

# Défis TextMine'23

EGC



emvista



# Objectifs des défis

Contribution tangible du groupe de travail TextMine

Proposer des nouveaux datasets

Proposer de nouvelles tâches

Mettre en regard l'état de l'art scientifique avec les défis proposés

# Principes des défis

Organisation des défis par des industriels et/ou des universitaires

- proposition d'un défi via un texte descriptif, idéalement un défi qui témoigne d'une difficulté/besoin des industriels
- proposition d'un dataset (déjà existant ou à créer)
- évaluation des résultats des participants

Equipe renouvelable chaque année

Lancement du défi avant l'été

Annonce des résultats pendant l'atelier TextMine (fin janvier/début février)

Publications dans le cadre de l'atelier TextMine sous l'égide de EGC

- sous forme d'actes en ligne...
- dans une collection HAL

Prix à la clé

**Proposez le défi #2 dès maintenant !**

**[textmine@emvista.com](mailto:textmine@emvista.com)**

# Défi TextMine #1



emvista



isahit.

## Organisateurs

Kévin Cousot, Emvista [kevin.cousot@emvista.com](mailto:kevin.cousot@emvista.com)

Cédric Lopez, Emvista [cedric.lopez@emvista.com](mailto:cedric.lopez@emvista.com)

Pascal Cuxac, CNRS-INIST [pascal.cuxac@inist.fr](mailto:pascal.cuxac@inist.fr)

Vincent Lemaire, Orange [vincent.lemaire@orange.com](mailto:vincent.lemaire@orange.com)

## Fournisseurs de données

Emvista, éditeur de logiciels basés sur du NLP

Isahit, solutions de labellisation éthique des données pour l'IA et le traitement des données

**Proposition du sujet :** Emvista

**Prix :** 300 euros

# Défi #1

## Reconnaissance d'entités d'intérêts dans les signatures d'e-mails

*par Kévin Cousot, Emvista.*



# Notre spécialité : l'intelligence artificielle appliquée au texte



Fondée en 2018  
à Montpellier



Dix personnes  
Équipe de recherche

**bpi**france



Nos partenaires  
institutionnels



# La mission de Prevyo Syndic Assistant ? Vous faire gagner du temps.

- 70%+ des e-mails sont traités automatiquement
- Récupérez une heure de temps par jour





Bonjour **Monsieur Garnier**,

Je suis **directeur commercial** de **Penbase** et vous contacte car nous souhaiterions avoir plus d'informations sur la solution **Prevyo** que vous proposez concernant la gestion des mails. Pour vous décrire notre situation, nous sommes une société d'environ **40 personnes** et passons **tous les jours** en moyenne **1 heure** à gérer nos mails.

Pourrions nous convenir d'un rendez-vous la **semaine prochaine** à **Montpellier** ?

Merci.

Cordialement,

**Nicolas Gagean Nicolas Gagean Directeur commercial**

**nicolas.gagean@penbase.com**

**Tél : +33 4 67 22 86**

**Penbase**

# Travaux précédents

- Organisation d'un stage de recherche
- Production d'un jeu de données et expérimentations de différents modèles d'apprentissage
- Tâche plus difficile que prévue et taille du jeu de données insuffisante

Modèles	Précision	Rappel	F-Score
Bert	0.49	0.21	0.29
Bi-LSTM	0.66	0.61	0.63
CRF	<b>0.8</b>	<b>0.78</b>	<b>0.79</b>
SVM	0.72	0.67	0.69

TAB. 3 – *Précision, Rappel and F-Score pour chaque modèle*

# Le défi

Le défi consiste à obtenir les meilleurs résultats sur la tâche de reconnaissance d'entités d'intérêt dans des signatures d'e-mails.

# Typologie

Classe	Définition
Human	Noms et prénoms des personnes qui figurent dans la signature. Ex : Martin Dupond
Organization	L'organisation à laquelle l'auteur de la signature est rattachée
Location	Bâtiments, bureaux, villes, numéros et noms des rues. Ex : 24 avenue Jean Jaurès, Montpellier
Phone_Number	Numéro de téléphone ou de fax. Ex : +33 01 23 45 67 89
Function	L'ensemble des fonctions assignées à la personne identifiée dans la signature. Ex : conseiller, assistant, professeur.
Email	Adresses e-mails. Ex : martin.dupond@bboite.com
Url	URL. Ex : www.anywaythewall.com
Social_Network	Nom des réseaux sociaux. Ex : LinkedIn, Facebook, Twitter.
Reference_User	Identifiant d'une personne ou d'une organisation sur un réseau social. Ex : @Kalypso.immobilier
Reference_Code_Postal	Code postal. Ex : 34080.
Project	Projet dans lequel la personne est impliquée. Ex : Comm & Partenariats, Master 2 Informatique, Direction de l'innovation.
Reference_CEDEx	Courrier d'entreprise à distribution exceptionnelle. Ex : Cedex 05
Reference_CS	Course spéciale. Ex : CS 39521.

Table 1: Classes annotées dans les jeux de données.

# Les données

## **Jeu de données authentique (JDA) :**

un jeu de données composé de signatures authentiques pseudonymisées

## **Jeu de données réaliste (JDR) :**

un jeu de données composé de signatures réalistes construites manuellement

## **Jeu de données factice (JDF) :**

un jeu de données composé de signatures factices créées automatiquement

# Constitution des jeux de données

## Jeu de données authentique (JDA)

- Données authentiques récoltées par formulaire web de “don”
- Pseudonymisation (manuelle + automatique)
- 606 signatures récoltées en 6 mois

Signature:

Mary Margho  
Directeur Général  
Téléphone : +33.(0)1.52.62.32.65  
6 Rue Jean-Paul Montagne  
75001 Paris  
[www.saveprogramma.com](http://www.saveprogramma.com)

Signature pseudonymisée :

Anna Dupont  
Directeur Général  
Téléphone : 01.55.52.12.96  
72, Rue Paul-Marie L'abbé  
34090 MONTPELLIER  
[www.anywaythewall.com](http://www.anywaythewall.com)

# Constitution des jeux de données

## Jeu de données réaliste (JDR)

- Produit par Isahit, plateforme de labellisation éthique des données pour l'IA
- Contraintes basées sur l'observation des données authentiques
  - nombre moyen de tokens
  - distribution des entités
- Validation
  - vérification automatique des contraintes imposées sur l'utilisation des classes
  - examen manuel d'un échantillon de 50 signatures
  - plusieurs aller-retours
- 473 signatures produites



# Constitution des jeux de données

## Jeu de données factice (JDF)

- Génération automatique
  - patrons
  - Fakenamgenerator : outil en ligne pour la génération de fausses identités
  - heuristiques pour altérer les entités, changer leur format
- Classes absentes :
  - Project, Url, Reference\_User, Reference\_CEDEx, Reference\_CS, Function
- 500 signatures générées

# Constitution des jeux de données

Classe	Quantité d'annotations	Classe	Quantité d'annotations	Classe	Quantité d'annotations
Human	1196	Human	971	Human	1023
Organization	1537	Organization	1023	Organization	943
Location	2680	Location	2533	Location	2150
Phone_Number	688	Phone_Number	473	Phone_Number	371
Function	1449	Function	567	Function	0
Email	344	Email	297	Email	371
Url	303	Url	227	Url	0
Social_Network	28	Social_Network	18	Social_Network	0
Reference_User	11	Reference_User	9	Reference_User	0
Reference_Code_Postal	349	Reference_Code_Postal	269	Reference_Code_Postal	367
Project	124	Project	98	Project	0
Reference_CEDEX	146	Reference_CEDEX	150	Reference_CEDEX	0
Reference_CS	33	Reference_CS	56	Reference_CS	0

Table 2: Nombre d'annotations dans JDA.

Table 3: Nombre d'annotations dans JDR.

Table 4: Nombre d'annotations dans JDF.

# Constitution des jeux de données

## Annotation

- Tokenisation commune :
  - découpage sur les espaces
  - certains caractères de ponctuation : , ? ! ' ( )
- Exceptions (non tokenisés) :
  - Phone\_Number
  - Url
  - Email
  - Reference\_User

# Constitution des jeux de données

## Format

- JSON
- Signature
  - *identifiant* : un identifiant numérique unique
  - *texte* : le texte brut
  - *annotations* : séquence d'annotations
- Une annotation
  - *forme* : la forme du token
  - *label* : le label de la classe associée
  - *begin / end* : les index début (inclus) / fin (exclus)

```
1  [
2  {
3    "identifiant" : 0,
4    "texte" : "Faustin Dupont",
5    "annotations" : [
6      {
7        "forme" : "Faustin",
8        "label" : "Human",
9        "begin" : 0,
10       "end" : 7
11      },
12      {
13        "forme" : "Dupont",
14        "label" : "Human",
15        "begin" : 8,
16        "end" : 14
17      }
18    ]
19  }
20 ]
```

# Évaluation

- Tâche : reconnaissance d'entités d'intérêt dans les signatures d'e-mails
- Tous les systèmes sont les bienvenus
  - symboliques
  - à base de connaissances
  - apprentissage...
- Données :
  - entraînement : JDF + JDR
  - évaluation : JDA (rendu public le jour de la remise du prix)
  - [https://github.com/Emvista/Challenge\\_TextMine\\_2023](https://github.com/Emvista/Challenge_TextMine_2023)
- Métrique de classement : F1

# Participation

- 5 équipes inscrites
- 4 participants
- 4 articles reçus

Merci à tous les participants  
pour leur critiques constructives  
à l'égard du dataset fourni !

**EDF**

Philippe Suignard

Meryl Bothua

**Univ. Caen**

Emmanuel Giguet

Tanguy Gernot

**Linkmedia (IRISA)**

Nicolas Fouqué

Cyrielle Mallart

Duc-hau Nguyen

Hugo Thomas

**Nextino**

Asceline Goudjo

Maelle Brassier