Classification des CV avec NLP

Aude, Jamal, Patricia

SOMMAIRE

Description et analyse des données

Interprétation du modèle

Présentation des modèles

Conclusion

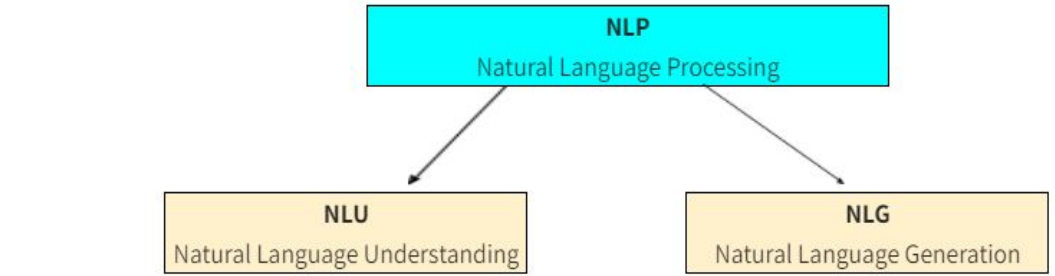
Contexte du projet

Embaucher les bons talents est un défi pour toutes les entreprises. Ce défi est amplifié par le volume élevé de candidats si l'entreprise est bien réputée. Dans une organisation de services typique, des professionnels possédant une variété de compétences techniques et d'expertise sont embauchés et affectés à des projets pour résoudre les problèmes des clients.

En règle générale, les grandes entreprises n'ont pas assez de temps pour ouvrir chaque CV. **L'idée du projet est d'utiliser des algorithmes d'apprentissage automatique pour la tâche de filtrage de CV.**

The Natural Language Processing

TALN (traitement automatique du langage naturel)



« compréhension » du texte

entrée texte

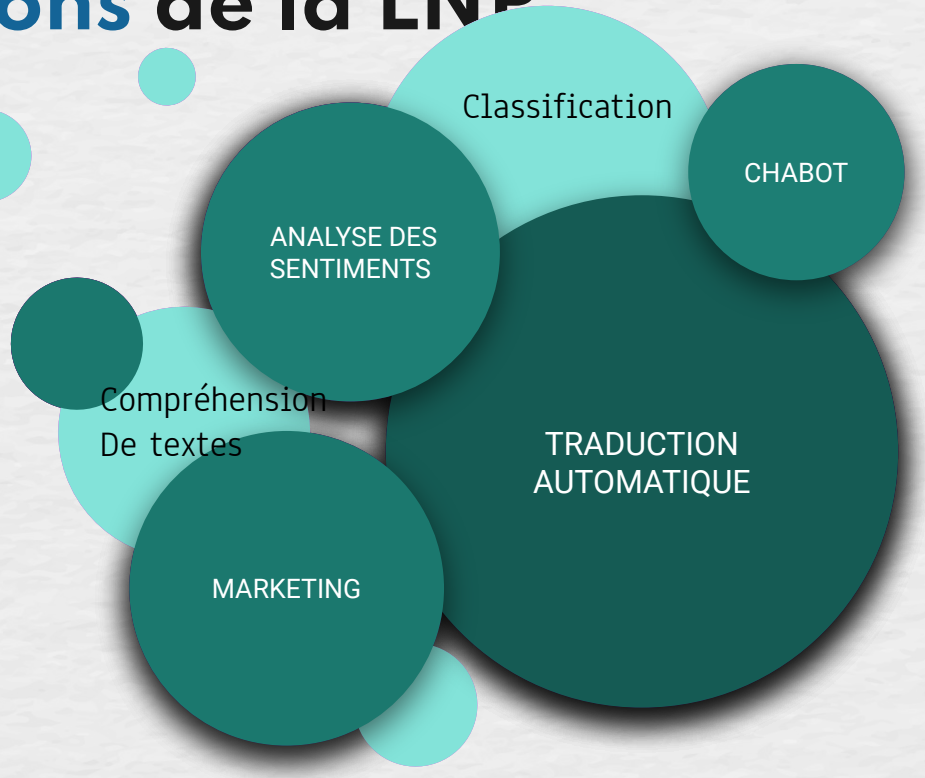
sortie données

« génération » du texte

données

construire des phrases cohérentes de manière automatique.

Applications de la LNP



-

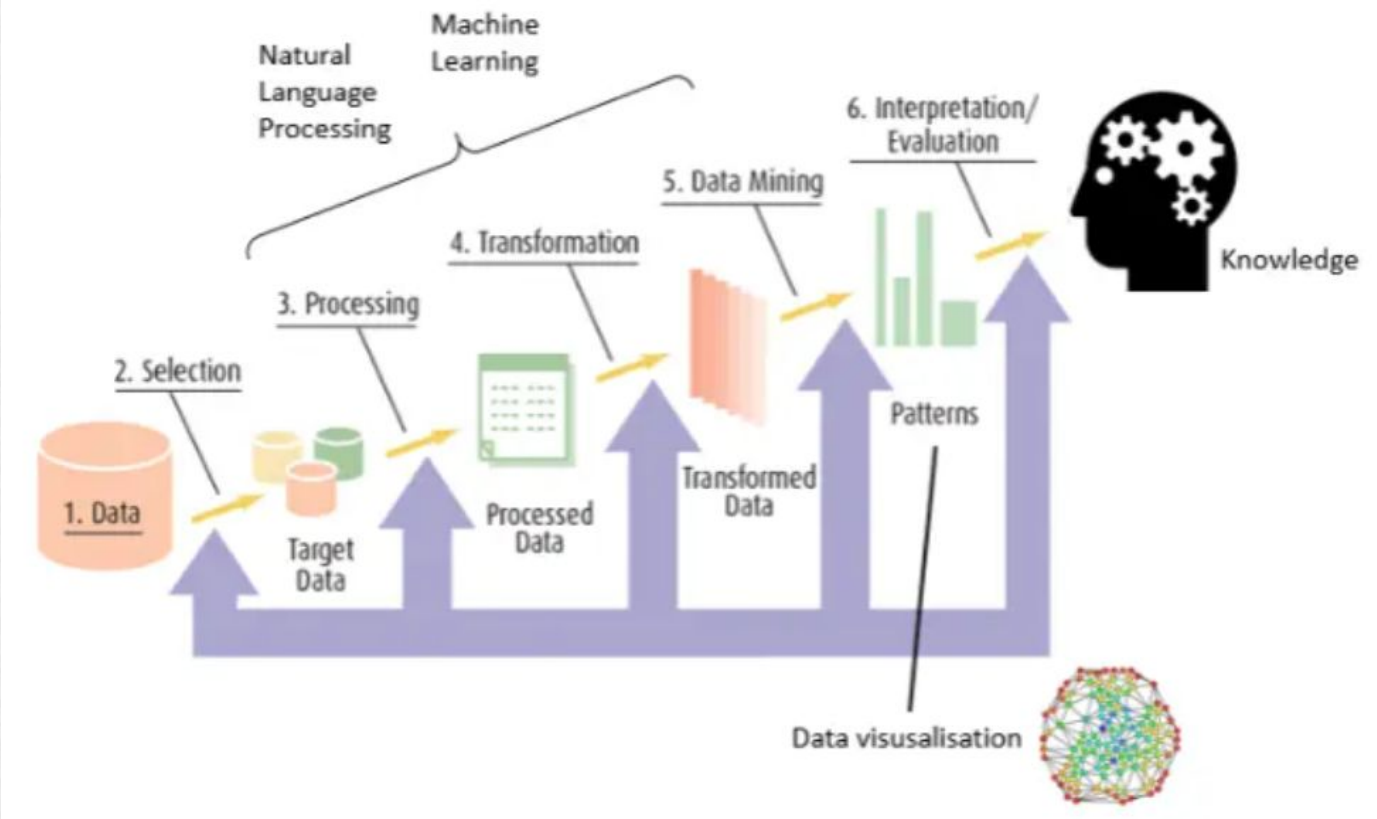
-

Principales méthodes utilisées de la NLP

- La **partie linguistique** = prétraiter et transformer les informations en données exploitables

- La **partie apprentissage automatique ou Data Science** qui porte à l'application de modèles de ML ou DL au jeu de données

-


Natural Language Processing
Machine Learning
6. Interpretation/ Evaluation
5. Data Mining
4. Transformation
3. Processing
2. Selection
Knowledge
Patterns
Transformed Data
Processed Data
Target Data
1. Data
Data visusalisation

01

Description et Analyse des données

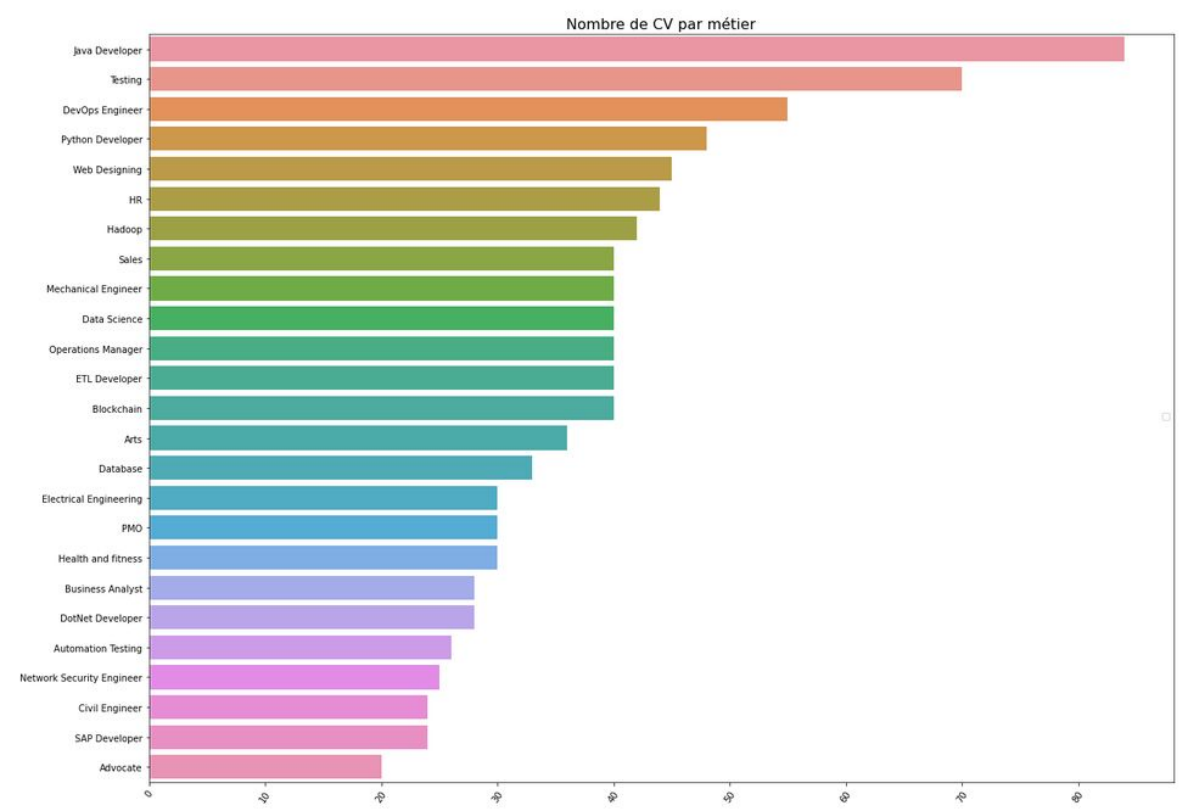
Description des données

	Category	Resume
0	Data Science	Skills * Programming Languages: Python (pandas...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...
2	Data Science	Areas of Interest Deep Learning, Control Syste...
3	Data Science	Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...
...	...	...
957	Testing	Computer Skills: â¢ Proficient in MS office (...
958	Testing	â Willingness to accept the challenges. â ...
959	Testing	PERSONAL SKILLS â¢ Quick learner, â¢ Eagerne...
960	Testing	COMPUTER SKILLS & SOFTWARE KNOWLEDGE MS-Power ...
961	Testing	Skill Set OS Windows XP/7/8/8.1/10 Database MY...

	Category	Resume
count	962	962
unique	25	166
top	Java Developer	Technical Skills Web Technologies: Angular JS,...
freq	84	18

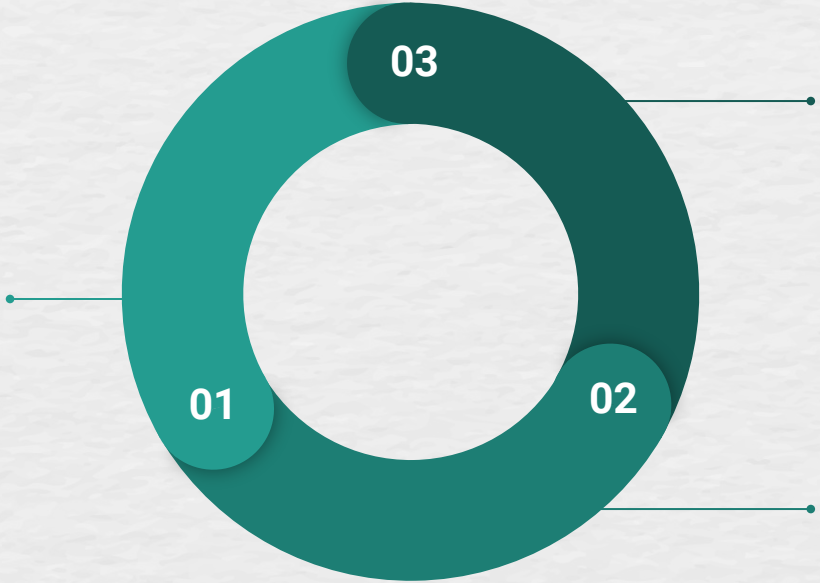
962 lignes - 2 colonnes

Nombre de cv par métier



Java Developer	84
Testing	70
DevOps Engineer	55
Python Developer	48
Web Designing	45
HR	44
Hadoop	42
Sales	40
Mechanical Engineer	40
Data Science	40
Operations Manager	40
ETL Developer	40
Blockchain	40
Arts	36
Database	33
Electrical Engineering	30
PMO	30
Health and fitness	30
Business Analyst	28
DotNet Developer	28
Automation Testing	26
Network Security Engineer	25
Civil Engineer	24
SAP Developer	24
Advocate	20

Prétraitement des données



Nettoyage des données
(suppression url, emoji, suppression des majuscules, des caractères spéciaux…)

Labellisation des données

Normalisation des données
Tokenisation
(« Vous », « trouverez », « en pièce jointe », « le document », « en question »)
Stemming
« trouverez » -> « trouv »
Lemmatisation
« trouverez » -> trouvez

Nettoyage des données et prétraitement de texte

'objectif : nettoyer le texte afin de faciliter l'apprentissage

	Category	Resume	clean
0	Data Science	Skills * Programming Languages: Python (pandas...	[skills, programming, languages, python, panda...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	[education, details, may, may, b, e, uit, rgpv...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	[areas, interest, deep, learning, control, sys...
3	Data Science	Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...	[skills, r, python, sap, hana, tableau, sap, h...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	[education, details, mca, ymcaust, faridabad, ...

Préparation des données

le tokenizer punkt de NLTK

	Category	Resume	clean	classe
0	Data Science	Skills * Programming Languages: Python (pandas...	[skills, programming, languages, python, panda...	6
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	[education, details, may, may, b, e, uit, rgpv...	6
2	Data Science	Areas of Interest Deep Learning, Control Syste...	[areas, interest, deep, learning, control, sys...	6
3	Data Science	Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...	[skills, r, python, sap, hana, tableau, sap, h...	6
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	[education, details, mca, ymcaust, faridabad, ...	6

Word2Vec attend des phrases uniques, chacune représente une liste de mots. (le format d'entrée est une liste de listes.)
Il est important de diviser phrase par phrase, pour une meilleure qualité de l'information avant de diviser en mots.

Nuage de mots

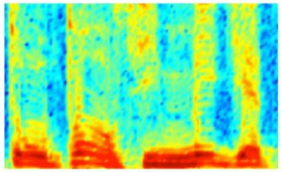
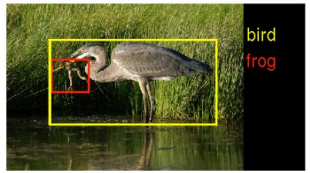
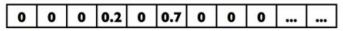
Objectif : faire apparaître visuellement les mots les plus utilisés


```
[('exprience', 3829),
 ('company', 3578),
 ('project', 3498),
 ('months', 3288),
 ('description', 3122),
 ('details', 3096),
 ('data', 2156),
 ('management', 1999),
 ('team', 1950),
 ('maharashtra', 1449),
 ('system', 1425),
 ('testing', 1349),
 ('year', 1344),
```

02

Présentation des modèles

Word2Vec - Le plongement des mots

AUDIO	IMAGES	TEXT
		
Audio Spectrogram	Image pixels	Word, context, or document vectors
DENSE	DENSE	SPARSE

Vecteurs de coefficients de densité spectrales.

Vecteurs de coefficients associées aux intensités de pixels.

word embeddings:

représentation d'un mot, ou un groupe de mots en vecteurs.

SENS = CONTEXTE

CONTEXTE -> VECTEUR

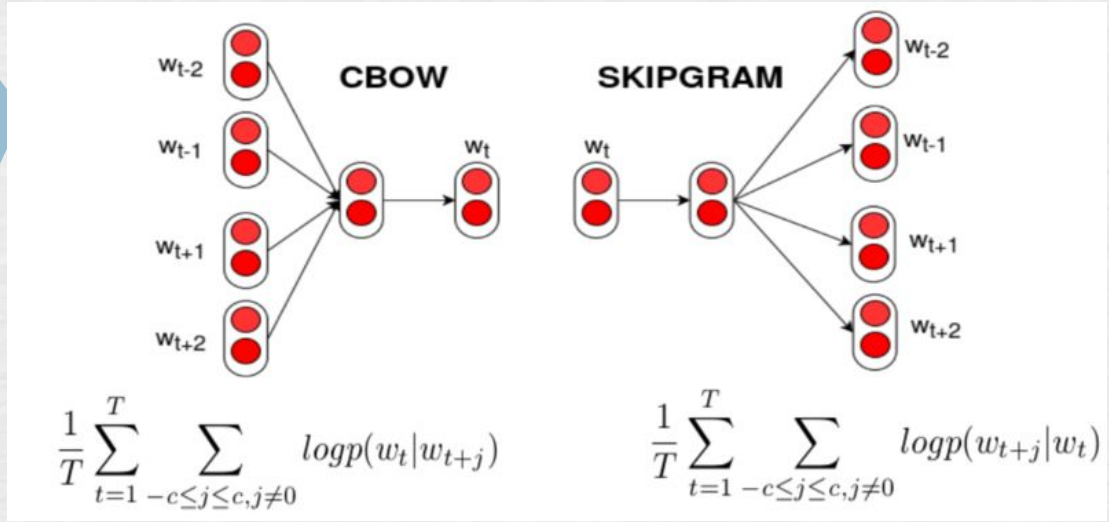
DISTANCE(VECT1, VECT2)

=DISTANCE(SENS1, SENS2)

Deux méthodes d'entraînement principales:

La **première** appelée « **Continuous Bag of Words** », qui entraîne le réseau de neurones pour **prédire un mot en fonction de son contexte**.

Dans la **seconde** méthode, on essaie de **prédire le contexte en fonction du mot.** C'est la technique du « **skip-gram** ».



Exemple de clusters

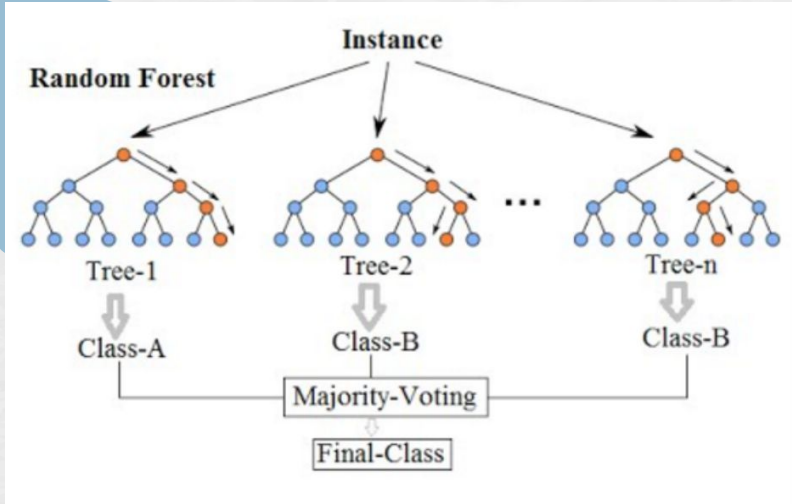
```
Cluster 0
['sending', 'emails', 'informing', 'deliveries']

Cluster 1
['created', 'maintained', 'repository', 'small']

Cluster 2
['front', 'called']

Cluster 3
['problems', 'appropriate', 'resolution', 'proactively']

Cluster 4
['building', 'relationship']

Cluster 5
['project', 'role', 'name', 'implementation', 'duration', 'main', 'detection', 'nestle', 'gateway']

Cluster 6
['skills', 'communication', 'ability', 'strong', 'learn', 'interpersonal', 'series', 'written', 'proficiency', 'inter', 'orient
ed', 'pressure', 'hardworking', 'strength', 'sound', 'verbal', 'foundry']

Cluster 7
['april', 'bsc', 'qualifications']

Cluster 8
['monitoring', 'maintenance', 'maintaining', 'troubleshooting', 'schedules', 'implementing', 'scheduling', 'needed', 'proxy',
'upgrading']

Cluster 9
['account', 'overseeing', 'logistics', 'dell', 'exports', 'receivables', 'johnson']
```

Random Forest

Un ensemble d'arbres de décisions entraînés individuellement.

Pour prédire une nouvelle valeur, on effectue la classification pour chaque arbre de cette forêt.

La forêt choisit la valeur ayant le plus de votes parmi tous ses arbres.



Créer notre forêt avec ***M** features*:

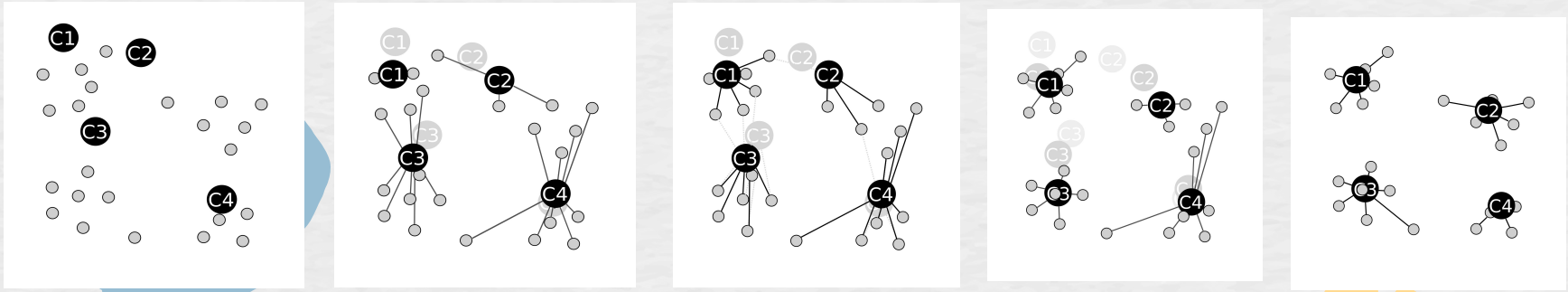
On choisit un nombre ***m* < *M*** tel qu'à chaque **nœud** de notre arbre on sélectionne **aléatoirement** ***m** features* parmi les **M** et le meilleur « **point de séparation** ».

La valeur **m** est gardée constante durant la création de la forêt aléatoire.

Le Sac de Centroïdes

L'algorithme **K-means** est un algorithme de partitionnement de données (clustering);

Il est capable d'identifier différents groupes homogènes de données au sein d'un ensemble hétérogène.



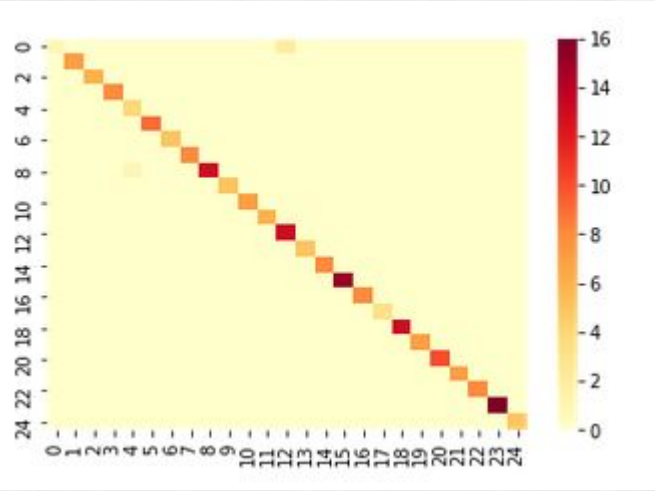
On crée des groupes de points en les "attachants" à la **centroïde** la plus proche;
Ensuite il ne reste plus qu'à répéter les étapes **d'assignation et de calibrage** jusqu'à ce que les groupes ne bougent plus.

03

Interprétation du modèle

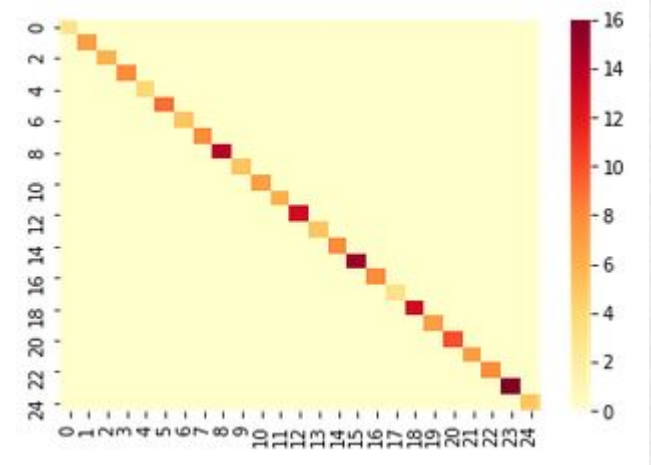
2 choix de modèles

random forest



Accuracy: 0.985

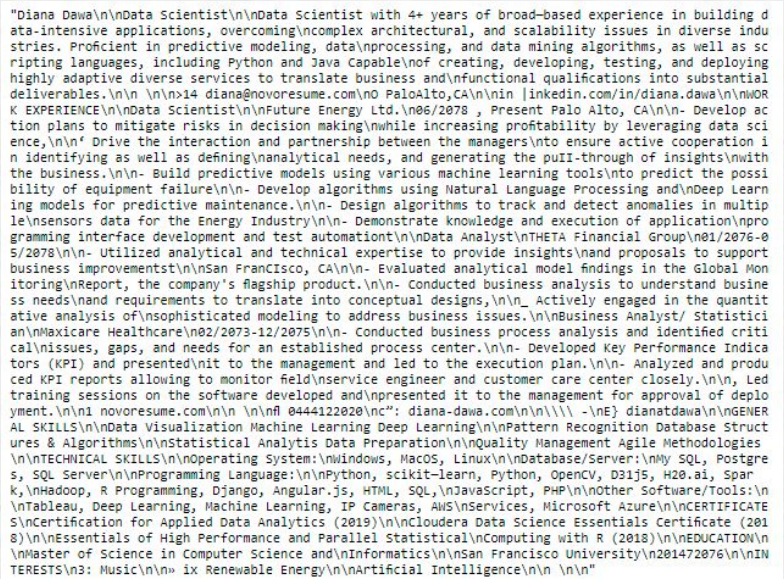
Bags of centroids



Accuracy: 1.0

Lecture des cv

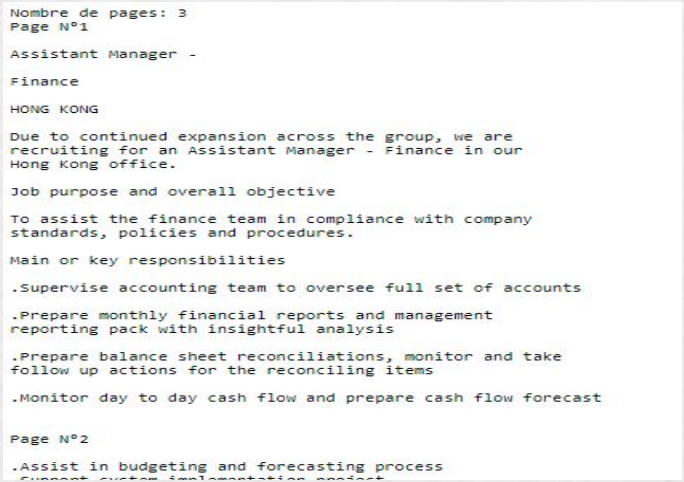
- Modèle en png -> tesseract-OCR



"Diana Dawa\n\nData Scientist\n\nData Scientist with 4+ years of broad–based experience in building d
ata-intensive applications, overcoming\ncomplex architectural, and scalability issues in diverse indu
stries. Proficient in predictive modeling, data\nprocessing, and data mining algorithms, as well as sc
ripting languages, including Python and Java Capable\nof creating, developing, testing, and deploying
highly adaptive diverse services to translate business and\nfunctional qualifications into substantial
deliverables.\n\n \n\n>14 diana@novoresume.com\nO PaloAlto,CA\n\nin |inkedin.com/in/diana.dawa\n\nWOR
K EXPERIENCE\n\nData Scientist\n\nFuture Energy Ltd.\n06/2078 , Present Palo Alto, CA\n\n- Develop ac
tion plans to mitigate risks in decision making\nwhile increasing profitability by leveraging data sci
ence,\n\n‘ Drive the interaction and partnership between the managers\nto ensure active cooperation i
n identifying as well as defining\nanalytical needs, and generating the puII-through of insights\nwith
the business.\n\n- Build predictive models using various machine learning tools\nto predict the possi
bility of equipment failure\n\n- Develop algorithms using Natural Language Processing and\nDeep Learn
ing models for predictive maintenance.\n\n- Design algorithms to track and detect anomalies in multip
le\nsensors data for the Energy Industry\n\n- Demonstrate knowledge and execution of application\npro
gramming interface development and test automationt\n\nData Analyst\nTHETA Financial Group\n01/2076-0
5/2078\n\n- Utilized analytical and technical expertise to provide insights\nand proposals to support
business improvementst\n\nSan FranCIsco, CA\n\n- Evaluated analytical model findings in the Global Mon
itoring\nReport, the company's flagship product.\n\n- Conducted business analysis to understand busine
ss needs\nand requirements to translate into conceptual designs,\n\n_ Actively engaged in the quantit
ative analysis of\nsophisticated modeling to address business issues.\n\nBusiness Analyst/ Statistici
an\nMaxicare Healthcare\n02/2073-12/2075\n\n- Conducted business process analysis and identified criti
cal\nissues, gaps, and needs for an established process center.\n\n- Developed Key Performance Indica
tors (KPI) and presented\nit to the management and led to the execution plan.\n\n- Analyzed and produ
ced KPI reports allowing to monitor field\nservice engineer and customer care center closely.\n\n, Led
training sessions on the software developed and\npresented it to the management for approval of deplo
yment.\n\n1 novoresume.com\n\n \n\nfl 0444122020\nc”: diana-dawa.com\n\n\\\\ -\nE} dianatdawa\n\nGENER
AL SKILLS\n\nData Visualization Machine Learning Deep Learning\n\nPattern Recognition Database Struct
ures & Algorithms\n\nStatistical Analytis Data Preparation\n\nQuality Management Agile Methodologies
\n\nTECHNICAL SKILLS\n\nOperating System:\nWindows, MacOS, Linux\n\nDatabase/Server:\nMy SQL, Postgre
s, SQL Server\n\nProgramming Language:\n\nPython, scikit–learn, Python, OpenCV, D31j5, H20.ai, Spar
k,\nHadoop, R Programming, Django, Angular.js, HTML, SQL,\nJavaScript, PHP\n\nOther Software/Tools:\n
\nTableau, Deep Learning, Machine Learning, IP Cameras, AWS\nServices, Microsoft Azure\n\nCERTIFICATE
S\nCertification for Applied Data Analytics (2019)\n\nCloudera Data Science Essentials Certificate (201
8)\n\nEssentials of High Performance and Parallel Statistical\nComputing with R (2018)\n\nEDUCATION\n
\nMaster of Science in Computer Science and\nInformatics\n\nSan Francisco University\n201472076\n\nIN
TERESTS\n3: Music\n\n» ix Renewable Energy\n\nArtificial Intelligence\n\n \n\n"

Notre model classe le cv de Diana en Data Science, soit dans la bonne catégorie

Modèle en pdf -> pumpler



Nombre de pages: 3
Page N°1

Assistant Manager -

Finance

HONG KONG

Due to continued expansion across the group, we are
recruiting for an Assistant Manager - Finance in our
Hong Kong office.

Job purpose and overall objective

To assist the finance team in compliance with company
standards, policies and procedures.

Main or key responsibilities

.Supervise accounting team to oversee full set of accounts

.Prepare monthly financial reports and management
reporting pack with insightful analysis

.Prepare balance sheet reconciliations, monitor and take
follow up actions for the reconciling items

.Monitor day to day cash flow and prepare cash flow forecast

Page N°2

.Assist in budgeting and forecasting process

Notre model classe le cv en HR, soit dans la bonne catégorie

04

Conclusion

avantages/inconvenients

Traitement d'un grand nombre de corpus de textes afin d'en extraire des probabilités et des règles de langage pour l'analyse de contenu. En plus de différencier le positif du négatif, cette technologie permet de distinguer l'intensité d'un propos.

Investissements de plus en plus importants dans cette technologie

Chaque **langue est unique**, nécessité de ré-entrainer les modèles pour chaque langue (complexe et couteux)

Nécessite **beaucoup de données**

Ambiguté des mots : un même mot peut avoir des sens différents (polysémie)

CONCLUSION

Word2Vec n'est pas la seule technologie existante (modèle prédictif, un réseau de neurones à anticipation qui apprend des vecteurs pour améliorer la capacité de prédiction).

D'autres solutions emergent :
GloVe (apprend des vecteurs ou des mots à partir de leurs informations de cooccurrence, c'est-à-dire à quelle fréquence ils apparaissent ensemble dans de grands corpus de texte),
ELMo (regarde la phrase entière avant d'assigner un embedding à chaque mot)
BERT(

La résolution de la plupart des problématiques de NLP passe par une étape de transformation de texte en vecteurs compréhensibles pour la machine.

Merci !

The end ...