Hackathon

« les champs de Sirene »

18 et 19 janvier 2018, Ensae

Venez aussi assister à une journée de préparation le 27 novembre ou le 4 décembre 2017 au cours de laquelle le sujet et les données une présentation du sujet, des données et des pistes techniques et une installation des outils informatiques vous permettront d'être au top pour le hackathon. Pour plus de détails, une note relative à ces journées est disponible sous le Github.

Le sujet :

Lors du recensement, les enquêtés indiquent où ils travaillent en déclarant une raison sociale, une adresse et une activité pour leur employeur, l'activité de cet établissement et l'adresse du lieu de travail. Identifier l'entreprise correspondante dans SIRENE à partir de ces informations déclarées n'est pas immédiat et représente un effort important.

L'Insee est responsable du **répertoire Sirene**. Ce répertoire est accessible en open data et contient au-delà de l'identifiant de nombreuses informations comme l'APE, la tranche d'effectif, la tranche de chiffre d'affaires, etc.

Dans de nombreux processus, il est nécessaire de pouvoir identifier des établissements ou des entreprises dans le répertoire Sirene (ou Sirus) à partir de variables identifiantes (n° Siren et Siret) ou quasi-identifiantes (raison sociale, adresse, etc.) afin de récupérer directement des informations plus précises (localisation, code d'activité dans la NAF). Dans notre cas, nous allons nous focaliser sur les **enquêtes annuelles de recensement**.

Pourtant, lorsqu'un individu répond à une enquête, ses déclarations peuvent être difficilement exploitables si elles ne sont pas standardisables. Face à ce problème récurrent, le hackathon vise à déterminer comment **identifier, pour les individus recensés, l'établissement employeur référencé dans Sirene à partir de trois variables d'identification non normalisées**. En effet, au cours du recensement, ces données sont librement renseignées par l'individu et potentiellement entachées d'erreur (orthographe, libellé flou ou imprécis, etc.) ou de différences de concept (activité perçue vs. APE de l'établissement dans Sirene) ce qui rend difficile un appariement automatique.

Trois variables pour l'identification :

- La raison sociale de l'entreprise
- L'adresse de l'entreprise
- L'activité de l'entreprise

À quelles techniques recourir ? 3 pistes proposées :

1 – Du ***text-mining*** afin d'évaluer la similarité entre la raison sociale déclarée et celle normalisée dans Sirene
2 – Des ***comparaisons géographiques*** à partir, notamment, de la base adresse nationale (BAN) afin de corriger l'adresse déclarée par un ménage au cours du recensement
3 – Du ***web-scraping*** afin de vérifier s'il est possible d'obtenir, malgré des erreurs ou approximations dans les libellés, une correspondance entre les deux bases de données à partir d'un « enrichissement » issu du web.

Ces pistes restent des propositions à partir desquelles trois groupes seront formés selon les appétences des participants. Si d'autres pistes veulent être explorées par les participants, les initiatives seront les bienvenues.

Un hackathon c'est quoi ?

C'est une séance de travail intense, ici sur 2 jours, et en équipe. Les participants cherchent à produire sur ce temps court un prototype opérationnel.

Le hackathon pour tous !

Pas besoin d'être un codeur fou pour participer : l'idée c'est d'apprendre et progresser ensemble. Le hackathon est ouvert à l'ensemble du SSP ainsi qu'à des organismes tels que Pôle Emploi, Etalab etc qui ont déjà signalé leur souhait de participer.

Comment ça se passe ?

Pendant deux jours, des équipes se forment pour proposer des solutions au sujet proposé. Les données sont mises à dispositions et les outils libres. Un planning des deux journées est en cours de préparation et sera par la suite diffusé.

Pour être fin prêt le jour J, on préparera l'environnement et débroussaillera le sujet lors des journées de préparation du 22 et 23 novembre 2017. Chaque participant pourra assister à une journée de préparation au cours de laquelle le matin sera consacré à des exposés techniques afin d'assurer une mise à niveau sur les différents modules nécessaires lors du hackathon (API Sirene, méthodes de text-mining, présentation des référentiels géographiques et principe du web-scraping). L'après-midi visera à gérer les problèmes matériels et les installations de logiciels qui seront nécessaires au bon déroulement des journées de janvier. Les journées de préparation seront un bon moyen, pour les participants, de découvrir le sujet et les personnes qui participeront mais elles nous permettront aussi de finaliser l'organisation et de détecter les éventuels problèmes.

À l'issue de ces journées, les présentations techniques seront mises à disposition sur Yammer et Github.

Pourquoi organiser un hackathon à l'INSEE ?

Cet événement est avant tout un outil d'**animation de réseau**. Dans l'optique de développer, entretenir et promouvoir le réseau de la datascience dans le SSP le hackathon doit permettre à ses membres de se rencontrer, d'échanger, d'apprendre, de partager leurs problématiques, de trouver une certaine stimulation et d'exercer leur créativité.
Par ailleurs, cette première version du hackathon reste à vocation opérationelle : le produit de ces deux jours, si les travaux ont été concluants, pourra déboucher sur un prototype d'amélioration de la chaîne de production du RP.

Pourquoi venir ?

- Le hackathon est aussi un mode de **formation** à des **outils de la datascience**. C'est pourquoi il est **ouvert à tous**, sans prérequis technique.

- C'est aussi l'occasion de développer des **modes de travail nouveaux, collaboratifs et créatifs**. Les participants travailleront en équipe, les compétences et les profils seront mélangés.

- ***Transposabilité du sujet.*** Le sujet d' « identification floue » est semblable à de nombreuses problématiques, de 'record linkage' par exemple. Une **note méthodo** devra en être issue pour faciliter l'utilisation sur d'autres sujets des méthodes explorées.

- ***Transposabilité des outils.*** Les **briques techniques** déployées sont utiles à de nombreux sujets : text mining, utilisation d'API, gestion de données géographiques, web scraping.

- ***Transposabilité de la démarche.*** Aujourd'hui le sujet du hackathon est une problématique INSEE, le prochain pourra porter sur un **sujet proposé par les participants**. La répétition de ces événements pourra constituer une **force de frappe mobilisable** facilement sur des sujets ouverts.

Quels sont les canaux de communication ?

- Sur Yammer
- Sur Github
- Par mail à info-hackathon@insee.fr