Développement d'algorithmes d'automatisation de processus web d'extractions de données

Céline Van Landeghem

Université de Strasbourg
Master 1 CSMI

2020-2021

Table des matières

Table des matières

Présentation de la start-up Tetrao

Figure – Tetrao [3]

La start-up

- Fondée en 2014
- Technologie basée sur l'IA
- Secteurs :
  - la gestion d'entreprise
  - la finance
- L'équipe s'agrandit de plus en plus

Présentation de la start-up Tetrao

Figure – Tetrao [3]

Le processus

- Extraction de toutes les données présentes sur Internet
- Modèles d'IA pour identifier automatiquement les données
- Projets :
  - Collecte de fonds d'investissement
  - Collecte d'obligations vertes

Table des matières

Fonds d'investissement

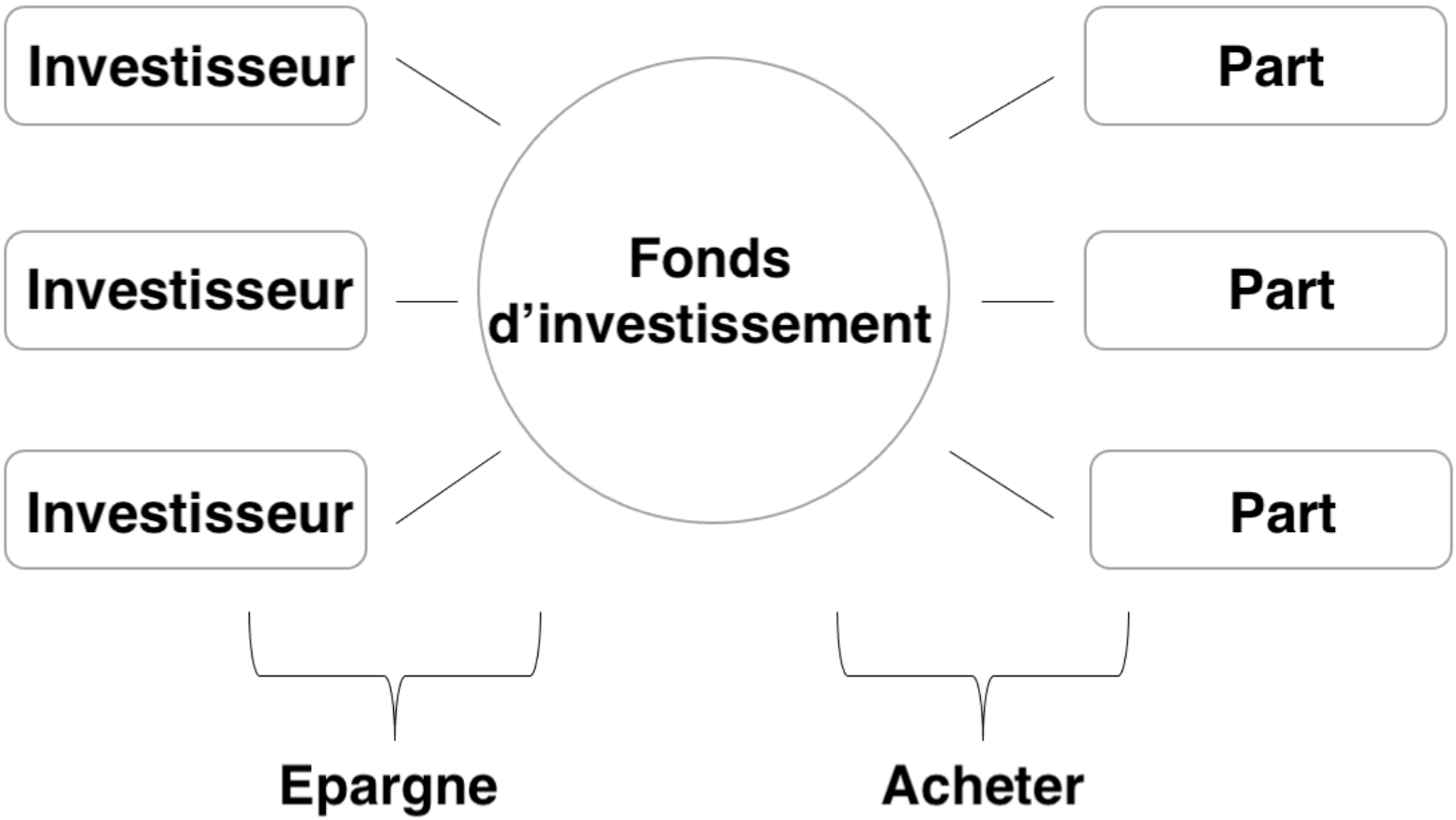

Figure – Structure des fonds d'investissement

Fonds d'investissement

Part

- Identifiée par un code : `ISIN`
- Informations sur le site web et dans les documents normés

But principal

- Collecter le code ISIN et toutes les informations de chaque part
- Augmenter la quantité de parts extraites
- Actuellement 120.000 parts collectées

Table des matières

Processus global

Figure – Processus global de Tetrao

Processus global

Extraction des données

- Étapes : analyses, scripts d'extraction
- Résultat visible sur Stratego : serveur central de Tetrao

Annotation des données extraites

- Réalisée sur la plateforme Cognita
- Tâche : indique quelle donnée doit être annotée

Processus global

Informations clés pour l'investisseur

LA FINANCIÈRE DE L'ÉCHIQUIER

Ce document fournit des informations essentielles aux investisseurs de cet OPCVM. Il ne s'agit pas d'un document promotionnel. Les informations qu'il contient vous sont fournies conformément à une obligation légale, afin de vous aider à comprendre en quoi consiste un investissement dans ce fonds et quels risques y sont associés. Il vous est conseillé de le lire pour décider en connaissance de cause d'investir ou non.

0: 12422415-1

ECHIQUIER ARTIFICIAL INTELLIGENCE - Action B (ISIN : LU1819480192)

Compartiment de la SICAV Echiquier Fund gérée par La Financière de l'Echiquier

Objectifs et politique d'investissement

Echiquier Artificial Intelligence est un compartiment dynamique recherchant la performance à long terme à travers l'exposition sur des valeurs de croissance des marchés internationaux. En particulier, le compartiment cherche à investir dans des valeurs qui développent l'Intelligence Artificielle et/ou des valeurs qui en bénéficient.

sur les notations proposées par les agences. Les titres obligataires concernés sont des titres réputés « Investment grade », à savoir notes au minimum BBB- par Standard & Poor's ou équivalent ou considérés comme tels par l'équipe de gestion.

Les instruments financiers à terme, négociés ou non sur des marchés

Figure – Exemple d'une annotation

Processus global

Annotation des données extraites

- Tâches validées : les samples
- La base des modèles d'intelligence artificielle

Modèles d'intelligence artificielle

- Entrainés par les samples
- Fonctionnels après quelques milliers de samples
- Capables de trouver automatiquement les données

Table des matières

Méthode de Tetrao

Différentes méthodes d'extraction

Méthode classique

- "Scrapping"
- Analyse le code HTML du site Web

Méthode de Tetrao

- Dépend que de la représentation visuelle
- Encore utilisable après un changement du code HTML

Méthode de Tetrao

Interface de programmation applicative de Tetrao

- Permet de parcourir le site web
- Extrait la position et la police des mots
- Détecte l'ensemble des éléments : le masque

Méthode de Tetrao

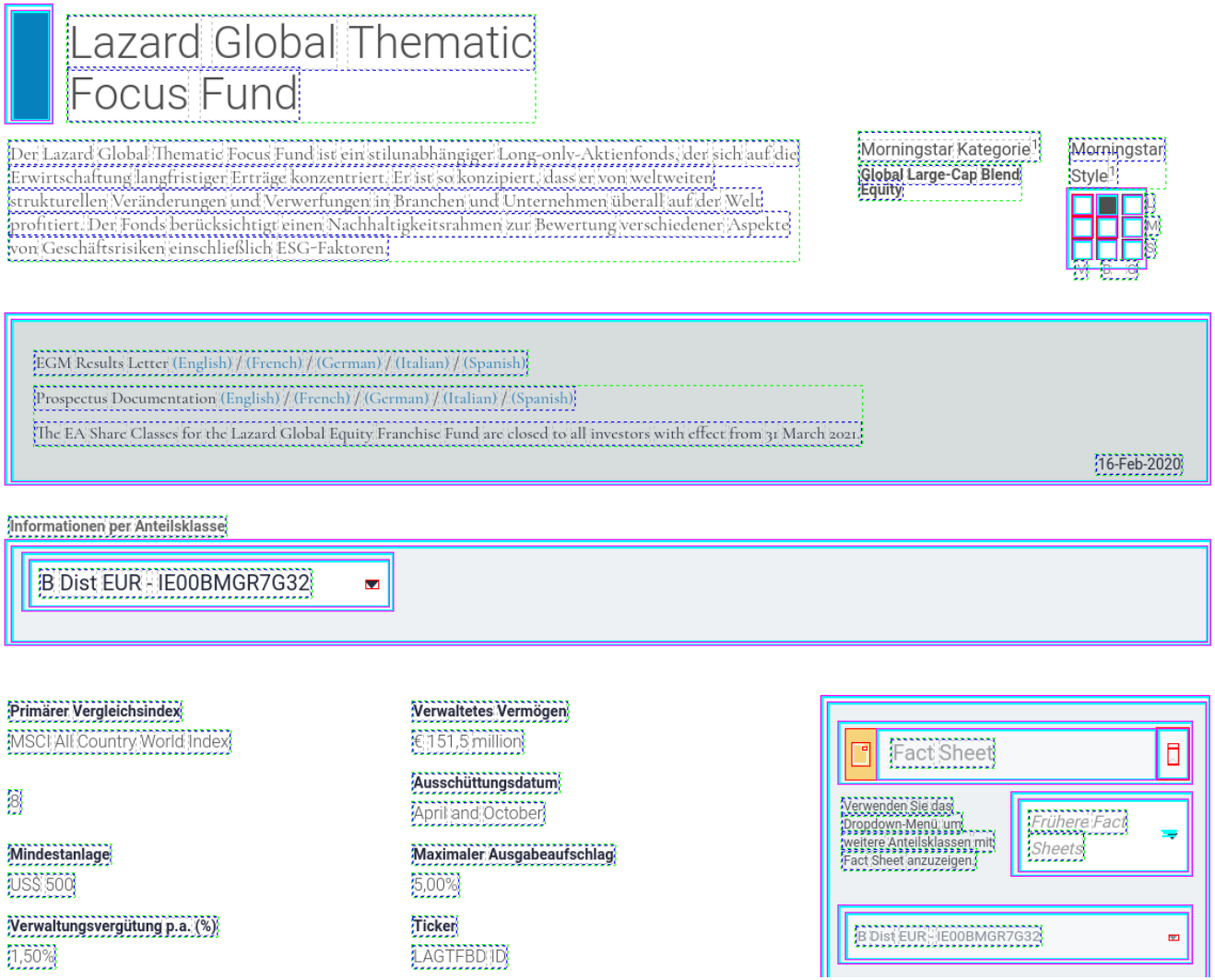

Figure – Affichage du masque

Deux tâches principales

- **Tâche de collecte** : collecte les informations d'une part

Task

Task ID	236056
Parent Task ID	
Workflow ID	28
Class Name	eu.tetrao.extractions.ffyn.allianz.AllianzNodeTask
Group Name	LU1363153583
Timeout	PT15M
Disabled	false
Created at	2018-11-27 12:59:24
Updated at	2018-11-27 12:59:24
country	lu
isin	LU1363153583
language	en
profile	professional
url	https://lu.allianzgi.com/en-gb/pro/our-funds/funds/list/allianz-global-opportunistic-bond-ct-eur

Figure – Les paramètres d'une tâche de collecte

- **Tâche de liste** : collecte les paramètres des parts d'un fonds

Analyses

- Description d'une méthodologie précise à réaliser
- Développeur doit qu'appliquer ces étapes

Figure – Différentes parties des analyses

Récherche d'émetteurs

- Trouver des sources appropriées

Écriture de l'analyse

- Nom de la société de gestion :
- Code :
- Profil d'investisseur :
- URL :
- Langue :
- Nombre de parts :

Cookies :

Identification du profil investisseur :

Liste des parts :

Collecte des données de parts :

Documents optionnels :

Figure – Template pour les analyses

Vérification

- Trouver la raison du manque de parts
- Raisons :
  - Existence d'un deuxième URL
  - Les parts ne sont pas publiées sur le web

Outils

Projet Earnestnet de Tetrao :

API Earnestnet

Navigateur Chromium

Figure – Scala [4]

Figure – Gitlab [5]

L'API Earnestnet

Agent : élément central

- Rôle : communiquer avec les éléments, faire les actions
- Méthodes :

Charger l'URL :

```
1    page = agent.get(URL)
```

Parcourir la page :

```
1    page = agent.snapshot()
```

Cliquer sur un élément :

```
1    page = agent.click(element)
```

Algorithmes d'extraction

Page : contient le masque

- Méthodes :

Récupérer les éléments d'un type :

```
1 mots = page.ive_mask.words
```

Spécifier la position :

```
1 mot_plus_haut = mots.uppermost
```

Filtrer :

```
1 mot = mots.text_matching(texte)
```

Récupérer les paramètres :

```
1 mot_text = mot.text
```

Table des matières

Tâche de collecte : « Hermitage Gestion Privée » [2]

Cookies : il faudra fermer le pop-up cookie en cliquant sur « Accepter »

```
1 page = agent.get("https://www.hermitagegestionprivee.com/")
2 bouton = page.ive_mask.words.text_matching("Accepter").lowermost
3 page = agent.click(bouton)
```

Tâche de collecte : « Hermitage Gestion Privée » [2]

Collecte des données de parts : il faudra parcourir toute la page et télécharger les documents PDF

```
1 page = agent.snapshot()
2 documents =  download_documents(page)
```

Vérifications

```
1 if (page.ive_mask.words.text_matching(ISIN).isEmpty) {
2   throw new IsinNotFoundException
3 }
4
5 check(page, MOTS)
```

Merci de votre attention !

Avez-vous des questions ?

Références

- Rigaud Etienne. *Développement de modèles d'apprentissage automatique pour la compréhension de documents*. Tetrao, 2020.
- Michels Théo. *Assistant-ingénieur*. Tetrao, 2021.
- Tetrao. `https://wiki.tetrao.eu/wiki/index.php/Main_Page`.
- Thierry Labro. *Comment Tetrao va disrupter l'industrie des fonds*. `https://paperjam.lu/article/comment-tetrao-va-disrupter-in`, 18.12.2019.
- Thierry Labro. *La Bourse s'invite chez Tetrao, un win-win intelligent*. `https://paperjam.lu/article/bourse-s-invite-chez-tetrao-wi`, 26.01.2021.

Références

- Mehdi Ouchallal. *Comment créer un fonds d'investissement ?*. https://www.legalplace.fr/guides/creer-fond-investissement/, 06.04.2021.
- Hermitage Gestion Privée. https://www.hermitagegestionprivee.com/gestion-actifs/.
- Image Tetrao. https://lu.linkedin.com/company/tetrao.
- Image Scala. https://medium.com/elp-2018/scala-a-la-conquête-du-big-data-49238d173a1f.
- Image Gitlab. https://developer.ibm.com/recipes/tutorials/deploying-a-multiarch-openshift-application-from-gitlab/.

Modèles d'IA

Modèles génériques

- Associés à une tâche précise
- Utilisés pour les données des documents normés

Modèles spécifiques

- Utilisées pour les données provenant des sites web
- Un modèle pour chaque société de gestion
- Plus que 1000 tels modèles

Les documents extraits de chaque société sont regroupés en corpus :

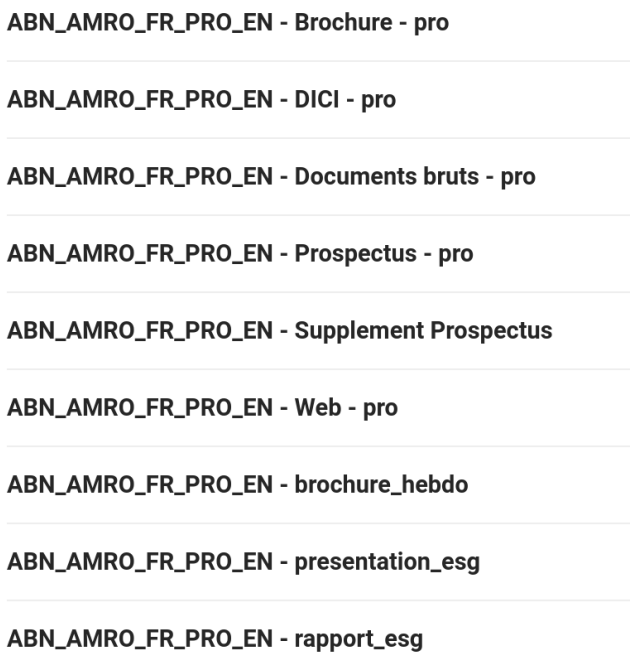

Figure – Différents corpus d'une société de gestion