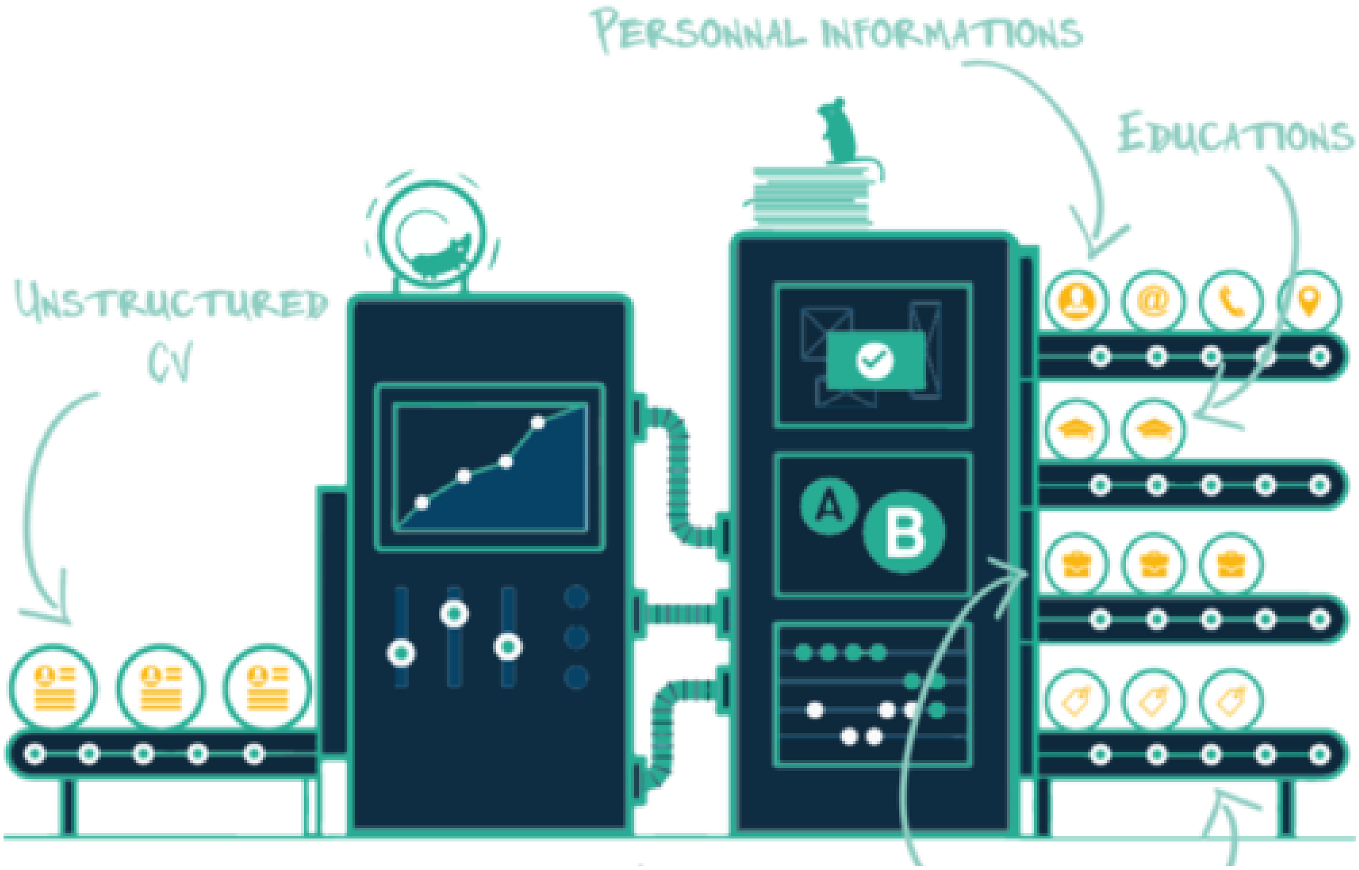
HR software companies? Why structuring your data is crucial for your business?

Benchmarking Resume Parsing Solutions : Daxtra, Sovren, Hireability, Textkernel and Segmentr (by Riminder)

Mouhidine SEIV Follow
Apr 16, 2018 · 4 min read

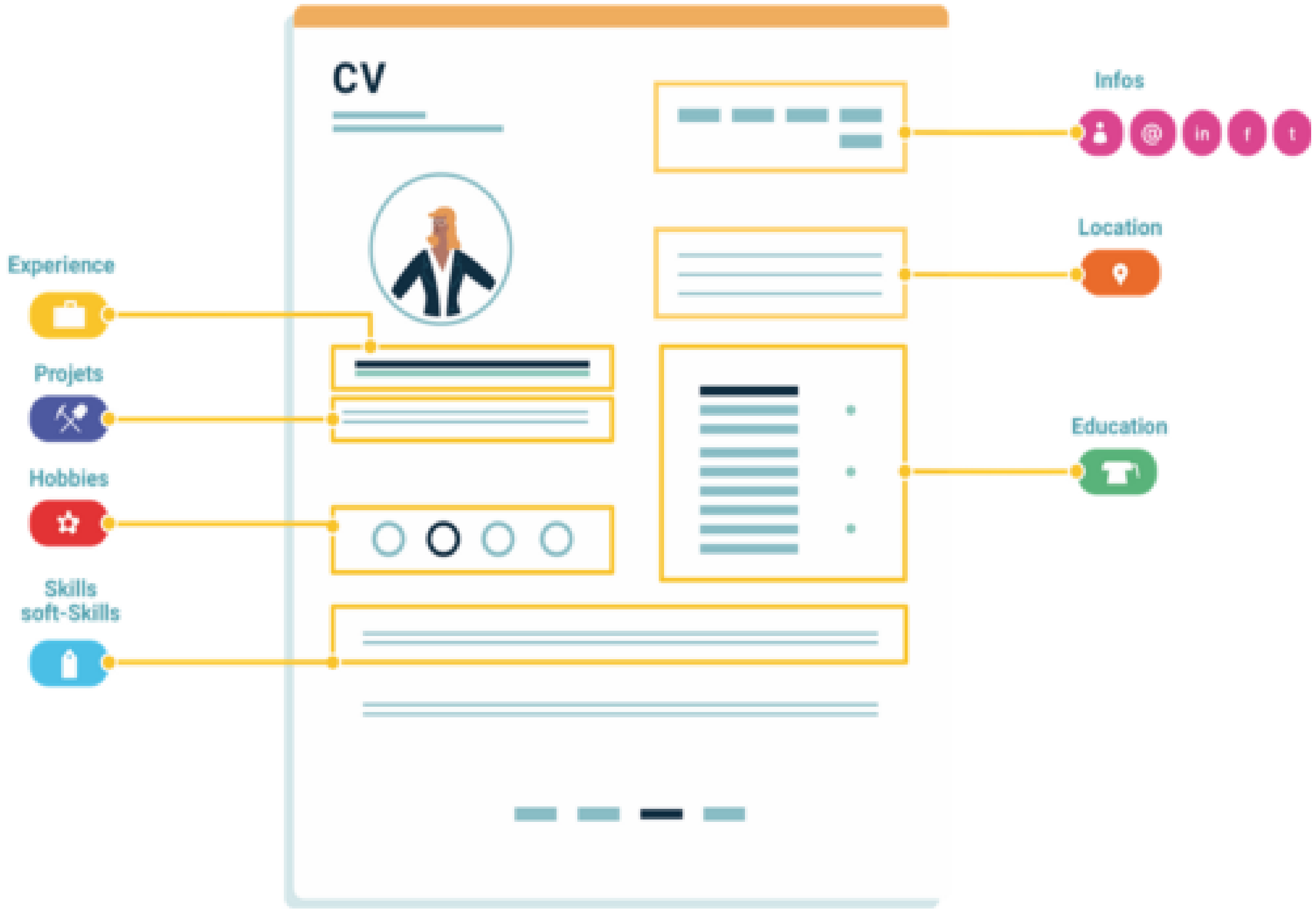
If you are working in the HR software industry or a member of an HR department, you definitely have thought about structuring your CV data before. Projects requiring resume structuring can be crucial to your business and can range from simple user experience improvement to strategic product roadmap advancements. These are some examples of the common use cases of CV structured data in the HR Software Industry:



Now you can parse all your resumes, no exception made!

- creating an efficient and relevant talent search experience
- getting market-relevant insights about your talent pools
- Building usable datasets for an AI based job matching tool.

Resume Parsing, the inevitable solution to your problem



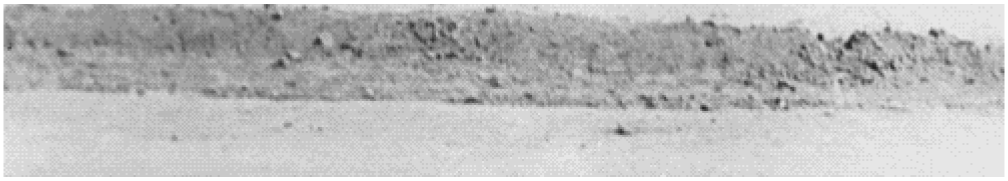
copyrights Riminder 2018

A resume parsing solution is a software that takes a resume as an input that can be in any media format (PDF, Word or image) or template, then convert it into a structured data format like — such as XML or JSON.

The information that is extracted by a resume parser usually includes the following:

> *personal information: name, email, address, phone*
>
> *list of experience: start date, end date, location, job title, company, description, …*
>
> *list of education: start date, end date, location, degree, university …*
>
> *list of skills, …*
>
> *list of interests*

Seems easy? but reality is hard!
No improvement for more +10 years



> ***Here are some few metrics:***
>
> **+1.4 Billion** *resumes are parsed every year.*
>
> **+40%** *of resumes have a complex layout (multi-column,etc.)*
>
> **+7%** *of resumes are either scans or images*

The first resume parsers were born in the late 90's to provide a data structuring technology to HR software companies that are looking for a stand-alone packaged solution in order to focus on their core business. Some of these first mover solutions are:

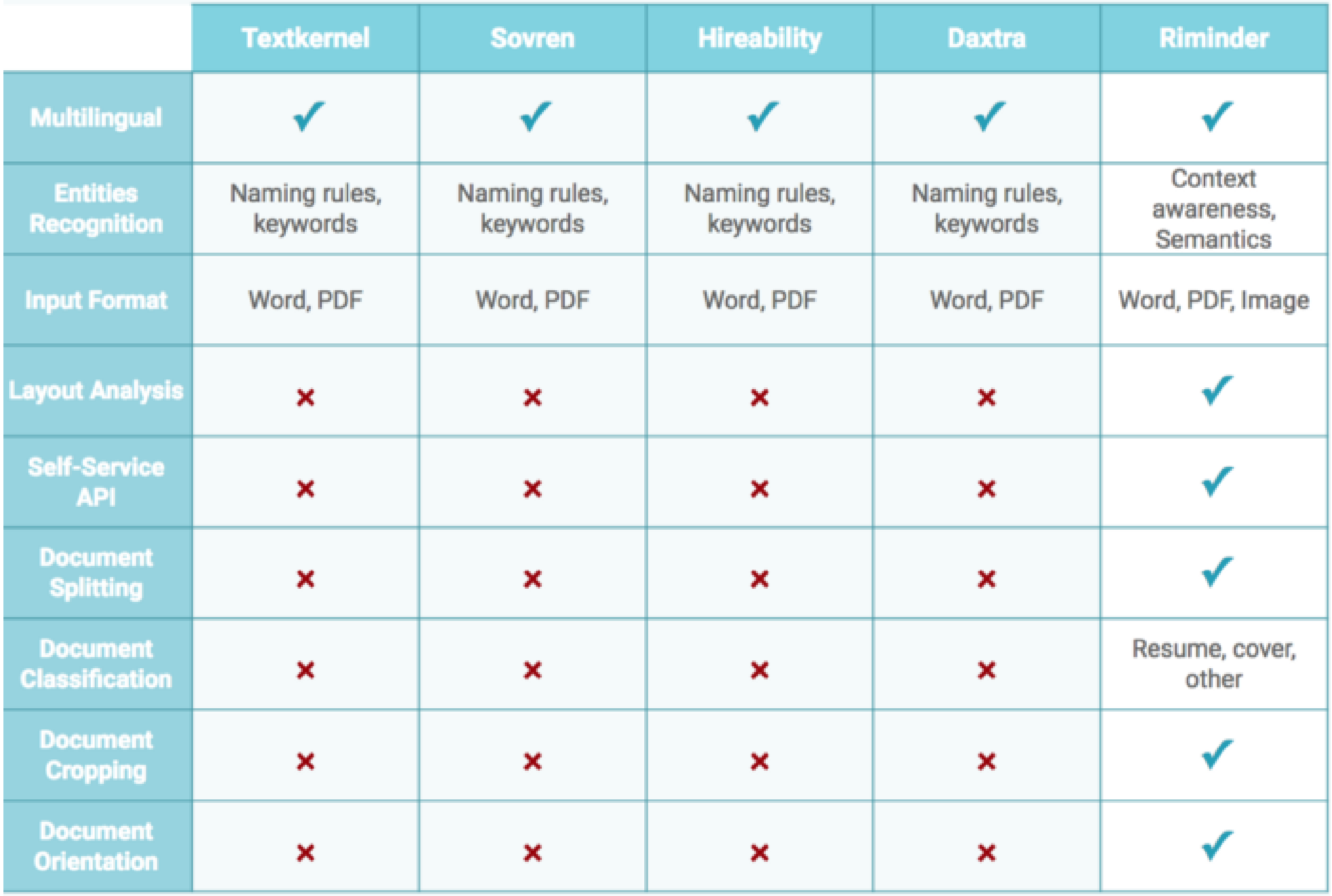
- Sovren (1996)
- TextKernel (2001)
- Daxtra (2002)
- …

How Daxtra, Sovren, Hireability, Textkernel and Segmentr (by Riminder) are doing at this task?

Building a general and reliable parser requiers many different blocks.
For instance, the system should be able to handle:

- complex layouts (ex: multi-column resumes, pictures with backgrounds, etc.)
- ambiguous entities (ex: Facebook, as a former employer vs. a social media skill)
- different media formats (PDF, Word, Image, etc.)
- multiple languages
- etc.

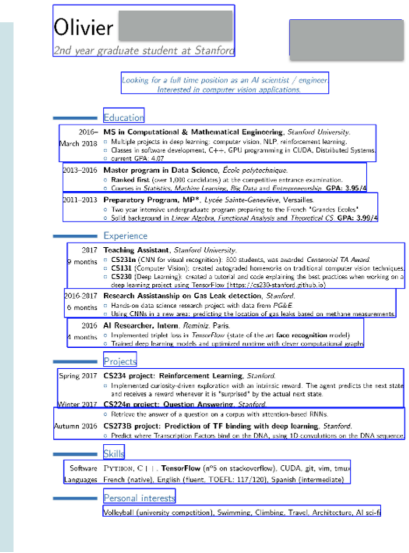
The following comparison between some of Segmentr's features and famous existing resume parsing solutions is the result of extensive validation tests we led at Riminder:



	Textkernel	Sovren	Hireability	Daxtra	Riminder
Multilingual	✓	✓	✓	✓	✓
Entities Recognition	Naming rules, keywords	Naming rules, keywords	Naming rules, keywords	Naming rules, keywords	Context awareness, Semantics
Input Format	Word, PDF	Word, PDF	Word, PDF	Word, PDF	Word, PDF, Image
Layout Analysis	✗	✗	✗	✗	✓
Self-Service API	✗	✗	✗	✗	✓
Document Splitting	✗	✗	✗	✗	✓
Document Classification	✗	✗	✗	✗	Resume, cover, other
Document Cropping	✗	✗	✗	✗	✓
Document Orientation	✗	✗	✗	✗	✓

Features benchmark

Segmentr (by Riminder) is the only Resume Parser able to handle such examples



Olivier [redacted]
2nd year graduate student at Stanford

[redacted]

Looking for a full time position as an AI scientist / engineer.
Interested in computer vision applications.

Education

2016– March 2018 **MS in Computational & Mathematical Engineering**, *Stanford University.*
- Multiple projects in deep learning: computer vision, NLP, reinforcement learning.
- Classes in software development, C++, GPU programming in CUDA, Distributed Systems.
- current GPA: 4.07

2013–2016 **Master program in Data Science**, *École polytechnique.*
- **Ranked first** (over 1,000 candidates) at the competitive entrance examination.
- Courses in *Statistics, Machine Learning, Big Data* and *Entrepreneurship*. **GPA: 3.95/4**

2011–2013 **Preparatory Program, MP***, *Lycée Sainte-Geneviève*, Versailles.
- Two year intensive undergraduate program preparing to the French "Grandes Ecoles"
- Solid background in *Linear Algebra, Functional Analysis* and *Theoretical CS*. **GPA: 3.99/4**

Experience

2017 9 months **Teaching Assistant**, *Stanford University.*
- **CS231n** (CNN for visual recognition): 800 students, was awarded *Centennial TA Award.*
- **CS131** (Computer Vision): created autograded homeworks on traditional computer vision techniques.
- **CS230** (Deep Learning): created a tutorial and code explaining the best practices when working on a deep learning project using TensorFlow (https://cs230-stanford.github.io)

2016-2017 6 months **Research Assistanship on Gas Leak detection**, *Stanford.*
- Hands-on data science research project with data from *PG&E.*
- Using CNNs in a new area: predicting the location of gas leaks based on methane measurements.

2016 4 months **AI Researcher, Intern**, *Reminiz*, Paris.
- Implemented triplet loss in *TensorFlow* (state of the art **face recognition** model)
- Trained deep learning models and optimized runtime with clever computational graphs

Projects

Spring 2017 **CS234 project: Reinforcement Learning**, *Stanford.*
- Implemented curiosity-driven exploration with an intrinsic reward. The agent predicts the next state and receives a reward whenever it is "surprised" by the actual next state.

Winter 2017 **CS224n project: Question Answering**, *Stanford.*
- Retrieve the answer of a question on a corpus with attention-based RNNs.

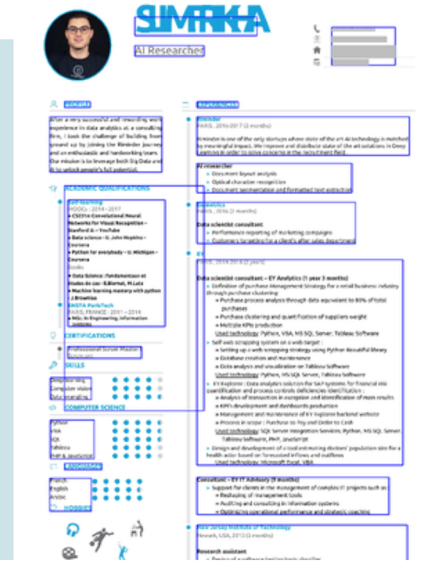
Autumn 2016 **CS273B project: Prediction of TF binding with deep learning**, *Stanford.*
- Predict where Transcription Factors bind on the DNA, using 1D convolutions on the DNA sequence.

Skills

Software	PYTHON, C++, **TensorFlow** (n°5 on stackoverflow), CUDA, git, vim, tmux
Languages	French (native), English (fluent, TOEFL: 117/120), Spanish (intermediate)

Personal interests

Volleyball (university competition), Swimming, Climbing, Travel, Architecture, AI sci-fi



SIMRIKA
AI Researcher

[redacted]

[illegible]

[illegible]

Riminder
PARIS, 2016-2017 [illegible]

[illegible]

AI researcher
- Document layout analysis
- Optical character recognition
- Document segmentation and formatted text extraction

ACADEMIC QUALIFICATIONS

Self-learning
MOOCs : 2014 - 2017
- CS231n Convolutional Neural Networks for Visual Recognition – Stanford U. – YouTube
- Data science - U. John Hopkins - Coursera
- Python for everybody – U. Michigan – Coursera

Books
- Data Science : Fondamentaux et études de cas - E.Biernat, M.Lutz
- Machine learning mastery with python - J.Brownlee

ENSTA ParisTech
PARIS, FRANCE : 2011 – 2014
- MSc. in Engineering, information [illegible]

CERTIFICATIONS

Professional Scrum Master I
Scrum.org

SKILLS

Deep learning
Computer vision
Data wrangling

COMPUTER SCIENCE

Python
VBA
SQL
Tableau
PHP & JavaScript

LANGUAGES

French
English
Arabic

HOBBIES

[illegible]
PARIS, 2016 (2 months)

Data scientist consultant
- Performance reporting of marketing campaigns
- Customers targeting for a client's after sales department

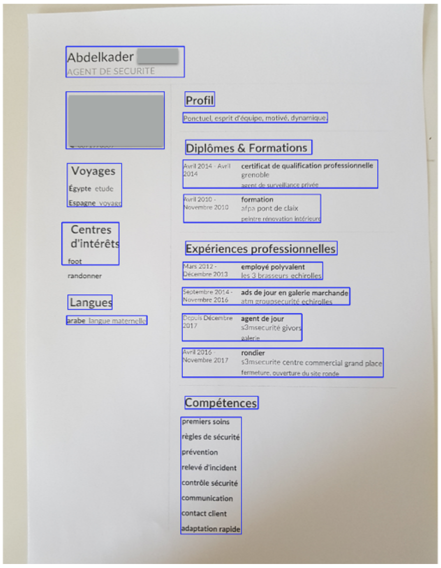
EY
PARIS, 2014-2016 (2 years)

Data scientist consultant – EY Analytics (1 year 3 months)
- Definition of purchase Management Strategy for a retail business industry through purchase clustering
  - Purchase process analysis through data equivalent to 80% of total purchases
  - Purchase clustering and quantification of suppliers weight
  - Multiple KPIs production
  - Used technology: Python, VBA, MS SQL Server, Tableau Software
- Self web scrapping system on a web target :
  - Setting up a web scrapping strategy using Python Beautiful library
  - Database creation and maintenance
  - Data analysis and visualization on Tableau Software
  - Used technology: Python, MS SQL Server, Tableau Software
- EY Explorer : Data analytics solution for SAP systems for financial risk quantification and process controls deficiencies identification :
  - Analysis of transaction in exception and identification of main results
  - KPI's development and dashboards production
  - Management and maintenance of EY Explorer backend website
  - Process in scope : Purchase to Pay and Order to Cash
  - Used technology: SQL Server Integration Services, Python, MS SQL Server, Tableau Software, PHP, JavaScript
- Design and development of a tool estimating doctors' population size for a health actor based on forecasted inflows and outflows
  - Used technology: Microsoft Excel, VBA

Consultant – EY IT Advisory (9 months)
- Support for clients in the management of complex IT projects such as :
  - Reshaping of management tools
  - Auditing and consulting in information systems
  - Optimizing operational performance and strategic coaching

New Jersey Institute of Technology
Newark, USA, 2013 (3 months)

Research assistant
- [illegible]



Abdelkader [redacted]
AGENT DE SECURITE

[redacted]

Voyages

Égypte etude
Espagne voyage

Centres d'intérêts

foot
randonner

Langues

arabe langue maternelle

Profil

Ponctuel, esprit d'équipe, motivé, dynamique.

Diplômes & Formations

Avril 2014 - Avril 2014	**certificat de qualification professionnelle** grenoble agent de surveillance privée
Avril 2010 - Novembre 2010	**formation** afpa pont de claix peintre rénovation intérieure

Expériences professionnelles

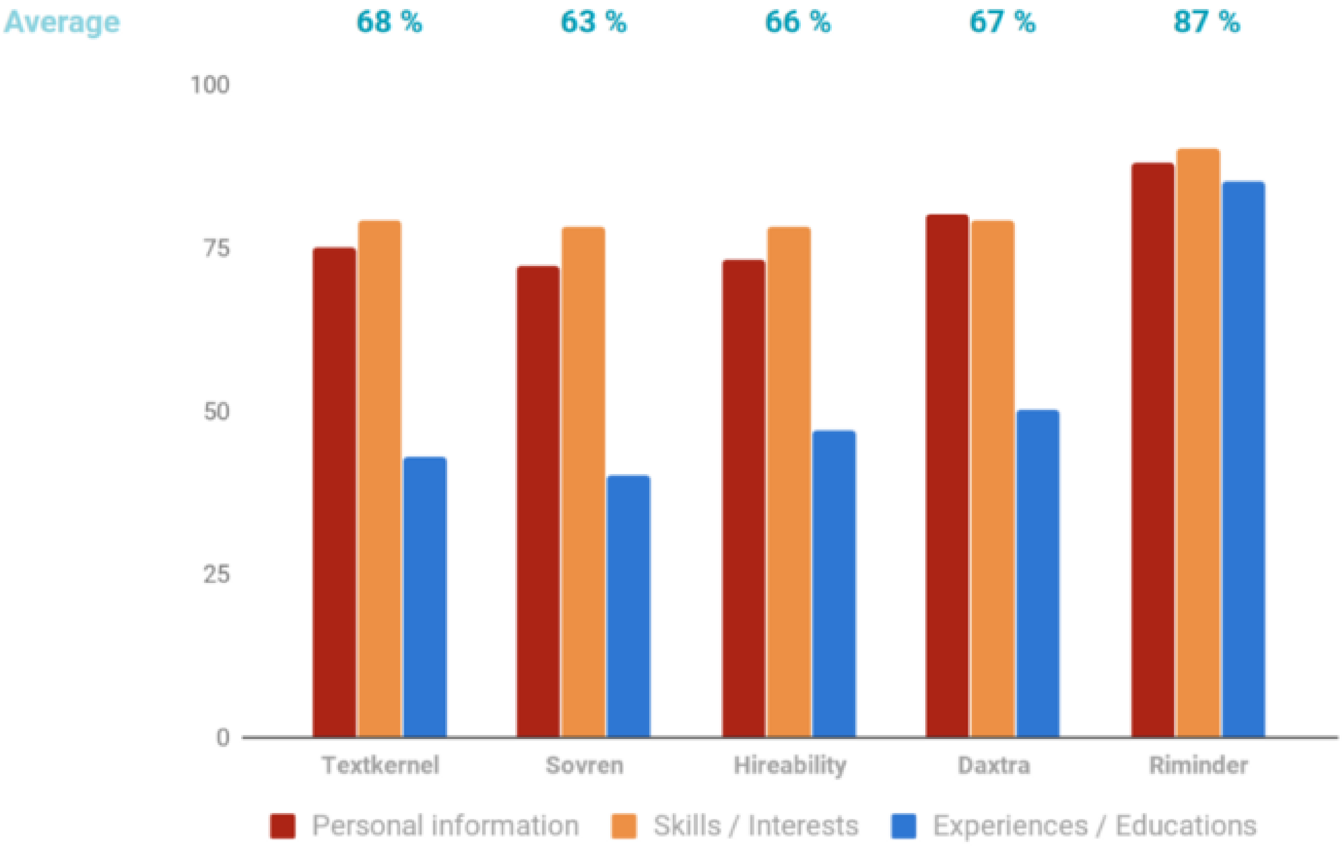
Mars 2012 - Décembre 2013	**employé polyvalent** les 3 brasseurs echirolles
Septembre 2014 - Novembre 2016	**ads de jour en galerie marchande** atm groupsecurité echirolles
Depuis Décembre 2017	**agent de jour** s3msecurité givors galerie
Avril 2016 - Novembre 2017	**rondier** s3msecurite centre commercial grand place fermeture, ouverture du site ronde

Compétences

premiers soins
règles de sécurité
prévention
relevé d'incident
contrôle sécurité
communication
contact client
adaptation rapide

Left: High density one-column resume
Center: Complex layout resume
Right: Picture of a multi-column resume with combined background, shadow and distorsion

We've also computed the performance of each solution over a validation dataset of around 100 resumes randomly sampled. For each output, we averaged the accuracy obtained across the multiple labels. Below is a graph summarizing the obtained results:



Extracted information accuracy overview

Segmentr example in python

First you have to post your data using a POST REQUEST on following the endpoint bellow (full documentation: https://developers.riminder.net/v1.0/reference#profile)

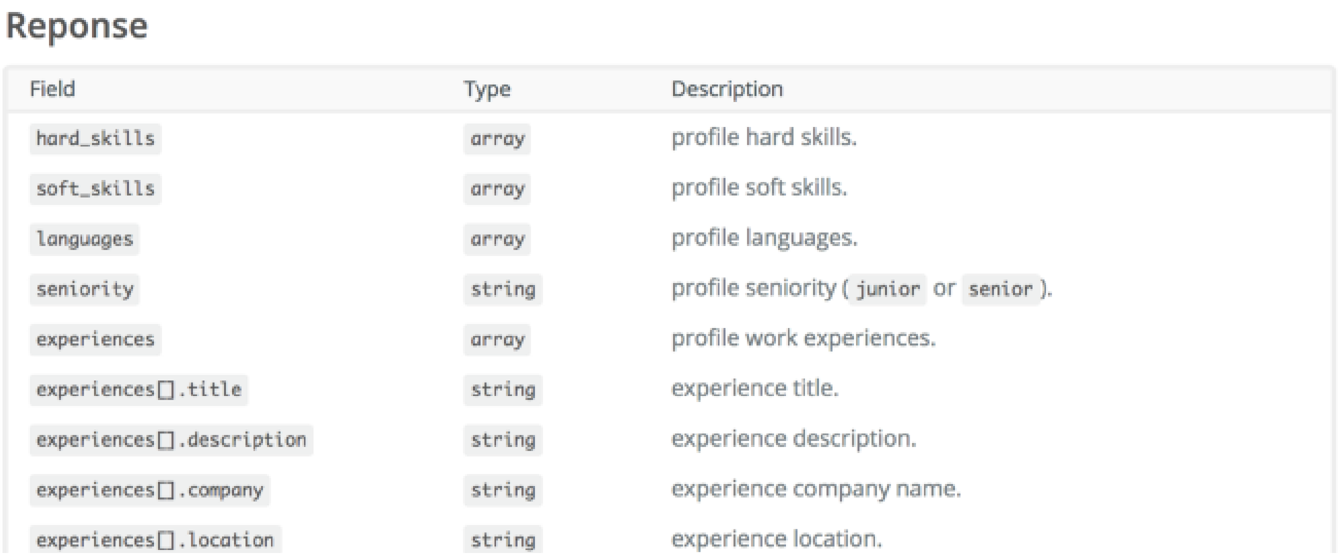
```
# data
import time
import random

rand_id     = random.randint(0, 99999)

```

```python
source_id   = "e67c715654bce8bfaa6605699026ec80d021be4d"
profile_ref = "id{}".format(rand_id)
timestamp   = time.time()
data        = {"source_id": source_id, "profile_reference": profile_ref, "timestamp_reception":

# file
import magic

mime = magic.Magic(mime=True)

file_path = ".test.pdf"
file_name = file_path.split('.')[-1]+profile_ref
files     = {"file": (file_name, open(file_path, 'rb'), mime.from_file(file_path))}

# request
import requests as r

url     = "https://www.riminder.net/sf/public/api/v1.0/profile"
headers = {'X-API-KEY': "ask_f06ac7631522b910878c543b9d05c19c"}
req     = r.post(url, data=data, files=files,  headers=headers)

# response
assert req.status_code == 201
req.json()
```

posting-resume.py hosted with ♡ by GitHub view raw

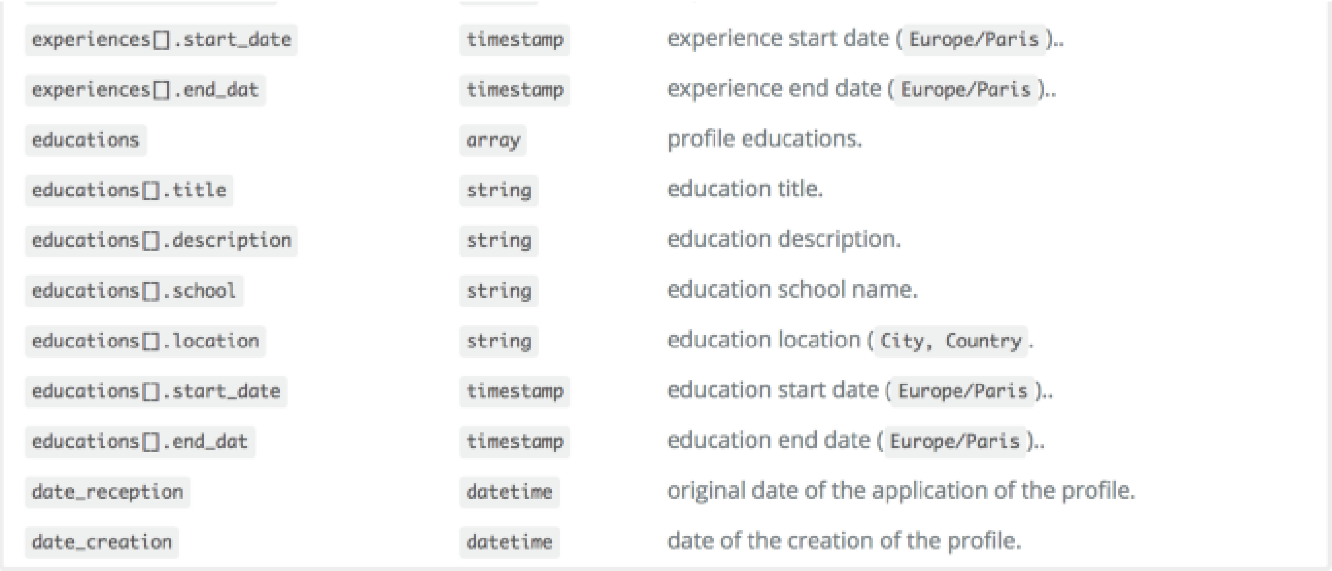
Python Example of how to post a resume on Segmentr (by riminder)

Here is the structure of the data that you'll get:



Reponse

Field	Type	Description
hard_skills	array	profile hard skills.
soft_skills	array	profile soft skills.
languages	array	profile languages.
seniority	string	profile seniority (junior or senior).
experiences	array	profile work experiences.
experiences[].title	string	experience title.
experiences[].description	string	experience description.
experiences[].company	string	experience company name.
experiences[].location	string	experience location.



experiences[].start_date	timestamp	experience start date (Europe/Paris)..
experiences[].end_dat	timestamp	experience end date (Europe/Paris)..
educations	array	profile educations.
educations[].title	string	education title.
educations[].description	string	education description.
educations[].school	string	education school name.
educations[].location	string	education location (City, Country .
educations[].start_date	timestamp	education start date (Europe/Paris)..
educations[].end_dat	timestamp	education end date (Europe/Paris)..
date_reception	datetime	original date of the application of the profile.
date_creation	datetime	date of the creation of the profile.

What's next for you?

Discover Segmentr Live

If you are interested to know more about Segmentr, you can book us for demo : https://riminder.net/book-us .

You can also visit our http://labs.riminder.net to see AI applied to HR in action.

Are you a Developer?

You can start now using our **self-service API** without any painful setups.
Get started in few minutes with our documentation:
https://developers.riminder.net/docs/can-it-parse-resumes

□ I agree to leave Medium.riminder.net and submit this information, which will be collected and used according to Upscribe's privacy policy.

If you enjoyed this article, it would really help if you hit recommend below :)

Follow us on Twitter @seivmouhidine , @othmaneizi1, and @riminderdotnet

If you want to work on AI + FAIRNESS, check our jobs page!

AI Resume Hrtech Recruiting HR

About Help Legal