Introduction au NLP

En français on dit TALN

Le NLP c’est …

Un ensemble de techniques et d’algorithmes qui vont nous permettre d’extraire des informations pertinentes contenues dans des textes.

Le NLP c’est …

(...), un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle. Il vise à créer des outils de traitement de la langue naturelle pour diverses applications. Il ne doit pas être confondu avec la linguistique informatique, qui vise à comprendre les langues au moyen d'outils informatiques.

Wikipedia

https://fr.wikipedia.org/wiki/Traitement_automatique_du_langage_naturel

Le NLP c’est …

Présent dans pleins d’outils que l’on utilise au quotidien :

- Moteurs de recherche
- Chatbot
- Détection de spam
- Articles de news générés automatiquement

Le NLP c’est …

Google

c'est quoi le nlp?

Tous | Actualités | Images | Vidéos | Shopping | Plus | Paramètres | Outils

Environ 162 000 résultats (0,44 secondes)

Qu'est-ce que le NLP (Natural Language Processing)? - Quora
https://fr.quora.com › Quest-ce-que-le-NLP-Natural-Language-Processing
1 réponse
4 juin 2018 - **Natural Language Processing** (**NLP**) autrement appelé en français “Traitement automatique du langage naturel” est une branche très ...

Intelligence Artificielle : quelle différence entre NLP et NLU ?
https://www.lemagit.fr › conseil › Intelligence-Artificielle-quelle-difference...
30 nov. 2018 - Néanmoins et bien **que le NLP** soit en développement depuis plusieurs ... **C'est** vrai, mais avec la montée en puissance de l'intelligence ...

Que signifie Traitement du langage naturel (TLN ou NLP ...
https://www.lemagit.fr › definition › Traitement-du-langage-naturel-TLN

Qu'est-ce que le NLP (Natural Language Processing) ?
https://www.linkedin.com › pulse › quest-ce-que-le-nlp-natural-language-pr...

Traitement automatique du langage naturel

Domaine d'étude

Le traitement automatique du langage naturel, ou traitement automatique de la langue naturelle, ou encore traitement automatique des langues, est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle.
Wikipédia

Le NLP c’est …


LYCOS
c'est quoi le nlp?
REHERCHER SUR LE WEB
NLP & NLU : comment extraire le sens des énoncés ? | Blog Clevy
blog.clevy.io/conversationnel/nlp-nlu-comment-extraire-le-sens-des-enonces
Le NLP, ou TALN (Traitement Automatique du Langage Naturel), c'est quoi ? Le NLP (Natural Language Processing), ou TALN (Traitement Automatique du Langage Naturel), est une technologie à la confluence de l'intelligence artificielle et de la linguistique. Il consiste à analyser automatiquement des phrases formulées par un humain, afin de ...
Qu'est-ce que la PNL ou programmation neuro-linguistique? - L ...
www.lexpress.fr/.../qu-est-ce-que-la-pnl-ou-programmation-neuro-linguistique_...
"Ce que nous visons, c'est l'effacement de la charge émotionnelle d'un souvenir désagréable et la mise en place de nouveaux modes de comportements", précise-t-elle, évoquant le cas d'une ...
La PNL, mais qu'est ce que c'est que ça? – Le petit coach
lepetitcoach.com/la-pnl-mais-quest-ce-cest-ca
Mais c'est quoi la PNL? La PNL (NLP en anglais) ou programmation neurolinguistique est un concept créé dans les années 70 par John Grinder et Richard Bandler, deux universitaires américains. Ils se sont inspirés de trois thérapeutes célèbres : Milton Erickson (fondateur de l'hypnose ericksonienne),
C'est quoi la PNL .wmv - YouTube
www.youtube.com/watch
En quelques minutes, je vous explique les origines de la PNL et les fondements sur lesquels elle

Le NLP c’est …

Date et heure	Les étapes de ma livraison	Complément
samedi 30/11/2019 10:06	Service d'avisage Destinataire informé par SMS ou mail	Type du message : Tournée Agence Média utilisé : E-mail
samedi 30/11/2019 07:45	BORDEAUX MERIGNAC - CHRONOPOST Tri effectué dans l'agence de distribution	
samedi 30/11/2019 07:35	MERIGNAC BORDEAUX OUEST Colis en cours de livraison	
samedi 30/11/2019 04:20	BORDEAUX MERIGNAC - CHRONOPOST Tri effectué dans l'agence de distribution	Commentaire : Colis prêt pour la livraison
vendredi 29/11/2019 21:01	Service d'avisage Destinataire informé par SMS ou mail	Type du message : Prise en charge Média utilisé : E-mail
vendredi 29/11/2019 20:38	BORDEAUX MERIGNAC - CHRONOPOST Tri effectué dans l'agence de départ	Commentaire : Colis pris en charge par Chronopost, en cours d'acheminement

Le NLP c’est …

Présent dans pleins d’outils que l’on utilise au quotidien :

- Moteurs de recherche
- Chatbot
- Détection de spam

Utilisé sur des sites e-commerce :

- Produits similaires
- Détection automatique de catégories produits
- Recherches connexes

https://www.cdiscount.com/

Le NLP c’est …

Ce qu’on va développer en 3 matinées :

1. Concepts et outils pour le NLP
    - + des concepts de Machine Learning
2. Comment capturer la sémantique ?
3. Des réseaux de neurones pour le NLP : let’s go deeper

\+ Un mini-projet avec soutenance en janvier

Concepts et outils pour le NLP

1. Comment faire entrer du texte dans mon algorithme?
2. Comment bien compter?
3. Comment éliminer les informations gênantes?
4. Comment passer d’une feature à une prédiction?
5. Comment s’assurer que ma prédiction est correcte?

Comment faire entrer du texte dans mon algorithme?

Un peu de vocabulaire.

On va travailler sur un **corpus** de textes : c'est l'ensemble des documents.

Un document est composé de plusieurs **phrases** et chaque phrase est composée de **tokens** qui peuvent être des mots ou de la ponctuation.

On va parler de **sentence boundary detection**, pour extraire les phrases d'un document.

Et de **tokenization** pour extraire chaque token d'une phrase.

On verra comment mettre ça en place en TP.

Comment faire entrer du texte dans mon algorithme?

L’approche historique en NLP est d’utiliser le modèle Bag Of Words.

Dans ce modèle, un texte est considéré comme un ensemble de mots non ordonnés.

On va alors pouvoir définir un dictionnaire de mots qui va représenter l’ensemble du corpus sur lequel on travaille.

Chaque document va finalement être décrit par un vecteur qui va avoir comme coordonnées un comptage des occurrences des mots du dictionnaire.

Différentes méthodes de comptages vont produire différents modèles.

Wooo : un exemple plutôt qu’un long texte

Le corpus:

1. Le chien aboie
2. Le chat miaule

Le dictionnaire

le
chien
aboie
chat
miaule

Wooo : un exemple plutot qu’un long texte

Le corpus:

1. Le chien aboie
2. Le chat miaule

Le dictionnaire

Les vecteurs de features

	Phrase 1	**Phrase 2**
le	1	1
chien	1	0
aboie	1	0
chat	0	1
miaule	0	1

Comment bien compter?

Le modèle le plus simple, utiliser la fréquence des mots dans le document : https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Comment bien compter?

Le TF-IDF : https://en.wikipedia.org/wiki/Tf%E2%80%93idf

The tf–idf is calculated as

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

with

- N: total number of documents in the corpus $N = |D|$
- $|\{d \in D : t \in d\}|$: number of documents where the term t appears

Wooo : un exemple plutôt qu'un long texte

Le corpus:

1. Le chien aboie
2. Le chat miaule

Le dictionnaire	**Les inverse document frequency (sans le log)**
le	1
chien	2
aboie	2
chat	2
miaule	2

Wooo : un exemple plutôt qu'un long texte

Le corpus:

1. Le chien aboie
2. Le chat miaule

Le dictionnaire

Les vecteurs de features TF-IDF

	Phrase 1	**Phrase 2**
le	1	1
chien	2	0
aboie	2	0
chat	0	2
miaule	0	2

Comment bien compter?

Modèle Okapi BM25 : https://en.wikipedia.org/wiki/Okapi_BM25

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

Comment bien compter?

Une des limitations majeures du modèle Bag Of Words est la perte de l'ordre des mots.

Pour pallier cela, on peut utiliser des n-grams.

Un n-gram est une suite ordonnée de "n" mots.

Par exemple, "le chat boit du lait" est composé de 4 "2-gram" : "le chat", "chat boit", "boit du" et "du lait".

On peut utiliser les différentes méthodes de comptage vues précédemment avec des n-grams.

Comment éliminer les informations gênantes?

Comment éliminer les informations gênantes?

Dans chaque langue il y a des mots qui sont présents très fréquemment dans tous les documents.

Ces mots apportent du liant dans le texte mais ne sont pas porteur de sens.

On appelle ces mots les “stop words”.

Par exemple en français : le, la, les, et, ou, mon, ton, son, …

Pour ne pas bruiter ces modèles simples de NLP on supprime les stop words.

Un corpus peut avoir ses “stop words” particuliers.

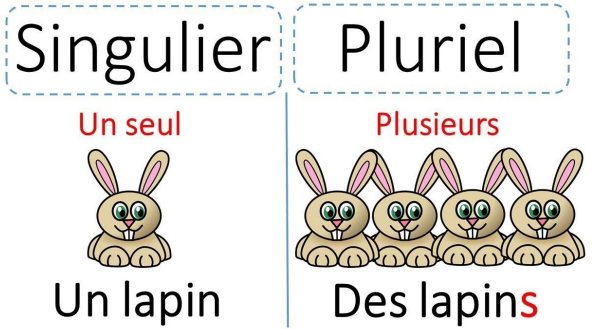
Comment éliminer les informations gênantes?

D’autres informations identiques sont présentes dans les textes sous des formes variées :

- Des conjugaisons
- Du singulier/pluriel
- Du féminin/masculin

Pour résoudre ces problèmes on utilise des stemmers et/ou des lemmatizers.

https://fr.wikipedia.org/wiki/Lemmatisation
https://fr.wikipedia.org/wiki/Racinisation



Comment passer de features à une prédiction?

On a transformé nos textes en vecteurs de features. Qu'est-ce que l'on peut en faire?

- Classification : Spam / Non Spam
- Clustering : Grouper des textes par thématique
- Ranking : Ordonner des textes par pertinences vis à vis d'un sujet
- Traduction : From english vers français
- Génération de texte : Des robots journalistes (ou votre autocorrect)
- …

On va commencer par de la classification binaire.

Comment passer de features à une prédiction?

On va se placer aujourd'hui dans le cadre d'un apprentissage supervisé.

On a un jeu de données dans lequel une partie des données a une valeur à prédire que l'on connaît, c'est les **targets**.

On va considérer que cette valeur est binaire. (ex : spam / pas spam)

On veut apprendre la fonction qui passe des features à la target.

On va partir sur un modèle simple de machine learning : La régression logistique

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

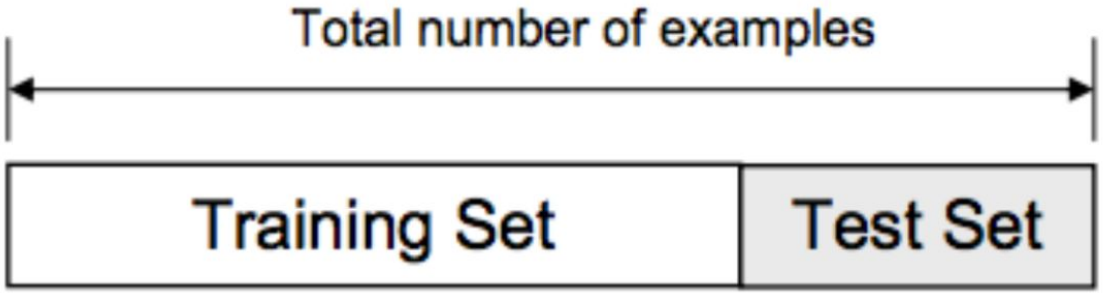
https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression wikipedia

Comment s’assurer que mon modèle est correct ?

On va d’abord entraîner le modèle, puis l’évaluer.

Il ne faut pas utiliser toutes les données pour l'entraînement.

On va évaluer le modèle sur un jeu de données non observé!



Mesure de l’erreur : Matrice de confusion

Dans une classification binaire, on peut prédire 2 choses et la réalité est l’une de ces 2 choses :

		Classe réelle	
		-	+
Classe prédite	-	True Negatives *(vrais négatifs)*	False Negatives *(faux négatifs)*
	+	False Positives *(faux positifs)*	True Positives *(vrais positifs)*

En pratique suivant le problème que l’on veut résoudre, chacun des termes de cette matrice peut avoir un impact différent. On va donc chercher à optimiser l’un de ces termes ou une combinaison de ces termes.

Mesure de l’erreur : Matrice de confusion

Exemple : test clinique

Pour comprendre un peu mieux à quoi toutes ces mesures correspondent, prenons l'exemple d'un test clinique. Il s'agit ici d'une étude conduite sur 4000 femmes, âgées de 40 ans et plus, apparemment en bonne santé. On leur a fait passer deux tests de dépistage du col de l'utérus :

- Un examen histologique, plutôt lourd, qui requiert d'être interprété par un expert, et qui servira de vérité terrain.
- Un frottis de dépistage, qui est un examen beaucoup plus simple et moins invasif, qui sera ici l'analyse de notre algorithme d'apprentissage.

Voici les résultats de l'expérience :

	Cancer	Pas de cancer	TOTAL
Frottis +	190	210	400
Frottis -	10	3590	3600
TOTAL	200	3800	4000

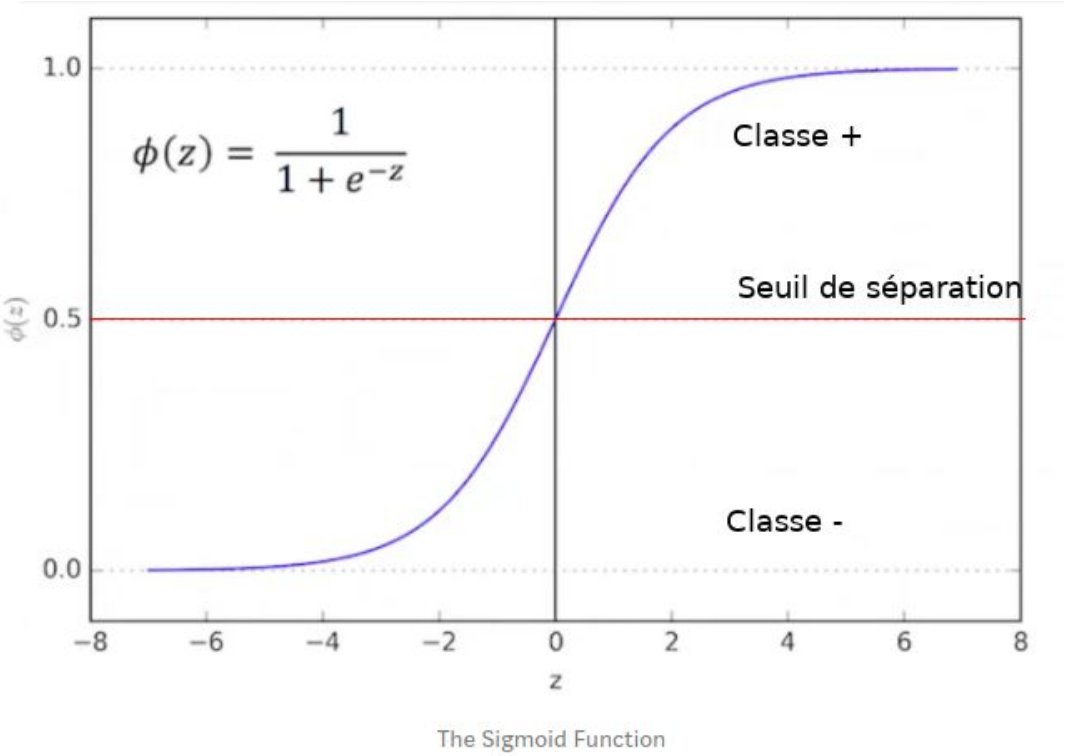
La probabilité de ne pas avoir de cancer quand le frottis est négatif est de 3590/3600 soit 99.7%, ce qui fait de ce test un bon outil de **dépistage**. À l'inverse, la probabilité d'avoir un cancer quand le frottis est positif est de 190/400 = 47.5%, ce qui en fait un très mauvais

https://openclassrooms.com/fr/courses/4297211-evaluez-et-ameliorez-les-performances-dun-modele-de-machine-learning/4308256-evaluez-un-algorithme-de-classification-qui-retourne-des-valeurs-binaires

Courbe ROC

La régression logistique va en fait nous donner une probabilité entre 0 et 1 que la classe est la classe + (ou -).

On va pouvoir identifier le seuil optimal qui sépare les 2 classes.



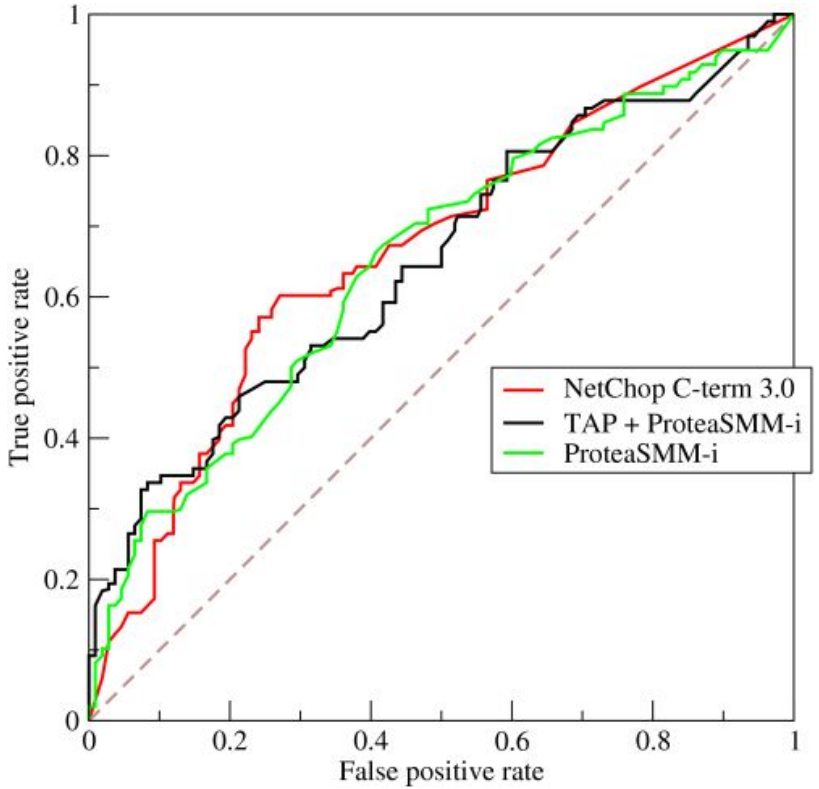
The Sigmoid Function

Courbe ROC

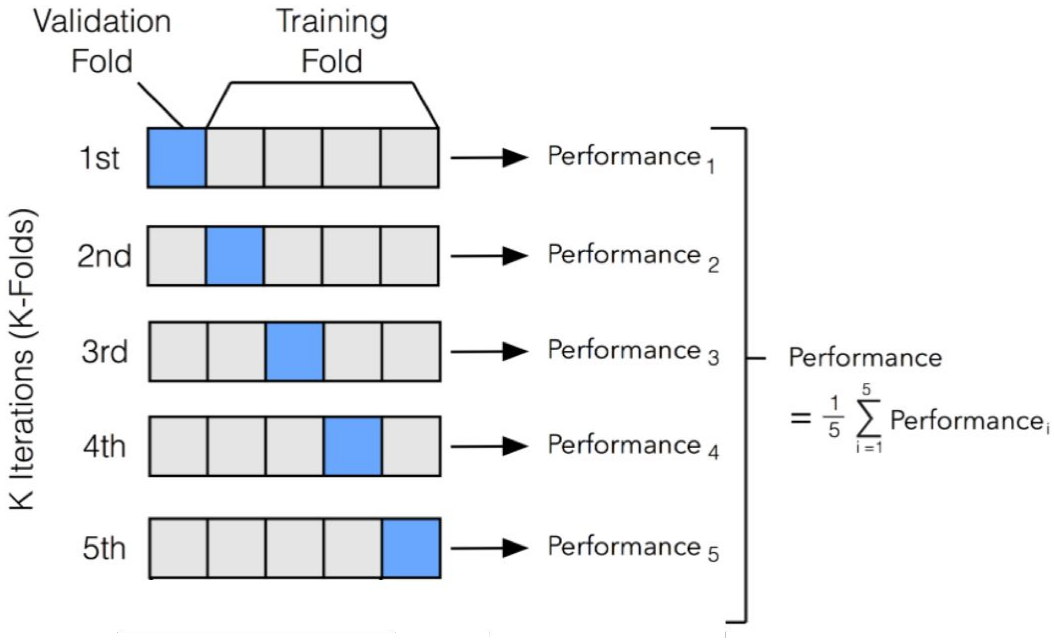
Pour Receiver Operatoring Characteristic :
https://fr.wikipedia.org/wiki/Courbe_ROC

2 utilisations :

- Mesure de performance en prenant l'aire sous la courbe : AUC (Area Under Curve)
- Identification du seuil de discrimination des classes



Comment s’assurer que mon modèle est correct ?



En résumé

- Le NLP permet de transformer du texte en features c’est du pre-processing de données
- Ces features servent à alimenter un algorithme de Machine Learning pour résoudre un problème donné
- Différents outils existent (et doivent être utilisés) pour s’assurer de la qualité des modèles appris sur nos données

En TP pour la mise en œuvre de ces techniques !