Le TAL au BRGM

Sharleyne Lefèvre

Master 2 Linguistique Outillée et Traitement Automatique des Langues

Sommaire

1. Parcours scolaire

- Licence en Sciences du Langage
  - 3ème année en parcours FLE
- Master 1 - Linguistique Appliquée aux Sciences et Technologies de l'Information et de la Communication
  - Stage au BRGM (4 mois)
- Master 2 - Linguistique Outillée et Traitement Automatique des Langues
  - Recherche d'un stage (6 mois)

2. Présentation du BRGM

- Établissement public à caractère industriel et commercial (EPIC) crée en 1959
- Établissement de référence dans les applications des sciences de la Terre pour gérer les ressources et les risques du sol et du sous-sol
- Service géologique national
- Placé sous la tutelle
  - du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation,
  - du ministère de la Transition écologique et solidaire,
  - du ministère de l'Économie et des Finances

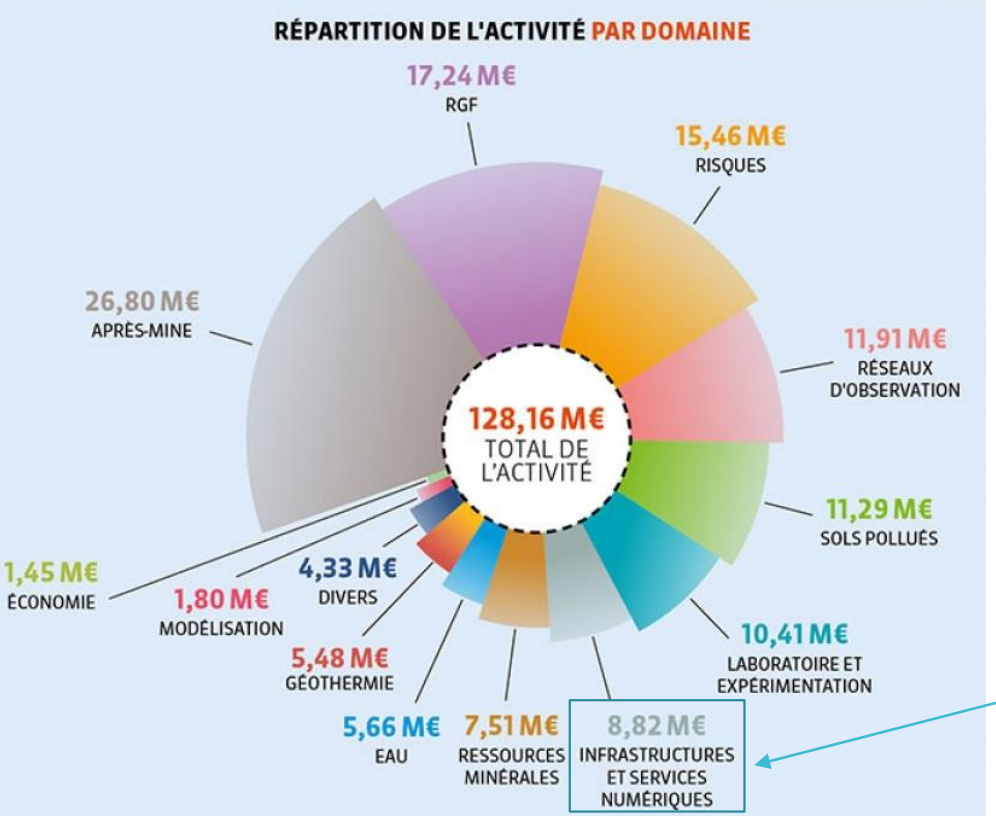
2. Présentation du BRGM

- Les actions du BRGM s'articulent autour de 5 missions :

- La recherche scientifique
- L'appui aux politiques publiques
- La coopération internationale
- La sécurité minière
- La formation

2. Présentation du BRGM

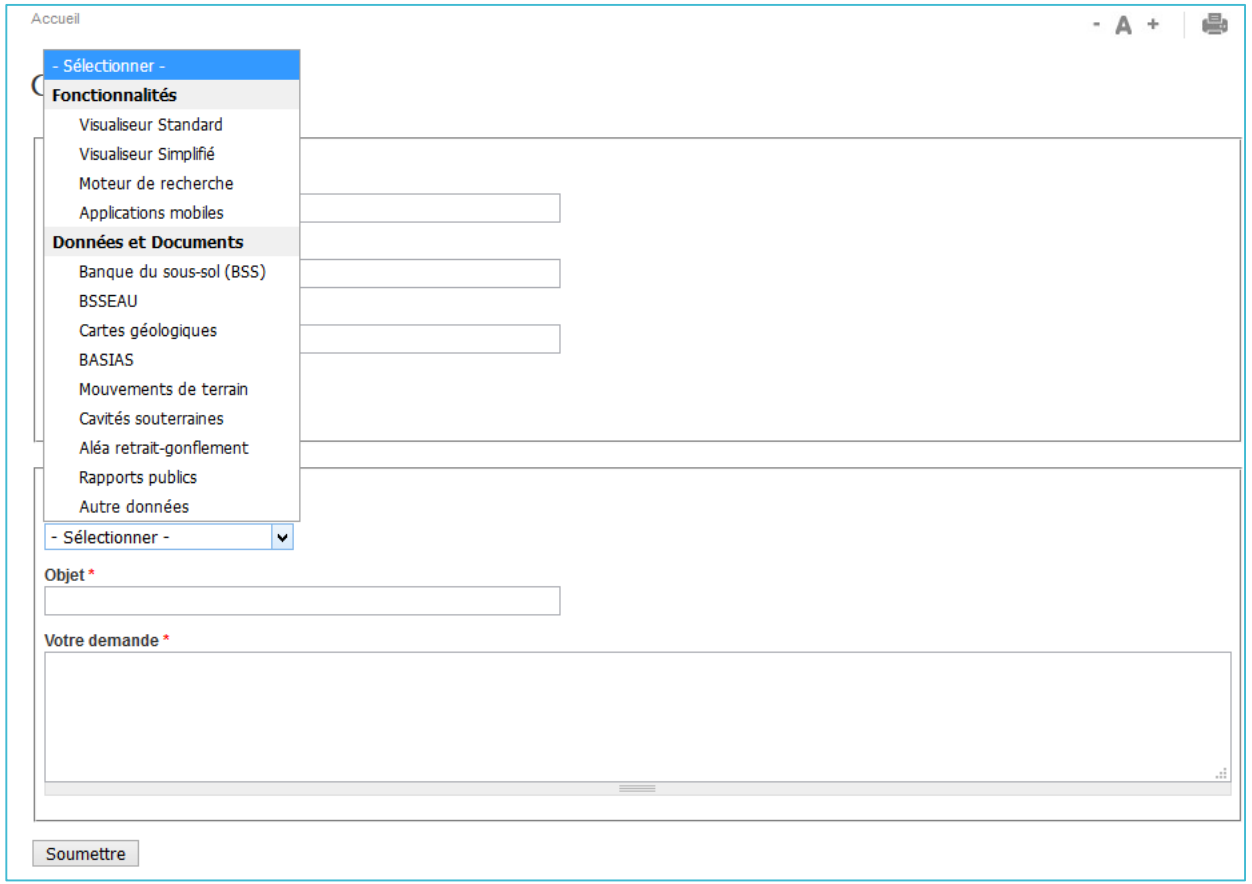
- Cette activité tourne autour de 10 grands domaines des géosciences :



Systèmes d'informations

2. Présentation du BRGM

- Direction des Infrastructures et Systèmes Numériques (DISN)
  - Développe et anime des sites web pour la diffusion des connaissances
- Accueil et Service aux Utilisateurs (ASU)
  - Support des demandes des utilisateurs (formulaires de contact avec thématiques existantes)



3. Missions de stage

- 2 composantes :
  - **Messages des utilisateurs** (4623 pour l'année 2017):
    - Enjeux :
      - 1 – Mettre en place des éléments de classification et aider le transfert des messages vers les agents qualifiés en vue d'un potentiel ChatBot
      - 2- Faciliter la constitution d'éléments globaux de statistiques
  - **Rapports scientifiques**
    - Enjeu : Économiser du temps d'agents pour indexer des rapports en PDF non produits par le BRGM

4. Réalisations

- Messages
  - Nettoyage (expressions régulières)
  - Classification (expressions régulières)
  - Lemmatisation (TreeTagger)
  - Indexation automatique avec thésaurus de mots-clés
  - Analyse du sentiment
- Rapports
  - Extraction des termes « essentiels »
- Langage de programmation utilisé : Python 3

4.1. Méthode d'indexation automatique : Les thésaurus

- Méthode des thésaurus (sur chaque message)
  - 3 thésaurus utilisés
    - Thésaurus Sciences de la Terre (Ortolang),
    - Thésaurus thématique interne (indexation des rapports du BRGM > 28k rapports),
    - Thésaurus catégorisation risques (Géorisques)
  - Comparaisons message lemmatisé ↔ termes lemmatisés des thésaurus

4.1. Méthode d'indexation automatique : Les thésaurus

Message d'origine :

Nom Prénom (adresseMail@yahoo.fr) a envoyé un message en utilisant le formulaire de contact suivant : http://www.georisques.gouv.fr/contact. Bonjour, Je suis à la recherche d'une carte indiquant les hauteurs d'eau dans ma commune de Mareuil sur Lay générées par une rupture du barrage le Marillet. Pouvez vous m'indiquer sur quel site rechercher ? Je vous en remercie par avance. Cordialement

Message nettoyé :

Bonjour, Je suis à la recherche d'une carte indiquant les hauteurs d'eau dans ma commune de Mareuil sur Lay générées par une rupture du barrage le Marillet. Pouvez vous m'indiquer sur quel site rechercher ? Je vous en remercie par avance. Cordialement.

Message lemmatisé :

*bonjour, je suivre|être à le recherche de un **carte** indiquer le hauteur de **eau** dans ma commune de Mareuil sur Lay générer par un **rupture** du barrage le Marillet . pouvoir vous me indiquer sur quel site rechercher ? je vous en remercier par avance . cordialement .*

Mots-clés : rupture - barrage – carte – eau

4.2. Méthode d'indexation automatique : RAKE

- Méthode RAKE – Rapid Automatic Keyword Extaction
  - Algorithme qui détermine les phrases clés dans des textes
  - Plus il y a de données meilleurs seront les résultats
  - Méthode implémentée par un développeur - Module « RAKE » sous Python 3
  - Prend en compte les cooccurrences (ce qu'il y a avant et après le terme pivot)
  - Se base sur :
    - la **fréquence** (nombre d'apparition des mots dans le texte)
    - le **degré (**relation des mots entre eux)
    - Calcul un score

4.2. Méthode d'indexation automatique : RAKE

- Méthode RAKE – Rapid Automatic Keyword Extaction
  - Inconvénient :
    - Mots-clés = phrases

Exemple d'un mot-clé :

« mouvements cisaillants présentant une composante horizontale nette »

- Analyse de la forme de mots-clés = structure syntaxique N-ADJ / N / N-ADJ-ADJ
- Utilisation du Chunk de TreeTagger (analyseur groupes syntaxiques)
  - Récupération des structures pertinentes

4.2. Méthode d'indexation automatique : RAKE

<NP>			
	mouvements		NOM mouvement
</NP>			
<AP>			
	cisaillants	ADJ	cisaillants
</AP>			
<VN>			
	présentant	VER:ppre	présenter
</VN>			
<NP>			
	une	DET:ART	un
	composante	NOM	composante
</NP>			
<AP>			
	horizontale	ADJ	horizontal
</AP>			
<AP>			
	nette	ADJ	net
</AP>			

« **mouvements cisaillants** présentant **une composante horizontale nette** »

- mouvements cisaillants
- une composante horizontale nette

→ Algorithme modulable

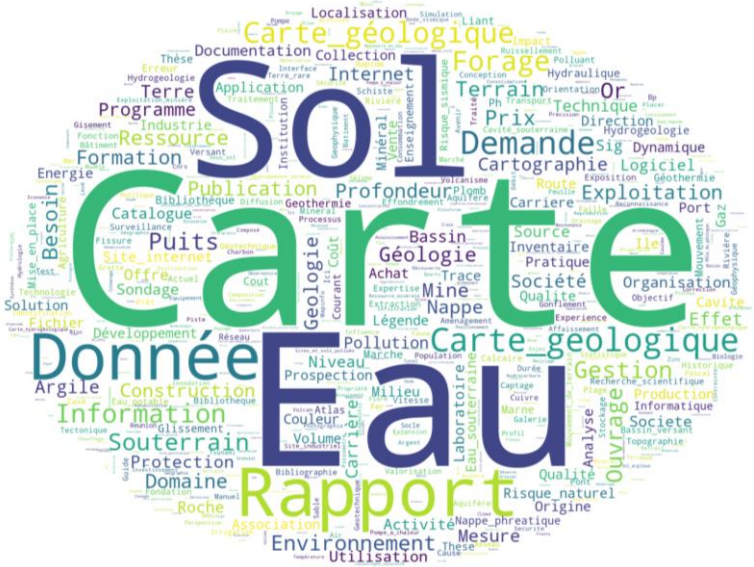
4.3. Analyse de sentiments (sur les demandes)

- Déterminer la proportion de messages positifs – négatifs – neutres
- Méthode 1 - liste de termes porteurs de sentiments (fichier XML)
  - Matching de termes
  - Récupération de la polarité
  - Faire une moyenne des polarités dans le message → donne polarité globale du message
  - Méthode pas adaptée, pas de contexte (négation)

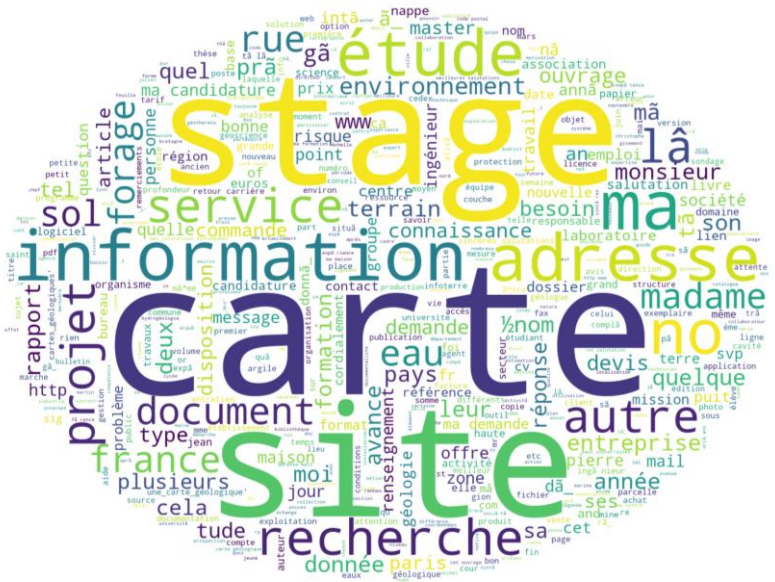
```
<sentiment language="fr" version="1.1" author="Tom De Smedt, Walter Daelemans, fabelier.org" license="PDDL">
 <word form="abandonné" pos="JJ" polarity="-0.30" subjectivity="0.40" intensity="1.0" confidence="0.9" />
 <word form="abandonnée" pos="JJ" polarity="-0.30" subjectivity="0.40" intensity="1.0" confidence="0.8" />
 <word form="abandonnées" pos="JJ" polarity="-0.30" subjectivity="0.40" intensity="1.0" confidence="0.8" />
 <word form="abandonnés" pos="JJ" polarity="-0.30" subjectivity="0.40" intensity="1.0" confidence="0.8" />
```

- Méthode 2 – Module TextBlob de Python

4.4 - Visualisations



Méthode des thésaurus



Méthode RAKE

Conclusion

- Indexation par thésaurus : méthode basique qui marche bien
- Indexation par RAKE : nécessite d'enrichir l'algorithme
- Analyse de sentiments : méthodes pas adaptées malgré une analyse morpho-syntaxique, l'apprentissage automatique aurait été préférable

Merci !

Des questions ?