« Question answering system » pour la langue française.

Un modèle de NLP utilisant la vectorization de phrase, le « sentence embedding ».

Auteurs : Brice **BROUSSEAU-RIGAUDIE**, Basile **ROTH**

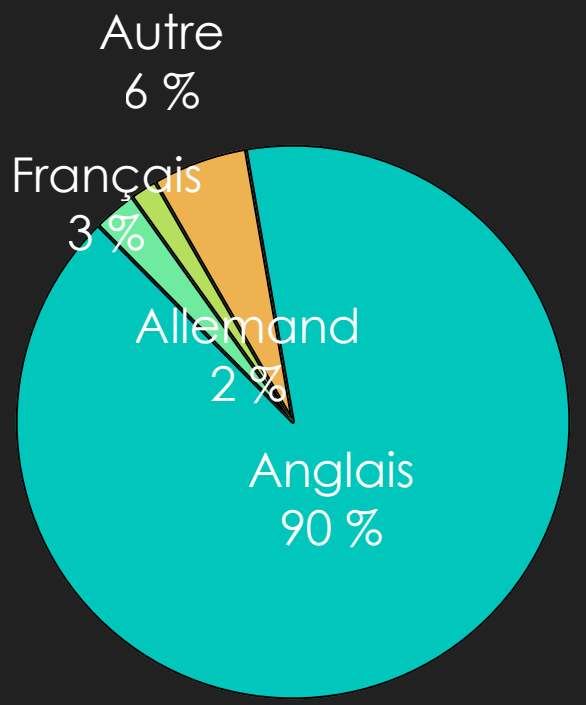
Professeur : Sylvie **RATTÉ**, PhD

MTI830 : Forage de textes et de données audiovisuelles

ÉTS
Le génie pour l'industrie
ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
Université du Québec

Contexte

- Système de question/réponse : un des plus gros sujet de recherche en NLP.
- Word2Vec, Bert, FastText : modèles de vectorization de mots pour la quantification de la langue anglaise.
- Possibilité d'adapter les techniques (ex : vectorization du language) utilisées pour le traitement de données anglaises au language français.



+ de 90% des travaux de recherche en NLP réalisés en langue anglaise. *

* **Dans le domaine du travail sur le texte clinique (étude par PubMed) :** Névéol, A., Dalianis, H., Velupillai, S. *et al.* Clinical Natural Language Processing in languages other than English: opportunities and challenges. *J Biomed Semant* **9**, 12 (2018)

Objectifs

- **Création d'un système de question/réponse en français** :
  - Trouver la **phrase** contentant la réponse à une question dans un texte.
  - Utilisation du **French Question Answering Dataset** :
    - **Format** : Question + Contexte + Réponse
  - Extraire les caractéristiques des données et construire les vecteurs pour chaque phrase (**sentence embedding**).
  - Tester des modèles supervisés sur les données préparées pour l'apprentissage.

FQuAD

- Créé en 2020
- Dataset français de compréhension le plus large
- Structure similaire au SQuAD

Dataset	Articles	Paragraphs	Questions
Train	117	4921	20731
Development	18	768	3188
Test	10	532	2189

Répartition des données dans le FQuAD 1.1

Tom Brendlé, Martin d'Hoffschmidt, Wacim Belblidia and Quentin Heinrich. FQuAD: French Question Answering Dataset. https://arxiv.org/pdf/2002.06071.pdf, 2020. [Online; accessed 01-July-2020]

Méthodologie

1. Sélection d'un modèle pour vectorizer les mots

WORD2VEC

fastText

2. Pré-traitement

- Enlever les mots communs (stop words)
- Minuscule
- Ponctuation, …

3. Extraction des caractéristiques : sentence embedding

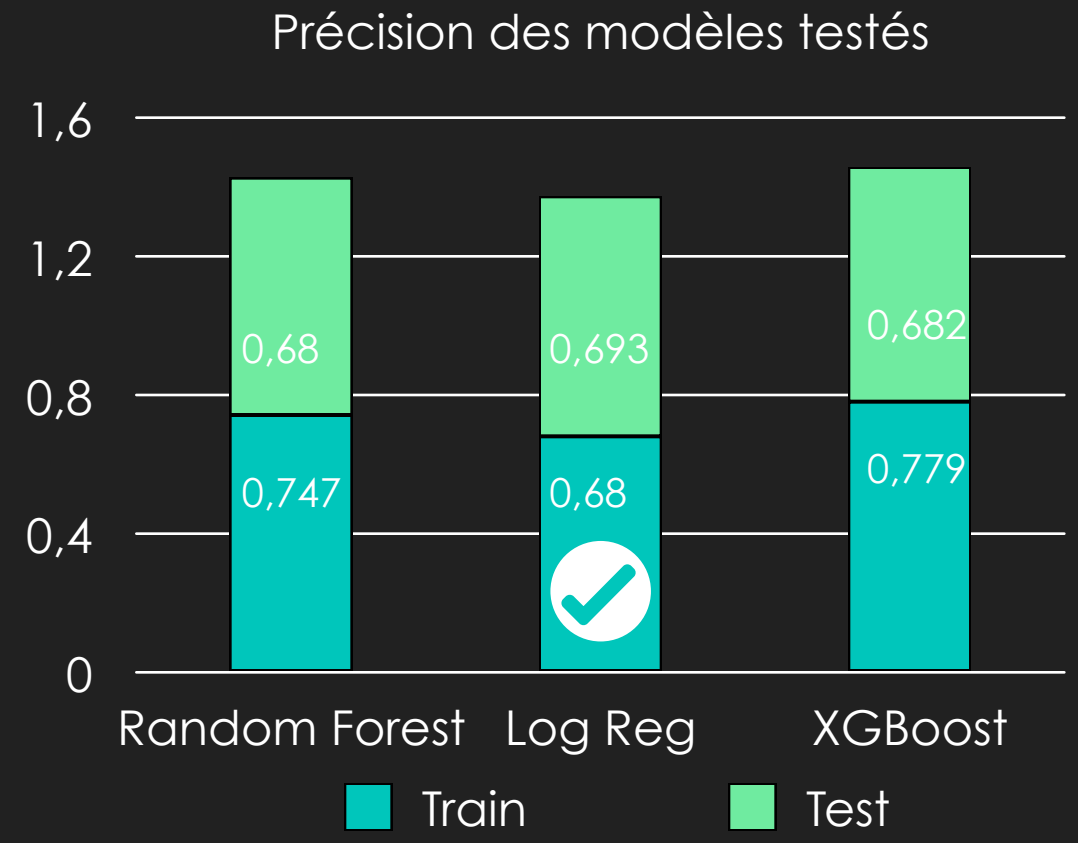
4. Sélection, entrainement et validation du modèle final

Résultat

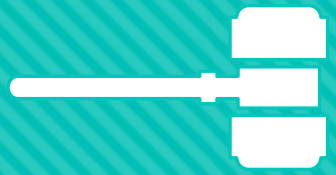
Sélection des features :

choix du modèle de vectorization du language

FastText Skip Gram	FastText CBow	Word2Vec Skip Gram	Word2Vec Skip Gram
67,8 %	54,2 %	63,03 %	55,04 %



Conclusion



- Modèle supervisé pouvant trouver la phrase contenant la réponse à une question dans un texte.
- Caractéristiques construites avec le modèle *fast*Text de Facebook AI Research.
- Performance : 69,3 % de précision.

Discussion

- Possibilité de rajouter de nouvelles caractéristiques pour entraîner le modèle …
- Tester des modèles non-supervisés pour en évaluer la performance.
- Trouver la réponse précise dans la phrase en utilisant la décomposition sémantique.