Le Traitement Automatique du Langage

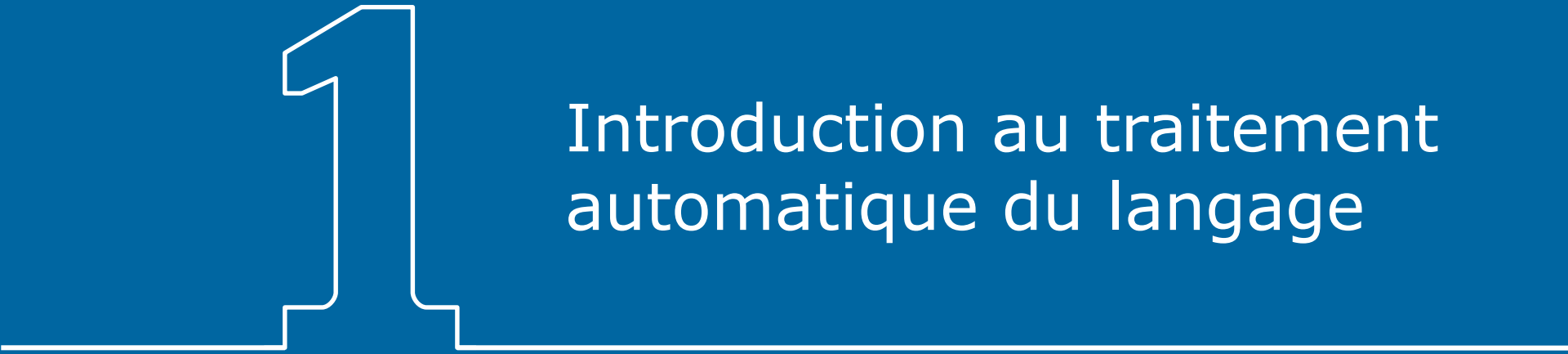
Au service des entreprises

Faïza Boulahya – BRGM
Stéphane Hudec - ATOS

16-10-2019

Géosciences pour une Terre durable
brgm

Atos

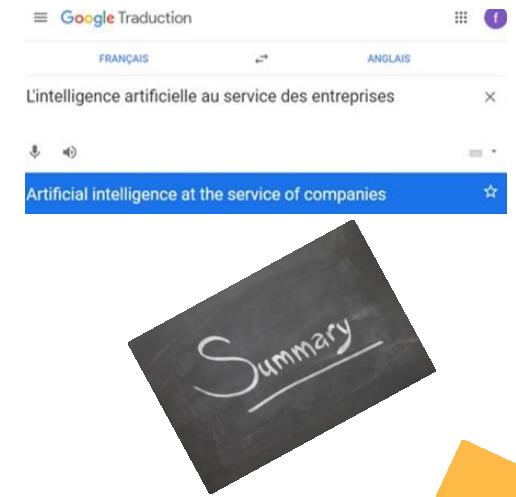


Introduction au traitement automatique du langage

Traitement automatique du langage

Pour quoi faire?

- Comprendre, manipuler et générer du langage naturel (vocal / textuel)
- Usages :
  - Traduction automatique
  - Aide à la rédaction : correcteur d'orthographe, saisie prédictive
  - Génération automatique : Résumés, articles de journaux
  - Traitement automatique des mails
  - Moteur de recommandation
  - Agents conversationnels, assistants vocaux, bots
  - Analyse de communauté, d'opinion/e-réputation
  - Extraction de connaissance (valorisation des fonds bibliographiques)
  - E-recrutement : Analyse automatique de CV - adéquation aux offres



Traitement automatique du langage

Comment ça marche ?

- Discipline qui fait appel à la linguistique, l'informatique et IA
- Les grandes étapes :
  - Prétraitements
    - Segmentation du texte (tokenisation)
    - Normalisation (minuscule – sans accent)

- Le premier Forum régional « L'intelligence artificielle au service des entreprises » se tient mercredi 16 octobre 2019 au LAB'O à Orléans.
- Le - premier - Forum - régional - « - L - ' -intelligence - artificielle - au - service – des - entreprises - » - se - tient - mercredi – 16 - octobre - 2019 - au – LAB - ' - O - à - Orléans - .
- le - premier - forum - regional - « - l - ' -intelligence - artificielle - au - service – des - entreprises - » - se - tient - mercredi – 16 - octobre - 2019 - au – lab - ' - o - a - orleans - .

Traitement automatique du langage

Comment ça marche ?

- Les grandes étapes :

  - Analyse /étiquetage morpho-syntaxique part-of-speech tagging (grâce au contexte + connaissances lexicales)
    - Lemmatisation/racinisation
    - Identification des catégories grammaticales (noms, verbes,...)
    - Constructions grammaticales

- Le premier Forum régional « L'intelligence artificielle au service des entreprises » se tient mercredi 16 octobre 2019 au LAB'O à Orléans.

- régional
  - Lemme : régional
  - Racine : région

- Le premier Forum régional « L'intelligence artificielle au service des entreprises » se tient mercredi 16 octobre 2019 au LAB'O à Orléans.

Déterminant
Préposition+Déterminant
Préposition
Ponctuation

Adjectif
Nom propre
Nom commun
Pronom réfléchi
Verbe

Traitement automatique du langage

Comment ça marche ?

- Les grandes étapes :

- Analyse sémantique, donner du sens pour
  - Identifier les entités nommées
  - Comprendre l'intention
  - Résumer

- Classification
- Identifier des topics
- Analyse de sentiments

- Le premier Forum régional « L'intelligence artificielle au service des entreprises » se tient mercredi 16 octobre 2019 au LAB'O à Orléans.

- Le premier Forum régional « L'intelligence artificielle au service des entreprises » se tient mercredi 16 octobre 2019 au LAB'O à Orléans.

Traitement automatique du langage

Quels outils ?

Open source

spaCy

Natural Language ToolKit

python

POLYGLOT

scikit learn

UNITEX 3.1

Stanford University NLP Natural Language Processing

TreeTagger

GATE

OPEN NLP

Textométrie

Commercial

Natural Language API 4

Google Cloud Platform

Amazon Comprehend

NL API

REFINITIV INTELLIGENT TAGGING

OPEN CALAIS

opentext

MonkeyLearn

Géosciences pour une Terre durable brgm

Atos

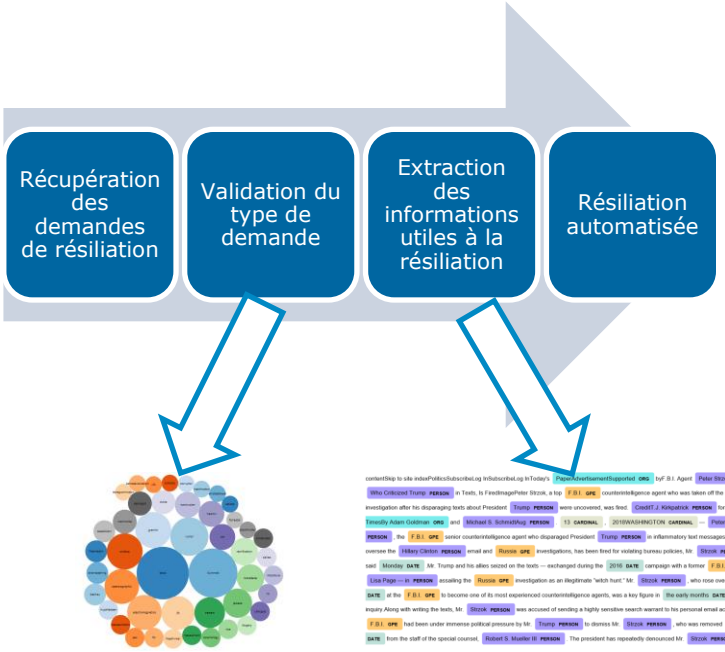
2

Exemples d’applications sur les données de l’entreprise

Traitement automatique du langage

Quelques mises en application

- **Traitement automatisé des demandes de résiliation de contrats d'assurance**
- Techniques mises en œuvre:
  - Reconnaissance de sujet afin de valider que le document est bien une demande de résiliation
  - Extraction d'entités nommées comme par exemple les noms, prénoms, n° de contrat pour résiliation par un robot
- Outils
  - Stanford Core NLP
- Limites
  - Meilleure performance des modèles en anglais pour les entités standards (nom, …)
  - Bon jeu de données nécessaire



Traitement automatique du langage

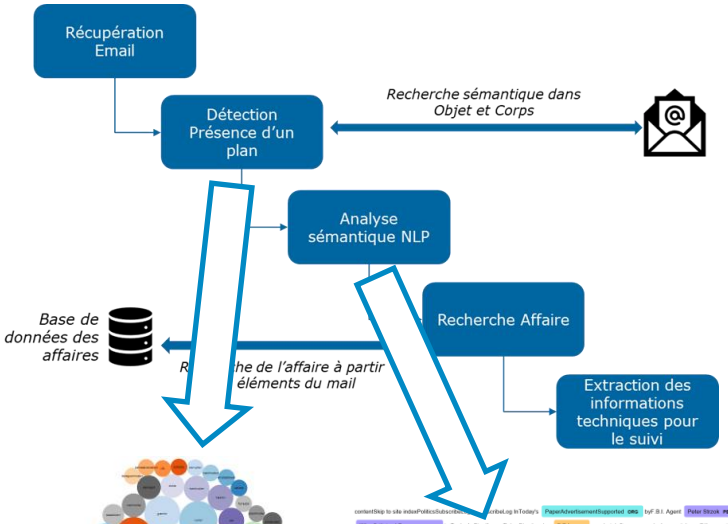
Quelques mises en application

- **Support utilisateurs**
  - meilleure connaissance des usages – satisfaction – recommandation de contenu
- Techniques mises en œuvre:
  - Annotation – indexation automatique
  - Classification – extraction de mots clés
  - Analyse de sentiments
- Outils
  - Gate
  - TreeTagger, python (TextBlob, RAKE)
- Limites
  - Travail encore en phase exploratoire
  - Savoir poser les objectifs : (indicateur global de satisfaction -> sujets qui impactent le plus la satisfaction)
  - Poids de la labellisation manuelle sur les résultats

Traitement automatique du langage

Quelques mises en application

- **Classification d'emails pour un cabinet d'architecture**
- Techniques mises en œuvre:
  - Reconnaissance de sujet afin de trier les emails contenant un plan d'architecture
  - Extraction d'entités nommées pour l'alimentation du dossier de suivi par un robot
- Outils
  - Stanford Core NLP
- Limites
  - Les messages peu « loquaces » ne sont pas détectés
  - Bon jeu de données nécessaire
  - Gestion des « faux négatifs »



Topic extraction

Named entity recognition

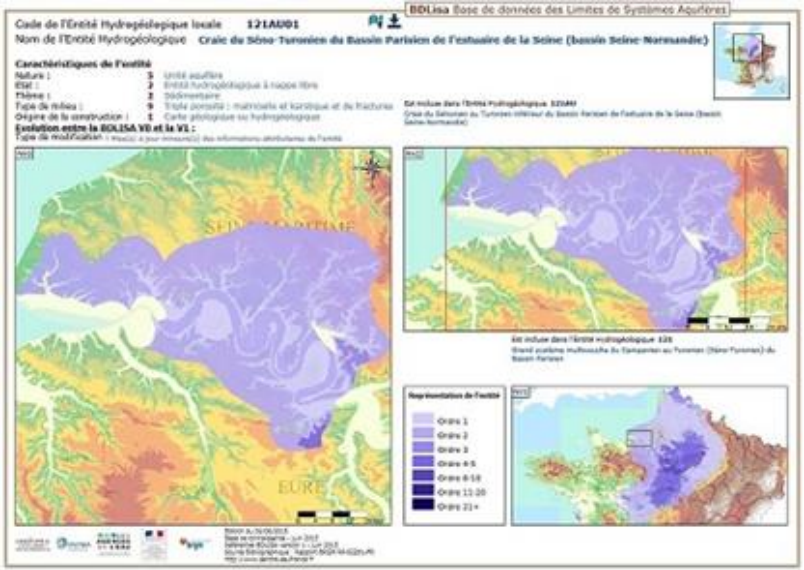
Géosciences pour une Terre durable brgm

Atos

Traitement automatique du langage

Quelques mises en application

- **Extraction de connaissance**
  - Identifier les entités de la BDLISA (Subdivision du sous-sol en réservoirs - aquifères / ou non réservoirs) dans les rapports brgm



Niveau 0 - Formations superficielles		
Niveau 1 - National	**Niveau 2 - Régional**	**Niveau 3 - Locale**
080AA72-- Formations des Limons des plateaux (code géol.: LP) dans l'extension de l'entité régionale : 107AK		
107-- Grand système multicouche de l'Oligo-Miocène du Bassin Parisien	**107AK**-- Calcaires de Brie du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie et Loire-Bretagne)	**107AK01**-- Calcaires de Brie du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie et Loire-Bretagne)
110-- Grand domaine hydrogéologique de l'Oligocène inf. à l'Eocène sup. (Sannoisien au Ludien) du Bassin Parisien	**110AA**-- Marnes vertes et supra-gypseuses du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie majoritairement et bassin Loire-Bretagne)	**110AA01**-- Marnes vertes et supra-gypseuses du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie majoritairement et bassin Loire-Bretagne)
	113BA-- Faciès marneux du Ludien moyen de l'Eocène sup. du Bassin Parisien (bassin Seine-Normandie)	**113BA01**-- Faciès de transition (marnes et calcaires) du Ludien de l'Eocène sup. du Bassin Parisien
	113AI-- Marnes Infra-gypseuses de l'Eocène du Bassin Parisien	**113AI01**-- Marnes Infra-gypseuses de l'Eocène du Bassin Parisien
		113AK01-- Sables de Monceau, de Marines, de Cresnes du Marinésien supérieur (Bartonien inf.) du Bassin Parisien
	113AK-- Sables, calcaires et grès du Bartonien (Eocène) du Bassin Parisien	**113AK03**-- Calcaires de Saint-Ouen du Bartonien inf. du Bassin Parisien
		113AK05-- Sables du Marinésien (sables de Mortefontaine, Calcaire de Ducy, Sables d'Ezanville) et de l'Auversien (Sables de BeauChamps, d'Auvers) du Bassin Parisien
113-- Grand système multicouche de l'Eocène du Bassin Parisien		

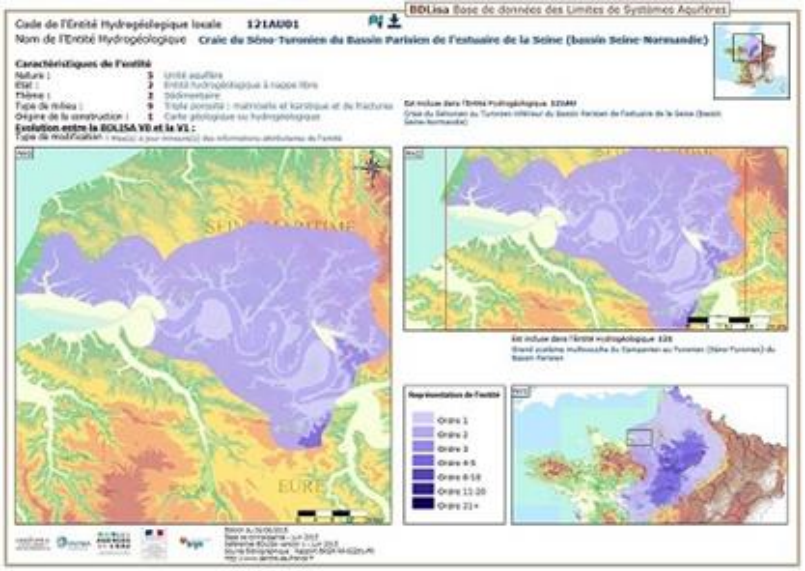
Ex. Calcaires de Brie du Rupélien (Oligocène inf.) du Bassin Parisien (bassin Seine-Normandie et Loire-Bretagne)

Géosciences pour une Terre durable brgm Atos

Traitement automatique du langage

Quelques mises en application

- **Extraction de connaissance**
- Techniques mises en œuvre:
  - Indexation des termes BD LISA et des lexiques lithologiques et âges géologiques
  - Algorithme de fenêtre glissante sur le texte avec indicateurs par rapport à l'index, considération de la localisation et de l'échelle des temps géologiques pour la parenté des périodes géologiques
- Outils
  - SolR + algorithmes dédiés
- Limites
  - Compléter la démarche (Espace vectoriel de mots, ML)



Géosciences pour une Terre durable brgm Atos

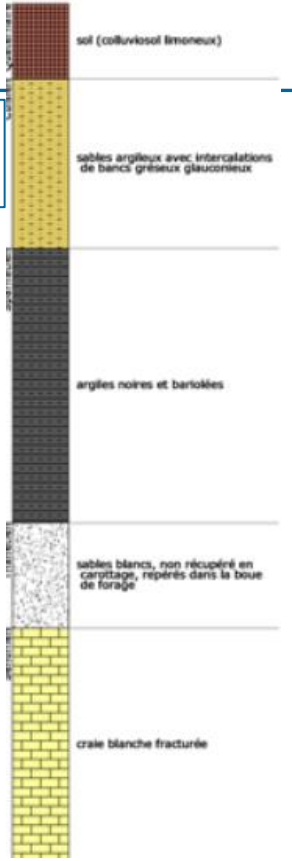
Traitement automatique du langage

Quelques mises en application

- **Extraction de connaissance**
  - Transformer les descriptions lithologiques complexes en langage naturel en une information structurée plus exploitable
- Techniques mises en œuvre
  - Analyse morpho-syntaxique
    - Tagging
    - Séquence
- Outils
  - Développement spécifique
- Limites
  - Formes plus narratives
  - Remplacer développement spécifique par SEM/Stanford NLP et l'usage de co-références

Alternance de calcaire argileux et de marnes vertes à nodules ferrugineux

Alternance
- calcaire
  - argile
- marne
  - vert
  - nodule
    - fer



Géosciences pour une Terre durable brgm Atos

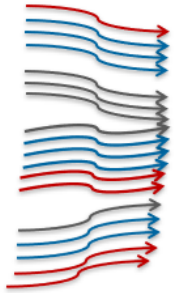
3

Exemples d’applications sur les données des réseaux sociaux

Traitement automatique du langage

Quelques mises en application

- **Waves – Détection d'anomalie et contextualisation (contexte distribution d'eau)**
- Techniques mises en œuvre:
  - Machine learning pour la détection d'anomalies
  - Reconnaissance de sujets afin de filtrer les informations
  - Génération de résumés
  - Analyse sentimentale
- Outils
  - Scikit-learn (Machine learning)
  - Python NLTK (Résumés)
  - Apache Open NLP
- Limites
  - Open NLP limité en français



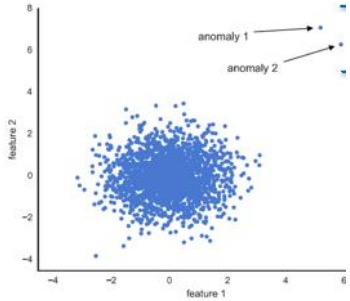
Données de capteurs

Open Data
Rss Feeds
Réseaux sociaux

Filtrer l'Open Data en se basant sur les coordonnées **spatio-temporel des anomalies détectées**

Détection d'anomalies

Explication d'anomalies



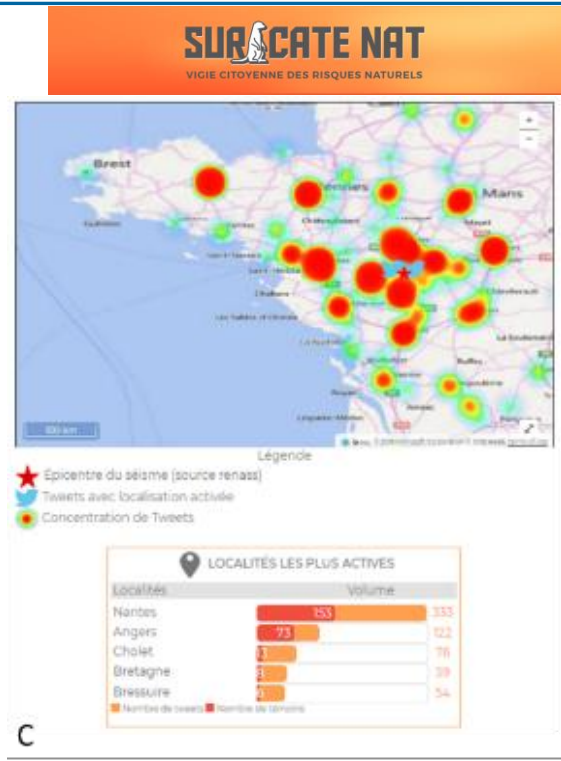
Géosciences pour une Terre durable brgm

Atos

Traitement automatique du langage

Quelques mises en application

- **Plateforme Suricate-NAT**
  - Dresser un paysage réaliste de la situation lors de catastrophes naturelles à partir des messages échangés sur Twitter
- Techniques mises en œuvre:
  - Extraction de sens par classification
    - Indications sur le phénomène : hauteur d'eau, intensité secousses sismiques, etc.
    - Quel type d'effets : effets sur les personnes, dommages matériels
    - Indicateurs de sévérité?
  - Géolocalisation / Emprise spatiale par extraction d'entité nommées



Traitement automatique du langage

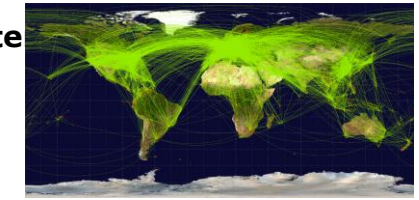
Quelques mises en application

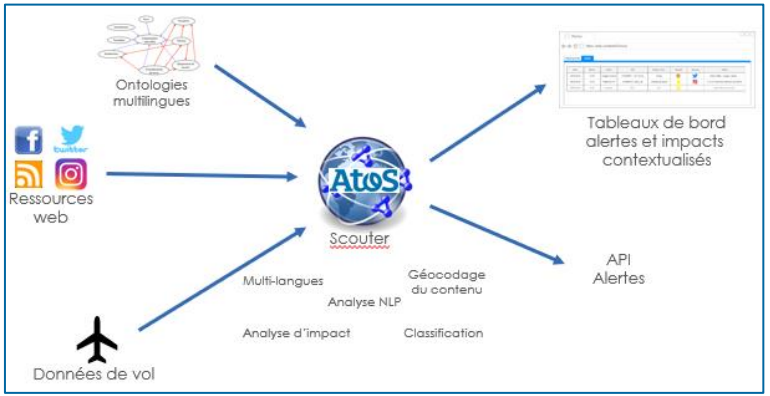
- **Plateforme Suricate-NAT**
- Outils
  - Python NLTK + scikit-learn (classification)
  - SEM (NER)
- Limites
  - Labellisation manuelle pour la classification couteuse
  - Désambiguisation des toponymes trouvés par l'analyse du texte en fonction du contexte
  - Géolocalisation à la commune insuffisante pour les inondations, submersions

Traitement automatique du langage

Quelques mises en application

- **Plateforme de détection d'événements pouvant impacte le trafic aérien**
- Techniques mises en œuvre:
  - Reconnaissance de sujet afin de filtrer les informations
  - Extraction d'entités nommées pour déterminer la géolocalisation du contenu
  - Détection de la langue
- Outils
  - Stanford Core NLP
  - Tensorflow (classification)
  - Azure Translator Text
- Limites
  - Bon jeu de données nécessaire pour la classification des événements
  - Gestion des langues





Questions ?

Géosciences pour une Terre durable
brgm

Atos