OCR Challenge

CRUNCH UTT 2024

INTRODUCTION

Ardian est le leader européen du private equity. Fondée en 1996, l'entreprise a accumulée depuis sa création un très grand nombre de documents confidentiels. Dans une démarche de valorisation de ces documents, l'équipe IT est à la recherche d'une solution pour extraire le texte de ces Documents. Pour mieux appréhender cette problématique, l'équipe stagiaire 100% UTTienne du projet ArdianBrowser vous lance ce défi.

DEFI

Enoncé

> « Extraire le maximum de texte brut de ce PDF, de la manière la mieux structurée possible. »

Explication

Le « texte brut » ciblé est l'ensemble des mots avec le maximum de ponctuation.

La récupération de la « structure » visée concerne la conservation de l'unité des paragraphes. À titre d'information, ceci est le 10e paragraphe de ce document

REGLES

Contraintes

1- Format

Votre solution sera sous la forme d'un projet python qui contiendra la fonction suivante :

```python
def extract_text(path):
    """
    Main function that extracts text from a file in a structured way to return the list of its paragraphs.
    :param path: path to the targeted file
    :return: list of paragraph strings
    """

    # TODO

    return raw_text
```

2- Compliance

L'intégralité de la documentation interne d'Ardian est confidentielle, pour des raisons évidentes de compliance, votre solution ne doit en aucun cas utiliser un service tiers qui aurait accès au document. C'est-à-dire pas d'appel à une API d'OCR (ilovepdf, adobe, etc..). Le fichier doit être intégralement traité en local sur votre machine.

Vous êtes libres d'utiliser n'importe quel outil ou logiciel tiers à condition de prouver que la compliance a été respectée (L'utilisation de l'outil n'entraine pas l'accès au fichier traité par un tiers). Si vous utilisez des outils tiers, vous êtes invités à décrire brièvement les aspects pertinents de leur installation et de leur configuration

3- Scalability

Votre solution devra être exécutable sur un corpus qui dépasse le million de fichiers, avec une évolution linéaire du temps d'exécution en fonction du nombre de fichiers à traiter. (3 points bonus si vous prouvez empiriquement la quasi-linéarité de votre solution sur une période d'une heure minimum)
Vous détaillerez les différents problèmes de mémoire qui pourraient potentiellement apparaître avec un corpus d'une telle taille, ainsi qu'une explication sur comment y remédier.
Vous indiquerez également le nombre de fois que votre fonction extract_text() réussi à traiter ce fichier en 10 minutes sur la meilleure de vos machines. (En précisant les caractéristiques de celle-ci)

Barème

Rien de vous y oblige, mais si vous êtes friand de challenges, essayez de récolter le maximum de points de notre barème :

Si vous parvenez à :	Récompense :
Récupérer le texte brut du paragraphe d'introduction	1 point
Conserver la structure du paragraphe d'introduction	4 points
Récupérer tout le texte de ce tableau	3 points
Conserver la structure de ce tableau (i.e. 1 cellule = 1 paragraphe)	4 points
Récupérer tous les pays européens de l'item 1	2 points
Pour chaque pays non-européen correctement récupéré dans l'item 1	2 points
Conserver la structure du paragraphe *« Market definition »* dans l'item 2	5 points
Récupérer le texte de toutes les cellules et de tous les libellés du tableau *« Market dynamics »* dans l'item 2	2 points
Conserver la structure de chaque cellule/libellé de ce même tableau (i.e. 1 cellule/libellé = 1 paragraphe)	5 points
Récupérer la totalité du texte du schéma *« ADE addressed market overview »* dans l'item 2	3 points
Conserver la structure des paragraphes dans l'item 3 (plus de 13 paragraphes)	7 points
Conserver la structure en 4 paragraphes dans la section « SNAPSHOT » de l'item 4	6 points
Conserver la structure en 3 paragraphes dans la section « M&A STRATEGY » de l'item 4 (6 paragraphes tolérés avec les dates)	6 points
TOTAL	**50 points**

Bonus

Parce qu'on est sympa :)

Si vous parvenez à récupérer au moins une fois le filigrane « CONFIDENTIEL » ➞ 5 points

Si vous parvenez à récupérer « ARDIAN » avec le logo du pied de page ➞ 2 points

Si vous étendez votre solution pour qu'elle prenne en charge d'autre formats de fichiers contenus dans la liste suivante ➞ 2 points par extension
['.doc', '.docx', '.ppt, '.pptx', '.png', '.jpeg', '.jpg']
(Il va sans dire que les fichiers Word & PowerPoint contiennent évidement des images avec du texte à récupérer)

Si vous prouvez empiriquement la quasi-linéarité de votre solution sur une période d'une heure minimum (un graphe nb de traitements en fonction du temps suffira) ➞ 3 points

Si vous n'utilisez que des outils open-source ➞ 3 point

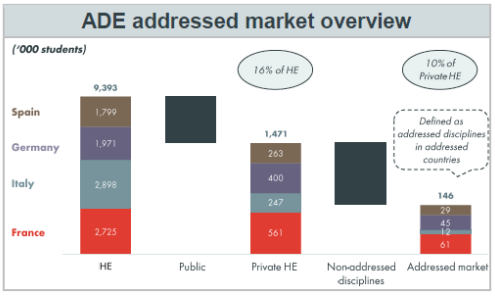
Real data

Item 1

Item 2

Market definition

The Higher Education ("HE") in Europe represents 23m students and ADE is present in France (63% of revenues), Italy (12% of revenues), Germany (12% of revenues) and Spain (13% of revenues), together representing **9.4m HE students (40% of the total HE market in Europe) and 1.5m students in private HE (50% of the private market)**. Out of these 1.5m students, **150k students (10%) are in ADE addressed disciplines of Creative Arts**, namely: (i) Design & Graphical Arts as well as Digital in France, (ii) Design in Italy, (iii) Creative Arts and Media & Communications in Germany and (iv) Audiovisual (professional curriculum) in Spain.



Market dynamics

	Demographics		GER		Share of discipline		Share of private HE		International students		Market volume		Average fee		Market value	
CAGR	16-20	20-25	16-20	20-25	16-20	20-25	16-20	20-25	16-20	20-25	16-20	20-25	16-20	20-25	16-20	20-25
France - Creative Arts	+0.3%	+0.7%	+0.7%	+0.7%	+1.4%	+1.3%	+1.8%	+0.9%	+0.5%	+0.2%	+4.6%	+3.8%	+1.5%	+1.5%	+6.2%	+5.3%
France - Digital	+0.3%	+0.7%	+0.7%	+0.7%	+9.6%	+4.2%	0.0%	0.0%	0.0%	0.0%	+10.5%	+5.6%	+1.5%	+1.5%	+12.2%	+7.2%
Italy- Design	-0.4%	+0.2%	+1.5%	+1.0%	+5.9%	+4.2%	+1.6%	+1.9%	+0.4%	+0.4%	+9.1%	+7.6%	+1.0%	+1.0%	+10.2%	+8.7%
Germany – Design & Media	0.0%	-1.3%	+0.4%	+0.7%	-0.6%	-1.3%	+4.8%	+5.7%	+0.7%	+0.7%	+5.3%	+4.5%	+1.0%	+1.0%	+6.3%	+5.5%
Spain – Audiovisual	-0.8%	+0.7%	+4.0%	+3.6%	+1.6%	+1.5%	+7.1%	+4.7%	+0.2%	+0.2%	+12.1%	+10.4%	+1.0%	+1.0%	+13.3%	+11.5%
Total	-0.2%	+0.4%	+1.7%	+1.6%	+2.9%	+2.0%	+3.4%	+2.6%	+0.3%	+0.2%	+8.2%	+6.9%			+9.5%	+8.2%
Total (weighted avg. ADE enrolments)	-0.1%	+0.3%	+1.4%	+1.2%	+2.6%	+1.7%	+3.0%	+2.3%	+0.4%	+0.3%	+7.3%	+5.8%			+8.6%	+7.1%

The addressed market for ADE has been growing in volumes at +8.2% p.a. over 2016-20 and in value at +9.5% p.a., mostly driven by increased share of private education in addressed disciplines (c.40% of vol. growth), increased share of

November 2020 – 2-Pager – AD Education p.5

Item 3

- We are in friendly territories with Jeausserand and Oloryn acting as KG's advisors (governance and Manpack);
- We have a strong value proposition with our geographical footprint in the Company's core geographies and potential proprietary cross-fertilization with Study Group's sourcing network;
- **We have further consolidated our set of experts and advisors:**
  - We have assembled an outstanding team of operating partners and advisors:
    - **Marc Adler** ("MA"): Former CEO of Eduservices (2013-2019 – French platform of c.€115m sales). Kevin Guenegan knows him well and validated him to work alongside us;
    - **David Lefevre** ("DL"): CEO of Insendi, digital platform acquired by Study group (spin-off from Imperial College);
    - **Jean-Marc Chamayou** ("JMC" - Lafayette Associés): Founder of Euridis, Head of Lafayette Associés, working with all education players in France, specifically on accreditations, employability and placement topics;
  - Our team is also comprised of a remarkable group of third-party advisors, including:
    - **Centerview** *(M&A advisor – Nicolas Constant & Pierre Pasqual)*: they have worked successfully in precedents in education and are both former Lazard executives, with a direct line and close relationship with the Lazard team;
    - **EY-Parthenon** *(Commercial DD)*: Parthenon is one of the very few consulting firms very knowledgeable on the education space (notably the Partner Anna Grotberg). They have performed a high-quality phase 1 strat. DD;
    - **KPMG** *(Finance DD - Guilhem Maguin):* KPMG has worked alongside us on Eureka and has extensive knowledge and experience in the education sector;
    - **ADIT** *(Business intelligence):* ADIT is performing a reputational and operational assessment of the management team, including school directors;
    - **Wilkie & Latham** *(M&A, Structuring & Financing – Eduardo Fernandez & Xavier Farde)*: Eduardo Fernandez is close to the personal advisor of Kevin Guenegan (Patrick Mousset) as they were former colleagues (at Wilkie).

Item 4

Kalbe Farma

KALBE

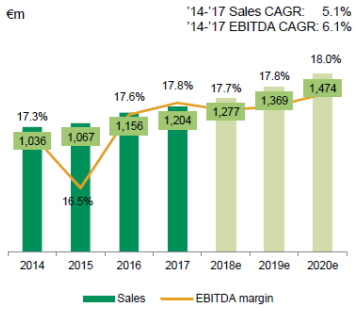
SNAPSHOT

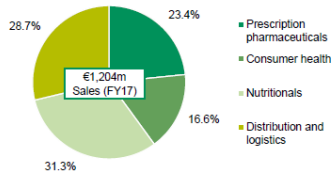
- Kalbe Farma is an integrated healthcare solution provider active in 4 business divisions: Prescription Pharmaceutical division (23%), Consumer Health division (17%), Nutritionals division (30%), Distribution and Logistics division (30%)
- The Company is active internationally with a footprint in Asian countries, Nigeria and South Africa
- Kalbe Farma has massively invested in an R&D department for generic drug formulation to ensure an ongoing development of innovative Consumer and Nutritional products
- It has over 17,000 employees worldwide

M&A STRATEGY

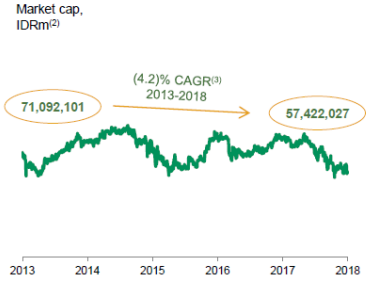
- **2016:** **Sale of Champion Pacific**, an Indonesian company active in the manufacturing of packaging for the pharmaceutical industry. Selling price 6.2x EBITDA
- **2012:** Kalbe Farma acquires **Hale International**, a manufacturer of beverages
- **2011:** **Acquisition of Enseval Putera Megatrading**, an Indonesian-based distributor of pharma products and medical equipment

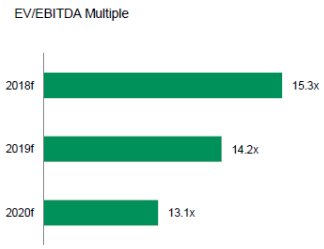
DRIVERS(1)





VALUATION





(1) FYE 31/12 (2) Indonesian rupiah (3) CAGR calculated from 10/09/2013 to 10/09/2018
Source: Company website, Broker reports, Annual Reports, MergerMarket, Factset

BNP PARIBAS The bank for a changing world DRAFT Project Demeter - September 2018 | 18