



Robots MALI

Le Centre National Collaboratif de l'Éducation en Robotique

Un Premier Cours sur le Traitement Automatique du Langage Naturel

Thème 1 : Multi-Classification de Textes

M. Leventhal, Directeur de RobotsMali
30 Octobre – 17 Décembre 2019



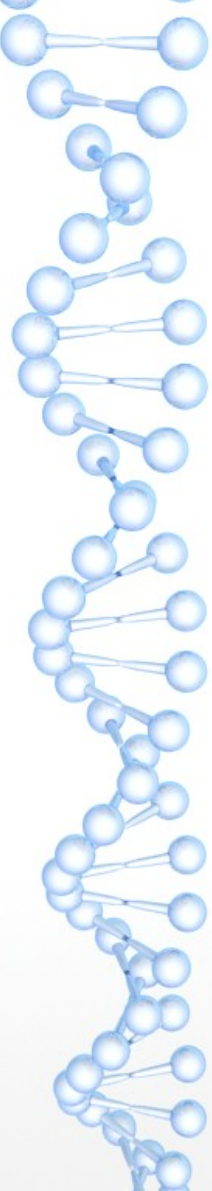


Préalablement ...

- Le cours ne sera pas plus difficile qu'il faut pour l'achever avec une connaissance utile.
- Le cours est technique. Les personnes qui n'ont pas l'habitude de voir les choses ordinaires par biais des maths et du logique vont s'ennuyer affreusement
- Il n'est pas nécessaire d'avoir programmé en Python, mais il sera nécessaire de programmer en Python du premier jour.
- Un ordinateur est requis. Je doute que moins que 8G de la RAM suffira et moins qu'un Core 5 sera, par fois, pénible. Nous allons faire le travail sur nos ordinateurs, mais une connexion pour télécharger les données et les exercices chaque semaine sera un peu incontournable.
- Je sais que beaucoup de vous sont surchargés avec les études et les autres responsabilités. Je compatis. Mais, une personne en train de dormir ou d'être assoupie décourage les autres et me décourage. Vous serez prié de trouver un meilleur endroit pour le repos.

Projet 1 : Multi-Classification de Textes

Autoclassification de questions de StackOverflow.com




stackoverflow Products Customers Use cases

Episode #125 of the Stack Overflow podcast is here. We talk Tide Club and mechanical keyboards. [Listen now](#)

Home

PUBLIC


 Stack Overflow

Tags

Users

Jobs

TEAMS [What's this?](#)

 First 25 Users Free

Search Results

[Advanced Search Tips](#) [Ask Question](#)

Results for nlp

[Search](#)

500 results

[Relevance](#) [Newest](#) [More ▾](#)

164 votes

[Q: Java Stanford NLP: Part of Speech labels?](#)

The Stanford **NLP**, demo'd here, gives an output like this: Colorless/JJ green/JJ ideas/NNS sleep/VBP furiously/RB ./.. What do the Part of Speech tags mean? I am unable to find an official list. Is ...

[java](#) [nlp](#) [stanford-nlp](#) [part-of-speech](#)

asked Dec 2 '09 by Nick Heiner

9 answers

21 votes

[Q: NLTK vs Stanford NLP](#)

I have recently started to use NLTK toolkit for creating few solutions using Python. I hear a lot of community



Données

Train

Entraînement. Questions avec les étiquettes de classification correctes.

Validate

Validation des résultats après entraînement avec les étiquettes de classification corrects.

Test

Pour le vrai ... les questions crues pour étiquetage



Environnement de Travail

- Compte de Github (github.com)
- Client Github
- Anaconda Distribution (anaconda.com/distribution/)
 - Python 3.7
 - Jupyter Notebook
 - Libraries
 - Numpy — scientific computing / calcul scientifique
 - Pandas — data structures and analysis / structure et analyse de données
 - scikit-learn — data mining and analysis / exploration et analyse de données
 - NLTK — Natural Language Toolkit / Traitement du langage naturel
- Editeur de texte (ex. Atom - atom.io/)



Comment télécharger des données à votre environnement Colab pour les utiliser avec votre Jupyter Notebook ?

Deux méthodes :

Clonez le dépôt BamakoNLP

Tout passe dans le Cloud sans utilisation de votre connexion. Vous téléchargerez tout le dépôt, mais normalement, c'est rapide.

Copiez les données de votre ordinateur

Vous utiliserez votre connexion pour télécharger les fichiers à votre environnement Colab. Vous téléchargerez juste les fichiers nécessaires, mais si les fichiers sont grands, il sera lent.

Clonez BamakoNLP



```
!git clone https://github.com/rusyazik/BamakoNLP.git  
%cd ../BamakoNLP/ClassificationTextesMultiEtiquette/
```



```
!git clone https://github.com/rusyazik/BamakoNLP.git  
%cd ../BamakoNLP/ClassificationTextesMultiEtiquette/
```



```
Cloning into 'BamakoNLP'...  
remote: Enumerating objects: 107, done.  
remote: Counting objects: 100% (107/107), done.  
remote: Compressing objects: 100% (87/87), done.  
remote: Total 107 (delta 22), reused 94 (delta 14), pack-reused 0  
Receiving objects: 100% (107/107), 4.96 MiB | 12.93 MiB/s, done.  
Resolving deltas: 100% (22/22), done.  
/content/BamakoNLP/ClassificationTextesMultiEtiquette
```




Copiez-les de votre ordinateur

```
[3] from google.colab import files

!mkdir data
%cd data
uploaded = files.upload()
%cd ..
```

```
from google.colab import files
```

```
!mkdir data
%cd data
uploaded = files.upload()
%cd ..
```

/content/data

Sélect. fichiers Aucun fichier choisi

Cancel upload

/content

« ClassificationTextesMultiEtique... » data

Rechercher dans : data

Organiser Nouveau dossier

Nom	Modifié le	Type
test.tsv	13/10/2019 01:12	Fichier TSV
text_prepare_tests.tsv	13/10/2019 01:12	Fichier TSV
train.tsv	13/10/2019 01:12	Fichier TSV
validation.tsv	13/10/2019 01:12	Fichier TSV

Week2

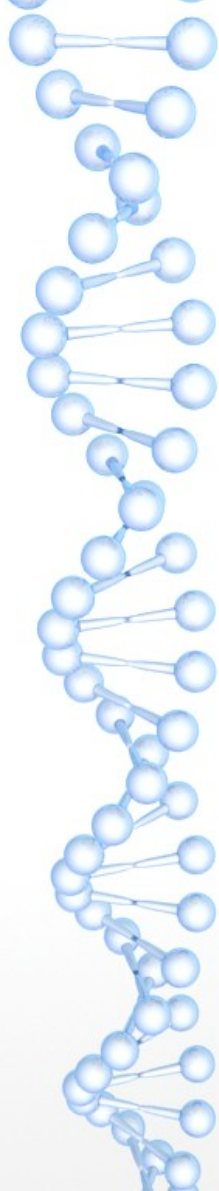
Ce PC

Bureau

Documents

Images

Musique



Sélect. fichiers 4 fichiers

- **test.tsv**(n/a) - 1041379 bytes, last modified: 13/10/2019 - 100% done
- **text_prepare_tests.tsv**(n/a) - 5091 bytes, last modified: 13/10/2019 - 100% done
- **train.tsv**(n/a) - 7196138 bytes, last modified: 13/10/2019 - 100% done
- **validation.tsv**(n/a) - 2166270 bytes, last modified: 13/10/2019 - 100% done

Saving test.tsv to test.tsv

Saving text_prepare_tests.tsv to text_prepare_tests.tsv

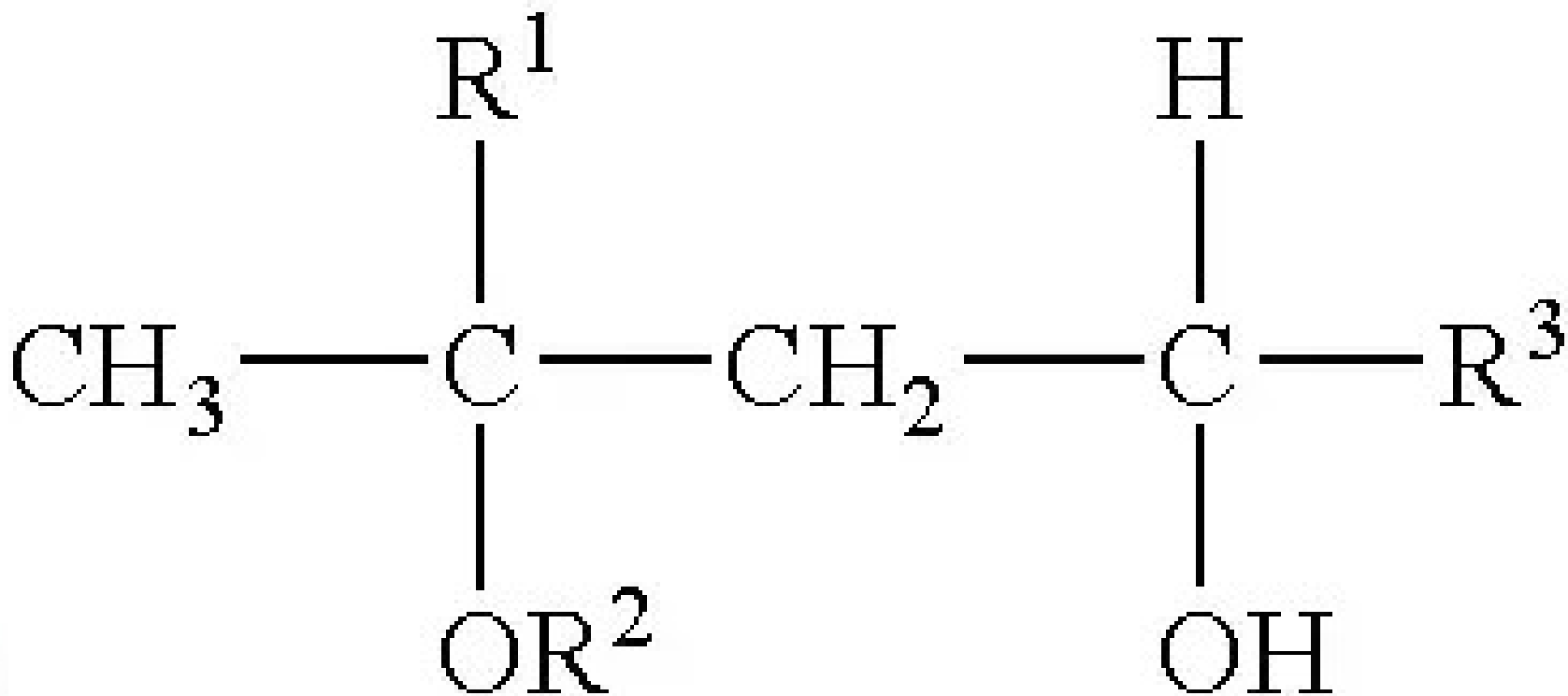
Saving train.tsv to train.tsv

Saving validation.tsv to validation.tsv

/content



Penser comme un
extraterrestre.





Outils de NLP

- Traitement basé sur des règles
 - Expression régulière / Regular Expression (Regex)
 - Grammaire non contextuelle / Context-Free Grammar (CFG)
- Modélisation probabiliste et statistique
 - Classifieur logistique & linéaire / Logistic & Linear Classifiers
 - Modèle de Markov
 - Maximum de vraisemblance / Likelihood Maximization
- Apprentissage Profound / Deep Learning
 - Réseau neuronal récurrent / Recurrent Neural Networks (RNN)
 - Réseau neuronal convolutif / Convolutional Neural Networks (CNN)
 - Long Short Term Memory (LSTM)



RegEx

- Devoir
 - Lire : crstin.com/fr/learn-regex
 - Jouer et apprendre : regexr.com



Texte

Qu'est-ce que le texte ?

- Une séquence d'un arrangement d'atomes avec certains caractéristiques communs (pour nos amis les extraterrestres)
- Glyphes
- Caractères
- Mots
- Locutions
- Phrases
- Paragraphes
- Chapitres, Articles, Clauses, Statuts, Livres ...



Mots

Qu'est-ce que un mot ?

ΠΟΙΗΕΝ ΤΗ ΒΑΣΙΛΕΙ
ΑΥΤΟΥ ΚΑΙ ΟΥΤΩΣ
ΠΑΣΑΙ ΑΙ ΓΥΝΑΙΚΕΣ
ΠΕΡΙΘΗΣΟΥΣΙΝ ΤΙ
ΜΗΝ ΤΟΙΣ ΑΝΔΡΑΣΙ
ΕΑΥΤΩΝ ΑΠΟΠΤΕ-
ΧΟΥΣ ΕΩΣ ΠΛΟΥΣΙΟΥ
ΚΑΙ ΗΡΕΣΕΝ Ο ΛΟ-
ΓΟΣ ΤΩ ΒΑΣΙΛΕΙ ΚΑΙ
ΤΟΙΣ ΑΡΧΟΥΣΙΝ ΚΑΙ
ΕΠΟΙΗΣΕΝ Ο ΒΑΣΙ-
ΛΕΥΣ ΚΑΘ' ΕΛΛΗΝ
ΣΕΝΟΜΑ ΜΟΥΧΟΣ



Mots

Qu'est-ce que un mot ?

日本語の先生

中西先生へのインタビュー

コロンビア国立大学には1999年からジャイカのボランティアとして日本人の先生が日本語を教えるために来られている。今まで来られた6人の先生のほとんどは2年間の滞在であったが、去年の12月に帰られた長岡順子先生は3年間も滞在された。今ここにいらつしやる中西不盡夫先生にコロンビア国立大学で教えられる経験とコロンビアの生活についてお聞きした。

中西先生はここに来られる前に韓国の自動車関係の会社で副社長をされていた。その前はメキシコ日産で働いておられた。その時にメキシコ人の仲間との相互理解のためにスペイン語の勉強を始めた。高校の時から日本語の勉強に興味をお持ちであった中西先生はメキシコに滞在された時の経験から、いつかチャンスがあれば日本語の先生になろうと思っておられて、韓国で仕事をされていた頃は会社の人に日本語を教えられたこともあった。更に日本語教師になるための勉強をされて、現在その資格をお持ちである。その後ジャイカのボランティアとしてコロンビアで日本語を教えられることになった。

それは中西先生にとつてとても面白い仕事だそうである。生徒が日本語を身につけるのを見ると非常にうれしいとおつしやった。なぜかという、コロンビアでの日本の存在感がとても薄いのに学生が日本語を勉強する動機を持ち、一生懸命勉強しようとしているのは大したものだと思うからだとおつしやった。

コロンビア国立大学について、とてもいい大学だとおつしやっていた。残念ながら、他の先生と話す機会があまりないが、今までの経験で、この大学の中ではばらしいと思うのは大学の自由さだとおつしやっている。

中西先生は仕事とその準備で忙しいながらも平穩にコロンビアで生活しておられて、困ることがあまりなく、少し困ることはボゴタの交通手段とお豆腐がないことだけだそうである。方で、コロンビアは気候がいいのがとても好きだとおつしやっていた。日本は春夏秋冬があり、その変化がきれいであるが、夏は暑くて、仕事をする気になくなるし、冬は寒くて、外に出る気もなくなるので、コロンビアのように、年中動ける気候の方がいいとおつしやっていた。

中西先生はコロンビアに後1年滞在される予定で、残りの時間にコロンビア人ともっと話したり、コロンビアのことをもっと知りたいとおつしやっていた。私も中西先生がコロンビアで残りの時間を楽しく過ごされることを願っています。

Mots

Qu'est-ce que un mot ?





Mots

Qu'est-ce qu'un mot ?

Il pleure dans mon coeur
Comme il pleut sur la ville

`/[^•]+•/g`

Text

Il·pleure·dans·mon·coeur·
Comme·il·pleut·sur·la·ville·



Mots

Qu'est-ce qu'un mot ?

On mange, les enfants !

On • mange, • les • enfants • !

```
/[^•!,]+/g
```

Text

On • mange, • les • enfants • !



Mots

Qu'est-ce qu'un mot ?

abat-jour

faites-moi

essuie-glace

2009-2012

occupe-t'en

aujourd'hui

tournevis

muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine

(vous parlez) comme si vous faisiez partie de ceux que nous n'avons
pas pu transformer en fabricant de personnes sans succès



Tokenization

- Devoir
 - Lire : `nltk.tokenize` et regarder un peu partout dans la documentation de nltk (nltk.org)
 - Jouer et apprendre : `tokenize.ipynb`



Racinisation (Stemming) & Lemmisation

La **racinisation** consiste à supprimer la fin des mots, ce qui peut résulter en un mot qui n'existe pas dans la langue

Ex : cheval, chevaux, chevalier, chevalerie,
chevaucher ⇒ « **cheva** » (mais pas « cavalier »)
marmaille, marmite ⇒ **marm**

La **lemmatisation** a pour objectif de retrouver le lemme d'un mot, par exemple l'infinitif pour les verbes.

Ex : ai, as, a, avons, avais, aurai, eûtes ⇒ **avoir**
Beau, beaux, belle, belles ⇒ beau
(cheval ≡ chevaux) ≠ chevalerie ≠ chevauche



Porter's Stemming

- 5 phases heuristiques de réduction de mots, appliquées de manière séquentielle
- Ex. Phase 1 réductions :

Rule	Example
SSSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

- Exemples : feet→feet, cats→cat, wolves→wolv, talked→talk
- Failles: échoue sur les formes irrégulières, production des non-mots



WordNet Lemmisation

- Recherche (lookup) de lemmas dans la base de données WordNet
- Exemples : feet→foot, cats→cat, wolves→wolf, talked→talked
- Failles: toutes les formes ne sont pas réduites

Réaliser des tests pour decider entre racinisation ou lemmisation ou une approche sur mesure



Racinisation/Stemming & Lemmisation

- Devoir
 - Jouer et apprendre :
RacinisationLemmisation.ipynb



Simple Pré-traitement de Texte

Lettres majuscules et minuscules

- « Je » et « je » n'ont pas le même encodage
- **Solution : mettre tous les lettres en minuscules**
- ... mais, les conventions orthographiques peuvent communiquer les informations utiles. ONT (Office National des Transports) n'est pas la 3ème personne plurielle d'avoir.



Simple Pré-traitement de Texte

- Caractères mieux traduit comme un espace

Ex : () { } [] | @ , ;

- Caractères mieux traduit comme nul

Ex : 0x00 - 0x1F



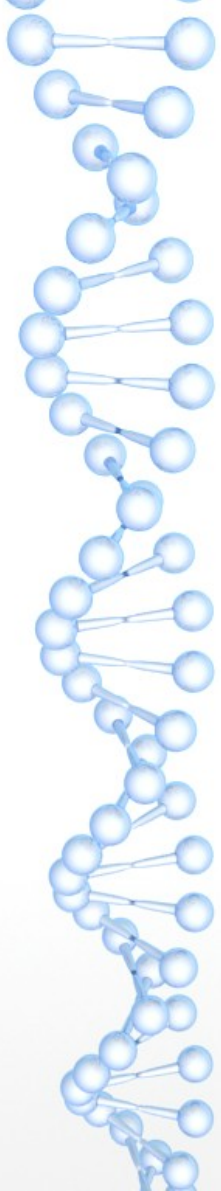
Simple Pré-traitement de Texte

STOP WORDS

Enlever les mots d'une fréquence élevée (STOP WORDS) qui ne seront pas utiles pour la prévision : avec, dans, mais, même ...

Il existe des listes préparées pour les différentes langues selon leur fréquence de mots établie.

- **Devoir**
 - Parcourez les listes de STOPWORDS dans le dossier stopwords dans le dépôt.

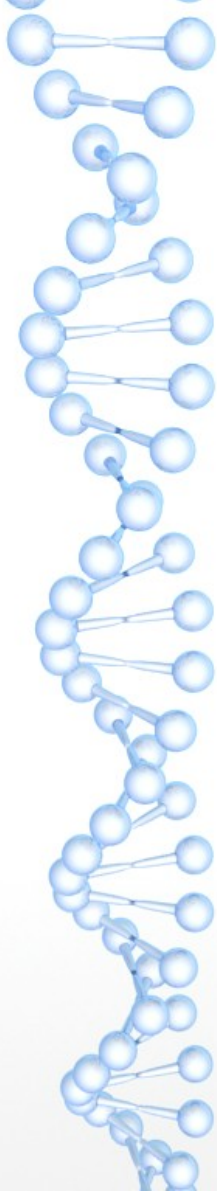


- Devoir

Tâche 1 TextPrepare dans le Notebook
Classification-MultiEtiquette

Un (trop) simple algorithme pour analyse des sentiments : **Bag of Words – BOW (Sac de Mots)**

Critiques (D documents)	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	Note
aucun intérêt	1	1	0	0	0	0	0	0	0	
sans intérêt	0	1	1	0	0	0	0	0	0	
un chef d'oeuvre	0	0	0	1	1	1	1	0	0	
sa meilleure oeuvre	0	0	0	0	0	0	1	1	1	



Critiques (D documents)	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	Note
	aucun	intérêt	sans	un chef d	oeuvre	sa	meilleure			

$$-0.5X_1 - 1.0X_2 - 0.5X_3 + 0.5X_4 + 0.5X_5 + 0.5X_6 + 1.0X_7 + 0.5X_8 + 0.5X_9$$



< 1.0 ≤



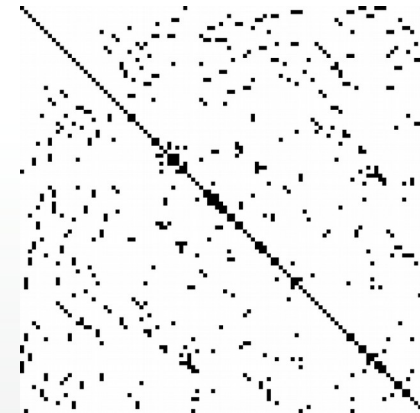
70 %

	Note	Prévision	Réel
ni plus ni moins GENIAL ! quel cycle ! quel talent !	0		
Ce livre a susité grand intérêt	-0.5		
Une oeuvre de celles que l'on relira toujours	2.0		
Un chef d'oeuvre du genre	2.0		
il signe sa meilleure œuvre	2.0		
à éviter de toute urgence	0.5		
Mal conçu, mauvaise exécution	0		
Nul, ennuyeux, mal écrit	0		
Vraiment aucun intérêt	-1.5		
considère comme un chef d'œuvre, il ne suffit pas	2.5		

Matrices Creuses

Une liste de fréquence des mots français a 129.000 éléments. Un Bag-of-Words peut avoir énormément des colonnes, la majorité avec une valeur 0. La taille de la matrice peut engendrer les problèmes de stockage et aussi ralentir le temps de calcul. Il existe plusieurs représentations pour les matrices creuses pour éviter les problèmes de stockage et performance.

$$\begin{pmatrix} 10 & 20 & 0 & 0 & 0 & 0 \\ 0 & 30 & 0 & 40 & 0 & 0 \\ 0 & 0 & 50 & 60 & 70 & 0 \\ 0 & 0 & 0 & 0 & 0 & 80 \end{pmatrix}$$

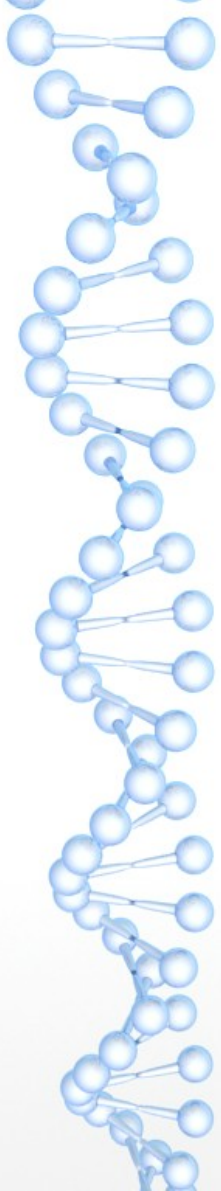




Structure des Données Matrices Creuses

$$\begin{pmatrix} 10 & 20 & 0 & 0 & 0 & 0 \\ 0 & 30 & 0 & 40 & 0 & 0 \\ 0 & 0 & 50 & 60 & 70 & 0 \\ 0 & 0 & 0 & 0 & 0 & 80 \end{pmatrix}$$

A	=	[10	20	30	40	50	60	70	80]	valeurs non-zéro
IA	=	[0	2	4	7	8]				cum nz par ligne
JA	=	[0	1	1	3	2	3	4	5]	colonne valeur non-zéro



- Devoir

Tâches 2 et 3 BagOfWords dans le
Notebook Classification-MultiEtiquette



Comment améliorer la précision de nos résultats ?

$$-0.5X_1 -1.0X_2 -0.5X_3+0.5X_4+0.5X_5 +0.5X_6+1.0X_7 +0.5X_8+0.5X_9$$

$$\sum W_n X_n$$

- Augmenter le **rélevance** des données d'entrée (**Feature Engineering**)
 - Choisir les X_n afin qu'ils pesent plus sur la prévision
- Améliorer notre formule afin qu'elle soit mieux ajusté aux données (par la technique de l'apprentissage automatique **Logistic Regression**)
 - Trouver les W_n afin que le plus grand nombre de cas d'entraînement donne la prévision correcte.



Feature Engineering

- Utiliser n-grammes
 - 1-gramme : aucun, intérêt, sans, un, chef, d, œuvre, sa, meilleure
 - 2-gramme : aucun intérêt, sans intérêt, un chef, chef d', d'oeuvre, sa meilleure, meilleure œuvre
 - 3-gramme : un chef d', chef d'oeuvre, sa meilleure œuvre
 - ...



Feature Engineering

Enlever n-grammes qui ne sont pas utiles pour les prévisions.

- Fréquence élevée (STOP WORDS) : avec, dans, mais, même ...
- Fréquence rare : faute d'orthographe, mots d'usage rare



Feature Engineering

Classement de features fréquence moyenne :TF-IDF

Term Frequency-Inverse Document Frequency
(Fréquence de terme-Fréquence Inverse du terme dans le document)

Remplacer feature valeur binaire (0,1) avec une valeur qui prend en compte l'importance du n-gram selon sa fréquence

Un TF-IDF sera élevé quand le TF (fréquence du terme dans le document) est élevé et l'IDF (inverse fréquence du terme dans tous les documents) est bas. On suppose que le terme a une signification particulière dans ce document dans ce cas.



TF-IDF

Term Frequency TF

$f_{t,d}$: comptages bruts du terme dans le document /
raw count of the term in the document

$$f_{t,d} / \sum_{t' \in d} f_{t',d}$$

Inverse Document Frequency IDF

N : nombre documents dans le corpus
 D_T : nombre documents où le terme se trouve

$$\log \frac{N}{D_T}$$

TF-IDF

$$\text{TF} * \text{IDF}$$



Quiz

Nous voulons supprimer quelques n-grammes en fonction de leur fréquence dans notre corpus de documents (combien de documents ont un n-gramme particulier divisé par le nombre total de documents). Qui peut être enlevé ?

- A N-grammes de haute fréquence ?
- B N-grammes de fréquence moyenne ?
- C N-grammes de fréquence basse ?

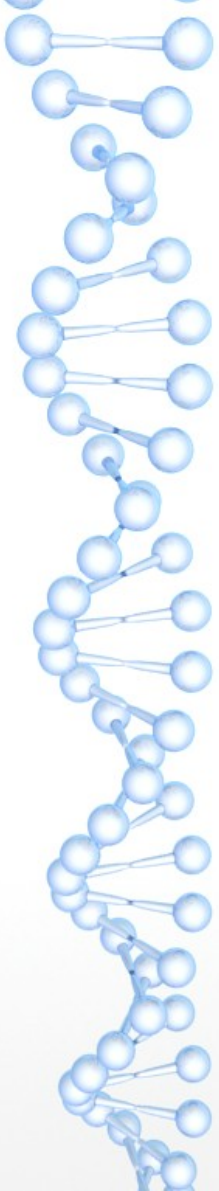
Python TF-IDF Exemple

```
from sklearn.feature_extraction.text import TfidfVectorizer
import pandas as pd
texts = [
    "good movie", "not a good movie", "did not like",
    "i like it", "good one"
]
tfidf = TfidfVectorizer(min_df=2, max_df=0.5, ngram_range=(1, 2))
features = tfidf.fit_transform(texts)
pd.DataFrame(
    features.todense(),
    columns=tfidf.get_feature_names()
)
```

Enlever termes de
fréquence élevé

Enlever termes de
fréquence bas

	good movie	like	movie	not
0	0.707107	0.000000	0.707107	0.000000
1	0.577350	0.000000	0.577350	0.577350
2	0.000000	0.707107	0.000000	0.707107
3	0.000000	1.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000



- Devoir

Jouer et apprendre : Notebook
ExempleTF-IDFVectorizer.ipynb. Changez
les texts, les n-grammes, les limites de
fréquence.

Multi-classification

- classification unique : sentiments
 - Exemples, 2 classifications positif, négatif ; 5 ; 12



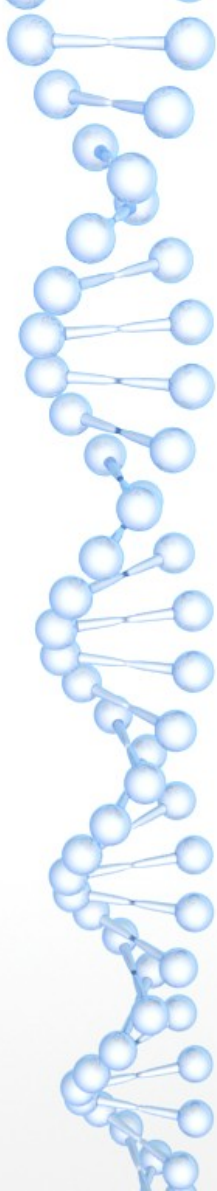
- classification multiple : étiquettes sujets
 - Exemples

How to create map from JSON response in Ruby on Rails 3?
['ruby', 'ruby-on-rails-3', 'json']
Eclipse C++ MinGW - Can not Lauch Program <Terminated>
['c++', 'eclipse']



Conséquences de MultiClassification

- Classification résultat max est nombre d'étiquettes totales par nombre des documents classifiés. Utiliser une structure binaire de données.
- Approche : 1 classificateur par étiquette, réduction du problème à un problème binaire avec beaucoup d'itérations.
- Précision ?



- **Devoir**

Continuez dans le Notebook Classification MultiEtiquette. Complétez le code dans les sections TF-IDF et Classificateur MultiLabel.



Comment améliorer la précision de nos résultats ?

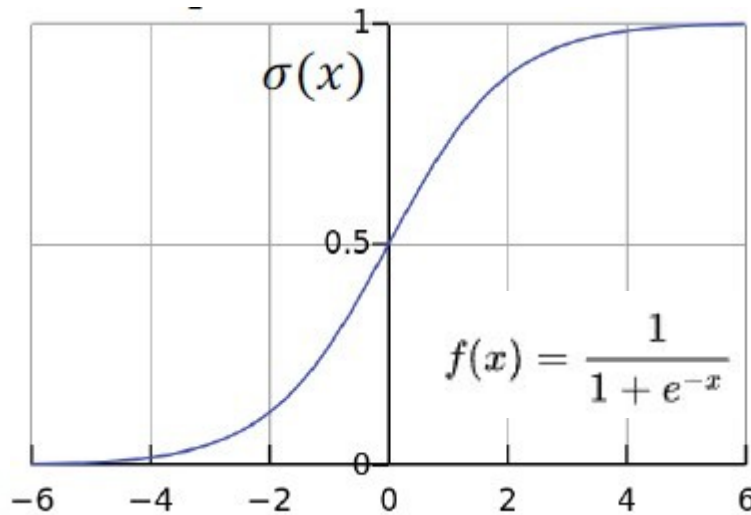
$$-0.5X_1 -1.0X_2 -0.5X_3+0.5X_4+0.5X_5 +0.5X_6+1.0X_7 +0.5X_8+0.5X_9$$

$$\sum W_n X_n$$

- Augmenter le rélevance des données d'entrée (**Feature Engineering**)
 - Choisir les X_n afin qu'ils pesent plus sur la prévision
- Améliorer notre formule afin qu'elle soit mieux ajusté aux données (par la technique de l'apprentissage automatique **Logistic Regression**)
 - Trouver les W_n afin que le plus grand nombre de cas d'entraînement donne la prévision correcte.

Logistic Regression

$$\sigma(\Sigma w_n x_n)$$



I. Fonction sigmoïde :
Pour classification,
valeur de sortie entre
0 et 1



Logistic Regression

II. Cost function (fonction de coût) : calculer le coût pour les classification erronées.

III. Gradient Descent : en apprentissage automatique, la technique la plus commune pour apprendre des erreurs

Gradient Descent est un technique iterative pour ajuster les weights (poids) afin de trouver le coût minimum (qui donne les prévision les plus précises)

Cost Function – Gradient Descent

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

Want $\min_{\theta} J(\theta)$:

Repeat {

$$\theta_j := \theta_j - \alpha \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

}
(simultaneously update all θ_j)

prévision

valeur
actuelle

Logistic Regression

Expérience : BOW 1-grams TF-IDF

Précision : 88.5 %

Poids appris / Learned weights :

ngram	weight
great	9.042803
excellent	8.487379
perfect	6.907277
best	6.440972
wonderful	6.237365
Top positive	

VS

ngram	weight
worst	-12.748257
awful	-9.150810
bad	-8.974974
waste	-8.944854
boring	-8.340877
Top negative	



Logistic Regression

Expérience : BOW 1,2-grams TF-IDF

Précision : 89.9 %

Poids appris / Learned weights :

well worth 13.788515

best 13.633200

rare 13.570259

better than 13.500025

Near top positive

bad -24.467648

poor -24.319746

the worst -23.773352

waste -22.880340

Near top negative

VS

Evaluation Metrics - Accuracy

tp : True Positive (Vrai Positif)

Patient reçoit un diagnostic du cancer. Patient a cancer.

fn : False Negative (Faux Negatif)

Patient reçoit un diagnostic qu'il n'a pas cancer. Patient a cancer.

fp : False Positive (Faux Positif)

Patient reçoit un diagnostic du cancer. Patient n'a pas cancer.

tn : True Negative (Vrai Positif)

Patient reçoit un diagnostic qu'il n'a pas cancer. Patient n'a pas cancer.

		prévision	
		yes	no
réel	yes	tp	fn
	no	fp	tn

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn}$$

Evaluation Metrics – Precision and Recall

tp : True Positive (Vrai Positif)

Patient reçoit un diagnostic du cancer. Patient a cancer.

fn : False Negative (Faux Negatif)

Patient reçoit un diagnostic qu'il n'a pas cancer. Patient a cancer.

fp : False Positive (Faux Positif)

Patient reçoit un diagnostic du cancer. Patient n'a pas cancer.

tn : True Negative (Vrai Positif)

Patient reçoit un diagnostic qu'il n'a pas cancer. Patient n'a pas cancer.

Precision

$$P = \frac{tp}{tp + fp}$$

Recall

$$R = \frac{tp}{tp + fn}$$



Evaluation Metrics – F1 Score

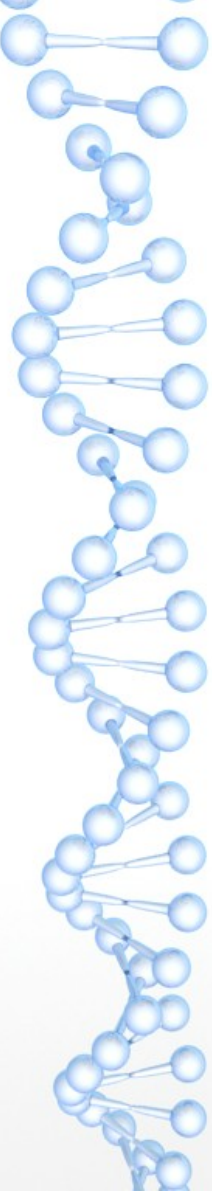
Precision

$$P = \frac{tp}{tp + fp}$$

Recall

$$R = \frac{tp}{tp + fn}$$

$$F_1 = \frac{2PR}{P + R}$$



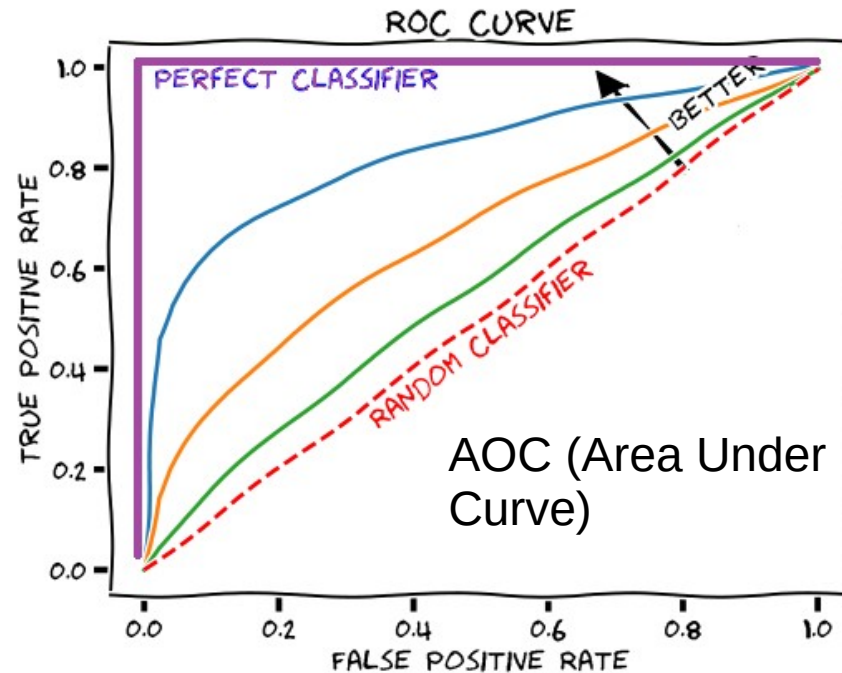
Evaluation Metrics – AUC – ROC Curve

La courbe AUC-ROC est une mesure de performance pour un problème de classification à différents réglages de seuil. ROC est une courbe de probabilité et l'AUC représente le degré ou la mesure de la séparabilité. Il indique combien de modèles sont capables de distinguer les classes.

Analyse de Modèles Multiclass avec AUC-ROC

Dans un modèle multi-classes, nous pouvons tracer le nombre N de courbes AUC-ROC pour les classes de nombres N à l'aide de la méthodologie One vs Rest. Ainsi, par exemple, si vous avez trois classes nommées X, Y et Z, vous aurez un ROC pour X classé contre Y et Z, un autre ROC pour Y classé contre X et Z et un troisième de Z classé contre Y et X. .

Evaluation Metrics – AUC – ROC Curve



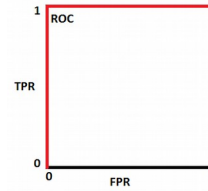
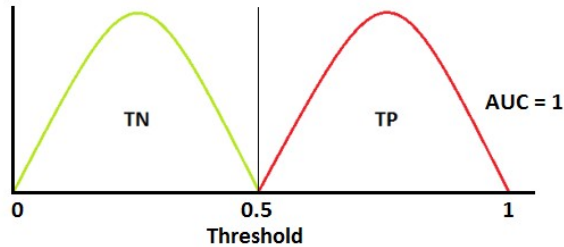
$$\text{TPR / Recall / Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

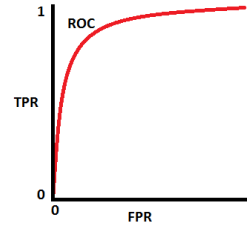
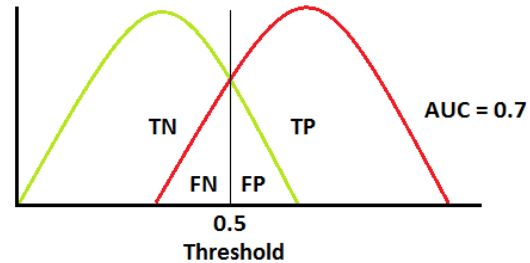
$$\text{FPR} = 1 - \text{Specificity}$$

$$= \frac{\text{FP}}{\text{TN} + \text{FP}}$$

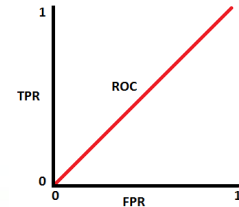
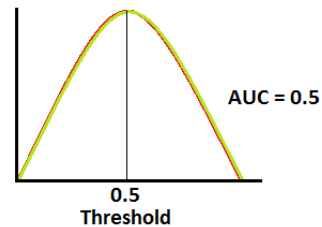
Evaluation Metrics – AUC – ROC Curve



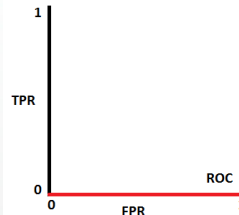
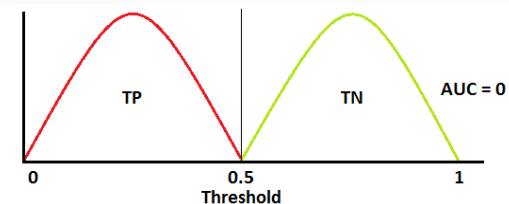
Le modèle est parfaitement capable de faire la distinction entre classe positive et négative.



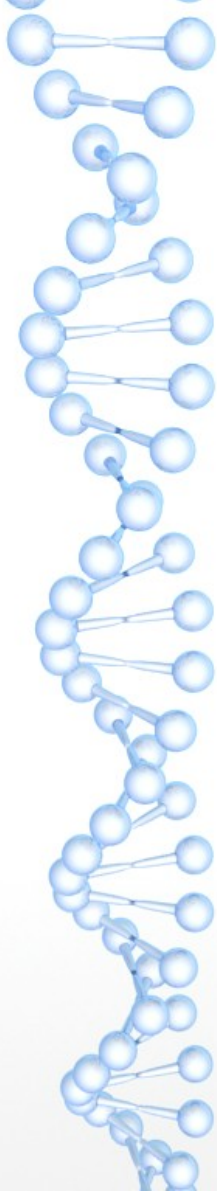
Il y a 70% de chances que le modèle puisse faire la distinction entre classe positive et négative.



Le modèle n'a aucune capacité de distinguer la classe positive de la classe négative.



Le modèle prédit que la classe négative est une classe positive et inversement.



- Devoir

Continuez dans le Notebook Classification MultiEtiquette. Complétez le code dans la section Evaluation. Arrêtez juste avant Tâche 4.

Régularisation en Logistique Régression

La régularisation est utilisée en logistique régression pour remédier à l'**overfitting** en améliorant la sélection de features. Elle peut créer un modèle moins complexe (parcimonieux) lorsque vous avez un grand nombre de features.

L1 (Lasso Regularisation)

L2 (Ridge Regularisation)

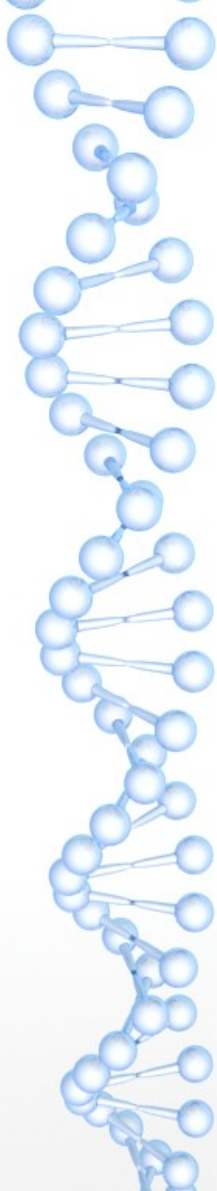
La différence principale entre ces deux est le penalty term (terme pénalisant) de la cost function (fonction de coût)

$$\sum_{i=1}^n (Y_i - \sum_{j=1}^p X_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

cost function

$$\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

cost function



- **Devoir**

Continuez dans le Notebook Classification MultiEtiquette. Complétez Tâche 4 et continuez jusqu'à la fin du Notebook.



Robots MALI

Le Centre National Collaboratif de l'Éducation en Robotique

Un Premier Cours sur le Traitement Automatique du Langage Naturel

M. Leventhal, Directeur de RobotsMali
30 Octobre – 17 Décembre 2019

