Marianne 5- Théo 4 - Nour 5- Valentine 5- Nathan 2

T : Diapo 1 :

Bonjour à tous, nous allons vous présenter notre projet tableau de bord. Vous, Mme Bahsoun et M. Mokadem êtes les clients et nous 5 sommes les fournisseurs. Nous nous sommes intéressés aux arnaques en ligne durant ces 20 dernières années.

Au cours de ce projet nous avons parfois travaillé ensemble, parfois en sous-groupe sur différentes tâches réparties équitablement et en fonction des compétences de chacun.

Notre objectif est de développer un système d’aide à la décision à partir de données textuelles issues d’articles en ligne.

T : Diapo 2 :

Indiquer le sommaire :

- Organisation du projet

- Diagramme de Gantt

- Présentation du sujet

- Démarche de développement

- Stockage des données

- Interrogation des données

- Résultats

- Discussion

- Bilan du projet

**Valentine - Diapo 3 :**

**1. Chef de projet (Théo) :** A su manager notre équipe en étant un bon leader et assurer le bon déroulement de chaque revue et de l’ensemble du projet.

**Sous chef (Nathan)**

**Gestion de configuration (Marianne) :** A su contrôler l’évolution de l’ensemble des processus permettant d’assurer la conformité du projet aux exigences du cahier des charges.

**Assurance et contrôle qualité (Nour) :** A su créer une véritable charte qualité à laquelle tout le groupe a pu se référer sans aucun problème & a rigoureusement contrôlé la qualité de chaque notebook, à chacune des versions.

**Responsable de rédaction (Moi) :** J’ai été chargée de rédiger le rapport ainsi que le PowerPoint en assurant la cohérence des contenus. J’ai également veillé à la relecture et à la validation des documents par les autres membres de l’équipe.

**2.** **Choix du sujet & création des équations de recherche :** Tout le groupe

**Préparation des données (codes Python) :**

Récupération, nettoyage et mise en forme des données : Marianne, Nour & Théo

Recherche d’anomalies dans les données : Marianne

**Valorisation des données (BD, codes SQL & fonction d’alimentation de la BD) :**

Création de la BD, stockage et interrogation des données : Valentine

**Visualisation des données :**

Analyse des données, résultats et discussion : Marianne, Nour, Théo & Valentine

**=>** Même si on avait des tâches différentes, rien ne nous empêchait de nous entraider. Bien que Nour & Théo étaient chargés de l’aspect programmation avec Marianne, ils sont venus m’aider à tour de rôle pour le côté BD/SQL. Marianne & moi, nous nous sommes spécifiées à nos tâches quant Nour & Théo ont été plus polyvalents.

**Valentine - Diapo 4 :**

**1.** Afin de mener à bien ce projet, nous avons réparti les tâches et les rôles à travers un diagramme de Gantt : modèle qui nous semblait le plus adéquat.

Diagramme de Gantt prévisionnel :

* Pensé en fonction de la suite logique des tâches de chacun.
* Trop optimiste sur le temps de travail.
* Prise de retard.

**2.** Nous avons donc ajusté le diagramme en fonction de notre avancement :

Diagramme de Gantt réel :

* Recherche et collecte des données beaucoup plus longue que prévue (problème sur l’équation de recherche).
* Préparation & valorisation des données retardées.
* Ajout de la semaine de vacances => impossibilité de trouver un créneau pour des revues
* Temps pour la rédaction du rapport & la préparation de l’oral respectés.

**M : Diapo 5 :**

Sujet : Evolution des arnaques en ligne (de 2000 à nos jours)

* Choix : Sujet avec assez de résultats (plus de 3000) et préférence parmi les sujet OK

Equation de recherche avec des termes anglais = nous avons étudié des articles anglais

Outils utilisés pendant le projet :

* Notebook Python (Jupyter, Google Collab)
* BD = Orcale SQL developper (serveur de l’université)
* Versionnage, partage de fichiers : Drive, peut-être moins efficace que Github. Mais Github pas assez maîtrisé
* Communication :
  + Zoom : Réunion Client/Fournisseur
  + Discord : Equipe

**M : Diapo 6 :**

Démarche de développement

A-0 : vue d’ensemble (Créer TBD en respectant cahier des charges)

Division en 4 étapes

A1 - Recherche des données : Choisir le sujet, Récupérer les données

A2 - Préparation des données : Nettoyer et mettre en forme données collectées

A3 - Valorisation des données : Concevoir, créer, peupler, interroger BD

A4 - Visualisation des données : Analyser les données, Sélectionner les résultats les plus pertinents, Présenter les résultats via TdB

T : Diapo 7 :

A préciser : date\_arnaque = date\_article

Lors de l’élaboration du schéma entité/association (créé grâce au logiciel [*mocodo*](http://mocodo.wingi.net/?fbclid=IwAR01syx-yYbG9opRVO5Nrq8gRaMgAjYgCK1WS1gyPiCdyrDOm9fFDZIsuyw) en ligne), il nous a été nécessaire de faire valider notre démarche et nos choix auprès du client. Une fois fait, nous avons pu commencer à créer la structure de notre base de données : création des tables (tout en faisant attention aux clefs primaires et étrangères), des procédures et des vues associées.

**Schéma relationnel correspondant :**

SUBIR (#id\_arnaque, #id\_acteur)

AUTEUR (id\_auteur, nom, affiliation, pays)

ECRIRE (#id\_article, #id\_auteur)

ACTEUR (id\_acteur, nom, demographie, nationalite)

CREER (#id\_arnaque, #id\_acteur)

ARNAQUE (id\_arnaque, type\_arnaque, date\_arnaque, plateforme\_utilisee, nb\_victimes, nb\_arnaqueurs, #id\_article)

MOTCLEF (#id\_article, mot\_clef)

CONTRER (#id\_arnaque, #id\_acteur)

ARTICLE (id\_article, date\_article, titre, theme)

**Nada - Diapo 8 :**

Pour la partie d'interrogation des données, on a choisi de les faire avec des vues. On en a proposé plusieurs, tel que :

* regarder la nationalité/démographique de l’acteur
* ou encore exploiter d’autres tables à savoir créer/contrer et subir.

Ces tables pouvaient être des axes intéressants à présenter afin d'exploiter au mieux nos données.

Mais, suite aux modifications du schéma E/A indiqué précédemment par Théo, nous étions dans l'obligation de laisser tomber la visualisation de ces vues.

Toutefois, nous avons quand même pu réfléchir à plusieurs vues qui pouvaient être intéressantes, comme l'évolution du type d’arnaque dans le temps, ou encore Mots clefs en fonction du temps.

Au cours de l’interrogation des données, nous nous sommes rendu compte que certaines anomalies n’avaient pas été détectées aux tâches précédentes.

Nous avons ainsi dû faire une partie du nettoyage des données via python (suppression des lignes - mise en forme des dates, .... ). Une seconde partie s'est faite sur la base de données qui était plus facile à manipuler pour effectuer ce type de changements. Par exemple, nous avons transformé “United Kingdom” en “UK” ou encore “Republic of China” en “China”.

**Nath : Diapo 9** :On va maintenant interpréter les mots clés dominants et essayer de les contextualiser

On constate qu’entre les années 2000 à 2010 l'intérêt des arnaqueurs pour les données personnelles s'ajoute à celle des données financières.

On assiste à cette époque à la « découverte » des risques liés à l’usage d’internet et à l’interconnexion des réseaux.

Cette époque connaît donc une explosion des attaques virales, face aux entreprises et aux particuliers, ce qui aura pour effet de placer la sécurité internet ‘the internet security’ au centre des priorités.

On peut prendre l’exemple du e-commerce qui prend son envol dans les années 2000, dont le plus grand risque au yeux des consommateurs est le paiement en ligne.

Entreprise et utilisateurs se confrontent donc à de nouvelles problématiques :

Comment garder l’intégrité, et la confidentialité de nos informations et de nos transactions ?

C’est dans le but de répondre à ces enjeux de sécurités que se multiplient les outils de lutte contre la fraude.

D’une part, la généralisation des logiciels antivirus ‘antivirus software’ devient indispensable.

Aussi interviennent des systèmes de lutte plus poussé tel que les systèmes de réputation ‘reputation system’ ainsi que les techniques d’analyse de données pour la détection des fraudes ‘fraud detection’ qui prennent leurs essor dans les années 2000.

En parallèle, la carte de crédit ’credit card’ devient le premier moyen de paiement en 2003, malgré que ce soit loin d’être un moyen de paiement garanti.

**Nath : Diapo 10** :En s’intéressant maintenant aux résultats datés de 2011 à 2020, la carte de crédit fait son retour dans notre nuage de mot, et c’est logique car 2011 est l’année de lancement du paiement sans contact, moyen de paiement non pas sans faille car très facilement piratable.

De plus, les années 2010 représentent l’âge d’or des réseaux sociaux, favorisant la profilation des arnaques, notamment les arnaques de fishing et de fraude d’identité’ identity theft’.

D’autre part, tout ce qui tourne autour de la protection des données personnelles et de la vie privée à l'ère des réseaux sociaux, devient un véritable risque pour les utilisateurs.

Risque accru par la commercialisation des divers objets connectés et l’apparition de la cryptomonnaie comme le ‘Blockchain’.

Concernant les années 2015 à 2020, l’impact des nouvelles technologies est donc beaucoup plus fort.

Et c’est en réponse à ces menaces que des nouvelles mesures se développent.

L’utilisation du ‘machine learning’ et du ‘big data’ dans la lutte contre les arnaque devient prédominante et est à l‘origine d’une nouvelle génération de détecteurs de fraude.

**M : Diapo 11 :**

Classification des différents types d'arnaques pas très efficace.

* Sur plus de 5000 articles, 4000 sont simplement placés en 'others'.

=> Difficultés techniques

=> Beaucoup d'articles parlent des arnaques de manière générale

=> Malgré l'équation de recherche : articles qui ne parlaient pas forcément d'arnaques.

Présenter les différents types : au fur et à mesure ; s’ils sont évoqués pendant les analyses

On remarque deux tendances:

* Augmentation progressive de l’intérêt pour les fraudes:
  + consumer scam = arnaque lié à des achats en ligne = bleu
  + bank scam = banque en ligne frauduleuse ou banque victime = gris
  + spam email = jaune
* La plupart des autres arnaques : assez bas, stagnation et/ou légère augmentation

=> signe que nous avons peut-être identifié quelque chose de trop précis

Arnaques impliquant des cryptomonnaies (virtual currency scam, marron) = comportement particulier

Articles les concernant quasi inexistants avant 2015, véritable montée en flèche ces dernières années

La baisse à la fin : nous avons gardé 2021 dans nos données. Et l’année n’est pas fini.

**M : Diapo 12 :**

Division de la période d’étude en 4

1999 - 2005 et 2006 à 2010 :

* Types les plus représentés : fraudes liées aux commerces (“consumer scam”, en bleu), fraudes liées aux banques (“bank scam”, en jaune), fraudes liées aux enchères (“auction scam”, vert pomme)
* spam: augmentation au cours de la décennie (vert pâle/turquoise)

**M : Diapo 13 :**

2011 - 2015 et 2016 - 2021 :

* Baisse de l’importance des fraudes aux enchères au cours de la décennie (“auction scam”, vert pomme)
* On remarque à nouveau l’arrivée des fraudes liées aux cryptomonnaies (“virtual currency scam”, vert foncé)
* Il semble y avoir au cours des dernières années une diversification des arnaques étudiées par la communauté scientifique => Précision des études dans le domaine ?

Interprétation à prendre avec précaution au vu de nos difficultés à identifier les types d’arnaques

**Nada - Diapo 14 :**

Un autre axe qui nous a semblé important à considérer est le thème des articles.

Nous avons choisi de faire une classification non supervisée avec le model lda qui a servi à modéliser des sujets dans l'ensemble des données.

on a choisi de conserver 2 thèmes qui se distinguaient avec des mot-clefs plutôt représentatifs : description et détection des fraudes.

Pour le thème description on remarque que les mots-clefs les plus courants sont :

* internet, cybercrime et fraude, ce qui semble cohérent vis-à-vis de notre sujet. Nous pouvons aussi distinguer les mots :
* -Credit card et identity qui sont fréquents dans les articles sur les fraudes en ligne.

Du côté des détections de fraude, il n’y a pas de mot-clef flagrant, qui se détache plus particulièrement des autres, contrairement aux descriptions.

Cela vient sûrement du fait que notre source comporte beaucoup moins d’articles parlant de ce thème.

Nous pouvons aussi distinguer un nombre important de mots-clefs liés aux nouvelles technologies (machine learning, deep learning, internet et Big data).

on peut donc en déduire qu'il y a un intérêt pour appliquer ces dernières à la détection des fraudes.

**Nada - Diapo 15 :**

Comme expliqué précédemment à partir des articles, nous avons dégagé deux thèmes :

- détection et description.

Le thème de la détection concerne les documents qui parlent de la façon de repérer les arnaques et de la lutte contre celles-ci.

Quant au thème de ‘description', ça concerne plutôt les articles qui parlent d’un type d’arnaque et qui le décrivent.

À partir de ces données, nous avons tracé un graphique qui représente l'évolution des thèmes dans le temps. On peut donc constater que notre source de données nous propose globalement plus d’articles parlant de descriptions d’arnaque que de détections.

Cela semble cohérent vis-à-vis de notre équation de recherche et du sujet abordé. Il est à noter que l’écart entre les deux thèmes s’intensifie à partir de l’année 2015. Les articles parlant de fraude montent en flèche en 2020 avec la pandémie et atteignent le sommet jusqu'à aujourd’hui.

**Valentine - Diapo 16 & 17 :**

**Tous les graphs**

* De + en + de nationalité d’auteurs différentes apparaissent au fil des années :
  + soit auteurs pas intéressés par les arnaques dans les pays non représentés
  + soit notre source étant exclusivement en langue anglaise => biais des résultats : les auteurs d’autres nationalités écrivaient peut être mais dans une autre langue
* Le nombre d’articles concernant les arnaques en ligne est en continuelle expansion : si on se concentre sur les USA par exemple : presque 90 articles écrits entre 1999 et 2010, plus de 250 entre 2011 et 2015 et enfin plus de 800 depuis 2016 sachant que la période n’est pas terminée, étant courant 2021.

**Graph 4 :**

* Une majorité de pays semblent peu écrire sur les arnaques en anglais, tels que la Belgique ou l’Afrique du Sud par exemple.
* Seuls la Chine, l’Inde, l’Allemagne, l’Australie, le Canada, le Royaume Uni, l’Italie et les Pays Bas ont une quantité d’auteurs d’articles conséquente.

**Valentine - Diapo 18 :**

Comme indiqué par Nour précédemment, il nous a été possible de dissocier les articles parlant de détection des arnaques, de ceux qui concernent une description d’arnaque.

* La majorité des pays écrivent plus sur les descriptions d’arnaques que sur leurs détections. Cela nous a semblé cohérent puisque lorsque l’on parle d’arnaques en ligne, on a plus tendance à chercher une description d’une arnaque précise. De plus, en fonction des arnaques, il peut ne pas y avoir de lutte qui la concerne. Seule l’Inde fait exception et présente plus de détections que de descriptions d’arnaques.
* Sur les 52 pays représentés, 88% d’entre eux ont écrit moins de 200 articles depuis 1999, dont la France. Peut-être sont-ils moins concernés par les arnaques en ligne que les USA, l’Angleterre, l’Inde, la Chine, l’Allemagne ou encore l’Australie ? Cela ne reste que des suppositions car nous n’avons aucun moyen de le vérifier.

**Nada - Diapo 19 :**

En conclusion, nous avons pu constater une prédominance des pays anglophones dans nos résultats et surtout des USA.

Cependant, nous, étant limités à la langue anglaise au niveau de la source et au niveau de l’équation de recherche, il nous semble cohérent que la majorité des affiliations des auteurs soient principalement de nationalité américaine.

De plus, au cours du temps, il est à constater que les nouvelles techniques sont sources de nouvelles menaces, via la crypto-monnaie comme le bitcoins.

Par exemple. De nouvelles solutions ont également tendance à apparaître simultanément tels que le machine learning ou encore le big data.

Au vu des résultats précédents nous pouvons aussi souligner le fait que dans 52 pays étudiés nous n’en retrouvons que 6 qui sont bien représentés. Cela est-il lié à notre source d’articles qui est biaisée ?

Nous pouvons remarquer que le thème description est le plus présent dans les articles que détection.Et que les arnaques liées au commerce et aux banques sont les plus présentes dans les articles.

**Nada - Diapo 20 :**

Toutes ses analyses précédentes nous font interroger sur le futur des arnaques à savoir :

* Le taux de fraude va-t-il continuer d’augmenter au fil des années ?
* La lutte finira-t-elle par prendre le dessus sur les arnaques ?
* Qu’en est-il de l’évolution de la technologie et du big data ?
* Si notre source n’avait pas été exclusivement en anglais, aurions-nous eu des résultats différents ?

T : Diapo 21 :

* **Difficulté à comprendre le but du projet**
* **choix du sujet / EQ recherche trompeuse => sources non pertinentes**
* **Schéma E/A beaucoup de changement**
* **Retard sur le diagramme prévisionnel (gantt)**
* **Bonne communication malgré la distance (merci valentine)**
* **Groupe complémentaire chacun son role**
* **(mises en pratique des cours)**

Diapo 22 : Blabla général qui ajuste les 20 minutes

Théo : Pour finir, ce projet nous a permis d’avoir une expérience professionnelle, pour ma part blablabla.

Théo

Valentine

Nour

Nathan

Marianne

Théo : Avez-vous des questions?