came to life.



Organization news

We are an active organization and you can keep updated with what we're doing by visiting our [News](#) page.



Our work

We want to raise awareness about how private data are handled online, what kinds of footprints people leave, and how they can control how they appear to the web.



Want to become a member ?

If you want to become a member, go to the [Contact](#) page and send us a message !

Website sources available on [GitHub](#)

FIGURE 6.2 – Page d'accueil du site web <https://sdipi.ch>.

Chapitre 7

Conclusion

7.1 Conclusion technique

Passons en revue les objectifs initiaux du projet afin d'évaluer les résultats obtenus. Reprenons donc les objectifs présentés au début du projet, au chapitre 1.2 :

Définir un profil utilisateur selon des critères de préférence, d'intérêt, d'habitude, d'opinion, etc. Nous avons accompli la définition initiale du profil d'un utilisateur en lui demandant de choisir ses intérêts parmi une liste hiérarchisée d'une centaine de centres d'intérêts. Cet objectif a été accompli de manière simple en implémentant le choix d'intérêts parmi une liste adaptée que nous avons repris d'une source externe. Nous nous sommes donc intéressés particulièrement aux intérêts de l'utilisateur, et avons laissé de côté les aspects plus personnels comme ses opinions et orientations.

Construire le profil d'un utilisateur en se basant sur sa navigation Internet ainsi que sur les métadonnées (durée de consultation des pages, heure de consultation, etc.). Des algorithmes de machine learning seront utilisés pour apprendre les profils en se basant sur des collections de profils annotées Pour commencer, nous n'avons pas utilisé de collections de profils annotées. Nous avons créé notre propre collection de profils en implantant un outil qui, en plus de récolter des informations sur l'utilisateur, lui demande certaines informations personnelles, et nous permet d'évaluer et valider notre propre modèle de profils.

La construction du profil de l'utilisateur se base en effet sur sa navigation Internet, ainsi que sur certaines données supplémentaires de sa navigation : Le nombre de consultations des pages web, le temps d'activité de l'utilisateur sur les pages web visitées ainsi que le contenu publique des pages Web visitées.

Nous utilisons plusieurs algorithmes de Machine Learning, dans la sous-catégorie de l'apprentissage automatique non supervisé. Nous faisons appel à un algorithme d'extraction de mots-clé (qui se nomme Term Frequency-Inverse Document frequency) pour reconnaître les mots importants de chaque page web, et également un algorithme de topic modeling (nommé Latent Dirichelet Allocation) dans le but d'apprendre les thèmes des pages visitées par l'utilisateur.

Nous avons donc répondu à l'objectif initial, mais avons utilisé une méthodologie légèrement différente que celle prévue initialement. Cet objectif a pris plus d'importance que l'idée initiale, et avons en effet généré des résultats supplémentaires comme l'interface de visualisation destinée à l'utilisateur.

Identifier des trackers qui ont la possibilité de construire des profiles utilisateurs en intégrant des données de plusieurs sources Nous avons enregistré et identifié des trackers qui ont des possibilités de construction de profils, mais n'avons pas intégré d'autres types de données dans l'analyse. L'importance de cette tâche s'est réduite durant le projet au bénéfice de l'objectif précédent, qui a vu des résultats grandissants et un potentiel intéressant plus élevé.

L'importance de cette partie s'est trouvée réduite durant le projet après une réévaluation des intérêts en jeu. L'objectif initial a donc été partiellement rempli, car celui-ci s'est trouvé aminci après un certain temps.

7.1.1 Réalisations

Au cours de ce projet, nous avons pu réaliser avec succès :

- Un état de l'art des techniques de tracking actuelles, ainsi que certaines méthodologies d'extraction de données permettant la génération de profils d'utilisateurs.
- Une solution complète de récolte, d'agrégation et d'analyse automatique de données de navigation web d'utilisateurs. Cette solution comprend également une interface permettant aux utilisateurs de consulter à tout moment les données que nous avons récoltées sur lui, ainsi que les plusieurs facettes du profil que nous avons généré à partir de ses données.
- Une analyse, alimentée par les données récoltées, révélant une parties des possibilités de génération de profils d'utilisateurs en se basant sur leur navigation web.
- Une association dont le but est de sensibiliser le public à l'importance de la gestion de son identité et de ses traces digitales

L'ensemble de ces réalisations s'est articulé autour de la volonté de mettre en lumière le potentiel de détection de traits personnels de profils d'utilisateurs naviguant sur le Web. Chaque partie énoncée du projet contribue à sa manière à cet objectif.

7.2 Travaux futurs

Les objectifs principaux du projet ont été atteints, mais les possibilités d'améliorations et de continuation du projet sont nombreuses et variées, à la fois sur le plan technique et sur le plan idéologique. Parmi ceux-ci, on peut citer principalement :

- **La distribution de l'extension à un public plus large.** Des optimisations dans l'implémentation ont permis au serveur actuel de supporter une base d'utilisateurs régulier bien plus large que celle qui a été utilisée pour cette étude. Sans changer d'architecture, projet est probablement capable de gérer une vingtaine d'utilisateurs concurrents sur la machine actuelle, soit probablement jusqu'à une centaine d'utilisateurs actifs totaux. Une idée peut donc être de développer un canal de communication pour toucher un public plus large, et ainsi amasser d'avantage de données. Une plus grande quantité de données va permettre d'améliorer l'évaluation des modèles actuels et d'accroître les révélations et les découvertes de trackers potentiels.
- **L'amélioration technique du nettoyage de données.** Actuellement, plusieurs techniques sont déjà mises en œuvre pour éloigner les données indésirables dans notre pipeline, mais ces étapes sont insuffisantes pour éliminer encore une bonne partie d'informations qui biaissent les algorithmes. Ici, de nouvelles techniques pourraient être introduites à différents endroits de la pipeline. Parmi les probablement plus impactantes et plus faciles à mettre en place, on peut citer :
 - L'amélioration de la liste d'URLs ignorées. Actuellement, une liste statique ignore certaines URL connues pour servir des données non voulues. Il serait possible d'améliorer grandement cette liste afin qu'elle contienne la plupart des sites web indésirables.
 - L'amélioration de la détection d'activité de l'utilisateur. Actuellement, l'extension considère une page comme étant active si l'utilisateur a effectué une action dans les 30 dernières secondes avec le clavier ou la souris. Ceci ne prend pas en compte des autres actions comme la lecture d'une vidéo, ou l'écoute de musique par exemple. Une meilleure détection de l'intérêt de l'utilisateur pour une page serait bénéfique.
 - L'amélioration de la détection des mots réellement affichés sur les pages web. La technique actuelle est d'émuler le lancement des URLs visitées dans un navigateur Chrome, et de récupérer le Document Object Model de la page puis d'en extraire le texte. Cette technique ne prend pas en compte si le texte présent dans les noeud DOM est effectivement affiché ou pas. Par exemple dans le cas d'une one-page app, il est fréquent que certains éléments soient simplement cachés de la page grâce à des direc-

tives CSS. Notre technique ne prend actuellement pas ceci en compte du tout, et conserve tous les mots présents dans les noeuds de la page.

- L'amélioration de la détection de langue sur les pages. La technique actuelle est de détecter une langue par page, puis de nettoyer les mots de cette page en ignorant les stopwords de cette langue. Cette technique a un point faible évident : Les pages qui contiennent du texte en plusieurs langues. Nous avons vu des exemples pratiques dans lesquels des pages multi-lingues ont donné des résultats aberrants lors de l'extraction de mots-clé, car les stopwords d'une ou plusieurs langues de la page n'ont pas été filtrés. Une amélioration pourrait être de lancer la détection la langue de la page par blocs de texte.
- Le stockage de plusieurs versions d'une page au cours du temps. Actuellement, seule la dernière version téléchargée d'une page fait foi pour tous les algorithmes. Ajouter la possibilité de conserver plusieurs copies d'une page dans le temps permettra de traiter correctement les données provenant de pages générées dynamiquement, par exemple les pages d'accueil de sites web de news.
- **Automatisation du calcul offline.** Actuellement, les scripts offline sont lancés ponctuellement par l'administrateur. Une automatisation du lancement de ces tâches réduirait le coût de maintenance du serveur grâce à un gain de temps.
- **Meilleure structures de données.** La structure de la base de données actuelle a été conçue pour un nombre d'utilisateurs réduit, ce qui convenait parfaitement pour le projet. Cependant si le nombre d'utilisateurs augmente, certaines structures de données et certains algorithmes devront être remaniés car ils ne scalent pas de manière suffisante. Par exemple, le pré-calcul de certaines informations commence déjà à prendre un temps considérable, de l'ordre d'environ une demi-heure.
- **Utilisations de modèles online.** Les deux modèles TF-IDF et LDA ne peuvent pas être mis à jour de manière online, ce qui signifie qu'il est nécessaire de lancer des exécutions ponctuelles de script périodiquement afin de rafraîchir certaines données de la base. Bien que ceci soit fonctionnel actuellement et que ces méthodes ont été choisies pour leur précision, il serait peut-être sage de se pencher vers des modèles pouvant être améliorés de manière continue afin de ne pas avoir à redémarrer le serveur.
- **Plusieurs méthodes de calcul des profils.** La méthode actuelle pour calculer les topics d'un utilisateur repose sur la multiplication intuitive des topics des pages de l'utilisateur avec le temps actif qu'il a passé dessus. Bien que fonctionnelle, cette méthode n'est certainement pas la seule valable, et il serait très intéressant de tester d'autres hypothèses et méthodes pour

la génération de topics d'un utilisateur. Par exemple, prendre plutôt le nombre de visites sur les pages au lieu du temps d'activité, ou alors de donner une importance non-linéaire au temps passé sur la page, par exemple exponentielle ou logarithmique. Des idées plus imaginatives sont également envisageables, comme par exemple assigner des poids différents aux mots contenus dans les pages web en fonction de la taille de leur police d'écriture.

- **Capture d'autres données utilisateurs** Bien que éthiquement discutable, il est toujours possible d'acquérir simplement plus de données différentes de l'utilisateur et de les prendre en compte dans la recherche. Par exemple, nous nous sommes volontairement limités dans cette étude dans le but de respecter au mieux les informations privées des utilisateurs. Nous avons par exemple exclu les paramètres de recherche dans une URL, alors qu'ils seraient possibles de les prendre en compte pour avoir des résultats plus précis. Il serait aussi envisageable d'envoyer simplement le contenu de chaque page visitée par le navigateur lui-même, au lieu d'accéder à une version publique de chaque document. Cette méthode, bien qu'elle arrangerait de nombreux problèmes énoncés précédemment, est évidemment très extrême car l'extension dévoilerait alors absolument toute information affichée à l'écran de l'utilisateur.

7.3 Conclusion personnelle

Mon choix pour ce projet s'est principalement fait car toutes les technologies du Web m'intéressent beaucoup. Je suis avide d'apprendre des nouvelles compétences qui me permettent d'élargir mon horizon de connaissances sur les dernières technologies, autant pour les utiliser en développant des outils utiles, que pour réfléchir sur leurs possibilités et leurs risques. Le projet m'a grandement motivé du début à la fin car il s'inscrivait tout à fait dans cette ligne de pensée. J'ai pu développer à la fois mes compétences en matière de recherche et de mise en contexte des technologies actuelles, que d'innovation et de développement de nouveaux outils dans le but de sensibiliser les utilisateurs d'aujourd'hui à l'utilisation de certaines technologies. J'ai également grandement appris de l'analyse finale de données, qui n'était pas une tâche facile au vu de la nature très subjective et peu cartésienne des informations que j'ai dû traiter, du moins par rapport à ce que j'ai l'habitude. Je suis d'avis que le grand public n'est généralement pas assez conscient des conséquences et des risques qu'ils prennent en publiant certaines informations en ligne, et j'ai l'impression d'avoir pu participer quelque peu à la sensibilisation globale en apportant ma modeste pierre à l'édifice grâce à ce projet. Loin d'être parfait, je suis tout de même très satisfait du résultat final qui me semble être un produit complet et adapté à l'objectif recherché en démarrant le projet. La

nature open-source du code développé ici a également été très motivant dans tout le processus. Au final, je suis très content d'avoir eu l'occasion de travailler sur ce projet, qui m'a enrichi dans de nombreux aspects.

Bibliographie

- [1] Michal Kosinski, *Dr Michal Kosinski*, <http://www.michalkosinski.com/>, Consulté en ligne en Septembre 2017, 2017.
- [2] Michal Kosinski, Yilun Wang, Himabindu Lakkaraju and Jure Leskovec, *Mining Big Data to Extract Patterns and Predict Real-Life Outcomes*, <http://psycnet.apa.org/fulltext/2016-57141-003.pdf>, Consulté en ligne en Octobre 2017, 2017.
- [3] Internet Society, *Digital Footprints*, <https://www.internetsociety.org/wp-content/uploads/2017/08/Digital20Footprints20-20An20Internet20Society20Reference20Framework.pdf>, Consulté en ligne en Novembre 2017, Janvier 2014.
- [4] Motherboard, *The Data That Turned the World Upside Down*, https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win, Consulté en ligne en Octobre 2017, Janvier 2017.
- [5] Michal Kosinski, *The End of Privacy, Keynote at CeBIT'17*, <https://www.youtube.com/watch?v=DYhAM34Hhzc>, Consulté en ligne en Octobre 2017, Mars 2017.
- [6] Kaggle, *Young People Survey*, <https://www.kaggle.com/miroslavabo/young-people-survey>, Consulté en ligne en Octobre 2017, 2013.
- [7] Michal Kosinski, *myPersonnality Project*, <http://mypersonnality.org>, Consulté en ligne en Octobre 2017, 2013.
- [8] Google, *Google Solutions Analytics*, <https://www.google.com/analytics>, Consulté en ligne en Octobre 2017, 2017.
- [9] BuiltWith, *Analytics Usage in Switzerland*, <https://trends.builtwith.com/analytics/country/Switzerland>, Consulté en ligne en Octobre 2017, 2017.
- [10] Chrome Web Store, *Extensions*, <https://chrome.google.com/webstore>, Consulté en ligne en Novembre 2017, 2017.
- [11] Modules Firefox, *Extensions*, <https://addons.mozilla.org/fr/firefox/extensions/>, Consulté en ligne en Novembre 2017, 2017.

- [12] Chrome Web Store, *timeStats*, <https://chrome.google.com/webstore/detail/timestats/ejifodhjoeenihgfpjjjmpomaphmah>, Consulté en ligne en Novembre 2017, 2017.
- [13] Ghostery, *Ghostery makes the Web Cleaner, Faster and Safer!*, <https://www.ghostery.com/>, Consulté en ligne en Novembre 2017, 2017.
- [14] Chrome Web Store, *Privacy manager*, <https://chrome.google.com/webstore/detail/privacy-manager/giccehglhacakcfemddmfhdkahamfcmd>, Consulté en ligne en Novembre 2017, 2017.
- [15] TheGoodData, *TheGoodData*, <https://thegooddata.org>, Consulté en ligne en Novembre 2017, 2017.
- [16] Noiszy, *Noiszy*, <http://noiszy.com>, Consulté en ligne en Novembre 2017, 2017.
- [17] Modules pour Firefox, *Privacy Badger*, <https://addons.mozilla.org/fr/firefox/addon/privacy-badger17/>, Consulté en ligne en Novembre 2017, 2017.
- [18] Kraken.me, *Home*, <http://www.kraken.me/#/home>, Consulté en ligne en Novembre 2017, 2017.
- [19] Electronic Frontier Foundation, *Defending your rights in the digital world*, <https://www.eff.org>, Consulté en ligne en Novembre 2017, 2017.
- [20] HTTP Archive, *Trends*, <http://httparchive.org/trends.php?s=Top1000&minlabel=Oct+15+2011&maxlabel=Oct+16+2017#numDomains&maxDomainReqs>, Consulté en ligne en Novembre 2017, 2017.
- [21] HTTP Archive, *Browser Statistics*, <https://www.w3schools.com/browsers/default.asp>, Consulté en ligne en Janvier 2018, 2018.
- [22] Google, *Chrome*, <https://www.google.fr/chrome>, Consulté en ligne en Janvier 2018, 2018.
- [23] Google Chrome, *JavaScript APIs*, https://developer.chrome.com/apps/api_index, Consulté en ligne en Janvier 2018, 2018.
- [24] Google Web Store, *Extensions*, <https://chrome.google.com/webstore/category/extensions>, Consulté en ligne en Janvier 2018, 2018.
- [25] Martin Degeling & Thomas Herrmann, *Your Interests According to Google – A Profile-Centered Analysis for Obfuscation of Online Tracking Profiles*, <https://arxiv.org/pdf/1601.06371.pdf>, Consulté en ligne en Novembre 2017, 2016.
- [26] Silvia Puglisi, David Rebollo-Monedero & Jordi Forne, *On Web User Tracking : How Third-Party Http Requests Track Users' Browsing Patterns*

- for Personalised Advertising*, <https://arxiv.org/pdf/1605.07167.pdf>, Consulté en ligne en Novembre 2017, 2016.
- [27] Opentracker Analytics, *How does user-tracking work?*, <https://www.opentracker.net/docs/resources/how-does-user-tracking-work>, Consulté en ligne en Novembre 2017, 2016.
- [28] Eli Weinstock-Herman, *Automated Keyword Extraction – TF-IDF, RAKE, and TextRank*, <http://www.tiernok.com/posts/automated-keyword-extraction-tf-idf-rake-and-textrank.html>, Consulté en ligne en Novembre 2017, 2016.
- [29] Feifan Liu, Deana Pennell, Fei Liu & Yang Liu, *Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts*, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.370.9534&rep=rep1&type=pdf>, Consulté en ligne en Novembre 2017, 2016.
- [30] Jayaprakash Sundararaj on Quora *What are the best keyword extraction algorithms - Quora*, <https://www.quora.com/What-are-the-best-keyword-extraction-algorithms-for-natural-language-processing-and-how-can-they-be-implemented-in-Python>, Consulté en ligne en Novembre 2017, 2016.
- [31] Let's Encrypt *Free SSL/TLS certificates*, <https://letsencrypt.org>, Consulté en ligne en Février 2018, 2018.
- [32] Brian Reavis & contributors, *Selectize.js*, <https://selectize.github.io/selectize.js/>, Consulté en ligne en Décembre 2018, 2016.
- [33] Panupong (Ice) Pasupat, *LSA / pLSA / LDA*, <https://cs.stanford.edu/~ppasupat/a9online/1140.html>, Consulté en ligne en Février 2018, 2016.
- [34] Joseph Turian on Quora, *What's the difference between Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA)? - Quora*, <https://www.quora.com/Whats-the-difference-between-Latent-Semantic-Indexing-LSI-and-Latent-Dirichlet-Allocation-LDA>, Consulté en ligne en Février 2018, 2012.
- [35] /u/punkmonk on reddit, *LSA vs pLSA vs LDA*, https://www.reddit.com/r/MachineLearning/comments/10mdtf/lsa_vs_plsa_vs_lda/, Consulté en ligne en Février 2018, 2013.
- [36] Radim Řehůřek, *gensim : Topic modelling for humans*, <https://radimrehurek.com/gensim/>, Consulté en ligne en Janvier 2018, 2018.

Glossaire

Document Object Model est l'arbre d'éléments qui compose un fichier HTML.
37, 38, 109

endpoint est le nom donné à un point autonome accessible d'une API. 49

Latent Dirichelet Allocation est le nom d'un modèle de topic modeling.
108

Natural Language Processing est une catégorie de techniques algorithmiques visant à comprendre un texte écrit dans une langue humaine. 22

open-source qualifie un logiciel dont le code initial est mis à disposition du grand public. 10

Rapid Automatic Keyword Extraction est le nom d'une technique de keyword extraction. 23, 24

stopword est le nom donné aux mots qui servent généralement de liaison, que nous cherchons à ignorer lors d'analyses. 83, 110

Term Frequency-Inverse Document frequency est le nom d'une technique de keyword extraction. 23, 24, 108

Remerciements

Je tiens à remercier plusieurs personnes. Tout d'abord, Madame Nastaran Fatemi pour m'avoir supervisé et avoir pris le temps de me guider lors des décisions à prendre, ainsi que Félicien Fleury pour m'avoir également soutenu et guidé tout au long de ce projet. Je tiens également à remercier mes parents pour leur soutien continu, ainsi que mes amis qui m'ont aidé dans les moments d'incertitude. Je remercie également tous les participants s'étant portés volontaires pour l'étude de cas dans ce travail.

Déclaration d'honneur

Je, soussigné, Kewin Dousse, déclare sur l'honneur que le travail rendu est le fruit d'un travail personnel. Je certifie ne pas avoir eu recours au plagiat ou à toutes autres formes de fraudes. Toutes les sources d'information utilisées et les citations d'auteur ont été clairement mentionnées.

Lieu

Date

Signature

Annexe A

Historique des versions

Voici l'historique des versions de ce document.

- 0.1 : Template du document
- 0.2 : Chapitre "Analyse"
- 0.3 : Correction, complétion du chapitre "Analyse", rédaction d'une partie du chapitre "Conception"
- 0.4 : Rédaction de la plupart de la partie "Design du système", anciennement "Conception"
- 0.5 : Rédaction de la partie "Développement des vues"
- 0.6 : Complétion du "Développement des Vues", réorganisation de la structure de parties précédentes, et début de la partie "Résultats"
- 0.7 : Fin de la rédaction du texte du rapport
- 1.0 : Correction, complétion du chapitre "Analyse", ajout des annexes

Annexe B

Cahier des charges

B.1 Activités

Le but de ce projet est de proposer un outil de visualisation pour sensibiliser le public à la question du profiling sur internet. Le projet propose de développer un plugin pour les navigateurs Mozilla Firefox et Google Chrome.

Le développement du projet peut se découper en plusieurs phases, qui elles-mêmes se divisent en plusieurs activités. Voici la liste de ces activités :

1. Analyse
 - (a) Recherche de sources de données
 - (b) Définition des objectifs
 - (c) Planification
 - (d) Accès aux sources de données
 - (e) Etude des données
2. Conception
 - (a) Design de la solution
 - (b) Design du système
 - (c) Etude de la faisabilité
3. Réalisation
 - (a) Mise en place de l'architecture
 - (b) Réalisation d'une pipeline fonctionnelle
 - (c) Plug-In navigateur
 - i. Implémentation de l'acquisition de données
 - ii. Implémentation de l'envoi des données

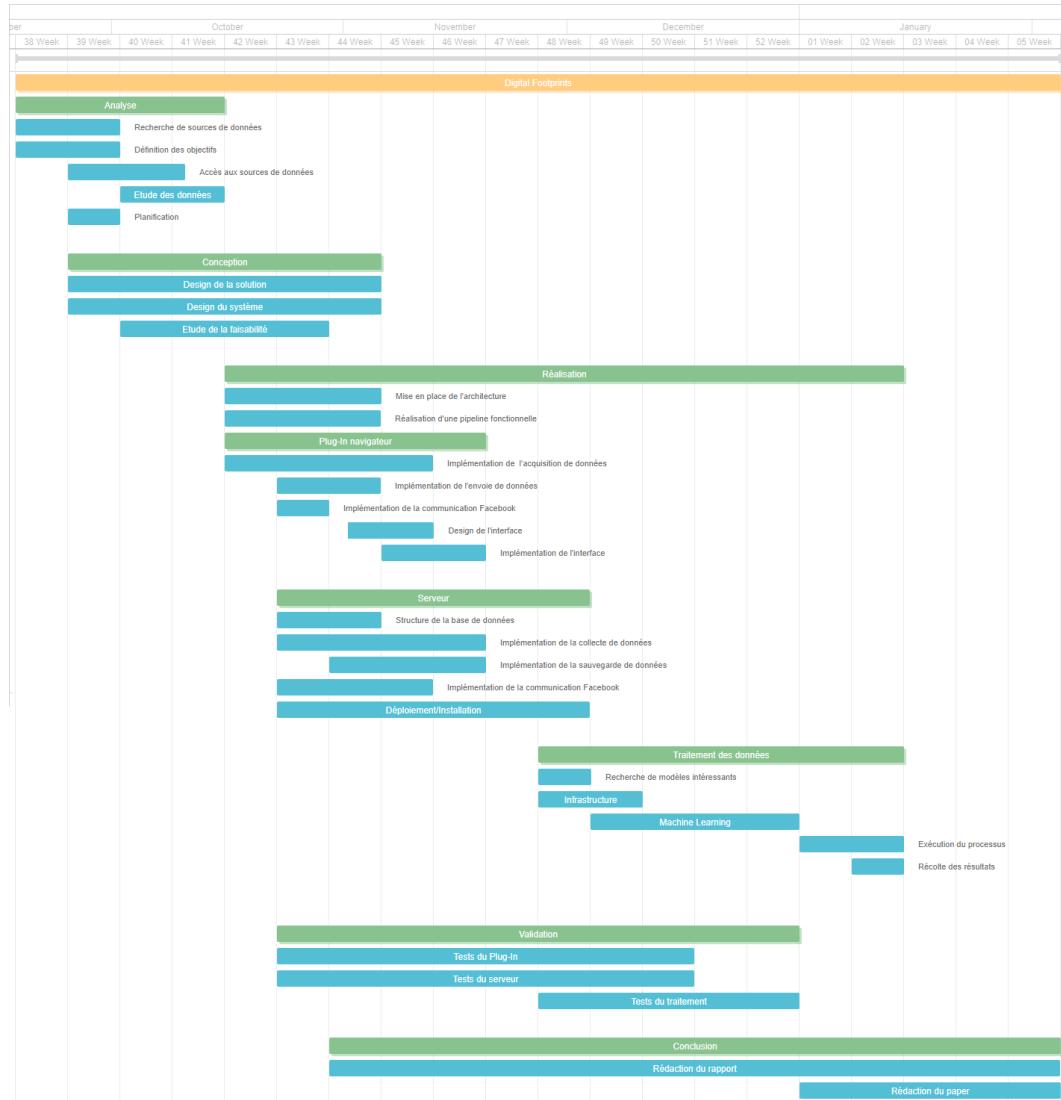
- iii. Implémentation de la communication Facebook
- iv. Design de l'interface
- v. Implémentation de l'interface
- (d) Serveur
 - i. Structure de la base de données
 - ii. Implémentation de la collecte de données
 - iii. Implémentetaion de la sauvegarde de données
 - iv. Implémentation de la communication Facebook
 - v. Déploiement/Installation
- (e) Traitement des données
 - i. Recherche de modèles intéressants
 - ii. Infrastructure
 - iii. Machine Learning
 - iv. Exécution du processus
 - v. Récolte des résultats
- 4. Validation
 - (a) Tests du plug-in
 - (b) Tests du serveur
 - (c) Tests du traitement
- 5. Conclusion
 - (a) Rédaction du rapport
 - (b) Rédaction du paper

B.2 Planification

Le projet comporte quelques dates clés qu'il est important de respecter :

| Date | Semaine | Tâche |
|-------------------------|-------------|------------------|
| Lundi 18 septembre 2017 | Semaine P1 | Début du projet |
| Vendredi 9 février 2018 | Semaine P15 | Dépôt du rapport |
| 26 février-9 mars 2018 | - | Défense orale |

B.3 Diagramme de Gantt



Annexe C

Documentation

C.1 Localisation

L'ensemble du code sources des parties du est disponible à l'adresse suivante :
<https://github.com/SDIPI>

C.2 Contenu

C.2.1 GitHub

L'association SDIPI présente sur GitHub contient les différentes parties du projet :

- <https://github.com/SDIPI/wdf-extension> contient le code source de l'extension de navigateur.
- <https://github.com/SDIPI/wdf-server> contient le code source du serveur recueillant les données.
- <https://github.com/SDIPI/wdf-frontend> contient le code source de l'interface utilisateur.
- <https://github.com/SDIPI/SDIPI.github.io> contient le code source du site de l'association SDIPI.

Annexe D

Procès-verbaux

Voici les documents des procès-verbaux réalisés.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



MASTER OF SCIENCE
IN ENGINEERING

PV de réunion

15 septembre 2017, de 9h05 à 10h55

Présent : Nastaran Fatemi, Félicien Fleury, Kewin Dousse

Rédaction du PV le 20 septembre

Compte-rendu

Points de discussion

- Objectif du projet

Il a été défini tout d'abord que l'objectif du projet était de répondre, dans un sens large, à la question suivante : Est-il possible de faire un profil d'un utilisateur en se basant sur sa navigation web ? Celui-ci aura un but informatif, sensibilisant.

- Sources de données

La question s'est posée sur quelles sont les sources de données à notre disposition pour ce projet. Leur utilisation spécifique n'est pas encore connue, mais nous aurons sans doute besoin de données d'utilisateurs à confronter à notre système. Afin de ne pas être bloqué par l'étape de la récolte de ces données plus tard, il est important d'y réfléchir tôt et d'entreprendre des démarches si nécessaires auprès d'organismes pouvant nous en fournir. Nous prévoyons donc de chercher un corpus de données d'utilisateurs assez tôt.

À court terme, il est donc nécessaire de rechercher quelles les sources de données possibles. Plus précisément, nous savons déjà que des organismes comme l'université de Cambridge peuvent détenir des données intéressantes et allons prendre contact avec eux.

- Plug-in

La question s'est posée : Est-ce que l'outil développé doit être utile après la fin de l'étude, ou est-ce que celui-ci n'est « qu'un » outil pour aider l'étude et atteindre des résultats finaux ? La question reste ouverte. Cette question en soulève également une autre : Quels sont les outils que nous nous autorisons éthiquement à utiliser pour celui-ci ?

- Organisation

Afin de communiquer et nous organiser efficacement, nous allons utiliser plusieurs outils dont Trello pour l'organisation des tâches, Slack pour la communication écrite, et Skype pour des communications audio. Une association sera créée afin de donner de la visibilité et de la légitimité à cette recherche. Son nom et ses statuts seront finalisés bientôt. La recherche visera également une publication, par exemple dans une conférence à définir.

Une réunion hebdomadaire est prévue le jeudi à Yverdon. Nastaran et Kewin y seront présents, et il est prévu que Félicien y participe alternativement sur place, ou par Skype.

Conclusion

La première phase du projet passe par une recherche et une compréhension des différentes sources d'informations disponibles.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

21 septembre 2017, de 13h00 à 10h55

Présent : Nastaran Fatemi, Kewin Dousse, Félicien Fleury (par Skype)

Rédaction du PV le 22 septembre

Compte-rendu

Points de discussion

- Recherches de Kosinski

Nous avons tout d'abord remarqué que les recherches de Michal Kosinski, diplômé de l'Université de Stanford, allaient être de précieuses sources d'informations. Nous allons pouvoir tirer des liens étroits entre les résultats de son étude sur la possibilité de deviner le profil psychologique d'une personne en se basant sur ses 'likes' Facebook.

- Google Analytics

Après avoir suivi le guide Débutant pour YouTube Analytics, Kewin a pu comprendre une partie de l'étendue des possibilités de l'outil. Enormément d'informations sont disponibles, et celles-ci peuvent être cachées/filtrées/triées etc. Cependant il serait intéressant de découvrir les possibilités avancées de l'outil pour se rendre compte jusqu'à quel point celui-ci peut tracker l'activité d'un utilisateur précisément.

- Alternatives à Google Analytics

Quelques alternatives à Google Analytics ont été découvertes, mais celles-ci ne présentent pas vraiment de concept intéressant à l'étude autre que le fait que certaines d'entre-elles sont open-source. Il semble que Google Analytics soit l'outil public le plus grand et le plus utilisé dans sa catégorie.

- Direction du projet

Après une discussion sur les différentes voies futures du projet, l'idée a été sur la proposition suivante : Il s'agira d'implémenter un plug-in pour navigateur qui va récupérer les informations de navigation de son utilisateur de manière automatique et transparente. L'utilisateur va devoir utiliser son compte Facebook afin de se connecter, pour que nous puissions lier les données de navigation avec les données présentes sur un profil Facebook. Toutes les données que nous récupérerons (à la fois par le plugin et par Facebook) seront anonymisées. L'utilisateur sera mis au courant de ce processus avant le début de l'utilisation du plug-in. Il y aura la possibilité d'activer le tracking par période de temps, par exemple à certaines heures de la journée. Nous centraliserons la récupération de ces données et appliquerons des algorithmes afin de déterminer si nous pouvons conclure des informations en se basant sur les données que nous avons récoltées nous-mêmes, et en les vérifiant avec les données que le profil Facebook nous donne en lui « appliquant » la méthode de M. Kosinski. Pour sa volonté de participer à cette enquête, nous allons mettre à disposition de l'utilisateur diverses métriques que nous calculerons en temps réel.

Conclusion

Nous avons désormais une idée bien plus précise du projet à réaliser, et les recherches pour le projet peuvent commencer en visant un but.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

28 septembre 2017, de 13h15 à 14h15

Présent : Nastaran Fatemi, Kewin Dousse, Félicien Fleury (par Skype)

Rédaction du PV le 29 septembre

Compte-rendu

Points de discussion

- **Association SDIPI**

La première discussion a été sur la création de l'association nommée « Swiss Digital Identity and Privacy Institute ». Cette association servira à encadrer le projet et lui donner de la légitimité/visibilité tout en montrant que le but n'est pas économique. Les statuts de l'association seront validés en principe la semaine prochaine.

- **E-mail à myPersonnality**

La source de données la plus importante pour le projet est <http://mypersonality.org>, un site web regroupant les données amassées par les études de Kosinski. Ces données ne sont pas accessibles publiquement, mais il est possible d'en demander un accès en envoyant un mail expliquant le but de notre recherche. Un mail sera écrit la semaine prochaine, une fois que l'association aura une certaine visibilité en ligne, afin de demander l'accès à ces données.

- **Site web SDIPI**

Il est nécessaire que l'Association ait une certaine présence et visibilité en ligne afin de montrer son but au public et de faire des demandes. La mise en place du site web discutée après la réunion de jeudi prochain.

- **Planning**

Une proposition de planning a été faite. Après quelques modifications, celui-ci semble être raisonnable pour le projet.

- **Pages web démonstratives pour GA**

Afin de bien se rendre compte des possibilités données par Google Analytics et également le montrer aux utilisateurs, il est décidé d'implémenter le plus de features possibles de Google Analytics sur un site web d'exemple.

Conclusion

Il est désormais primordial que l'association ait une visibilité en ligne et une certaine visibilité afin de pouvoir demander l'accès à la base de données de Kasinski, qui sera probablement la principale source de données utile au projet.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

5 octobre 2017, de 11h10 à 12h40

Présent : Nastaran Fatemi, Kewin Dousse, Félicien Fleury

Rédaction du PV le 6 octobre

Compte-rendu

Points de discussion

- Google Analytics et site d'exemple

Un site web présentant plusieurs pages de contenu factice imitant un shop en ligne a été implémenté ainsi qu'une intégration avec Google Analytics et certains des concepts de tracking avancés, comme les évènements. Il s'est avéré qu'aller plus loin dans l'implémentation de certaines mesures n'était pas une priorité car la seule limite aux données qu'il est possible de récupérer est en réalité une limitation technique : Il s'agit des informations que les navigateurs peuvent potentiellement révéler à un script, ou à un serveur distant.

- Informations des utilisateurs

Il sera donc intéressant de se poser la question « Quelles sont les informations qu'une page peut potentiellement envoyer à un serveur distant ? ». Ces informations doivent passer par le net pour Google Analytics, et donc il faudra se renseigner non seulement sur les moyens possibles qu'un client a de contacter un serveur (par exemple avec une requête AJAX, ou même avec une tentative d'accès à un fichier comme une image sur le serveur), ainsi qu'aux types d'informations qu'a accès un navigateur web classique.

- Création de l'association SDIPI

La fin de la réunion formelle a porté sur le review des statuts de l'association prochainement créée : « Swiss Digital Identity & Privacy Institute ». Quelques changements ont été faits ; Les status seront donc définitivement validés plus tard.

- Objectifs du projets

Une discussion sur les objectifs du projets a également eu lieu. Nous avons décidé que le projet allait viser à chercher une correspondance entre les URL visitées par une personne et son profil psychologique. Ce lien se fera à l'aide des données de Kosinski, qui nous aidera à lier les likes Facebook d'une personne et son profil psychologique. Le remplissage du questionnaire psychologique par les volontaires de notre projet sera facultatif.

Conclusion

L'association va terminer de se créer afin d'envoyer une lettre de demande à Kosinski, et pendant ce temps les recherches sur les possibilités techniques de divulgation des informations va continuer.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

10 octobre 2017, de 9h30 à 10h05

Présent : Nastaran Fatemi (Skype), Kewin Dousse (Skype), Félicien Fleury (Skype)

Rédaction du PV le 11 octobre

Compte-rendu

Points de discussion

- **Statuts de l'association**

Les statuts de l'association doivent subir quelques changements mineurs avant d'être définitifs. Félicien va effectuer les modifications nécessaires, puis les statuts seront lus, imprimés et signés par tous les membres de l'association. Un PV de l'assemblée constitutive déroulée sera également rédigé.

- **Site web**

Un début de site web a été présenté. La structure générale et le thème seront conservés. Celui-ci ne contient que peu de contenu, il sera étoffé pour jeudi dans le but d'être présentable et mis en ligne.

- **Lettre à myPersonnality**

Le template de l'e-mail à envoyer à myPersonnality reste à compléter par quelques détails : Un enregistrement du projet sur le site <https://osf.io> est nécessaire. Les détails du projet et des membres seront complétés pour jeudi également. Le but est d'avoir en main tous les éléments nécessaires pour écrire le mail définitif jeudi et l'envoyer.

Conclusion

La complétion des informations et contenus pour envoyer l'e-mail de demande d'accès aux données à Kosinski est actuellement la priorité, et cette tâche devrait arriver à son terme jeudi.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

12 octobre 2017, de 13h00 à 13h40

Présent : Félicien Fleury (Skype), Kewin Dousse (Skype)

Rédaction du PV le 12 octobre

Compte-rendu

Points de discussion

- Site web**

La première discussion a porté sur les détails du site web. La plupart du contenu a été ajouté, quelques corrections ont été effectuées, et la mise en ligne officielle du site s'est terminée quelques heures après la fin de la réunion.

- Comité d'éthique**

Après la complétion d'informations à la fois sur le site web officiel de l'association et sur la page OSF requise du projet, il a été remarqué que dans le template d'e-mail pour Kosinski se trouve une ligne faisant référence à l'IRB (Institutional Review Board). Ceci n'avait pas été mis en avant jusqu'ici, et signifie probablement qu'une approbation d'un comité d'éthique est nécessaire pour continuer le projet, car Kosinski s'attend à le recevoir par e-mail. Cette étape sera discutée avec Nastaran car l'école d'ingénieurs est probablement compétente pour ce problème.

Conclusion

La question du comité d'éthique est à traiter au plus vite car il s'agit d'une étape non anticipée qui pourrait considérablement ralentir l'obtention des données

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

16 octobre 2017, de 15h00 à 15h20

Présent : Félicien Fleury (Skype), Kewin Dousse (Skype)

Rédaction du PV le 17 octobre

Compte-rendu

Points de discussion

- Comité d'éthique

Bien que la question de l'acceptation du projet par un comité d'éthique soit en suspens, nous allons pour l'instant avancer tout de même dans la partie technique du projet

- Architecture

Il y eut ensuite une discussion sur l'architecture de l'application de base à réaliser pour la récupération des données des utilisateurs. L'idée initiale d'extension de navigateur est bonne, mais demande de développer une extension par navigateur différent. Bien que la plupart du code soit le même, le développement partira sur un « userscript » dans un premier temps : Il s'agit d'une extension avec des fonctionnalités réduites, n'utilisant que du JavaScript pur (sans utiliser d'API navigateur) et ayant l'avantage d'être compatible sur plusieurs navigateurs. De même, le développement de la partie serveur va également commencer, suite à la mise en fonction d'une machine virtuelle pour accueillir le software serveur.

Conclusion

L'acquisition des données se trouve retardée, mais le projet avance tout de même du point de vue développement pendant ce temps.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

26 octobre 2017, de 13h00 à 13h45

Présent : Félicien Fleury (Skype), Kewin Dousse, Nastaran Fatemi

Rédaction du PV le 28 octobre

Compte-rendu

Points de discussion

- Comité d'éthique**

Etant donné les délais attendus par non seulement la réponse espérée de Kosinski, mais surtout par celui du comité d'éthique, la décision a été prise de passer cette idée au second plan et de chercher un autre axe de développement pour le projet.

- Idées**

Le but de la discussion suivante a été de chercher de nouveaux axes de développement pour le projet, en partant de l'idée que nous n'aurons pas accès aux données de la base de données de Kosinski.

Plusieurs idées ont vu le jour ici, dont celle de développer un produit en partenariat avec une entreprise externe. Mais l'idée qui a été retenue au final est différente, mais reste en cohésion avec le développement technique effectué jusqu'ici : Le but sera de développer dans un premier temps une extension de navigateur pour :

- Récolter des données utilisateurs concernant leur fréquentation des sites web
- Renseigner les utilisateurs sur leur utilisation du web, et les informer en leur montrant la manière dont ils apparaissent au web, par exemple en générant un avatar leur ressemblant, ou en leur montrant des statistiques sur leur navigation et les dangers potentiels

Cette récolte d'information donnera lieu dans un deuxième temps à un jeu de données sur la navigation des utilisateurs qui sera mis en relation avec leur profil Facebook. Les données seront ensuite analysées afin d'y trouver par exemple des corrélations intéressantes.

Conclusion

La direction du projet change, mais la partie technique qui a été faite jusqu'ici n'est pas perdue : Nous changeons de vision et d'objectifs à moyen terme, mais le développement continue dans le même sens.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

1 novembre 2017, de 14h05 à 14h55

Présent : Félichen Fleury (Skype), Kewin Dousse (Skype), Nastaran Fatemi (Skype)

Rédaction du PV le 3 novembre

Compte-rendu

Points de discussion

- Planning

Le planning du projet a été réadapté en fonction des modifications dans les objectifs à moyen terme. Nous n'allons donc pas nous baser sur les données de l'étude de Kosinski, et par conséquent n'allons pas attendre sa réponse pour continuer le projet.

- Plug-In Chrome

Nous allons changer les objectifs du projet ainsi : Le but ne sera pas de trouver des corrélations entre les URLs visitées par un visiteur et son profil psychologique (déduit par son profil Facebook + données de Kosinski). Nous allons à la place :

- Donner à l'utilisateur une interface montrant des statistiques sur ses habitudes de navigation du web sous plusieurs formes. Images, graphiques, et nombres.
- Récolter des données sur la navigation des utilisateurs afin d'en trouver des statistiques intéressantes.

- Stratégie

Il est nécessaire de savoir comment positionner le plug-in et l'étude par rapport aux concurrents. Des plug-ins avec existent déjà proposant des fonctionnalités similaires, et un état de l'art est nécessaire afin de savoir dans quelle direction va continuer le développement.

Conclusion

Nous devons savoir comment se positionner par rapport aux plug-ins similaires afin de pouvoir développer des fonctionnalités attrayantes pour les nouveaux utilisateurs.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

9 novembre 2017, de 14h05 à 14h55

Présent : Félicien Fleury (Skype), Kewin Dousse, Nastaran Fatemi

Rédaction du PV le 9 novembre

Compte-rendu

Points de discussion

- Rapport**

La partie d'analyse du rapport a été rédigée en grande partie. Les quelques parties manquantes seront complétées par la suite. La comparaison des extensions existantes sera étoffée afin d'en tirer une conclusion pouvant nous renseigner sur la place que prendra notre extension par rapport à celles existantes, et en quoi les fonctionnalités seront novatrices.

- Fonctionnalités**

La discussion centrale a été les fonctionnalités que le plug-in allait proposer, ainsi que l'intérêt pour les statistiques que nous allions tirer à la fin de l'étude. Nous allons devoir nous baser non seulement sur les données Facebook des utilisateurs, mais nous allons également analyser le contenu des pages que celui-ci visite, et pas seulement leur URL. La discussion a porté sur les méthodes d'analyse de contenu de pages web ; Lesquelles utiliser, que stocker comme données et comment les utiliser au mieux. Nous allons procéder par étapes ; la première d'entre elles sera d'enregistrer le contenu des pages web dans la bases de données.

- Techniques envisagées**

Nous avons réfléchi à des algorithmes à appliquer lors de la récolte de données dans le but d'obtenir des statistiques plus intéressantes sur la navigation des utilisateurs. Le principal intérêt que nous voyons dans l'analyse de contenu des pages est d'effectuer de la reconnaissance de topics sur les pages. Ainsi, nous pourrons – par exemple - tirer des parallèles entre les sujets visités par un utilisateur et ses informations démographiques, ou ses « likes ».

Conclusion

Les fonctionnalités principales du plug-in se définissent, et le développement de la récolte de données progresse en parallèle. Restera à discuter de la stratégie de « publicité » pour le plug-in.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

13 novembre 2017, de 16h00 à 16h30

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype)

Rédaction du PV le 15 novembre

Compte-rendu

Points de discussion

• Méthodes de User Profile Tracking, et de Topic Recognition

Une série de liens et papers ont été synthétisés dans le but d'en apprendre plus sur les techniques actuelles de deux objectifs différents : Premièrement, reconnaître les traces d'un utilisateur et reconstituer son profil en utilisant plusieurs sources de données. Deuxièmement, être capable de définir un ou plusieurs mot-clés représentant le sujet discuté sur une page web/un document.

La conclusion de ces études est la suivante : Les techniques pour tracker un utilisateur sur le web sont déjà connues, et le rapport de James Nolan est toujours intéressant quant à certaines techniques à utiliser. Une nouvelle information est cependant la performance des algorithmes permettant d'extraire le sujet d'une page web : Il semblerait d'après plusieurs sources indépendantes que la méthode de TF-IDF, en conjonction avec certaines autres techniques, donne les résultats les plus probants pour notre cas. Nous allons donc probablement l'implémenter.

Conclusion

Nous sommes à présent au clair sur les techniques à utiliser pour la suite d'outils, particulièrement au niveau de la reconnaissance des sujets d'une page. Ceci pourra désormais être implémenté.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

16 novembre 2017, de 14h15 à 15h00

Présent : Kewin Dousse, Nastaran Fatemi, Félicien Fleury (Skype)

Rédaction du PV le 19 novembre

Compte-rendu

Points de discussion

- Fonctionnalités de l'Extension**

Après avoir passé en revue une liste des plug-ins existants, nous allons pouvoir nous concentrer sur l'implémentation du nôtre au travers de deux axes principaux. La liste actuelle est lacunaire et sera complétée par la suite par d'avantage d'explications sur certaines extensions.

Nous allons nous focaliser sur montrer des informations à l'utilisateur concernant : 1) Les trackers sur la page et vers qui les informations sont envoyées, et 2) Comment le profil reconstitué de l'utilisateur apparaît vu par le web.

- Implémentation**

La méthode de TF-IDF sera initialement utilisée pour reconnaître les topics d'une page web. Il s'agit de la fonctionnalité qui sera implémentée au plus vite.

Conclusion

L'implémentation des fonctionnalités continue, avec une vision plus claire sur les techniques à utiliser ainsi que

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

23 novembre 2017, de 13h35 à 14h40

Présent : Kewin Dousse, Nastaran Fatemi, Félicien Fleury (Skype)

Rédaction du PV le 24 novembre

Compte-rendu

Points de discussion

- **Maquettes de l'interface**

Des maquettes papier de l'interface du plug-in ont été discutées. Deux pages principales seront présentées : La page « Trackers » montrant des informations et visualisations sur les différents trackers rencontrés sur les pages, et la page « Profile » montrant des informations sur le profil reconstitué de l'utilisateur. En plus de ces deux pages, se trouveront une page « Général » montrant un résumé de l'état du plug-in et de la connexion de l'utilisateur, et une page « Stats » montrant des informations générales sur l'utilisation du projet, tous utilisateurs confondus.

- **Implémentation**

Le TF-IDF fonctionne. Pour lundi sera implémenté un début de l'interface de la page « Profile ».

Conclusion

Bien que non définitive, la liste des fonctionnalités de l'interface client est assez bien définie pour prodécer à un début d'implémentation et de tests.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

30 novembre 2017, de 13h30 à 14h15

Présent : Kewin Dousse, Nastaran Fatemi

Rédaction du PV le 1 décembre

Compte-rendu

Points de discussion

• Interface

L'avancement de l'interface de la page 'Profile' a été discuté. Celle-ci contient les visualisations avec les graphiques en barres comme les sites/domaines les plus vus/regardés ainsi que la liste des pages et les mots-clés associés, mais il manque encore le wordcloud, le graphe des intérêts, et les graphiques des mots-clés sur la durée ainsi que la sélection de l'intervalle de dates. Plusieurs modifications seront à effectuer pour la qualité des informations affichées sur l'interface, comme la détection de la langue lors du retrait des stopwords des pages, ainsi qu'une meilleure détection du temps passé sur les pages par un utilisateur en comptant tout type d'interaction avec celle-ci. D'autres changements purement sur l'affichage de l'interface seront aussi effectués, comme la combinaison de plusieurs tableaux en une seule visualisation.

• Dates

Quelques dates clés ont été définies pour la suite à court terme :

- Mardi 5 déc. : Fin de la page Profile
- Vendredi 15 déc. : Fin de la page Trackers
- 15 – 22 déc. : Tests de l'interface en interne + retrait du login Facebook pour un login personnalisé

• Données utilisateur

La question de l'intérêt de la récolte des données utilisateurs s'est également posée. Les quelques idées proposées vont dans le sens d'une publication scientifique, et visent à articuler le contenu principalement autour de deux axes : La présentation de statistiques concernant les données récoltées, et la validation que les profils détectés par le plug-in correspondent à la réalité vue par les utilisateurs. On pourra par exemple émettre un questionnaire à ceux-ci afin de chercher une corrélation entre les informations recueillies, et les informations que ceux-ci délivrent volontairement.

Conclusion

Avec les réponses à quelques questions touchant sur le but final du projet, nous sommes au cœur de la phase d'implémentation de l'interface et des fonctionnalités qui lui sont relatives.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

5 décembre 2017, de 8h30 à 9h05

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 6 décembre

Compte-rendu

Points de discussion

• Interface

Un point a été fait sur la page « Profile ». Celle-ci comprend les sections supplémentaires « Wordcloud » montrant un nuage des mots les plus vus par l'utilisateur, et « History » affichant un graphique des sites les plus visités sur un intervalle de temps. De plus, les sections « Most visited » et « Most watched » ont été remaniées : Le tableau des keywords par page a été intégré à chacun des autres tableaux montrant les sites et les domaines de la section. Bien que l'interface soit fonctionnelle, plusieurs facteurs rendent les résultats affichés peu fiables (pas de JavaScript exécuté sur les pages, améliorations possibles dans la phase de cleaning des données). Ceci sera remédié.

• Objectifs

Les prochaines tâches à effectuer ont été définies : Jusqu'à la fin de la semaine, l'accent sera mis sur la page « Profile » afin de la terminer et de rendre plus fiables les résultats montrés, notamment les keywords et les intérêts de l'utilisateur. La semaine suivante, la page « Trackers » sera implémentée.

Conclusion

Avec les réponses à quelques questions touchant sur le but final du projet, nous sommes au cœur de la phase d'implémentation de l'interface et des fonctionnalités qui lui sont relatées.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

7 décembre 2017, de 15h05 à 15h40

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 8 décembre

Compte-rendu

Points de discussion

- **Interface**

La page « Profile » a été discutée. Tous les onglets sont implémentés, mais certains méritaient encore une discussion. Ainsi, l'objectif de l'onglet « Interests Graph » a été plus précisément décidé et celui-ci subira quelques modifications, ainsi que l'onglet « History » qui servira à montrer des tendances de keywords, plutôt que de sites web. Le choix d'un intervalle de dates reste à implémenter. Ces changements sont prévus pour mardi matin.

Conclusion

La page « Profile » arrive à la fin de son implémentation, et le focus devrait être sur la page « Trackers » dès mardi prochain.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



MASTER OF SCIENCE
IN ENGINEERING

PV de réunion

12 décembre 2017, de 8h35 à 9h05

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 12 décembre

Compte-rendu

Points de discussion

- **Filtrage par date**

Le filtre par date est ajouté sur la page : Il est possible de choisir une date de début et une date de fin pour l'affichage de toutes les données. Des changements conséquents ont été faits sur la manière de calculer les données afin que l'interface soit réactive à ces changements : La plupart des données sont pré-calculées sur le serveur.

- **Graphique « History »**

La deuxième version du graphique « History » a été mis en place, mais ne semble pas assez concluant pour être définitif. Les résultats visuels obtenus ne sont pas toujours représentatifs et visuellement intéressants des données que nous souhaitons afficher, et nous rediscuterons de cette partie jeudi prochain.

- **Graph « Interests »**

La page du graphe des intérêts a suscité des questions sur son fonctionnement. Après discussion, il sera plus intéressant de lier les topics et les mots-clé trouvés, aux intérêts de l'utilisateur que lui-même aura défini lors de l'inscription. Il sera donc nécessaire de lui demander ses intérêts parmi une hiérarchie de centres d'intérêts lors de l'inscription, et cette page permettra de faire un lien entre les intérêts décrits par l'utilisateur, et les intérêts que nous trouverons nous-même.

Conclusion

Des discussions sont encore en cours sur des aspects de la page « Profile », mais de plus en plus d'entre-eux approchent une version finale.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

21 décembre 2017, de 8h35 à 9h05

Présent : Kewin Dousse (Skype), Félicien Fleury (Skype)

Rédaction du PV le 21 décembre

Compte-rendu

Points de discussion

- Points restants avant le lancement de l'extension**

Parmi les quatre points restants à résoudre énoncés la dernière fois, le refactoring de la partie communication serveur et la logique d'envoi de l'extension est terminée. L'extension stocke les messages et ne les envoie qu'une fois toutes les 30 sec.

L'authentification Facebook est enlevée mais un nouveau système à mettre en place a été discuté : Lorsque l'utilisateur installe l'extension, un identifiant lui sera associé et communiqué. Il pourra ensuite le réutiliser sur d'autres machines si il le souhaite. Ceci évite à l'utilisateur une phase d'inscription.

La vue des Trackers et la fin de l'implémentation des intérêts reste à terminer. Comme le temps restant est probablement insuffisant jusqu'aux vacances de Noël, un ou deux jours seront pris entre le 26 et le 28 décembre pour terminer complètement l'extension afin de la proposer à une dizaine d'utilisateurs.

Conclusion

La phase d'implémentation arrive à son terme, et l'extension sera bientôt prête pour une utilisation réelle.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

9 janvier 2017, de 9h05 à 9h30

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 9 janvier

Compte-rendu

Points de discussion

• Résultats récoltés

Les modifications requises de l'extension pour une utilisation par plusieurs utilisateurs ont été effectués durant la première partie des deux semaines de pause : Remplacement du login Facebook par un login instantané à l'installation, et finition du système de centres d'intérêts.

L'extension a été utilisée par 7 utilisateurs différents pendant une période d'environ une semaine. Les résultats récoltés ont commencé à être traités, mais la nouvelle taille de ceux-ci pose des problèmes techniques au serveur qui était jusqu'ici suffisant. La résolution de ces problèmes est en cours.

Conclusion

Les prochaines tâches sont la résolution des problèmes techniques dûs à la quantité de données, et le début de l'analyse des résultats obtenus en plus de l'ajout de la page Trackers dans l'interface.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

18 janvier 2017, de 9h05 à 9h30

Présent : Kewin Dousse, Nastaran Fatemi, Félicien Fleury

Rédaction du PV le 23 janvier

Compte-rendu

Points de discussion

• **Interface**

L'implémentation de la partie Trackers de l'interface touche à son terme. Il est désormais possible de lister les domaines envoyant et recevant le plus grand nombre de domaines, ainsi que de cliquer sur l'un deux pour avoir les détails de quels domaines ont communiqué avec celui cliqué. Quelques améliorations sont discutées, comme la possibilité d'afficher le nombre de domaines contactés directement sur les premières pages sans avoir à cliquer sur un domaine particulier.

Pour la partie Profile, l'utilité du « Topics Graph » a été rediscutée : Nous nous en servons principalement pour demander des informations de l'utilisateur sur sa reconnaissance des centres d'intérêts dans les topics proposés. Un graphe n'est donc plus nécessaire : La vue sera désormais une liste, où l'utilisateur peut entrer un centre d'intérêt par ligne (topic). Une révision de la structure du backend est nécessaire afin que ces opérations puissent être faites en cohésion avec un changement de modèle LDA.

Conclusion

L'implémentation de l'interface se termine cette semaine. Une fois les dernières modifications effectuées, le focus sera mis sur le rapport écrit.