

ECAM - Institut Supérieur Industriel

Rapport de stage

Stage chez EDITx

Année académique : 2018-2019
17 septembre - 26 octobre 2018

Nom et prénom de l'étudiant :.....Kessels Julien (16214)
Email :.....julien.kessels@hotmail.com
Section :.....5MIN

Nom et prénom du maître de stage :.....Combéfis Sébastien
Email :.....scombefis@editx.eu

Nom et prénom du superviseur :.....Taymans Rémy
Email :.....r.taymans@ecam.be

Lieu du stage :.....Etterbeek, Rue des Francs, 79

The logo for EDITx, featuring the word "EDIT" in blue and "x" in grey.

Table des matières

1	Prélude	3
2	Présentation de l'entreprise	3
3	Buts et objectifs du stage	5
4	Projets	5
5	Phase 0 - Etude	6
6	Projet 1 - Tutoriel intelligent	8
6.1	Idée	8
6.2	Analyse de la DB	8
6.3	Algorithme & Migration vers Neo4j	9
6.4	Application Web	11
6.5	Déploiement sur Docker	12
6.6	Pour après	12
7	Projet 2 - Question Matcher	13
7.1	Idée	13
7.2	Outils	13
7.3	Application	16
7.4	Pour plus tard	16
8	Autres Projet	17
8.1	Design & Web Development	17
8.2	School Scraper	18
8.3	Concours en Allemagne	19
9	Conclusion	20
10	Bibliographie	21
11	Annexes	21

Avant-propos

J'aimerais avant tout remercier toute l'équipe d'EDITX pour son accueil chaleureux, les nombreux conseils et son souhait de me montrer les divers aspects de leur travail.

1 Prélude

Pour obtenir son diplôme de Master en ingénierie à l'ECAM, tout étudiant doit réaliser un stage d'une durée de six semaines en entreprise. Ce stage d'immersion vise à mettre l'étudiant en situation et à le faire participer aux projets de l'entreprise dans laquelle il réalise son stage. Idéalement, le lieu et la thématique du stage servent comme introduction au travail de fin d'étude. Ayant déjà acquis une bonne connaissance du monde de l'applicatif et du Web de part mes études et mes projets personnels, je me suis penché vers une branche de l'ingénierie en informatique qui m'intéresse particulièrement en ce moment : Le Data Science.

Mon professeur monsieur Combéfis m'a parlé d'une opportunité de faire du data science chez EDITx qui depuis quelques années récolte des données à travers ses concours informatiques.

2 Présentation de l'entreprise

EDITx est une plateforme en ligne sur laquelle un utilisateur peut s'inscrire et participer à des concours informatiques ou bien s'entraîner sur des petits Quiz de quelques questions.

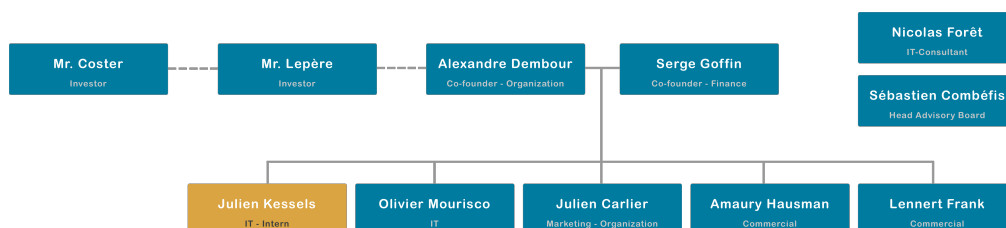
EDITx vise deux types de personnes différents :

- Les étudiants et professionnels qui peuvent tester leurs skills et gagner des prix à travers des concours ludiques.
- Les éducateurs, dans le but de contribuer à la plateforme et élargir le nombre de questions proposées par le site. Ainsi chaque contributeur a accès à des statistiques à propos des questions qu'il a mis en ligne.

Organigramme

EDITx (SPRL) est une société informatique créée en 2016 avec un capital de 18600€ subdivisé en 100 parts réparties de manière égale (50/50) à Serge Goffin, directeur financier et Alexandre Dembour, directeur opérationnel.

Depuis décembre 2017, EDITx a augmenté son capital et son nombre d'actions. Ces actions ont été réparties entre deux actionnaires, M. Coster et M. Lepère ainsi que M. Forêt, consultant IT et M. Combéfis, nommé Head of Advisory Board.



Produit

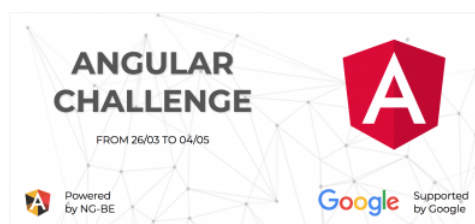
Les challenges organisés par EDITx sont toujours sponsorisés par des entreprises. En plus de la visibilité qu'apporte le challenge en terme de publicité, l'entreprise montre son intérêt pour une certaine technologie comme par exemple Angular pour le Challenge sponsorisé par Google ou CSharp pour le challenge organisé par Microsoft.

De plus, à la fin du concours, l'entreprise a accès aux résultats des différents participants et leur permet d'envoyer des offres de recrutement auprès des candidats qui leur semblent intéressants.

L'organisation de A-Z d'un concours dure environ 15 semaines. Dans un premier temps, entreprise et EDITx se mettent d'accord sur la technologie du challenge, pour qu'EDITx puisse ensuite récolter une liste de questions fournies par des professionnels ou professeurs dont certains pourront également faire parti du jury.

Pendant les prochaines 4 à 6 semaines, les utilisateurs vont pouvoir participer au concours en ligne et tenter d'avoir le meilleur score. Les règles sont simples : 15 minutes, 15 questions à choix multiple, 3 jokers (cinquante cinquante, choix du public, call a friend).

A la fin de cette période de concours en ligne, le sponsor organise une finale avec les 10 à 20 meilleurs participants, mais cette fois sur papier. Les meilleurs de la finale peuvent gagner des prix allant jusqu'à 1500€.



3 Buts et objectifs du stage

Les buts de ce stage s'étalent sur deux dimensions. Premièrement un but personnel qui consiste à me former sur la théorie et les outils de machine learning en vue de la réalisation de mon TFE, deuxièmement à intégrer une équipe et d'utiliser mes connaissances en informatique pour réaliser des outils pour améliorer la plateforme d'EDITx.

- Self learning (machine learning)
- Tutoriel intelligent V0
- Question Matcher V0
- Participation active aux projets d'EDITx

4 Projets

Le business modèle d'EDITx se positionne sur deux axes : D'une part les étudiants et professionnels participant aux concours et quiz, et d'autre part les entreprises qui sponsorisent les challenge en échange de repérer d'éventuels talents.

Etant deux stagiaires sur ce projet, nous nous sommes chacun penchés sur une des deux axes. Amaury tentera de créer des badges de compétence que les entreprises pourront consulter après leur challenge.

De mon côté je me focalise sur les questions des quiz de la plateforme d'EDITx. Il s'agira d'une part d'ajouter de nouveaux modules pour booster l'intérêt de la plateforme pour les utilisateurs et en attirer des nouveaux et d'autre part de faciliter l'interaction entre contributeurs et la plateforme lors de la création de questions.

5 Phase 0 - Etude

En 5e année, le stage a pour but de préparer l'étudiant à son travail de fin d'étude. Ayant choisi un sujet qui me passionnait, mais dont je n'avais pas beaucoup de connaissances, j'ai passé une grande partie du stage à m'instruire. Pour pouvoir appliquer les modèles et algorithmes de machine learning il faut aussi bien connaître la théorie derrière ces modèles qu'avoir un sens de où et quand appliquer quel modèle. Est-ce qu'on a à faire avec des données "labeled" ou "non-labeled" ? Dans un cas, il faudra utiliser un algorithme supervisé et dans l'autre cas un algorithme non supervisé comme on le verra plus tard dans le rapport.

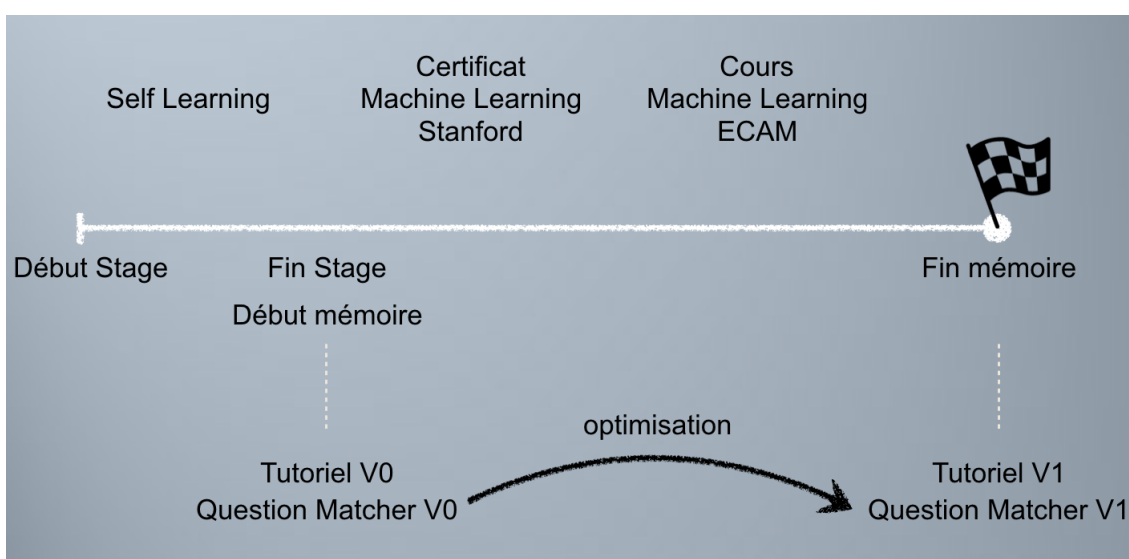


FIGURE 1 – Timeline

Ressources

- **Machine Learning from Scratch** de O'Reilly ¹
- Cours & Conseils de Nicolas (Ingénieur en informatique UCL ; EDITx)
- 18 heures Machine Learning Course - youtube (CALTECH) ²
- **Inscription et participation aux 12 semaines de cours de ML en ligne (Coursera) organisé par l'université de Stanford** ³

1. <http://shop.oreilly.com/product/0636920033400.do>
2. <https://www.youtube.com/watch?v=mbyG85GZ0PI&list=PLD63A284B7615313Aindex=1>
3. <https://www.coursera.org/learn/machine-learning/home/welcome>

Python ML libraries & Preliminary Stats

Avant de me lancer dans les projets je me suis un peu plongé dans les données d'EDITx et généré quelques statistiques et graphes pour avoir un aperçu des données. Sans trop de surprise, on voit que la plupart des utilisateurs sont des hommes.

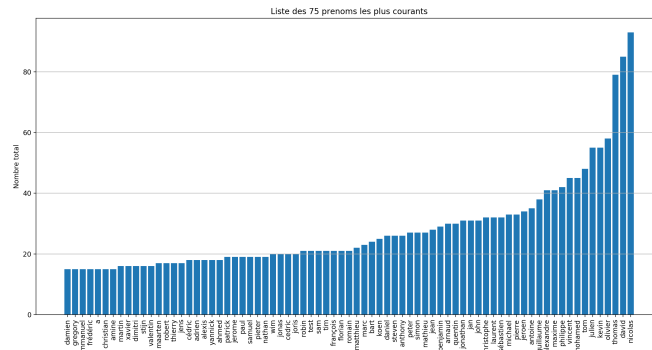


FIGURE 2 – Répartition des noms

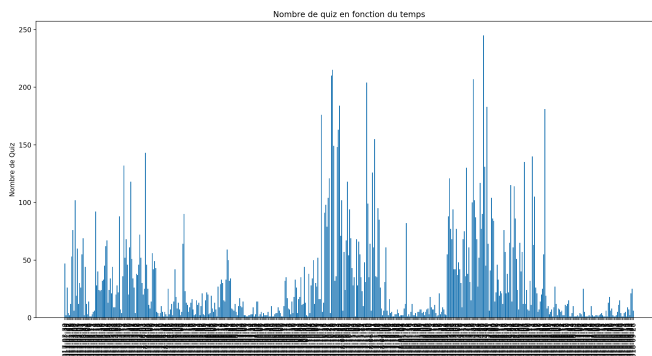


FIGURE 3 – Nombre de Quiz % temps

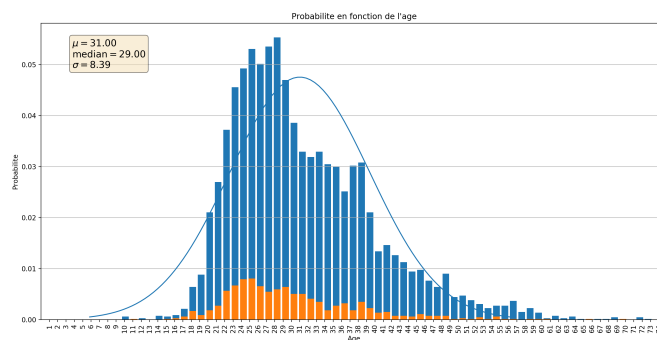


FIGURE 4 – Probabilité âge - (M, bleu /F, orange)

6 Projet 1 - Tutoriel intelligent

6.1 Idée

Une application intéressante serait d'étudier les questions auxquelles les utilisateurs de la plateforme répondent. Le but est de créer un tutoriel proposant une suite de questions adaptées au niveau et intérêts de chaque utilisateur pour essayer de le faire progresser.

6.2 Analyse de la DB

Première étape importante est l'analyse des données existantes. Nicolas Forêt, l'informaticien consultant chez EDITx nous a envoyé un dump de la base de donnée. Très vite nous nous sommes rendu compte de la complexité de la base de donnée. En effet, le site web d'EDITx a été créé avec Drupal, un CMS (content management system) largement répandu. Comme le nom l'indique, Drupal gère lui-même le backend et la base de donnée. Un système bien réfléchi mais pas très lisible (1000 tables) et compréhensible à l'œil humain.

Il était absolument primordial de se faire une fiche récapitulative des tables intéressantes et de faire le lien entre les différentes tables.



Il est évident que pour certaines applications, ce genre de base de donnée n'est pas très efficace.

6.3 Algorithme & Migration vers Neo4j

Il existe plusieurs façons d'implémenter un système de tutoriel.

On peut d'ailleurs comparer ce système à un système de suggestion comme la suggestion d'articles sur Amazon. Et pour ces applications, on peut appliquer des algorithmes de machine learning comme les neural networks. Malheureusement je n'ai pas encore le bagage pour mettre en oeuvre de tels algorithmes, mais en regardant le problème sous un autre angle, on peut s'imaginer un autre système pour mettre en place un système de suggestion de questions.

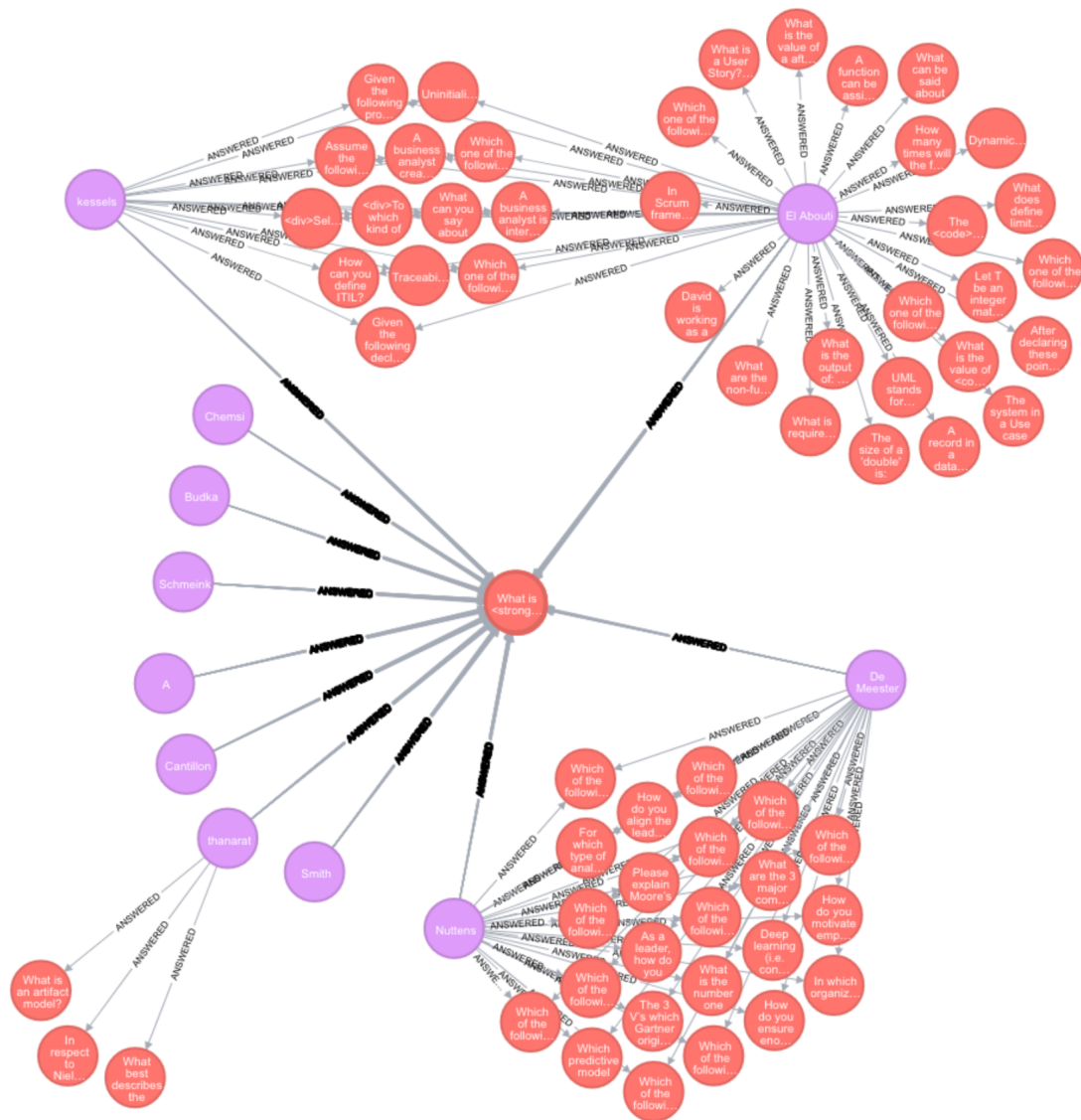
Si on représente les questions et les utilisateurs comme des entités, on peut créer des relations entre ces entités : Un utilisateur peut avoir répondu ou non à une question. Ce genre de représentation est celui de bases de données graphes comme Neo4j.

En effet, le point fort de neo4j est de traverser rapidement des graphes. Pour réaliser des requêtes de ce type en relationnel, il faudrait faire des requêtes avec beaucoup de jointures entre les tables mais le monteur n'est pas assez performant le faire en pratique. Selon Stackoverflow⁴, pour une requête avec 12 jointures, l'ordre qu'on appelle encore "Bushy tree" peut monter jusqu'à 28,158,588,057,600. À priori, le seul facteur limitant pour le nombre de jointures sont les ressources disponibles du serveur(mémoire) mais on voit bien qu'on monte très vite à des valeurs pharamineuses.

Tables	Left-Deep Trees	Bushy Trees
1	1	1
2	2	2
3	6	12
4	24	120
5	120	1,680
6	720	30,240
7	5,040	665,280
8	40,320	17,297,280
9	362,880	518,918,400
10	3,628,800	17,643,225,600
11	39,916,800	670,442,572,800
12	479,001,600	28,158,588,057,600

4. <https://stackoverflow.com/questions/2558667/what-is-the-maximum-number-of-joins-allowed-in-sql-server-2008>

La procédure de migration de la base de donnée relationnelle gérée par Drupal vers la nouvelle base de donnée Neo4j est un travail fastidieux mais nécessaire pour notre application.



Les noeuds représentés en rose sont les utilisateurs et ceux en rouge les questions. L'arête entre les deux correspond au fait qu'un utilisateur a répondu à une question ou pas.

6.4 Application Web

Le critère du choix des questions proposées à l'utilisateur pour le tutoriel est la **ressemblance entre l'utilisateur et un autre utilisateur**.

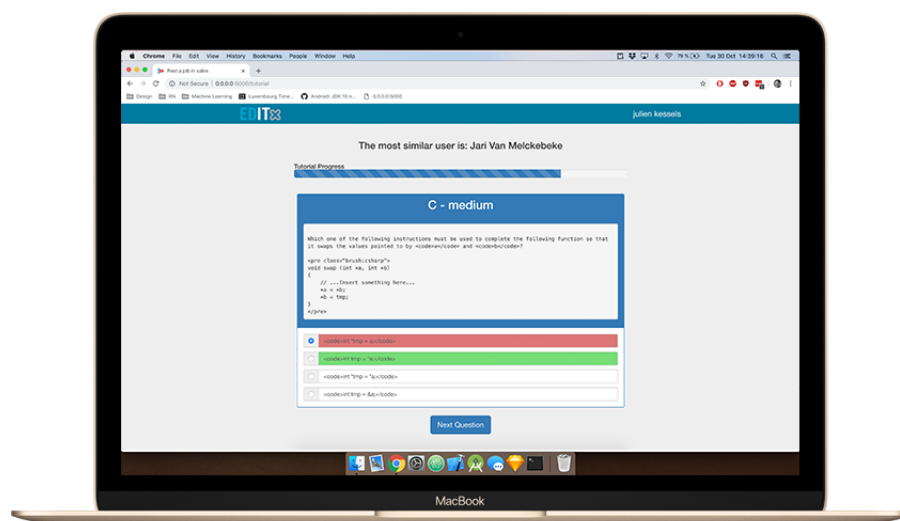
Voici les étapes suivies pour le choix des questions posées pour un certain utilisateur.

- Récupération des questions réalisées par l'utilisateur.
- Récupération des utilisateurs qui ont répondu aux questions ci-dessus

En neo4j, ces deux étapes correspondent simplement à la requête :

```
(me:User {id: myID})-(:HAS_ANSWERED)->(questions:Question) RETURN questions.id
(q:Question {id: question_id})<-(:HAS_ANSWERED)-(other:User) RETURN other.id
```

De ces requêtes on peut facilement retrouver l'utilisateur qui ressemble le plus à la personne suivant le tutoriel. Nous verrons après que dire qu'un utilisateur est similaire à un autre juste par le fait que ces deux personnes ont le plus de questions répondues en commun n'est pas la méthode la plus sophistiquée.



Après avoir récupéré les différentes réponses de chacune des 20 questions du tutoriel, l'application Web présente les questions une à une. En cas d'une mauvaise réponse, l'application met en surbrillance la bonne réponse et passe à la prochaine question.

6.5 Déploiement sur Docker

Docker est un logiciel permettant de facilement et rapidement déployer une application sur un serveur qui ne tourne pas forcément sur le même OS. Son système de conteneurisation au lieu de virtualisation est une forme plus légère qui s'appuie sur certaines parties de la machine hôte pour son fonctionnement. Cette approche permet d'accroître la flexibilité et la portabilité d'exécution d'une application, laquelle va pouvoir tourner de façon fiable et prédictible sur une grande variété de machines hôtes.

Les deux fichiers importants de Docker sont **Dockerfile** et **docker-compose.yml**. Docker-compose.yml est un fichier de configuration qui contient une liste de tous les conteneurs à monter sur le serveur ainsi que de les lier entre eux. On retrouve ici par exemple les conteneurs comme mariadb et neo4j à installer.

Le Dockerfile est un fichier qui contient une liste de commandes à exécuter lors de la création de l'image. Dans ces commandes à exécuter on a par exemple les bibliothèques à installer avec apt-get.



6.6 Pour après

Le tutoriel V0 donne déjà un résultat convaincant, mais il y a moyen de pousser la recherche bien plus loin. L'indice de similitude entre utilisateurs actuel ne se repose que sur le fait que deux utilisateurs ont répondu à un maximum de mêmes questions. Mais cela n'est pas très précis. Pour le mémoire il y a quelques caractéristiques qu'il faudrait absolument ajouter aux critères de similitude :

- niveau de l'utilisateur
- centres d'intérêts
- difficulté des questions
- ...

7 Projet 2 - Question Matcher

M'occupant principalement au sujet des questions j'ai souhaité créer un outil pour faciliter la création de nouvelles questions par les contributeurs.

7.1 Idée

L'idée est de créer un classifieur de questions pour avoir une librairie structurée et pouvoir retrouver des questions facilement.

Une manière de faire est d'attribuer des tags à chaque question, comme un tag de difficulté, un ou plusieurs tags pour les skills d'une question et un tag pour le langage de la question (Python, JAVA, C++, ..).

C'est le dernier point que j'ai choisi pour commencer. Comment automatiquement attribuer un tag de langage pour une nouvelle question ?

7.2 Outils

Il y a beaucoup de manières différentes pour attaquer le problème. Une manière très connue est l'utilisations d'algorithmes NLP (Natural Language Processing) qui analysent la structure d'une phrase, des suites de mots ou d'autres caractéristiques pour créer et/ou déchiffrer une phrase le plus humainement possible.

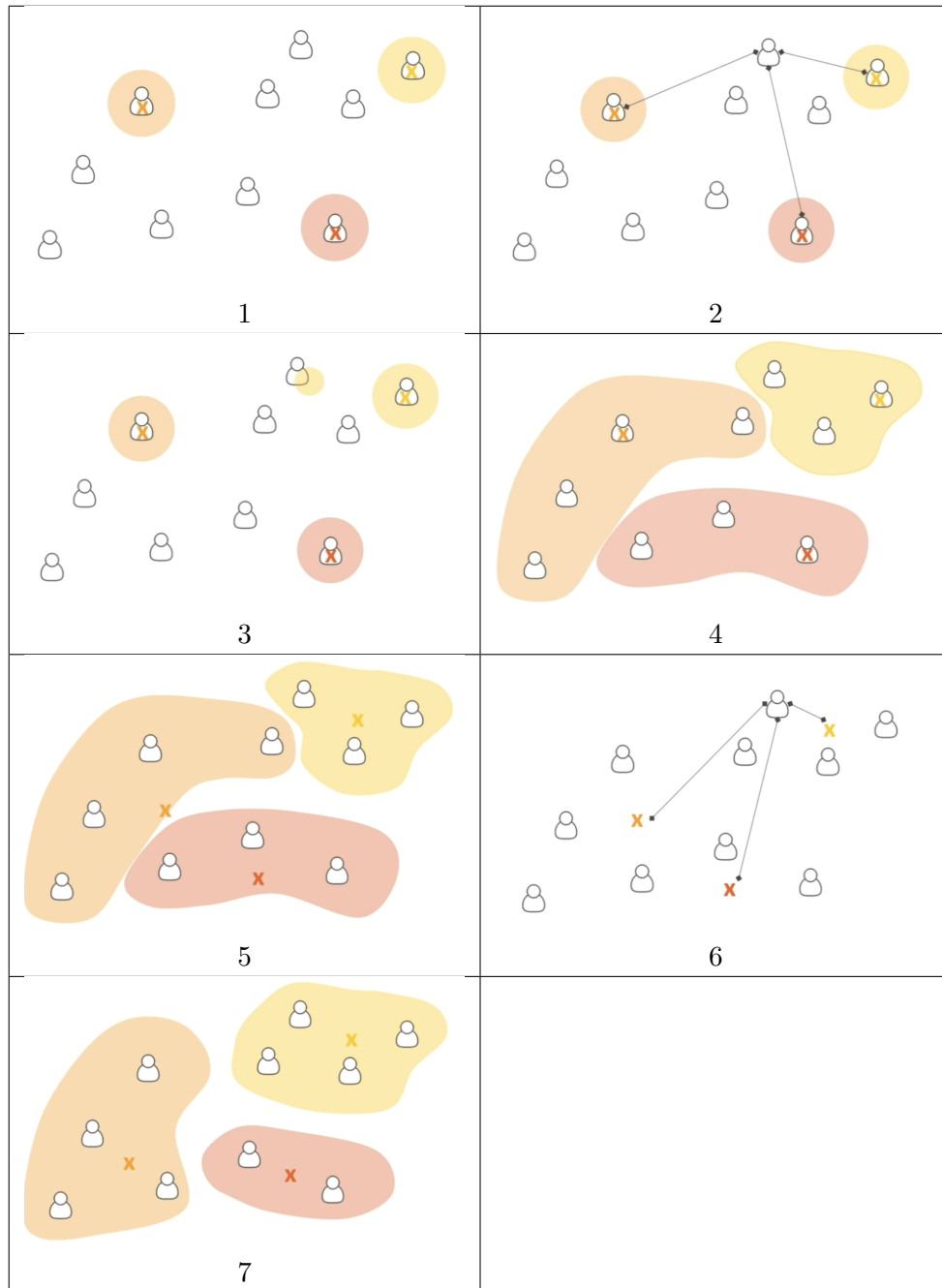
Dans mon application j'ai utilisé un algorithme de clustering très connu dans le monde du machine learning : **K-Means**.

K-Means

K-Means est un algorithme de clustering non supervisé c'est à dire qu'il s'applique sur des données dites "non-labeled" et permet dans notre cas de former des groupes de questions qui se ressemblent. Chaque groupe ainsi constitué représentera un langage spécifique. Ceci est exactement ce qu'il nous faut quand un contributeur ajoute une nouvelle question. L'algorithme attribuera automatiquement un tag de langage à la question créée.

L'algorithme

L'algorithme K-Means est illustré ci-dessous :



1. Choix de N centres initiaux

On commence par choisir au hasard N (ici N=3) centres appelés centroïdes. Les centroïdes se dénotent c_1, c_2, \dots, c_k . On a :

$$C = c_1, c_2, \dots, c_k$$

avec C, le set de tous les centroïdes.

2-4. Première catégorisation

Ensuite on assigne les entrées au centre le plus proche en calculant la distance euclidienne entre les points et chaque centre.

$$\arg \min_{c_i \in C} \text{dist}(c_i, x)^2$$

avec $\text{dist}(\cdot)$, la distance euclidienne.

5. Recalcul des centres

Ensuite on recalcule les nouveaux centres qui correspondent au centre de gravité pour chaque set.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

avec S_i le set de tous les points attribués au centroïde c_i .

6-7. Recatégorisation

On recommence l'étape 2-5 jusqu'à ce que plus rien ne bouge. Nous avons alors trouvé les centres optimaux et catégorisé les entrées selon ces centres.

7.3 Application

L'algorithme K-Means est implémenté dans scikit-learn, une librairie Python qui regroupe une multitude d'algorithmes de machine learning.

Les entrées de l'algorithme sont un dump d'environ 300 questions des langages : "Bebras", "Java", "Python", "PHP", ".NET", "UX/Usability".

295 de ces questions vont faire parti du learning set, donc les données pour entraîner le modèle. Les 5 autres questions sont une question de chaque langage pour tester notre modèle.

Après avoir configuré le modèle avec le nombre d'itérations (1000) et le nombre de clusters(5), on lance l'entraînement et on obtient une liste de 5 clusters générés avec les mots représentatifs des différents clusters comme on peut le voir ci-dessous.

```
Top terms per cluster:
Cluster 0: console writeline code output foo following var println int new 10 lt exception bar add echo null php write class
Cluster 1: print execution data following does program code int def instruction len hold range variables lambda values length value
Cluster 2: following method class net true use correct used value type string does statement statements int code een object new
Cluster 3: public int string static main void class println args new code return output following test hello program result private
Cluster 4: le la les et une il est des en dans castor sur pour par que au deux se castors sont
Cluster 5: gt lt list integer stream string public print final arrays aslist map add filter does new static args main foreach

Prediction
Question is: Bebras. Would put in cluster [4]
Question is: Java. Would put in cluster [5]
Question is: Python. Would put in cluster [1]
Question is: PHP. Would put in cluster [0]
Question is: .NET. Would put in cluster [3]
Question is: UX/Usability. Would put in cluster [2]
```

On peut ensuite tester notre modèle avec les 5 questions restantes et les résultats sont concluants comme on peut le voir.

7.4 Pour plus tard

Le Question Matcher V0 est très puissant et la précision surpasse ce à quoi je m'attendais. Hors pour l'instant, au lieu de décomposer la question en une série de tags, le programme ne donne que le langage. Il reste encore du chemin à faire pour le mémoire. Dans la catégorie d'analyse de phrases, il y a quelques algorithmes connus que j'aimerais bien étudier et appliquer pour le mémoire.

- Latent Dirichlet allocation
- Bag of Words
- Word2Vec

8 Autres Projet

Lors du stage chez EDITX, j'ai également eu l'opportunité d'appliquer d'autres skills qui n'ont pas forcément de rapport avec le machine learning.

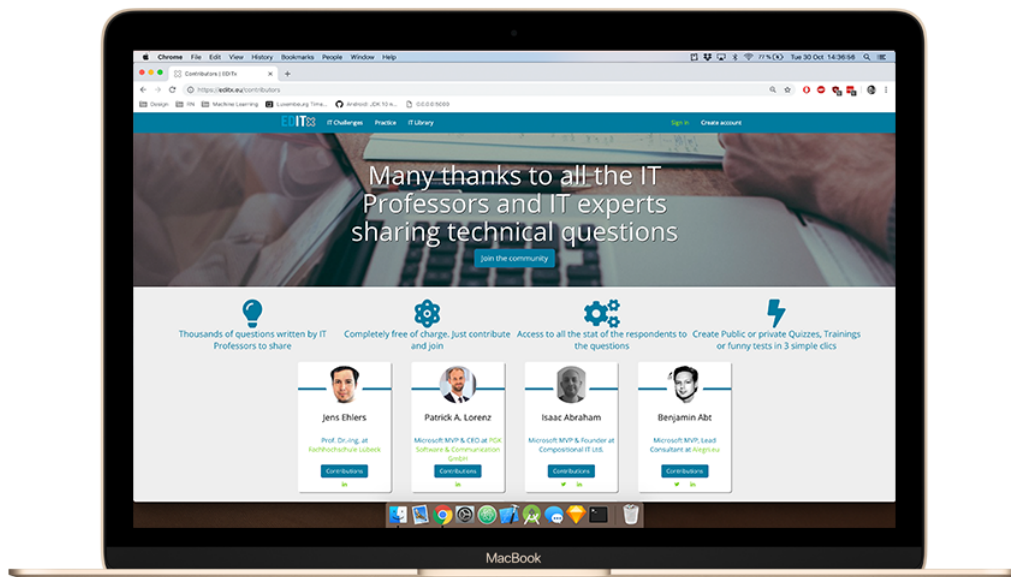
8.1 Design & Web Development

Ayant déjà fait beaucoup de Frontend auparavant, j'ai participé au développement d'une nouvelle page sur le site d'EDITx dédié aux collaborateurs et d'une autre sur le site d'OnlySalesJob pour le pricing.

editx.eu/contributors

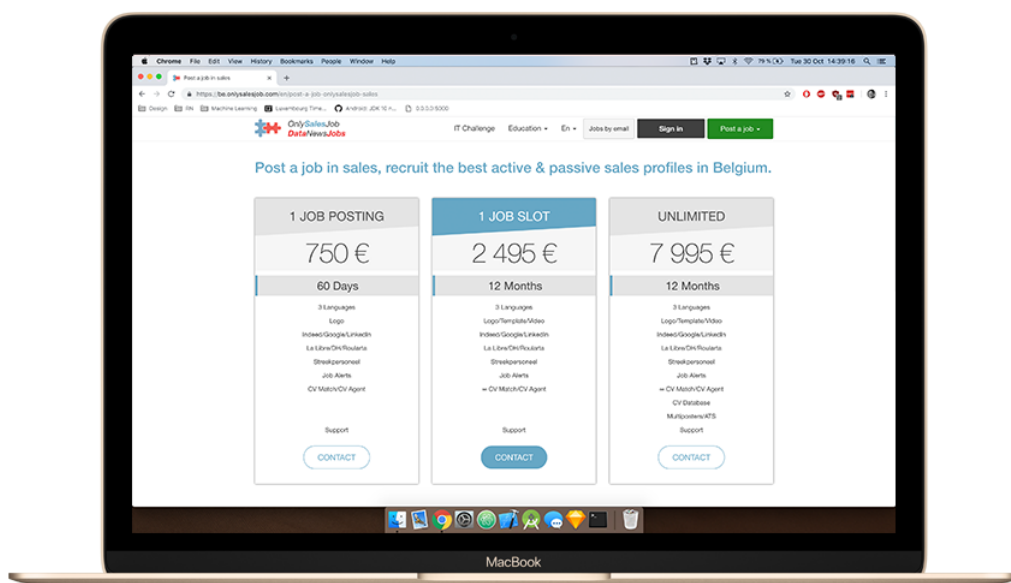
Avant de me lancer dans le HTML nous avons établi un cahier des charges. La page manquait un peu d'organisation et ne mettaient pas en avant les contributeurs. Le design de la nouvelle page a été élaborée avec Sketch App, un outil de design qui permet de travailler en SVG (Scalable Vector Graphics) qui permet de travailler en vectoriel et de garder une résolution parfaite même en changeant la taille. Un point important du cahier des charges était d'avoir une grande bannière avec un text de remerciement.

Le design a ensuite été traduit en HTML/CSS en profitant du Framework Bootstrap pour rendre le contenu "responsive".



onlvsalesjob.com/en/post-a-job-onlysalesjob-sales

Pour la page de pricing d'OnlySalesJob le plus important était d'avoir une grille de pricing claire, qui pour chaque modèle regroupe les options incluses pour ce modèle spécifique.



8.2 School Scraper

Une des "side-missions" consistait à trouver une façon intelligente et efficace pour récolter emails et numéros de téléphones d'écoles en Belgique dans le but de les contacter et faire la promotion de Bebras, une initiative qui vise à initier les jeunes à l'informatique.

Pour ce faire j'ai directement pensé à PhantomJS. PhantomJS est un navigateur web navigateur "headless" (sans interface graphique) scriptable utilisé pour automatiser des interactions avec des pages web. C'est un outils que j'avais déjà utilisé auparavant pour des applications personnelles. PhantomJS n'est plus utilisé depuis que Google lancé Headless Chrome qui est plus stable et plus performant. Mais le principe reste le même.

En collaboration avec Selenium, qui permet d'interagir avec les éléments de la page web, il est possible d'automatiser des séquences de navigation sur une page web, le tout en arrière plan sans interface graphique. Cela implique une réduction en temps et en mémoire tout en augmentant les performances.

Le site "www.enseignement.be" regroupe toutes les écoles catégorisées en régions.

Pour chaque école on retrouve le lien vers le site web et parfois le numéro de téléphone et email. Il y a plus de 400 écoles répertoriées et il faudrait un temps inimaginable pour récolter pour chacune d'entre elles le numéro de téléphone et adresse email.

Le script python que j'ai créé parcourt cette liste d'écoles, retrouve les champs (divs, ...) contenant la page web et/ou numéro de téléphone et email si présents. De ça je crée un objet école que je stocke dans une liste.

Pour avoir plus de résultats, le script navigue à chaque fois aussi vers la page web de l'école et fait une recherche regex pour retrouver emails et numéros de téléphone en plus de ceux indiqués sur la page qui regroupe toutes les écoles.

Voici à quoi ressemble le résultat après avoir exporté et formaté la liste d'objets écoles en fichier .txt :

```
Name: ECOLE AVE MARIA
Number: 081-731234
Websites: http://www.ecoleavemaria.org
Emails: avemaria.direction@gmail.com  avemaria.mons@sec.cfwb.be
-----
Name: ECOLE PROFESSIONNELLE
Number: 081-729011
Websites: http://www.asty-moulin.be
Emails: dir.generale@asty-moulin.be
-----
Name: INSTITUT DE LA PROVIDENCE
Number: 081-208500
Websites: http://www.providencechampion.be/
Emails:
-----
Name: INSTITUT SAINT-BERTHUIIN
Number: 081-447230
Websites: http://www.isbm.be
Emails: info@isbm.be
-----
Name: INSTITUT DES TECHNIQUES ET DES COMMERCES AGRO-ALIMENTAIRES
Number: 081-569064
Websites: http://www.itca.be
Emails: itca.direction@skynet.be  itca.secretariat@skynet.be
-----
Name: ATHENEE ROYAL GEMBOLOUX
Number: 081-611413
Websites: http://www.argembloux.be/
Emails: info@atheneegembloux.be
```

8.3 Concours en Allemagne

Depuis le 8 octobre et jusqu'au 11 novembre, EDITx organise un challenge JAVA en Allemagne. Ce challenge est sponsorisé par Consorsbank et attire des participants à travers toute l'Allemagne. Il fallait évidemment traduire certaines parties du règlement ou faire la promotion du concours sur les réseaux sociaux, en allemand. Pratiquant l'allemand depuis le primaire, je m'en suis occupé.

9 Conclusion

Le stage d'immersion chez EDITx fut une expérience incroyable. J'ai appris beaucoup de choses allant plus loin que ce à quoi je m'attendais.

Toute cette phase de self learning m'a bien préparé pour la suite. J'ai déjà un bon aperçu de ce qu'est le machine learning et j'ai même eu l'occasion de déjà l'appliquer pour les versions V0 des projets.

Le mémoire va me permettre d'approfondir mes connaissances et d'améliorer les modèles utilisés dans les V0 en vu de sortir en mai les V1 qu'on pourra ajouter sur le site d'EDITx.

J'aimerais conclure sur le fait qu'à mes yeux le stage n'aurait pas pu se passer mieux. L'expérience que j'en tire est très enrichissante. L'équipe d'EDITx est, je pense, très satisfaite de mon travail.

10 Bibliographie

- www.editx.eu
- www.onlysalesjob.be
- <https://duo.com/decipher/driving-headless-chrome-with-python>
- <https://fr.wikipedia.org/wiki/PhantomJS>
- <https://mubaris.com/posts/kmeans-clustering/>

11 Annexes

/