Reconnaissance de contexte

Nour ZEROUALI

Contexte

- Classification du contexte d'un texte

Ces champs d'application peuvent être séparés en deux branches, l'une, de nature théorique, qui concerne la définition de concepts et modèles, et l'autre, de nature pratique, qui s'intéresse aux techniques concrètes de mise en œuvre. Certains domaines de l'informatique peuvent être très abstraits, comme la complexité algorithmique, et d'autres peuvent être plus proches d'un public profane. Ainsi, la théorie des langages demeure un domaine davantage accessible aux professionnels formés (description des ordinateurs et méthodes de programmation), tandis que les métiers liés aux interfaces homme-machine sont accessibles à un plus large public.

Il faudra prédire le sujet de ce paragraphe (informatique dans cette exemple)

Scraping

- ## Objectif

  Récupérer des données textuelles des sites proposés par google avec un mot clé en entrée :
  - **les mots entrant ils s'agissent de notre label**
  - **Informatique, hotel, science et technologie.**
  - **Scraper le texte brute en format .txt**

- ## Méthode

  - **Collecter les liens des sites web proposés.**
  - **Scraper le texte brute de chaque site web**
  - **Stocker les données**

- ## API utilisé

  - **BeautifulSoup**
  - **Selenium**

DataSet

- ## Pré-traitement
  - Suppression des caractères spéciaux
  - Suppression des mots vide
  - Normaliser le texte
  - Stocker les données en format CSV

- ## API utilisé
  - nltk
  - regex
  - pandas
  - csv

DataSet

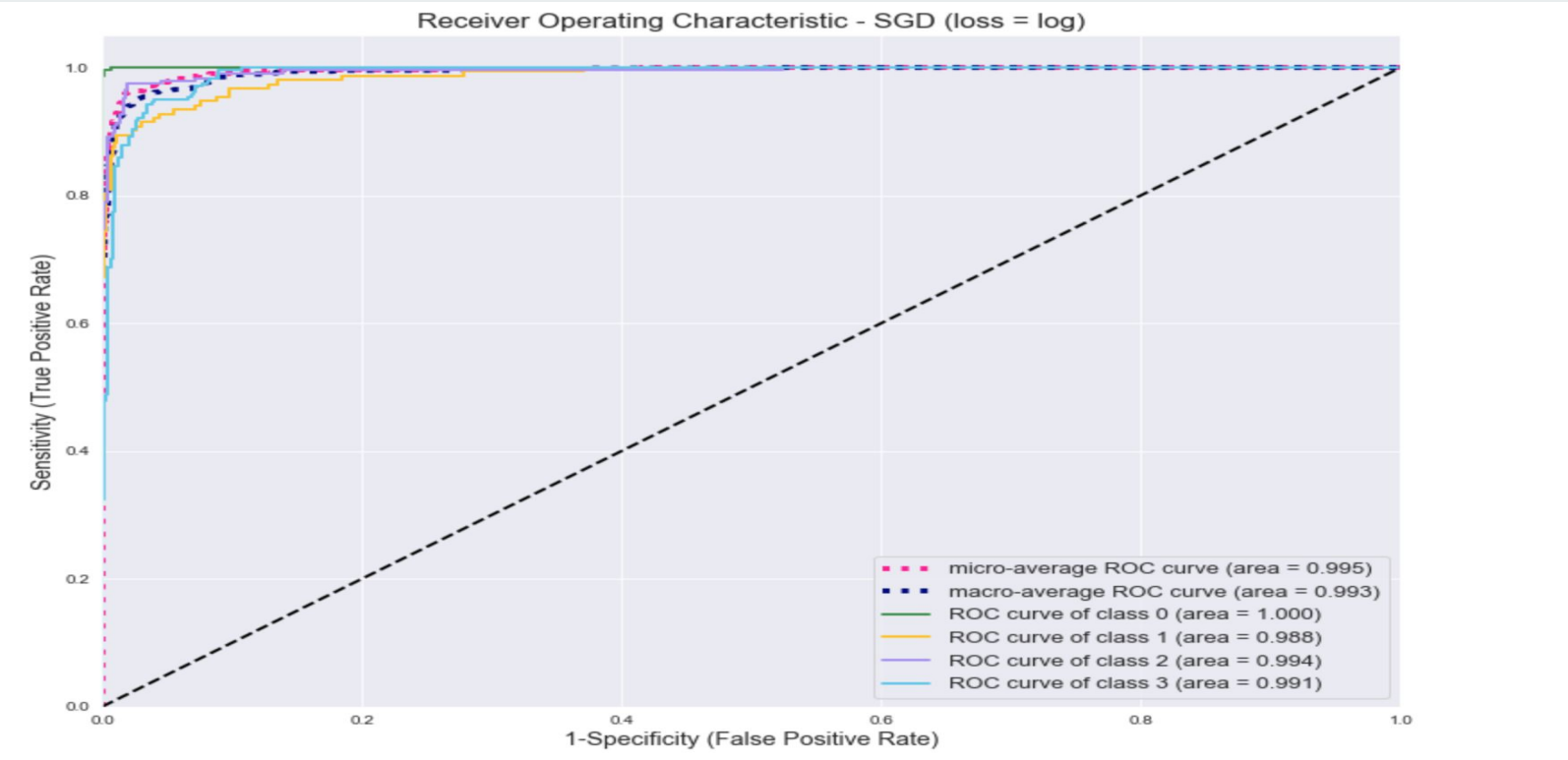
description	label	label_num
créer vidéos faire profiter savoir expérience cliquez jaquette accéder formation clics informatique cédric vidéos partage objets amusants utiles trouve amazon découvrirez exemple multiprise	informatique	1
peut incité analyse comparant phénomènes commun caractères peuvent définir esprit aide pourrait idéalement reconstruire idées genre peut vérifier physique tenter mathématique induire all	science	2
agents voyage bénéficierez commissions avantageuses remises spéciales ensemble tarifs avantages entreprise voyage affaires programme meliá hotels international réservez séjours hôtels l	hotel	0
extraordinaire photographie infrarouge centre galaxie permettant distinguer régions ainsi position trou noir supermassif heures apporter aide australie bungie lancé campagne destiny heures	technologie	3
aujourd hui animé commerces atypiques restaurants équipe hôtel attend faire vivre expérience unique personnalisée niché ancien quartier pêcheurs emplacement idéal visiter ville découvrir c	hotel	0
expertise remarquable bénéficier disponibilité norme équipe images technologie amélioré grandement réseau existant fourni conseils solutions importantes grandement aidé période crise ima	technologie	3
moustapha bons conseils séjour versailles merveilleux chambre extrêmement propre équipé conforme réservation faite direct hôtel petit déjeuner copieux frais justine utilisons cookies offrir r	hotel	0
tamisés directionnels mise valeur objets décoration led type smd surface mounting device utilisent composants fins épaisseur montés surface circuit imprimé composant désigné forme smd	technologie	3
lille chercheurs laboratoire océanologie géosciences log pris large novembre rejoindre archipel kerguelen grande expédition baptisée enviker objectifs étudier cellulaires observer effets produ	science	2
analyse algorithme αβ logiciel entièrement gardant séparation logique métier aspect robotique implentant design pattern mvc aide fonctions projets encadré nicolas delestre robot segway réa	informatique	1
ttc partir disponibilités soumis conditions voir conditions applicables offre accessibles site edreams droits réservés vacaciones edreams société soumise droit espagnol inscrite registre comn	hotel	0
sciences exactes pur raisonnement déduction sciences nature inductives partir axiomes simplicité évidents esprit combler conssructions nécessaires priori réduire proche proche complexes	science	2
spéculation wundt appelle sciences normatives distinction découle définition science idée peuvent ordonner logiquement abstraites rejet métaphysique philosophie proprement dite idée form	science	2
agents voyage bénéficierez commissions avantageuses remises spéciales ensemble tarifs avantages entreprise voyage affaires programme meliá hotels international réservez séjours hôtels l	hotel	0
garance pastel indigo graisses colles suif chandelles base déchets animaux os acides alcools produits fermentation vinaigre cuirs fourrures fibres laine lin coton chanvre locomotives machine	technologie	3
is supported by for scientific research by japan society for the promotion of science the national institutes of natural sciences astrobiology center japan the joint research program of the instit	science	2
fontaine jacques offenbach jules verne marc chagall artistes créateurs entrepreneurs célèbres véritable voyage plongera coeur histoire puis allez découvrir paris hôtel parcourez ville lumières	hotel	0
démontrer dieu existe remontant causes raisonnement reste croyance repose émotion splendeur univers revient prendre émotion axiome peut échapper axiomes raison pure premier connaiss	science	2
automobile kilomètres parcourus données échangées stockées largement supérieure gains unitaires trafic internet multiplié vaut amélioration dizaines points efficacité énergétique solution te	technologie	3
and the amount of acceleration deceleration the rider puts in based the slope measurement combined with the other factors the system is able to automatically adjust motor output accelerom	technologie	3
effectuer travaux réparation amélioration petit robot baptisé stimey utilise intelligence artificielle développé cadre projet européen horizon itu telecom world tient jusqu septembre budapest h	technologie	3
avenir métaphysique etc gallimard découvertes bacon oeuvres descartes discours méthode passim ampère philosophie sciences comte cours passim claude bernard médecine expérimentale	science	2
savez comment trouver hôtel bon marché panique existe astuces infaillibles permettront trouver bon hôtel prévu passer amoureux hôtel huppé dernier moment confronté situation contraignar	hotel	0

Machine Learning - Scikit-learn

Tester sur plusieurs algorithmes de classification grâce au Scikit-learn : K Nearest Neighbor, Gaussian Naive Bayes, AdaBoost, Decsision Tree, Random Forest, Stochastic Gradient Descent et Dummy

	model_name	accuracy_score	precision_score	recall_score	f1_score
1	Stochastic Gradient Descent	0.94	0.94	0.93	0.93
2	Random Forest	0.94	0.93	0.93	0.93
6	K Nearest Neighbor	0.94	0.93	0.93	0.93
3	Decsision Tree	0.90	0.89	0.89	0.89
4	AdaBoost	0.88	0.88	0.87	0.87
5	Gaussian Naive Bayes	0.60	0.69	0.63	0.59
0	Dummy	0.26	0.25	0.25	0.25

Coubre ROC


Receiver Operating Characteristic - SGD (loss = log)
Sensitivity (True Positive Rate)
1-Specificity (False Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
micro-average ROC curve (area = 0.995)
macro-average ROC curve (area = 0.993)
ROC curve of class 0 (area = 1.000)
ROC curve of class 1 (area = 0.988)
ROC curve of class 2 (area = 0.994)
ROC curve of class 3 (area = 0.991)

Conclusion et perspectives

- Compression de langage naturel
- Traduction plus précis