Insee
Mesurer pour comprendre

Big Data et Statistique Publique

Lettre d'information
N°7 – avril 2018

Coopérer pour innover, retour sur le hackathon les Champs de Sirene

À l'image d'Eurostat en mars 2017, l'Insee s'est lancé, à l'initiative d'une petite équipe de datascientists de l'informatique et de la direction de la méthodologie et de la coordination statistique et internationale (DMCSI), dans l'organisation d'un hackathon, ouvert au Service Statistique Public (SSP) et à ses partenaires proches. Pour cette première, le sujet retenu était directement en lien avec la production statistique. Les objectifs du hackathon étaient à la fois d'aboutir à un prototype expérimental permettant d'améliorer la chaîne de production du recensement mais aussi d'animer le réseau de la datascience au sein du SSP en développant des modes de travail innovants et décloisonnés. Et le résultat est plus qu'encourageant ! En trois demi-journées de travail intense prolongées par une courte séance de restitution, plusieurs pistes prometteuses ont été identifiées. Et le fort engagement des équipes dont certaines se sont constituées au début du hackathon sans se connaître au préalable, témoigne de l'appétence du SSP pour ces modes de travail collaboratifs. Il reste maintenant à prolonger l'essai ! Le SSP Lab, unité nouvelle de la DMCSI, qui vient d'être mise en place pour animer et impulser l'innovation statistique au sein du SSP en collaboration étroite avec l'unité innovation et stratégie du système d'information, va désormais reprendre le flambeau. Il s'agira d'approfondir les pistes techniques identifiées pendant le hackathon en lien avec les équipes du recensement de la population et du répertoire Sirene, pour aboutir, si les résultats sont concluants, à un prototype opérationnel. Il faudra aussi faire vivre dans la durée le réseau de datascientists du SSP, en poursuivant différents modes d'animation : sprints, hackathons, séminaires, formations… Pour en savoir plus, rendez-vous en juin pour le séminaire de lancement du SSP Lab !

Syvie Lagarde, Directrice de la DMCSI

Un hackathon ça ressemble à quoi ?

Un hackathon est une séance de travail d'un format particulier. Il s'agit de se consacrer, toutes affaires cessantes, à des travaux définis et circonscrits pendant un intervalle de temps restreint, avec une intensité certaine. Les participants sont réunis en petites équipes, autre particularité des hackathons, que ceux-ci soient compétitifs ou pas.

« Les champs de Sirene », première édition de hackathon organisé en interne à l'Insee a réuni, sans prérequis de niveau, plus de soixante personnes réparties en une douzaine d'équipes de cinq ou six participants les 18 et 19 janvier derniers. Organisé dans le but d'animer un réseau étendu d'acteurs et de curieux des datasciences au sein du SSP l'événement a rassemblé des agents de l'Insee, des SSM culture, éducation, travail, défense, environnement, agriculture et justice, et de partenaires tels que la Cnam, Etalab, Tracfin, la Banque de France ou Pôle Emploi.

Des journées de préparation en amont du hackathon ont permis aux participants de se familiariser avec le sujet, complexe, et la démarche encore inédite à l'Insee. Différentes présentations ont explicité la question posée par le Recensement de la population, présenté les bases de données fournies, et fait la démonstration de quelques outils et techniques (API Sirene, géocodage, webscraping, techniques de text mining…) que les participants pourraient utiliser. Lors des après-midi de ces journées, les participants ont mis ces techniques en pratique sur quelques cas de test pour arriver prêts au hackathon.

Les deux jours qu'a duré le hackathon « Les champs de Sirene » ont été un beau succès par l'implication et la volonté des équipes, à l'œuvre pour certaines jusque tard dans la soirée dans les anciens locaux de l'Ensae : l'intention d'en découdre avec les données du recensement et du répertoire Sirene était bien là. La restitution des solutions proposées devant le directeur général et un large public, puis une *poster session* ont consacré les efforts fournis par les équipes.

Les champs de Sirene pourquoi ? Mieux identifier l'employeur dans le recensement

Le hackathon « Les champs de Sirene » avait comme objectif d'améliorer le codage automatique de l'établissement employeur dans le recensement, à l'aide en particulier du répertoire Sirus ou de l'API Sirene. Cela pourrait, à terme, réduire la charge de reprise manuelle par les établissements régionaux.

Ce codage est notamment nécessaire pour déterminer le secteur d'activité de l'emploi exercé par les personnes recensées. Dans les questionnaires auto-administrés du recensement, les personnes déclarent le nom de l'établissement qui les emploie, l'activité de cet établissement, et l'adresse de leur lieu de travail. Pour autant, ces variables ne sont pas normalisées et peuvent comporter des erreurs (orthographe, libellé flou ou imprécis, inversion des champs). Elles peuvent aussi s'éloigner du concept recherché (activité perçue vs. APE de l'établissement dans Sirus, confusion entre l'entreprise et l'établissement...). Elles doivent donc être confrontées à des informations externes (répertoire Sirus en particulier) pour identifier précisément l'employeur. Une première identification de l'établissement employeur s'effectue de manière automatique mais elle n'aboutit que dans 45 % des cas. Il est donc nécessaire de recourir à une reprise manuelle qui sollicite, pour toutes les variables à coder (établissement, activité, profession) environ 70 personnes en établissements régionaux pendant 5 mois chaque année.

Un hackathon, pour quels résultats ? Les solutions proposées par les équipes

Si le temps de l'épreuve était assez bref, un jour et demi seulement pour coder, et les données à la fois complexes et volumineuses, les différentes équipes (BadQOP, Eklekgeek, Elastic, GeoCodeurs, Iemanja, Illumines, Noname, PoleEmploi, Qokka, Rmax, Scraping19 et Scrapules) ont su proposer autant de solutions originales qui tracent des pistes et une série de briques pouvant inspirer le développement futur d'un prototype plus complet.

Plusieurs équipes ont proposé de procéder par étapes en cherchant d'abord une liste d'échos possibles de Siret à travers un moteur de recherche, qu'il s'agisse de l'API Sirene dont une version avancée a pu être déployée en test, ou bien en s'appuyant sur les barres de recherche de sites tels que societe.com et en scrapant[1] les résultats, ou même de moteurs de recherche web renvoyant ensuite vers une page d'un tel site. Ensuite, plusieurs critères étaient combinés afin de sélectionner le meilleur écho possible : la proximité dans les champs textuels entre raison sociale déclarée et enregistrée dans Sirene, la distance géographique à partir du géocodage des établissements. En effet, des coordonnées

1 Cf lettre Big Data n5 sur le scraping

géographiques avaient été ajoutées aux données mises à disposition grâce à l'application Geoloc. Certaines équipes ont également proposé d'autres enrichissements de données en récupérant sur internet une raison sociale normalisée par le moteur des pages jaunes par exemple, en affinant la géolocalisation avec la BANO (Base Adresse Nationale d'OpenStreetMap) ou en ajoutant l'adresse proposée par Mappy.com.

Une autre piste explorée par plusieurs équipes a été de réduire préalablement le nombre d'établissements candidats sur la base d'une proximité géographique et/ou d'activité, puis de choisir, dans cet ensemble, l'établissement ayant la raison sociale la plus proche au sens d'une distance adaptée (Levenshtein[2] par exemple). Ce choix pouvait être raffiné, une autre équipe a ainsi réalisé une analyse en composantes principales (ACP) s'appuyant sur trois variables synthétiques que sont le rang d'apparition dans le moteur de recherche, la distance textuelle entre les variables de Sirus et celles du recensement et la distance géographique. La position sur le premier axe de l'ACP fournit ainsi un indicateur de qualité de l'identification établie. D'autres équipes ont suivi des voies plus spécifiques : construire un pipeline capable d'accueillir différents modules d'enrichissement, de requêtage des API et d'évaluation des résultats, ou encore concentrer les efforts sur l'emploi public, pour lequel le taux d'identification automatique est actuellement le plus faible. Cette dernière équipe a ainsi proposé un prétraitement spécifique des informations déclarées lors du recensement. Enfin, une autre approche a consisté, de manière originale, à rechercher le code NAF se rapportant à l'activité déclarée plutôt que l'établissement exact par projection des termes déclarés dans l'espace du vocabulaire de la nomenclature.

Au fil de leurs avancées, les équipes pouvaient jauger de leur progression grâce à un module d'évaluation mis à leur disposition qui indiquait leur taux de codage et le taux de concordance entre leurs résultats et ceux de l'application de mise en concordance automatique, aujourd'hui utilisée. Les performances étaient comparées à la fois sur l'enquête de 2017 et sur un échantillon de 2014 passé par une double phase de reprise manuelle (opération Récap Qualité) dans le but d'évaluer la qualité de cette reprise. Afin de stimuler positivement les équipes, ce module d'évaluation leur permettait aussi de situer les performances de leur approche à celles des autres participants.

Pour en savoir plus : le repo github du hackathon

On trouvera sur ce site à la fois toutes les ressources proposées aux équipes et les codes et présentations qu'elles ont produits.

En direct de l'équipe Ekklegeek ! Le témoignage de Rémi Pépin, développeur au centre national d'informatique d'Orléans

Les 18 et 19 janvier 2018 a eu lieu le hackathon des «Champs du Sirene » auquel j'ai participé. Le sujet était d'automatiser la codification du Siret de l'employeur des personnes recensées à partir des informations déclarées. C'était mon premier hackathon, et pour être honnête, j'y allais pour participer, plus que pour performer. Le hackathon se composait d'une journée de préparation pour nous présenter le sujet et commencer à former les équipes, 2 jours de travail et 2 heures de présentation des résultats des équipes. Pour titiller notre esprit compétitif tout en conservant une bonne ambiance, nous disposions d'une petite appli web pour voir les résultats des équipes en direct. Sans être trop présente, elle permettait de suivre l'avancée de tout le monde et se motiver à mieux faire. Mon équipe se composait de 4 statisticiens et 2 informaticiens, et à cause de la diversité de nos spécialités nous nous sommes baptisés les « Ekklegeek ». Pour poursuivre dans cette diversité, nous avons décidé d'explorer les 3 pistes proposées par les organisateurs. À savoir distance textuelle entre déclaration et base Sirene, distance géographique entre adresse déclarée et base Sirene, et webscraping.Tout devait bien se passer, mais ce ne fût pas le cas. Pour faire simple, nous ne savions pas coder en Python, alors que nous avions fait le choix de cette technologie, et nos ordinateurs manquaient de puissance. En plus, au cours de la nuit entre les deux jours mon ordinateur a été mis en veille alors qu'un script devait tourner toute la nuit. Le lendemain, nous avons dû improviser.

C'était donc deux journées laborieuses où rien ne se passait comme prévu, mais riches d'enseignements : apprendre à toujours aller de l'avant et à trouver des solutions rapidement. La durée courte d'un hackathon permet d'être constamment sur la brèche (bon par contre vous allez en sortir épuisés), et en groupe, les idées fusent rapidement. Comme il faut se concentrer sur le fonctionnel, on arrive vite à produire quelque chose. Et c'est très valorisant.

Un hackathon est également un formidable outil d'apprentissage : je n'ai jamais autant appris que pendant ces deux jours. Aussi bien en technique, qu'en organisation.

« Pour moi, c'était une formidable expérience ! Nous sommes arrivés à réaliser une solution qui produit de bons résultats ». Et même si elle n'est peut-être pas utilisable à grande échelle, nous étions très contents d'être arrivés à une solution fonctionnelle.

Pour conclure, je pense que tout le monde a sa place dans un hackathon, et qu'il ne faut pas se brider. Chacun peut apporter quelque chose à son groupe : des idées, de la technicité, de l'organisationnel, etc. Le seul réel pré-requis est l'envie de faire quelque chose. Donc si un prochain hackathon est organisé par l'Insee, n'hésitez pas à y participer !

Actualités / Brèves

- L'expérimentation continue : suite à ce hackathon, un groupe de travail va être formé afin d'explorer les pistes et d'aboutir à un prototype expérimental.
Si vous souhaitez participer à ce travail collaboratif, n'hésitez pas à nous contacter à info-hackathon@insee.fr !

- Le **SSP Lab** se crée à partir de la division MAEE, plus d'informations sur les objectifs, collaborations, modes de travail de cette nouvelle unité lors du **séminaire de lancement le 5 juin à 14h en salle Malinvaud à l'Insee**

- Prochaines **formations** : « **analyse textuelle** » le 25 juin
« **machine learning** » les 10 et 11 septembre

- Les 14 et 15 mai prochains, la **Big Data for European Statistics conference**, dressera le bilan de **l'ESSnet BigData** : webscraping des offres d'emploi, des caractéristiques des entreprises, compteurs communicants, données SIA (système d'identification automatique), téléphonie mobile, estimations précoces, plus d'informations sur
https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata

Ont participé à ce numéro Yves-Laurent Bénichou, Elise Coudin, Julie Djiriguian, Pauline Givord, Sylvie Lagarde, Rémi Pépin, Benjamin Sakarovitch et Lucile Vanotti.

Cette lettre est une occasion d'informer largement et d'échanger. N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à julie.diriguian@insee.fr ***et*** benjamin.sakarovitch@insee.fr
Les archives de cette lettre sont disponibles sur :
http://www.agora.insee.fr/jahia/Jahia/site/dmcsi/SiteDMCSI/DMSaccueil/*DMAEEaccueil*/BigData
Demande d'inscription individuelle à la lettre : dg75-l101@insee.fr

2 https://fr.wikipedia.org/wiki/Distance_de_Levenshtein