TheCodingMachine
TCM://

Le Natural Language Processing,
au cœur de l'interaction Humain-IA

—

Ce livre blanc vous propose de vous expliquer ce qu'est précisément le NLP, mais aussi de vous présenter ses possibilités et ses limites, tout en évoquant ses différents champs d'application.

En partenariat avec Axel de Goursac,
Directeur des Opérations et Chief Data Scientist de Myriad

Avril 2018

GLOSSAIRE

Big Data /
Données, structurées ou non, dont le très gros volume requiert des outils d'analyses adaptés.

Deep Learning /
Famille de méthodes de Machine Learning permettant un apprentissage automatique différent par niveau de détail, en utilisant des réseaux de neurones artificiels.

Small Data /
En opposition aux Big Data, ces « petites » données concernent la vie de tous les jours (ex : données des tickets de caisse, nombre de clients ayant pénétré dans le magasin...). Dès lors que celles-ci sont exploitées et que leur utilité est avérée, ces petites données devenues « intelligentes » sont souvent appelées Smart Datas.

Machine Learning /
Technique combinant la combinaison de l'efficacité des modèles statistiques à décrire la réalité avec la puissance de traitement et d'automatisation de l'Informatique. La machine va donc « apprendre » son propre modèle prédictif en s'entraînant sur des données d'apprentissage.

Intelligence Artificielle /
Elle correspond à des techniques informatiques permettant de réaliser des tâches, nécessitant des capacités de réflexion ou de calcul avancées pour des humains. Celle-ci contient notamment le Machine Learning, le Natural Language Processing, la Computer Vision...

Introduction

L'Intelligence Artificielle connaît un essor impressionnant sur grand nombre de secteurs d'activité et ses cas d'application se multiplient.

Un des grands axes de développements de l'IA est constitué par la prise en compte et la mise en valeur des données non structurées et donc autrefois peu utilisées, telles que :

- Le contenu des sites internet et leurs métadonnées
- Les messages des réseaux sociaux
- Les logs des machines
- Les emails
- Les articles, documents, présentations au sein d'une entreprise
- Les livres publiés

Aujourd'hui, l'Intelligence Artificielle tire parti de ces données textuelles, grâce à une technique appelée **« Natural Language Processing »** ou Traitement Automatique du Langage Naturel.

Selon le rapport[1] rendu par McKinsey en mars 2017 sur le sujet, les entreprises (tous secteurs confondus) ayant expérimenté cette technologie ont pu constater des résultats impressionnants :

- 50 à 70 % des tâches automatisés
- 20 à 35 % d'économies annuelles réalisées
- 50 à 60 % du temps de traitement réduit

De manière simplifiée, on peut dire du Natural Language Processing (NPL) qu'il correspond à la partie textuelle/linguistique de l'Intelligence Artificielle. Or le texte/langage est justement le moyen d'expression naturelle de l'homme.

C'est en cela que le NLP constitue la technique permettant une interface interactive entre l'humain et la machine.

ce livre blanc vous propose de vous expliquer ce qu'est précisément le Natural Language Processing, mais aussi de vous présenter ses possibilités et ses limites, tout en évoquant ses différents champs d'application.

[1] *https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/intelligent-process-automation-the-engine-at-the-core-of-the-next-generation-operating-model*

1 Qu'est-ce que le Natural Language Processing ?

Petite histoire du NLP

Le Natural Language Processing, ou Traitement Automatique du Langage, n'est pas une discipline nouvelle. Son origine remonte à la fin de la deuxième guerre mondiale, avec des recherches portant principalement sur la traduction automatique entre différentes langues.

En 1954, un ordinateur réussit à traduire automatiquement 60 phrases du Russe à l'Anglais. La publication en 1957 du livre *Syntactic structures* par Noam Chomsky fut une révolution pour le domaine. Il y montra notamment qu'il existe des caractéristiques communes à tous les langages et inventa un type de grammaire qui convertit le langage naturel en une forme compréhensible par des ordinateurs.

A partir des années 80, l'augmentation de la capacité de traitement des ordinateurs, puis le développement d'Internet et de la communication textuelle numérisée (sms, emails, réseaux sociaux...), ainsi que plus récemment l'émergence d'infrastructures Big Data et d'algorithmes d'Intelligence Artificielle ont permis une explosion des capacités et des applications du Natural Language Processing.

Définition et angle d'approche

Le Traitement Automatique du Langage ou Natural Language Processing en Anglais, correspond à un **cycle automatisé par l'informatique de lecture/correction/analyse de données textuelles** pour en tirer différents types d'information.

Une de ses déclinaisons fréquemment utilisées **pour la recherche de données** s'appelle le **« Text Mining »** (ensemble de méthodes, de techniques et d'outils pour exploiter les documents non structurés que sont les textes écrits). De plus, le Traitement Automatique du Langage est de nos jours souvent supporté par des algorithmes d'Intelligence Artificielle ou Machine Learning.

Type d'information considéré dans le texte

Un **texte est une donnée très riche**, qui contient beaucoup d'informations sous la forme synthétique d'une chaîne de caractères que sont les lettres, les chiffres ou autres symboles.

Il est en effet **rédigé par une personne dans un style particulier**, il dépend d'un **contexte** donné ainsi que de **la langue** et de **la culture** de la personne, et il vise à exprimer et transmettre un contenu objectif sur des événements extérieurs, ce contenu étant souvent accompagné d'une opinion plus subjective de cette personne.

Ainsi, suivant le type d'information que l'on souhaite extraire d'un texte, le NLP va permettre d'acquérir de la connaissance sur :

- Le langage utilisé en lui-même : orthographe, grammaire, sens et connotations des mots…
- Le contenu du texte : le message que la personne veut faire passer
- La personne à l'origine du texte : style, sentiments …
- La réalité extérieure : fiabilité et adéquation du message à l'environnement décrit

Mode d'application du traitement effectué

L'objectif du Natural Language Process est d'arriver à **extraire les différents types de connaissance** décrits ci-dessus **de manière automatique grâce à l'informatique**.

Cette extraction peut se faire suivant deux modes :

- **L'automatisation peut être déterministe**

  Des règles métiers bien définies pour le traitement et l'analyse des données textuelles sont ainsi implémentées. Cela peut concerner la correction orthographique ou l'identification de mots-clés à l'aide d'un référentiel prédéfini et par des techniques dites « d'expressions régulières ».

- **Ou l'automatisation peut être statistique**

  Ce mode utilise alors des algorithmes « auto-apprenants », c'est-à-dire qui vont apprendre les règles métiers de traitement et d'analyse eux-mêmes grâce à des lois statistiques. Ces algorithmes relèvent du Machine Learning, du Deep Learning et plus largement de l'Intelligence Artificielle.

Niveau de compréhension du texte

Un texte formant un tout homogène et logique, on peut s'intéresser à différents niveaux dans le degré de cohérence de celui-ci.

- **L'analyse lexicale**

  Il s'agit de trouver pour chaque mot sa nature grammaticale.
  Par exemple nom, déterminant, adjectif, verbe…

- **L'analyse syntaxique**

  On s'intéresse à la structure des groupes de mots et des phrases.
  Par exemple, le regroupement des mots dans des groupes nominaux (« une petite souris »), groupes verbaux (« a mangé ») puis, des liens entre les différents groupes de mots au sein d'une phrase (sujet, compléments, propositions subordonnées…).

- **L'analyse sémantique**

  Elle cible le sens des mots et des groupes de mots.
  Ce type d'analyse comporte plusieurs niveaux : le regroupement de mots ou groupes de mots dans des concepts, la gestion des synonymes ou de la proximité de sens, l'identification et la classification de mots dans des catégories plus vastes que des concepts (par exemple la catégorie entreprise, lieu, ou personne…), la détermination de mots-clés dans un texte et des sujets dominants…

- **L'analyse logique**

  Comment les concepts sont-ils reliés entre eux au sein du texte ? Quel sont les liens logiques qui les associent ?
  Ce type d'analyse s'appuie sur l'analyse syntaxique mais va plus loin en apportant une classification logique des relations entre concepts.

- **L'analyse de sentiments**

  Quelle est la tonalité du texte ou d'une de ses parties ? en terme de jugement positif ou négatif, d'objectivité ou subjectivité, d'optimisme ou pessimisme, de bienveillance ou malveillance, de détente ou nervosité, d'intéressement ou d'ennui… ?

2 Les modèles du Natural Language Processing

Après avoir présenté les différentes façons de considérer les données textuelles, nous pouvons maintenant examiner les technologies, ou les classes de modèles, liées au Natural Language Processing.

On divise celles-ci en fonction de la représentation du texte utilisée :

Modèles	Expressions régulières	Importance relative TF-IDF	Classification	WordNet	Plongements prédictifs	Statistiques de langage	TextRank
Représentation du texte	Chaine de caractères	Poids chiffré par mot	Tags lexicaux, syntaxiques, sémantiques	Concepts et synsets	Vecteurs de prédiction	Distribution de probabilité	Graphes de similarité

Non exhaustive, cette liste rassemble un échantillon représentatif des modèles utilisés à l'heure actuelle.

Les expressions régulières

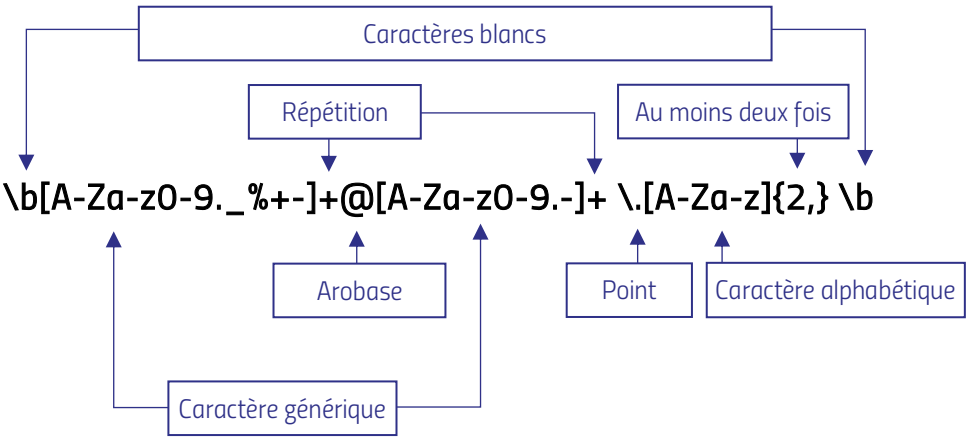
Les expressions régulières constituent un mode d'application déterministe du Natural Language Processing. En effet, celles-ci fournissent une méthode puissante, flexible et efficace mais déterministe pour le traitement du texte.

La notation étendue de correspondance de motifs d'expressions régulières permet notamment :

- d'analyser rapidement de grandes quantités de texte pour trouver des motifs de caractères spécifiques
- de découper une chaîne de caractères en paragraphes, phrases, mots
- de valider le texte afin de s'assurer qu'il correspond à un modèle prédéfini, par exemple une adresse électronique
- d'extraire, modifier, remplacer ou supprimer des sous-chaînes de texte
- d'ajouter des chaînes extraites suivant des règles prédéfinies à une collection afin de générer un rapport

Par exemple, pour vérifier si une chaîne de caractères correspond bien à un fichier.txt, on utilisera l'expression régulière «^.*\.txt$ », ce qui se décompose en :
à partir du début de la chaîne (^), autoriser tout type de caractère un certain nombre de fois (.*), puis le texte doit se terminer par .txt (\.txt$).

Pour trouver toutes les adresses email d'un texte, on implémente l'expression régulière suivante :



Cette expression régulière signifie :

Après un caractère blanc (\b), autoriser tout caractère alphanumérique (avec ._%+-) un certain nombre de fois mais au moins une ([A-Za-z0-9._%+-]+), puis le caractère arobase (@), puis des caractères suivis d'un point ([A-Za-z0-9.-]+\.), puis au moins deux caractères alphabétiques ([A-Za-z]{2,}) correspondant à la terminaison (.com ou .fr) et enfin un blanc (\b).

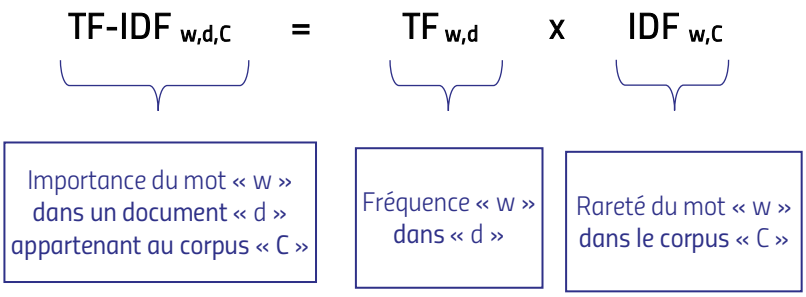
Extraire des mots-clés avec TF-IDF

« Term Frequency-Inverse Document Frequency » correspond au premier niveau d'analyse statistique des mots d'un texte, mais reste efficace.

Ce poids est une mesure statistique utilisée **pour évaluer l'importance d'un mot dans un document faisant partie d'un ensemble de documents (ou corpus).** L'importance du mot augmente proportionnellement au nombre de fois où ce mot apparaît dans le document mais est compensé par la fréquence globale du mot dans le corpus.

En effet, plus un mot apparaît dans un document, plus il va caractériser celui-ci ; mais plus sa fréquence est importante dans l'ensemble des documents, (par exemple les mots « et », « le », « à » ...), moins son apparition dans un document précis ne va caractériser celui-ci.

Le poids TF-IDF associé à un mot est donc d'autant plus élevé que le mot est fréquent dans le document considéré et que le mot est rare dans le corpus. Cela permet de détecter facilement les mots-clés (avec TF-IDF élevé) d'un document.



Des variantes de ce système de pondération TF-IDF sont utilisées par les moteurs de recherche pour évaluer et classer la pertinence d'un document en fonction des requêtes des utilisateurs.

Modèles de classification de texte

Les modèles de classification sont des modèles de Machine Learning à **apprentissage supervisé**. Ces modèles fournissent des **prédictions de catégorie basées sur un historique d'exemples** déjà classifiés, appelé ensemble d'apprentissage.

Par exemple, une banque souhaite prédire si des emprunteurs particuliers vont pouvoir rembourser leur emprunt à terme ou non, c'est-à-dire à les classifier en deux catégories « solvable » ou « non solvable ».

Elle dispose d'un certain nombre d'informations sur ses clients : caractéristiques démographiques, profils financiers, profils professionnels… ainsi que d'un historique important sur des emprunts passés indiquant si ces emprunts ont été totalement remboursés ou non.

La machine considère itérativement l'historique de l'emprunteur, puis à l'aide d'un modèle de classification, tente une prédiction « solvable » ou « non solvable » en fonction de ses caractéristiques.

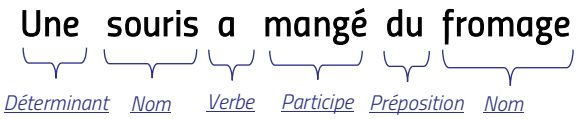
Suivant la justesse de la prédiction, la machine ajuste les coefficients du modèle de classification, et recommence le processus pour un autre emprunteur. **Plus le modèle est entraîné sur un nombre important de données (de bonne qualité), meilleures seront les prédictions du modèle,** qui peuvent ensuite s'appliquer à de nouveaux emprunteurs, dont on ne sait pas encore la catégorie « solvable » ou « non solvable ».

- ## Pour l'analyse lexicale

On peut utiliser des modèles de classification « **Part-Of-Speech Tagging** ».

Ils permettent à partir des caractéristiques des mots, comme leur place dans la phrase, les mots précédents et suivants, la casse… de **classifier ces mots suivant leur nature grammaticale**.

Ainsi une phrase comme « une souris a mangé du fromage » devient alors :



On peut faire remarquer qu'il suffirait d'utiliser un mode déterministe grâce à la constitution d'un référentiel préenregistré donnant pour chaque mot sa nature.

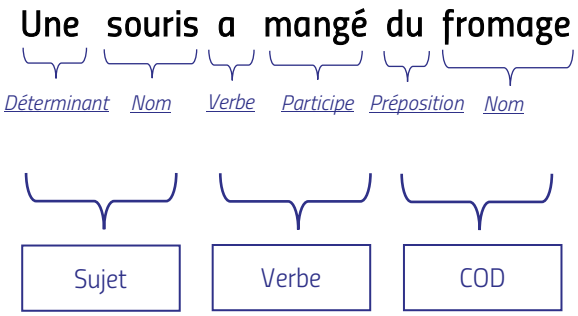
Mais, construire un référentiel complet serait trop difficile et certains mots peuvent avoir une nature différente suivant le contexte et qui ne sera détecté que par un modèle probabiliste.

Par exemple, « faible » peut être un nom ou un adjectif, « bien » un nom, un adjectif ou un adverbe…

- ## Pour l'analyse syntaxique

On peut utiliser des **modèles de parsing**, qui permettent, à partir du POS Tagging des mots d'une phrase, de les regrouper les mots en groupes et de leur donner une fonction.

L'exemple précédent devient alors :



- **Pour l'analyse sémantique**

On peut utiliser des **modèles de classification en « entités nommées » ou Named Entity Recognition**, qui représentent des catégories importantes de mots pour l'analyse du texte.

Cette détection se fait sur des groupes de mots à partir de leurs caractéristiques syntaxiques obtenues à l'étape précédente. Par exemple, il y a la catégorie des entreprises, celle des noms propres, celle des dates, celle des pays, celle des villes…

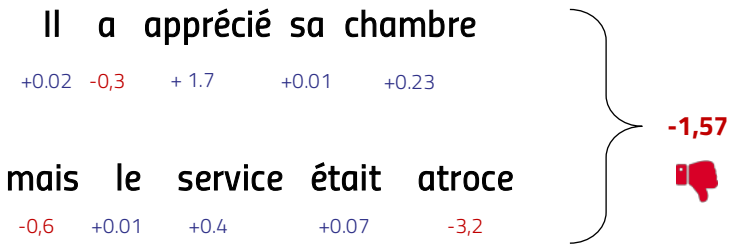
Ici encore, on peut utiliser des référentiels préenregistrés pour faire cette détection, mais les modèles Named Entity Recognition permettent de détecter aussi des mots inconnus au référentiel et de classifier des mots pouvant appartenir à plusieurs catégories suivant le contexte.

- **Pour l'analyse de sentiment**

On utilise des **modèles de classification binaire** par rapport à un sentiment du type positivité du propos, objectivité, optimisme, nervosité, ou intérêt.

Ces modèles permettent durant l'apprentissage **d'associer à chaque mot un coefficient de positivité ou négativité par rapport au sentiment considéré**, puis de prédire le sentiment global d'un nouveau texte en sommant les coefficients de tous les mots présents dans ce texte.

Par exemple, si l'apprentissage d'un modèle sur un corpus a permis de déterminer les coefficients pour le sentiment positif/négatif associés aux mots de la phrase « il a apprécié sa chambre mais le service était atroce » comme ci-dessous, le sentiment global de cette phrase sera négatif.



Modèles de concepts – WordNet

Pour aller plus loin dans la compréhension du langage, des linguistes ont recensé les **attributs sémantiques des mots du vocabulaire de différentes langues dans des bases de données lexicales.**

Une des plus connues s'appelle WordNet, construite par l'Université de Princeton. L'identification d'un concept dans un texte se fait alors de manière déterministe par consultation d'une base de données de ce type.

A partir d'un mot – gardons toujours la souris – **le modèle permet de déterminer quels sont les différents concepts associés à ce mot, avec leurs différents sens.**

Chaque concept est codé par un « synset » (synonym set), c'est-à-dire un ensemble de synonymes décrivant chacun le concept. L'intersection des sens de ces synonymes permet de caractériser de manière univoque le concept.

Ici, le mot « souris » est associé à 4 concepts différents, chacun étant listé avec son « synset » et sa définition :

[Souris, pointeur] : Dispositif de pointage pour ordinateur

[Souris, mus musculus] : Rongeur de petite taille au museau pointu, aux oreilles rondes, au pelage gris-brun et une queue relativement longue et mince.

[Souris, muscle] : Morceau de viande constitué par le tibia de la patte arrière de l'agneau, en bas de cuisse.

[Souris, jeune fille] : Terme familier pour désigner une jeune fille.

Ces synsets, représentant des concepts, sont regroupés dans des catégories plus générales et abstraites, qui forment ainsi une **hiérarchie de concepts appelée ontologie**. Il existe d'autres bases de données proposant des ontologies similaires ou plus spécialisées, certaines répertorient aussi les relations pouvant avoir lieu entre des concepts de type différents.

Modèles de prolongement prédictif de mots – représentation vectorielle

Les modèles de plongement prédictif de mots, ou « Word Embedding », dont le plus emblématique est Word2Vec, utilisent des **réseaux de neurones artificiels pour apprendre statistiquement une représentation vectorielle de chaque mot présent dans le texte.**

Le vecteur associé à chaque mot prend en compte le contexte dans lequel est apparu ce mot tout au long du texte, ce qui permet d'avoir une représentation numérique encodant des propriétés grammaticales et sémantiques : l'hypothèse est que deux mots seront d'autant plus proches de sens qu'ils apparaissent dans des contextes similaires.

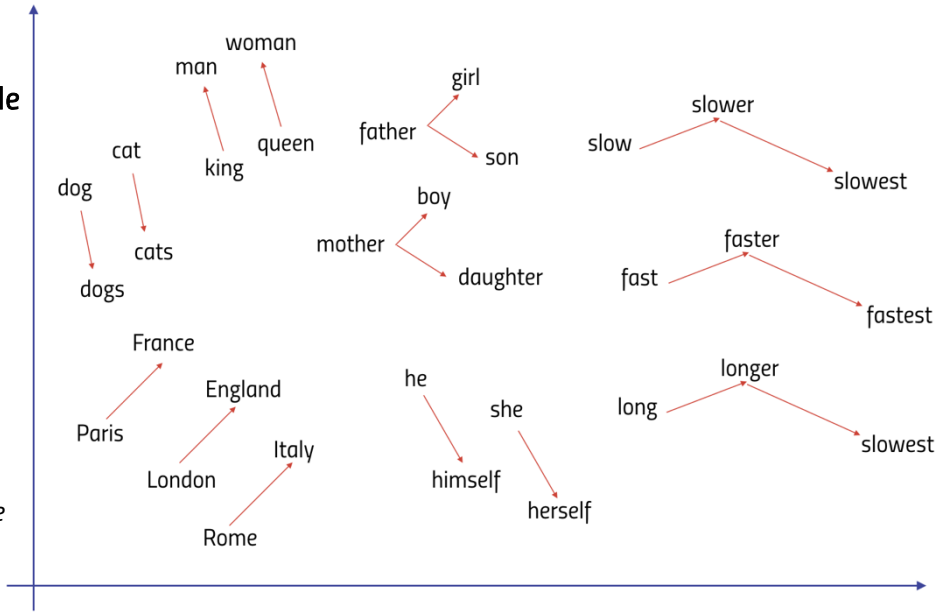
La pertinence de cette représentation dépend évidemment de l'entraînement des modèles – taille du corpus textuel d'apprentissage, qualité de ce corpus, optimisation du paramétrage via des métriques de tests – mais certains résultats sont étonnants de précision.

En effet, la représentation vectorielle permet de faire des opérations algébriques, comme l'addition ou la soustraction, aux vecteurs V[w] de chaque mot w.

Par exemple, V[roi]-V[homme]+V[femme] a pour plus proche voisin V[reine], l'interprétation basique de ce fait peut être la suivante :

Au concept de roi, on a soustrait le concept de masculinité, et on obtient le concept de royauté. En lui ajoutant le concept de féminité, on obtient alors le concept de reine. Cette représentation par plongement prédictif de mots permet ainsi une compréhension sémantique étonnante des mots d'un texte.

Projection d'un Espace sémantique pour un modèle de plongement prédictif



Source : Site de S. Zafrany sur le Machine Learning

Modèles statistiques de langage

Un des buts des modèles statistiques de langage est de **construire un modèle qui peut estimer la distribution du langage naturel** de manière aussi précise que possible.

Un des avantages de ce type de modèles est de fournir un moyen simple de traiter le langage naturel et qui peut s'adapter à des textes très différents. Un autre avantage réside dans le fait d'utiliser de l'apprentissage non supervisé, c'est-à-dire qu'il n'y a pas besoin de fournir des réponses à l'algorithme, pas de tagging de corpus textuels à faire, mais **l'algorithme apprend grâce aux associations statistiques entre les mots.**

Les modèles de **thèmes probabilistes** constituent des cas particuliers très utilisés de modèles statistiques de langage. Ils **ont vocation à trouver les thèmes dominants de textes donnés.**

Les thèmes sont eux aussi représentés par des distributions de probabilités qui modélisent la fréquence d'apparition de tel ou tel mot dans un texte associés à ces thèmes, et permettant :

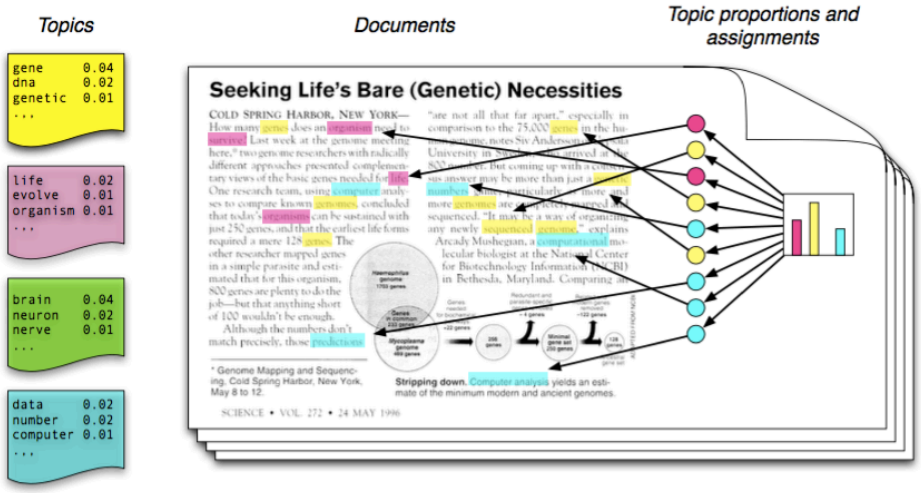
- D'inclure dans un thème non seulement les mots qui le décrivent précisément mais aussi les mots reliés
- De permettre à un mot en particulier d'appartenir à plusieurs thèmes, notamment si ce mot a plusieurs connotations ou sens

Chaque distribution de probabilité associée à un thème est estimée par des algorithmes statistiques sur des textes que l'on souhaite classifier, puis que l'on peut visionner afin d'évaluer leur pertinence, et enfin trouver la proportion de ces thèmes dans d'autres textes.

Ces modèles sont donc très intéressants pour avoir une vision globale d'un corpus et en produire une classification thématique.

Fonctionnement d'un modèle de thème probabiliste

Source :
David Blei, Commun. ACM (2012) 77-84



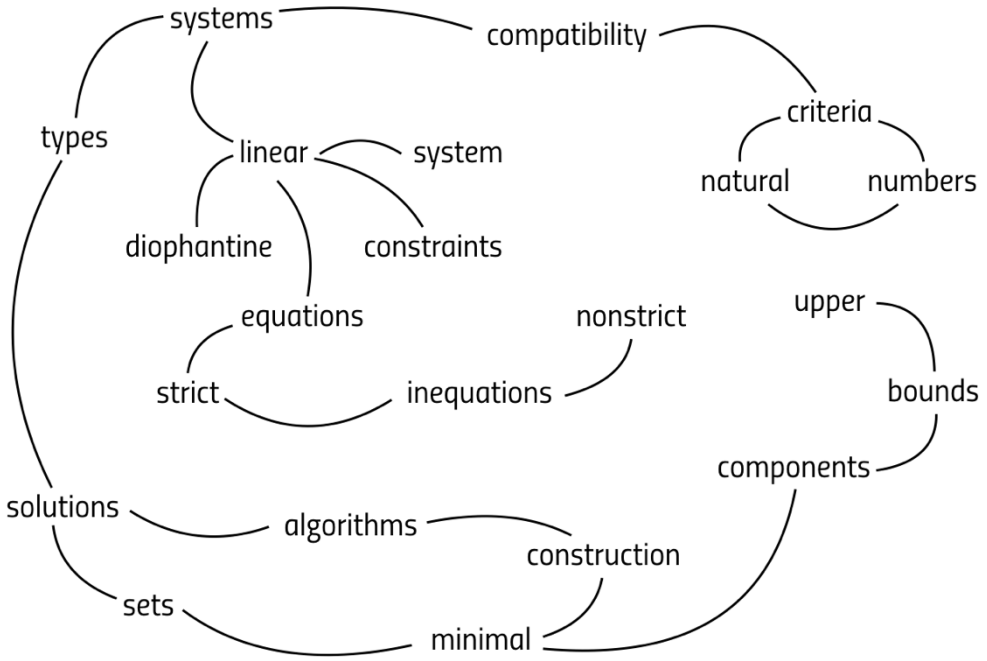
TextRank – Graphes de similarité

Ce type de modèles, inspiré de l'algorithme PageRank de Google, permet de **détecter des groupes de mots caractérisant un texte**.

Un texte est modélisé par un graphe constitué de nœuds représentant des groupes de mots du texte et d'arêtes représentant la fréquence à laquelle les deux groupes de mots apparaissent ensemble dans le texte.

Exemple de fonctionnement de TextRank

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given.
The criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types.



Keywords assigned by TextRank:
Linear constraints; linear Diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds.

Keywords assigned by human annotators:
Linear constraints; linear Diophantine equations; minimal generating sets; nonstrict inequations; set of natural numbers; strict inequations; upper bounds.

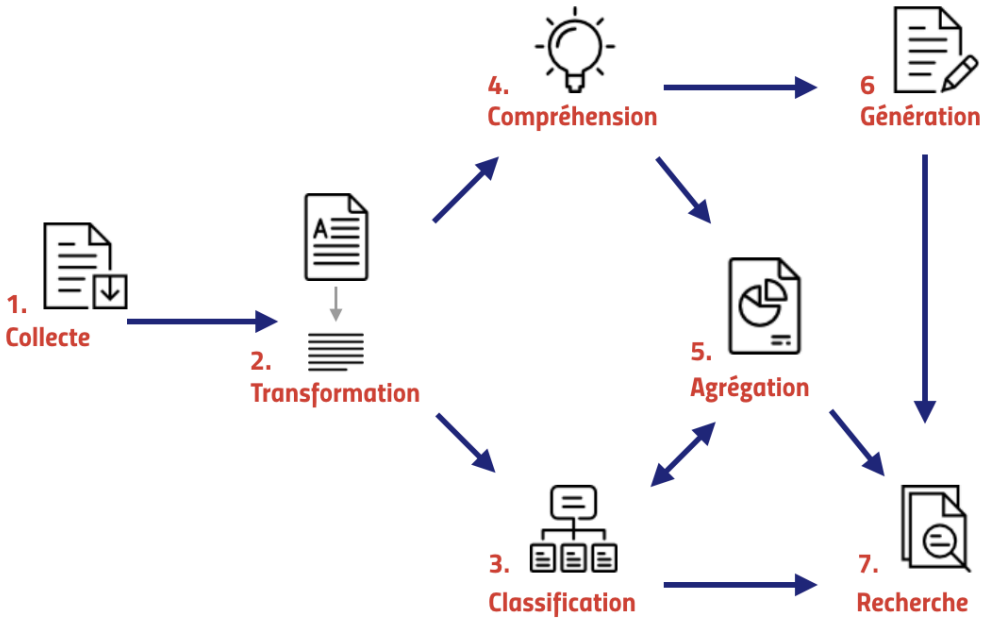
Source : R. Mihalcea and P. Tarau, « TextRank: Bringing Order into Texts »

2 Possibilités et limites du Natural Language Processing

Compte-tenu des modèles décrits dans la section précédente et de bien d'autres encore, la question suivante est naturelle : comment mettre tous ces modèles bout à bout et que permettent-ils de faire ?

Les possibilités du Natural Language Processing peuvent être génériquement regroupées dans le workflow ci-dessous. Chacune de ces possibilités correspond à des tâches qui sont réalisables par des humains à petite échelle, mais la puissance du NLP réside dans l'automatisation de ces tâches et le déploiement à grande échelle.

Workflow du Natural Language Processing



Collecte

La collecte de données textuelles se situe plutôt en amont du Natural Language Processing, mais elle constitue une étape primordiale pour les modèles qui suivent.

Celle-ci inclue des techniques très différentes comme :

- Le Web Scraping ou le web crawling, pour la récupération de données textuelles disponibles sur internet
- L'extraction du texte contenu dans des fichiers informatiques comme les formats PDF, MS Word, MS PowerPoint…
- La reconnaissance optique de caractères, dans des fichiers de type image, qui utilise aujourd'hui du Deep Learning
- La conversion de fichiers audio en texte (SpeechToText) qui utilise aussi des techniques de réseaux de neurones artificiels
- L'intégration dans un système GED de gestion électronique des documents au sein d'une entreprise
- La consommation de données textuelles provenant d'API de réseaux sociaux ou de sites d'information (news)

Transformation

La transformation des données textuelles regroupe des réalités très diverses.
Cela concerne tout changement des données par rapport à un but particulier et connu (règle métier), notamment :

- Les prétraitements, nettoyage et standardisation de textes
- La correction orthographique et syntaxique
- La structuration des données par rapport à une structure prédéfinie
- La traduction, qui fait partie des problèmes difficiles appelés « AI-complets », c'est-à-dire nécessitant différents types de connaissances que possèdent les humains (grammaire, sémantique, faits extérieurs sur le monde réel, etc. .)

Classification thématique

La classification de documents textuels intervient à un niveau assez haut de compréhension des textes. Il s'agit d'identifier les thèmes globaux de ces textes pour permettre une classification pertinente par rapport à des objectifs métiers.

- Détection de thèmes principaux d'un texte et des mots représentatifs de ces thèmes
- Identification de titres et de mots-clés importants dans un texte
- Regroupement des textes d'un corpus par similarité thématique. Ceci peut être effectué avec un niveau de détails arbitrairement élevé en fonction des besoins

Compréhension

Dans la partie compréhension des textes, on peut inclure beaucoup de types d'analyses que nous avons déjà décrites :

- Analyse lexicale, syntaxique et logique pour une compréhension de la forme
- Analyse sémantique et conceptuelle pour une compréhension du fond, permettant notamment une gestion des synonymes
- Analyse conceptuelle et/ou reconnaissance automatique de catégories (NER) de mots

Cependant, une compréhension fine, et de bout en bout, de textes longs composés de multiples phrases, avec un enchaînement logique complexe, est actuellement hors de portée des algorithmes et constitue un champ de recherche actif et encore ouvert.

Agrégation

L'agrégation concerne les informations générales que l'on peut extraire à partir d'un texte. Elle se nourrit de la partie Classification pour la détection des thèmes, et de la partie Compréhension pour la détection des entités nommées (NER).

Cela permet notamment :

- De produire des résumés automatisés de textes, avec une vision directe des thèmes importants et/ou des mots-clés, sous la forme de tableaux de bord ou de textes grâce à la partie Génération de texte
- De construire des indicateurs sur les textes grâce à de l'analyse de sentiment

Génération

Ici, la machine génère du langage naturel pour faire passer une information ou répondre à une question de manière directement compréhensible par les humains. Cette technique est en particulier très utilisée dans les chatbots.

Ceci peut se faire principalement de deux manières :

- **Mode déterministe**

  Les phrases générées par la machine sont pré-écrites avec certains trous à combler par la machine avec des informations dans sa base de données et elles sont sélectionnées en fonction de ce qui est demandé.

  Ceci est particulièrement performant dans un périmètre restreint, sur un domaine particulier ou pour répondre à des « FAQs ».

- **Mode statistique**

  Les phrases sont générées par la machine grâce à des modèles probabilistes de langage. En fonction des thèmes considérés, qui sont repérés grâce à la partie Compréhension, une loi de probabilité est déterminée pour la réponse à apporter, à partir de corpus d'apprentissage écrits par des humains, et des mots sont ainsi générés suivant cette loi de probabilité.

  Cette technique relève encore du domaine de la recherche pour ce qui est de la génération de textes longs et structurés.

  D'autre part, pour répondre à une question spécifique, cette technique peut être couplée avec la partie Recherche d'informations, et à partir des mots-clés détectés dans la question, la machine recherche des documents, isole la partie spécifique des documents pertinente pour la question, et peut la restituer/la modifier au demandeur.

Recherche d'information

La recherche d'information permet à un utilisateur de formuler une requête (en langage naturel ou non) et que lui soient restituées des informations pertinentes. Cette requête est, après une partie Compréhension du contenu de la requête, comparée à une base de données propre au système de recherche d'informations.

Cette base de données est donc cruciale pour la pertinence des informations retournées.

Elle est constituée de documents qui ont été analysés (cf parties Compréhension et agrégation), dont notamment les mots clés ont été extraits par NER ou TextRank, puis ont été intégrés dans la base de données en tant que métadonnées associées au document considéré.

Lorsqu'une requête est effectuée, un score de pertinence avec chaque document de cette base est effectué et les documents les plus pertinents sont retournés à l'utilisateur.

3 Cas d'application

Le Natural Language Processing, ainsi que nous l'avons analysé, permet l'interaction directe homme-machine, et donc d'automatiser nombre de processus et traitements portant sur des données textuelles ou générées par l'homme.

Voyons ici quelques cas d'application.

/ RESSOURCES HUMAINES

- Automatisation de recherche de profils
- Matching Cv VS. Offres d'emploi
- Construction de référentiels de compétences
- Identification de compétences manquantes et proposition de formation

/ RESEAUX SOCIAUX

- Études de tendances (buzz, influences…)
- Étude d'image de marques ou de produits
- Analyse d'opinions, réactions à l'actualité

/ E-COMMERCE

- Recherche des acheteurs basée sur le sens, gestion des synonymes
- Analyse de sentiments dans les commentaires
- Évaluation de la satisfaction client
- Personnalisation de la recommandation produit

/ INDUSTRIE

- Structuration des rapports écrits de maintenance / renforcement des capacités GMAO
- Analyse de marché et concurrence
- Suivi e-réputation
- Suivi des règles de conformité produit

/ CHATBOTS

- Recherche des acheteurs basée sur le sens, gestion des synonymes
- Analyse de sentiments dans les commentaires
- Évaluation de la satisfaction client
- Personnalisation de la recommandation produit

Juridique,

Transport,

Luxe,

Comptabilité …

Focus sur des applications

Pour mieux comprendre la valeur ajoutée du Natural Language Processing et la combinaison possible des fonctionnalités et algorithmes, focalisons-nous sur des **retours d'expérience.**

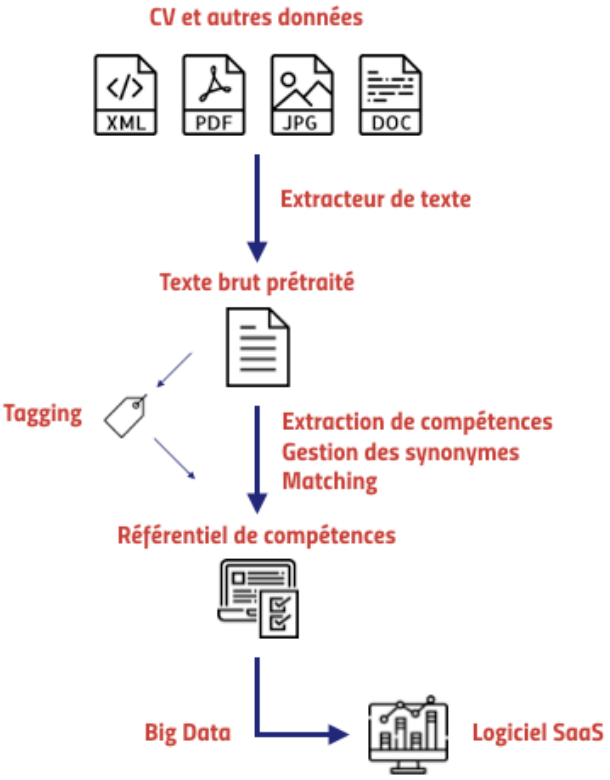
- **Focus 1 :**
  La gestion des compétences en Ressources Humaines

Un grand groupe a souhaité tester les possibilités du Natural Language Processing pour aider ses RH et automatiser certaines tâches dans la gestion interne des talents et dans la partie recrutement.

ENJEUX :

- Grand nombre de collaborateurs aux données non structurées
- Plusieurs milliers de postulants chaque année
- Grand nombre d'offres de postes à saisir
- Trop de saisie manuelle pour l'exploitation des données non structurées

Expérimentation :



1. Extraction du texte depuis tout types de formats
2. Prétraitements : nettoyage, gestion de la langue
3. Tagging des données
4. Modélisation mathématique du texte des CV par NLP
5. Détection automatique des compétences
6. Matching : identification des CV les plus pertinents
7. Structuration des CV
8. Construction des indicateurs agrégés (niveau d'expertise, durée de formation...)
9. Référentiel de compétences complet
10. Scalabilité Big Data

Le pipeline utilisé pour l'expérimentation chez ce client suit ce schéma et combine certaines possibilités du Natural Language Processing :

- La collecte des données textuelles : extraction
- La transformation de ces données : prétraitements, tagging, structuration
- Leur compréhension : modélisation sémantique, détection des compétences
- L'agrégation : construction d'indicateurs
- La génération : matching
- Et la recherche d'informations : mise en place du référentiel de compétences

RESULTATS : Statistiques de performance

? Pour aller plus loin …

Le Natural Language Processing permet d'augmenter considérablement l'efficacité des fonctions RH. Il existe aussi beaucoup d'autres cas d'usage possibles du NLP, combiné à d'autres techniques de l'Intelligence Artificielle, pour ce domaine.

Le schéma[2] suivant permet d'en avoir un aperçu global :

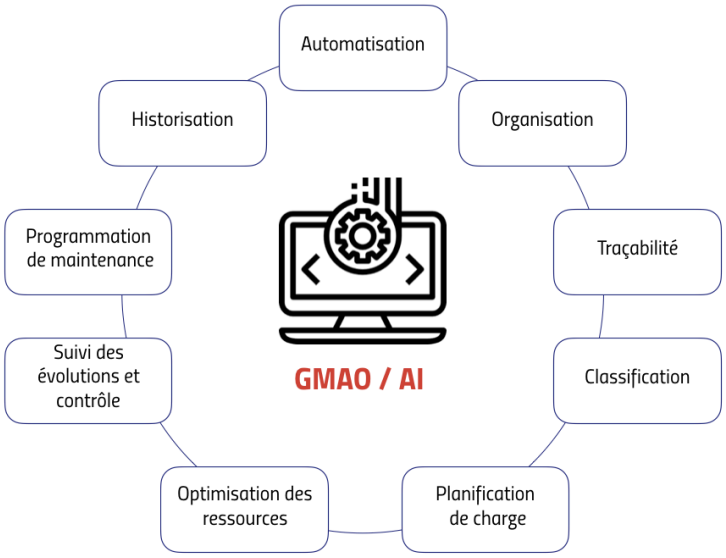
[2] *www.skillminer.com*

- **Focus 2 :**
  La gestion de la maintenance assistée par Intelligence Artificielle

Dans le secteur de l'Industrie et de l'Énergie, une des grandes possibilités offerte par l'Intelligence Artificielle est celle de la maintenance prédictive, et qui permet notamment :

- De diagnostiquer en temps réel l'état de fonctionnement de la chaîne de production ou d'un équipement à surveiller
- D'analyser les risques futurs de défaillance technique et identifier les pièces concernées
- De déterminer en conséquence un plan de maintenance prédictive de l'équipement en minimisant son temps d'arrêt

Pour alimenter les algorithmes de prédiction, on remonte les données de capteurs connectés, mais une possibilité d'enrichissement de ces données consiste à connecter la GMAO (Gestion de Maintenance Assistée par Ordinateur). Cette base de données contient notamment des données textuelles concernant les événements d'intervention/maintenance sur l'équipement.



Le Natural Language Processing permet de

- Repérer et classifier les mots-clés d'un texte décrivant un événement de maintenance : le matériel concerné et son état, le technicien intervenant sur le matériel, le type d'événement, la date…
- Pouvoir rechercher facilement et corréler ces données une fois classifiées
- Utiliser ces données, avec d'autres de type capteurs, pour des algorithmes de Machine Learning déterminant un plan de maintenance prédictive

- **Focus 3 :**
  L'analyse automatisée de la satisfaction client

A l'heure où l'expérience utilisateur prend tant d'importance et où les méthodes agiles préconisent le co-développement des solutions avec les utilisateurs, la **compréhension de la satisfaction client est un enjeu clé pour bon nombre de secteurs d'activité.**

Le Natural Language Processing permet une automatisation de la détection, l'analyse et la surveillance de cette satisfaction client, de la manière suivante :

- Détermination de sites web clés pour l'étude de la satisfaction client et chargement des contenus via APIs/crawling
- Détection de messages concernant tel produit ou telle marque
- Analyse automatisée du contenu pour déterminer la tonalité et le sentiment du message
- Analyse approfondie pour cibler les caractéristiques appréciées/critiquées
- Agrégation de l'analyse au sein d'un tableau de bord permettant d'avoir une vision globale du produit/de la marque sur les sites ciblés

- **Focus 4 :**
  Le suivi de la conformité produit

Le Natural Language Processing peut aussi être utilisé pour **suivre et comprendre de manière automatique l'évolution des textes légaux**, règlementaires, et les normes concernant un produit, ou un secteur d'activité. Ce cas d'usage peut se décliner comme suit

- Collecte des textes légaux ou règlementaires sur le périmètre concerné (produit, secteur), par exemple sur des sites comme Legifrance ou AFNOR.
- Analyse par l'humain des modalités d'expressions des obligations dans ce texte et de leur structure logique
- Choix de textes de référence en tant que corpus d'apprentissage des algorithmes, et tagging de ces textes, en particulier des mots-clés, des catégories importantes de mots, des coordinations logiques
- Implémentation des algorithmes de Natural Language Processing et constitution d'une base de données s'appliquant au périmètre considéré

A PROPOS

AXEL DE GOURSAC est directeur des opérations de Myriad.

Après avoir été diplômé de l'École Polytechnique et de l'École Normale Supérieure de Paris, il a soutenu en 2009 une thèse de doctorat en Mathématiques et Physique aux Universités de Paris-Sud et de Münster en Allemagne.

Puis, il a obtenu un poste de manager de projets de recherche à l'Université Catholique de Louvain et au Fond National de la Recherche scientifique (FNRS, Belgique) en Mathématiques et applications.

Passionné de science et de technologie, il est également un chercheur internationalement reconnu et un expert en Machine Learning et Natural Language Processing. Il dirige maintenant le département opérationnel de Myriad.

MYRIAD est une société de service qui assure le conseil et le déploiement de solutions d'Intelligence Artificielle et de Big Data.

Myriad offre aux Entreprises une véritable expertise dans les domaines analytiques, de Science des Données et d'Architecture. Aujourd'hui, les sociétés reconnaissent qu'il est nécessaire de s'affranchir d'une organisation des données en silos pour en révéler la valeur. Cela suppose de mettre en place une source unique pour les données de l'entreprise, qu'elles soient ou non structurées.

Myriad accompagne ses clients dans le cadre de leur transformation vers le Big Data et leur fournit une assistance globale pour l'implémentation de cette transformation, allant d'une stratégie de données claire au Machine et Deep Learning.

SUIVEZ-NOUS SUR LINKEDIN

Retrouvez notre **actualité**, nos **articles,**
nos **livres blancs**, nos **innovations**
et bien plus encore !

+33 (0)1 85 08 34 98
communication@thecodingmachine.com

56 rue de Londres
75008 Paris

TheCodingMachine
TCM://