Hackathon Insee
Les champs de Sirene

Un hackathon expliqué avec Star Wars

Il y a pas si longtemps, dans une direction générale pas si lointaine ...

HACKATON INSEE

Episode I : La menace fantôme

- 1 journée de préparation
  - Sujet
  - Pistes
  - Configuration des postes
  - Equipes
- Ouverture de Hackathon
- Actuellement, on sous estime un peu la difficulté du sujet

Une équipe hétéroclite

Christophe Alviset (DSE)

Maxime Bergeat (DARES)

Christophe Michel (DARES)

Gwendoline Volat (Culture)

Elise Hamelin (CNIN)

Rémi Pépin (CNIO)

Episode 2 : La guerre des clones

Division de l’équipe en 3 groupes de 2

- Team web scraping
- Team textmining
- Team géographie

Exploration de solutions mais sans vision global

Episode 3 : La revanche des Siths

Premières grosses désillusions

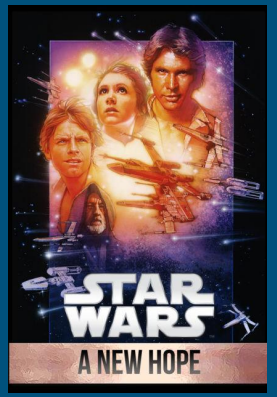
- Connaissance trop légère du langage Python (perte de temps considérable)
- Problème pour scraper
- Problème pour réaliser la distance textuelle
- Proxy Insee
- Fichier trop gros (ou machine trop faible)

Réalisation de la complexité du sujet

- Une application s’occupe déjà du recodage
- Comment nous sans expérience, on va faire mieux en 2 jours ?

L’absence de résultat est quand même un résultat

Episode 4 : Un nouvel espoir



Solutions trouvées pour certains problèmes

- Ralentissement du scraping pour ne pas se faire détecter comme robot
- Profiter de la nuit pour coder un fichier de 10 000 lignes
- Bases tronquées avec Excel
- Solution globale trouvé

Utiliser le scraping pour coder les numéros siret, et utiliser position de la recherche sur la page et la distance géographique et textuelle comme critère de qualité du codage

Episode 5 : L’empire contre attaque

- Ordinateur mise en veille pendant la nuit (350 lignes codé sur les 10 000)
- Distance textuelle pas encore au point
- Plus qu’une demi journée pour terminer (gros pic de stress)

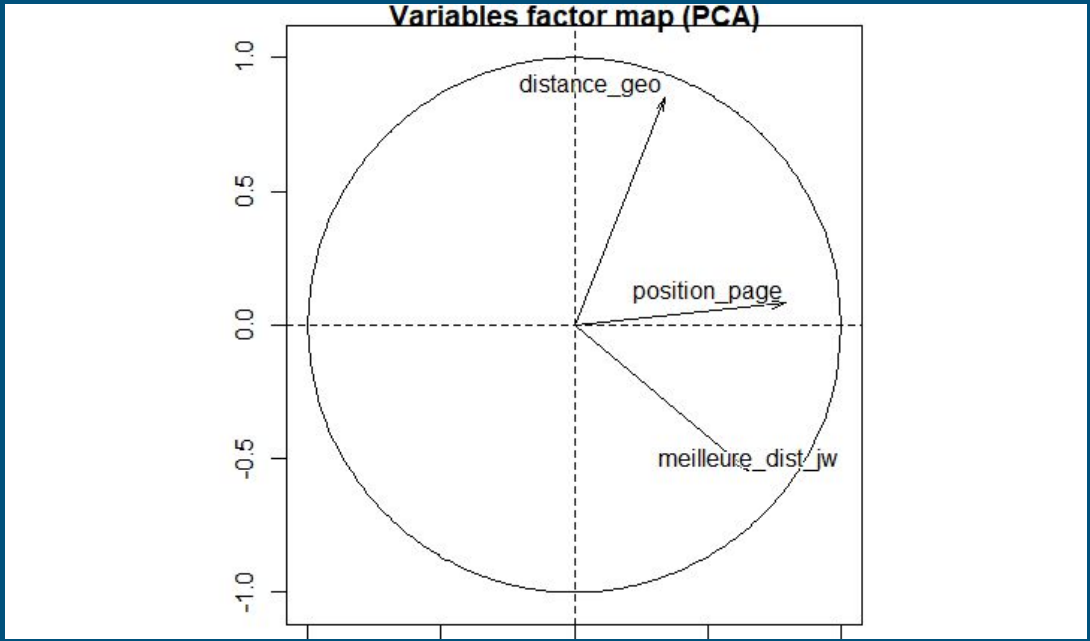
Episode 6 : Le retour du jedi

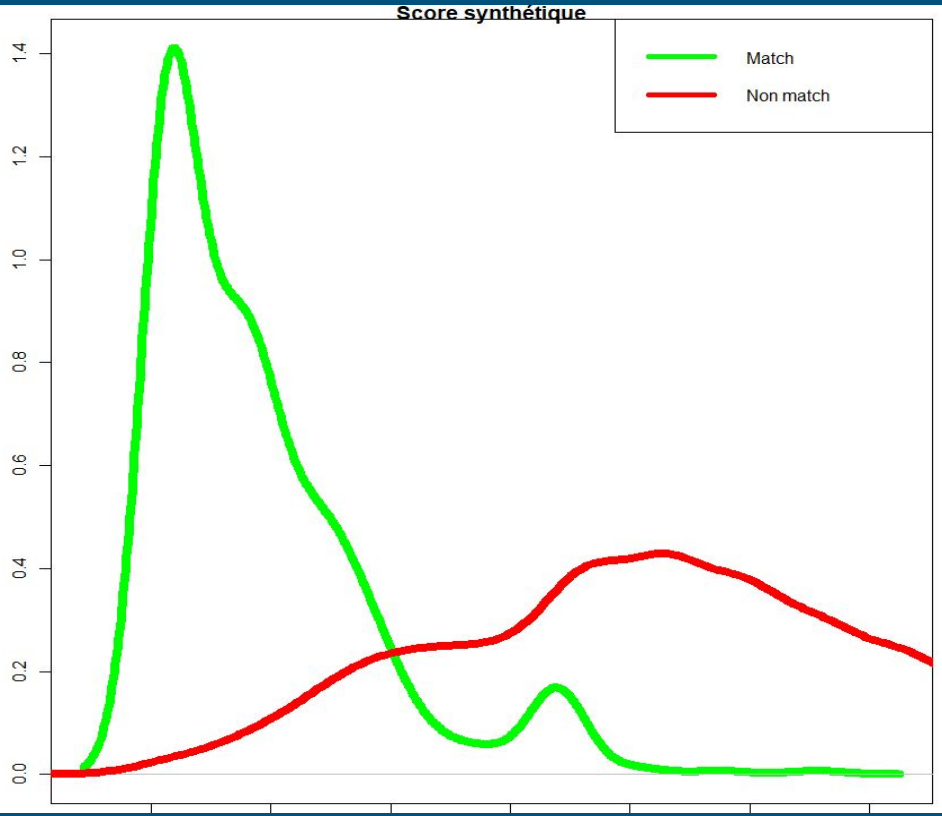
- Parallélisation du processus de scraping avec plusieurs sources
- Distance géographique ok
- Distance textuelle ok

Il ne reste plus qu’à faire tourner les distances sur notre fichier final

Episode 7 : Le réveil de la Force

Réalisation de l’acp




Score synthétique
Match
Non match
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Nos résultats

Méthode	Concordance
1er résultat scraping	43%
Scraping + matching raison sociale	45%
Scraping + géographie	35%
Scraping + indicateur synthétique (ACP)	47%
Scraping + contrôle qualité + indicateur synthétique (ACP)	79%

Pour résumer

Web-scraping

- Échantillon de 2000 individus aléatoires.
- Interrogation moteurs de recherche (Qwant, Google, Bing)
- Limite aux dix premiers résultats sur le site societe.com.

Calculs de distance

- Appariement avec fichier SIRUS.
- Distance “textuelle” entre 5 variables du fichier Sirus et RS de recensement.
- Distance euclidienne entre les x,y “RP” et ceux de Sirus.

Mesure de la qualité

ACP sur les variables pour avoir un “score” synthétique de la qualité: choix entre les plusieurs SIRET.

Ressentis

- “Montagne russe” émotionnelle
- Oblige à aller chercher le meilleur de soi
- Toujours réagir au changement
- Progression considérable
- Gestion du temps importante
- On se concentre sur les fonctionnalités

Hackathon = session de travail intensive ludique