La Rochelle Université

D'ici, on voit + loin !

Projets

Mise en situation professionnelle

Ahmed Hamdi

4 Topics ➔ 5 sujets

+ Classification (2 sujets)
+ Analyse de documents (1 sujet)
+ Fouille de processus (1 sujet)
+ Big Data (1 sujet)

Travail demandé

+ Rapport (20 pages maximum sans les références)

- Etat de l'art :
  travaux existants, étude comparative, synthèse …
- Méthodologie :
  Votre méthode, Code, screen shots …

+ Soutenance (~20 minutes de présentation + 10 minutes de questions)

- Partie théorique
- Partie pratique
- Préparer bien votre discours

Evaluation

	Rapport	Soutenance
Qualité de présentation	3 points	5 points
Qualité de rédaction	5 points	--
Réponses aux questions	--	3 points
Code	4 points	--

Sujet 1
Détection d'émotions

+ Objectif

- Assigner l'émotion convenable à un commentaire

Sujet 1
Détection d'émotions

+ Dataset

- 6000 commentaires
- 3 labels (anger, fear, joy)
- Format : csv

Lien : https://www.kaggle.com/datasets/abdallahwagih/emotion-dataset

+ Méthodologie

- Explorer et appliquer des méthodes de machine learning (SVM, Naïve Bayes, arbre de décision…)
- Diviser les données en ensembles de train et de test
- Evaluation des résultats et étude comparative
- Analyse de résultats

Sujet 2
Classification de texte (fake/real news)

+ Objectif

- Détecter la classe associée (**fake** ou **real**) à un texte

+ Dataset

- 21k real data
- 23k fake data
- Format : csv
- Data ➔ title, text, subject, date

Lien : https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Sujet 2
Classification de texte (fake/real news)

+ Méthodologie

- Analyse statistique de données (sujets/ date)
- Pré-traitement de données (stop words, encodage…)
- Application de méthodes de machine learning
- Division les données en ensembles de train et de test
- Evaluation des résultats et étude comparative
- Analyse de résultats

Sujet 3
Séparation de flux documentaires

+ Objectif

- Retrouver la structure initiale d'un flux de documents

+ Dataset

- 9 articles scientifiques fusionnés en 1 pdf
- 81 pages
- 1 fichier csv pour la structure

Lien : https://ao.univ-lr.fr/index.php/s/gRiAFar47iPzR4b

Sujet 3
Séparation de flux documentaires

+ Méthodologie

- Considérer le flux comme une séquence de paires de pages
- Extraire le texte à l'aide de l'OCR
- Identifier des descripteurs de continuité/rupture
- Appliquer un vote majoritaire

Descripteurs de continuité

> Pagination

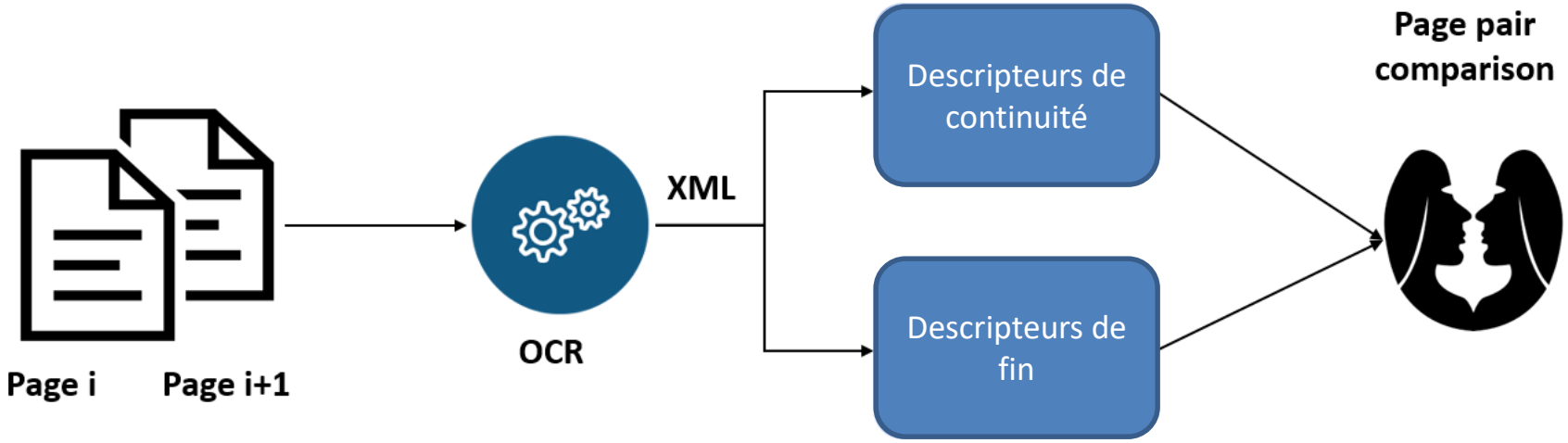
> Entités nommées

Descripteurs de rupture

> Signe de début : *adresses mail, abstract…*

> Signe de fin : *bibliographie*

Sujet 3
Séparation de flux documentaires



+ Résultats attendus

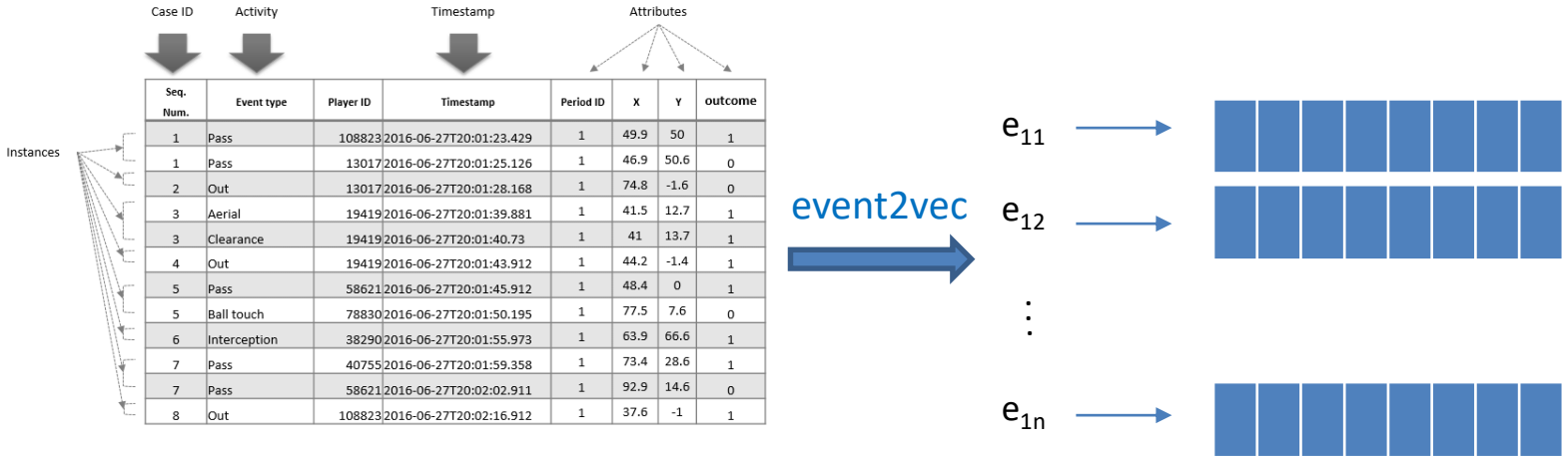
- Fichier csv (séquence de pages – séquence de valeurs booléennes)
- Évaluation : précision (documents, pages), erreurs (segmentation, fusion)

Sujet 4
activity2vec

Objectif

- Convertir des activités en vecteurs à l'aide d'un modèle word2vec que vous devez entraîner sur quelques millions de traces

Event log



Seq. Num.	Event type	Player ID	Timestamp	Period ID	X	Y	outcome
1	Pass	108823	2016-06-27T20:01:23.429	1	49.9	50	1
1	Pass	13017	2016-06-27T20:01:25.126	1	46.9	50.6	0
2	Out	13017	2016-06-27T20:01:28.168	1	74.8	-1.6	0
3	Aerial	19419	2016-06-27T20:01:39.881	1	41.5	12.7	1
3	Clearance	19419	2016-06-27T20:01:40.73	1	41	13.7	1
4	Out	19419	2016-06-27T20:01:43.912	1	44.2	-1.4	1
5	Pass	58621	2016-06-27T20:01:45.912	1	48.4	0	1
5	Ball touch	78830	2016-06-27T20:01:50.195	1	77.5	7.6	0
6	Interception	38290	2016-06-27T20:01:55.973	1	63.9	66.6	1
7	Pass	40755	2016-06-27T20:01:59.358	1	73.4	28.6	1
7	Pass	58621	2016-06-27T20:02:02.911	1	92.9	14.6	0
8	Out	108823	2016-06-27T20:02:16.912	1	37.6	-1	1

Sujet 4
activity2vec

Travail demandé

- Construction d'un « multidomain » event log
- Étude statistique (domaine, activités, variants)
- Apprendre un modèle word2vec sur les activités
- Calculer la similarité entre les activités

Links to event logs

Domaine	Liens de téléchargement
Education	MoocData
	Open Learning Analytics | OU Analyse | Knowledge Media Institute | The Open University
	SDP@US1.0 - Dataset for Student Dropout Prediction (figshare.com)
	https://github.com/riiid/ednet
	Educational Process Mining (EPM): A Learning Analytics Data Set - UCI Machine Learning Repository
Système de recommandation	https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset?select=events.csv
Librairie numérique	https://ao.univ-lr.fr/index.php/s/3tYqfkg4RrB7JrN
Aachen université	https://www.pads.rwth-aachen.de/cms/PADS/Forschung/Event-Logs/~bcbswu/-Un-Fair-Event-Logs/

Sujet 5
Big Data

PART -I- vers un big data responsable

+ Impact environnemental du big data

- 2% de la consommation de l'électricité mondiale
- Grandes émissions de consommation électrique
  - 80g de CO_2 par kwh en France
  - 460g de CO_2 par kwh en Allemagne

+ Objectifs

- Détailler l'impact du big data sur l'environnement
- Proposer des solutions envisageables (matérielles et logicielles)

Sujet 5
Big Data

PART -II- SKACK is the new SMACK

+ Spark Mesos Akka Cassandra Kafka ➔ S Kubernetes ACK

- Dans ce projet on se propose de construire une étude théorique pour éprouver l'intérêt de l'architecture SKACK vs SMACK.

+ Objectifs

- Décrire l'architecture SKACK
- Décrire l'architecture SMACK
- Etude comparative

Répartition des étudiants

Sujet	Etudiants
1	
2	
3	
4	
5	

Deadline et ordre de passage

Rapports - date limite de soumission : mercredi 08/11/2023 à 13h00

Présentation – ordre de passage : vendredi 10/11/2023

Sujet	Heure de passage
1	13h15
2	13h45
3	14h15
4	14h45
5	15h15

D’ici, on voit + loin !

La Rochelle Université

univ-larochelle.fr