Participation d'EDF R&D à EGC 2023 TextMine

P. Suignard, L. Hassani, M. Bothua

Atelier du 17/01/2023

Plan de la présentation

- EDF R&D
- Pourquoi participer à ce concours ?
- Les méthodes utilisées
- Résultats obtenus

EDF R&D

- Structure

✓EDF R&D au service de toutes les entités du groupe EDF

✓Basée à Saclay, le nouveau centre de recherche

✓Environ 2000 personnes

EDF R&D

- Travaux sur le texte
  - ✓En appui aux différents métiers : EDF Commerce, Hydraulique, Eolien, Nucléaire, Enedis, RH, IT…
  - ✓En permanence : ~8 personnes, 1 doctorant, 2 stagiaires
  - ✓Thèmes : classification, clustering, orthographe, annotations, web sémantique, résumé, anonymisation de données, détection de nouveauté, plongements mots/documents…
  - ✓Sujets : mails, réclamations, comptes-rendus d'interventions, documents techniques, manuscrits anciens, conversations téléphoniques, réseaux sociaux, chatbots…
  - ✓Type de prestations : veille, développement, conseil, méthode, étude.

Pourquoi participer à ce concours ?

- Contexte :
  - ✓Nous travaillons à la détection de données à caractère personnel afin d'en assurer la pseudonymisation et ainsi respecter le RGPD
  - ✓Donc détecter des entités d'intérêts dans les signatures d'e-mails était intéressant pour nous
- Participer à ce défi est l'occasion d'échanger sur des méthodes de traitement automatique du langage avec des universitaires et des industriels
- Contribuer à la communauté
- Reconnaissance interne
- Les résultats contribuent directement à EDF Commerce et à d'autres entités du groupe EDF

Méthode 1

- Approche :
  - ✓Extraction de features
  - ✓Entraînement d'un classifieur
- Features
  - ✓Par token :
    - Forme : longueur, nb blancs, nb points, dem par maj, tout en maj, numérique, tiret, arobase
    - Contenu : cedex, RS, localisation, fonction, organisation, projet, mail, adresse (1 xx 75100 YY)
    - Position : n° ligne, longueur de ligne, nb mots dans ligne, pos dans ligne
  - ✓Nb lignes
  - ✓20 features pour le mot précédent
  - ✓20 features pour le mot suivant
- Classifieur : Random Forest

1. Faustin Chabot
2. Adresse : 19 rue Descartes 94370 Sucy-en-Brie (France)
3. Cedex 9 CS 12468
4. Data Engineer / Algorithm XZ Project
5. faustinchabot@teleworm.com / Tel : +33 0134354919
6. Linkedin : https ://fr.linkedin.com/in/fauchab
7. Teleworm France
8. teleworm.france.com

Méthode 2

- FLAIR
  - ✓Framework Open Source proposé par Zalendo
  - ✓basé sur la librairie PyTorch
  - ✓Embeddings positionnels (WordEmbeddings) + CharacterEmbeddings + embeddings contextuels (FlairEmbeddings)
  - ✓Couche de CRF

Résultats

- Run :
  - ✓Run 1 : méthode 1 entraînée sur le JDR ;
  - ✓Run 2 : méthode 1 entraînée sur le JDR et le JDF;
  - ✓Run 3 : méthode 2 (Flair).

	Run1	Run2	Run3
Accuracy	68,17	68,32	49,64
Précision	69,50	69,78	72,11
Rappel	68,17	68,32	49,64
F-mesure	67,21	67,24	42,77

Résultats

- Analyses

✓Forte chute des scores (/apprentissage), en partie due au formatage des données du JDA, beaucoup moins propre que JDR

✓Très faible écart entre run1 et run2, apport du JDF très faible

✓Certaines catégories très correctes (Human, Location, url ou phone_number), d'autres plus difficiles (function, organization ou project)

✓Pour FLAIR, le modèle semble avoir été entraîné sans prendre en compte le contexte. Manque de temps