Appariement des FAQ: BoW vs sBERT

Chengwanli YANG
Mustapha LAZREG

22/02/2022

Plan

Introduction

- Objectif : repérer la bonne réponse correspondant à la question.
- Données : quatre sous-corpus:
  - 1_faq_dmc: 5 questions/réponses
  - 2_questions_sorbonne: 42 questions/réponses
  - 3_faq_syp: 11 questions/réponses
  - 3_syp: 52 questions/réponses
- Méthode : vectorisation + similarité cosinus.
  - Configuration 1 : utilisation de CountVectorizer (BoW)
  - Configuration 2 : utilisation de SentenceTransformer (sBERT)

Plan

Corpus

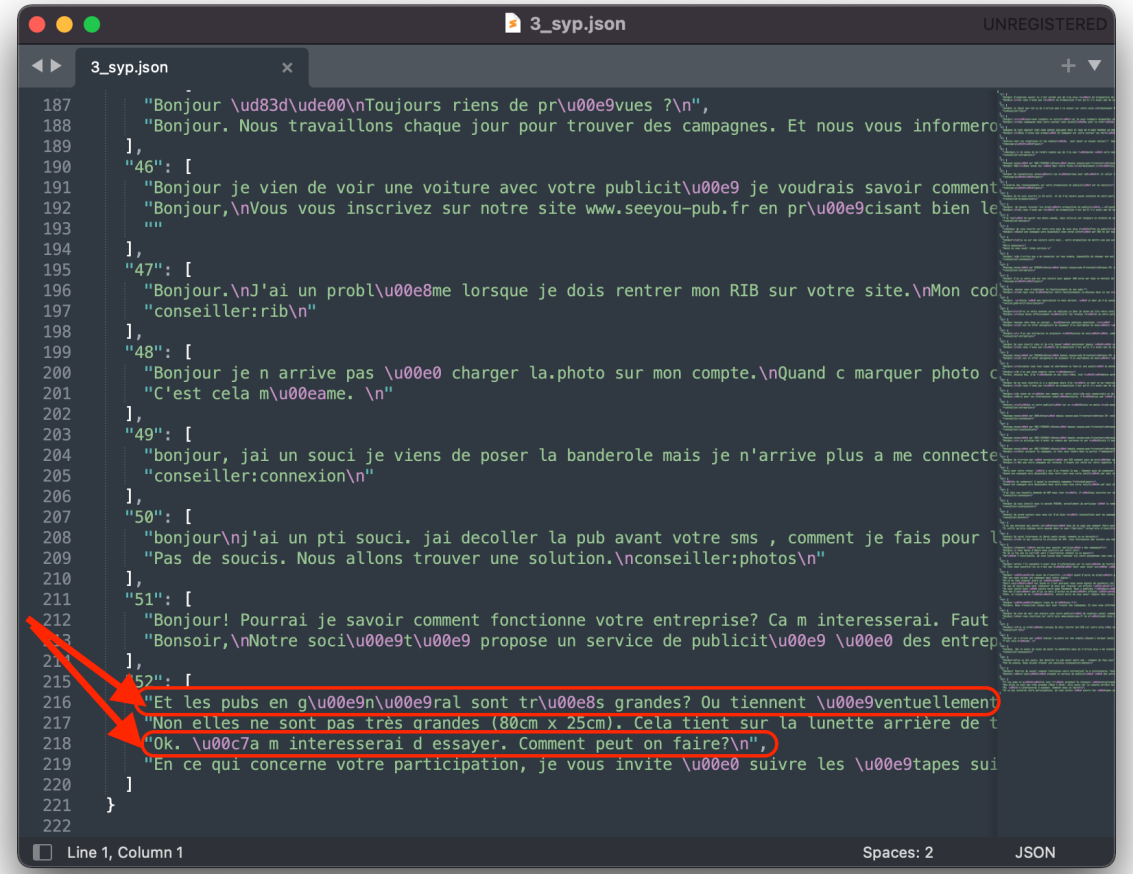

Figure: Un aperçu du corpus

Plan

SentenceTransformers

SentenceTransformers (Reimers, N., Gurevych, I., 2019)

- Performance plus forte que la similarité cosinus de BERT
- Vitesse de calcul plus rapide que celle de BERT(5s vs 65h pour clustering hiérarchique de 10 000 phrases)
- Modèle multilingue (jusqu'aux 50 langues)

Évaluation: Groupe de contrôle : CountVectorizer (Pedregosa, F et al., 2011)

Trois modèles pré-entraînés:

- "all-mpnet-base-v2" (1 billion training pairs, 420Mo) – meilleure qualité
- "all-MiniLM-L6-v2" (moins de dimensions, 80Mo) – 5 fois plus rapide et bonne qualité
- "distiluse-base-multilingual-cased-v1" (le français ,480Mo) – s'adapter bien à notre corpus

Plan

Résultat

```
La réponse 36 est la plus proche de la question 34, avec un score de 0.4090895652770996 : la réponse est mauvaise
La réponse 15 est la plus proche de la question 35, avec un score de 0.26917025446891785 : la réponse est mauvaise
La réponse 36 est la plus proche de la question 36, avec un score de 0.5493180751800537 : la réponse est bonne
La réponse 4 est la plus proche de la question 37, avec un score de 0.2893078625202179 : la réponse est mauvaise
La réponse 38 est la plus proche de la question 38, avec un score de 0.3671208322048187 : la réponse est bonne
La réponse 41 est la plus proche de la question 39, avec un score de 0.40827494859695435 : la réponse est mauvaise
La réponse 30 est la plus proche de la question 40, avec un score de 0.26767799258232117 : la réponse est mauvaise
La réponse 41 est la plus proche de la question 41, avec un score de 0.4880366921424866 : la réponse est bonne
La réponse 10 est la plus proche de la question 42, avec un score de 0.2844419777393341 : la réponse est mauvaise
------------------------------------------------------------
| fichier : ['1_faq_dmc'], model utilisé : distiluse-base-multilingual-cased-v1|
------------------------------------------------------------
La réponse 1 est la plus proche de la question 1, avec un score de 0.453148752450943 : la réponse est bonne
La réponse 2 est la plus proche de la question 2, avec un score de 0.37164175510406494 : la réponse est bonne
La réponse 3 est la plus proche de la question 3, avec un score de 0.347391813993454 : la réponse est bonne
La réponse 4 est la plus proche de la question 4, avec un score de 0.2033480852842331 : la réponse est bonne
La réponse 3 est la plus proche de la question 5, avec un score de 0.18484939634799957 : la réponse est mauvaise
------------------------------------------------------------
le temps d'exécution: 1671.7331938743591 secondes
```

Figure: Résultat obtenu

Évaluation

	Modèle\Corpus	3_faq_syp	3_syp	2_questions_sorbonne	1_faq_dmc		
sBERT	all-mpnet-base-v2	0.27	0.11	0.26	0.6	0.31	6900.68 sec
	all-MiniLM-L6-v2	0.25	0.15	0.26	0.4	0.265	2243.56 sec
	distiluse-base-multilingual-cased-v1	0.54	0.13	0.28	0.8	0.4375	1671.73 sec
BoW	CountVectorizer	0.25	0.07	0.26	0.4	0.245	0.69 sec

Table: Mesures de performance : précision et rappel

Evaluation

	Modèle	Fichier	Rang BR
1	Sentence Transformer (modèle multilingue)	faq_dmc (5 paires)	1.2
		questions_ sorbonne (42 paires)	17.83
		syp (51 paires)	28.25
		faq_syp (11 paires)	2
		Moyenne	**12.320**
2	Count Vectorizer	faq_dmc (5 paires)	1.8
		questions_ sorbonne (42 paires)	19.35
		syp (51 paires)	26.7
		faq_syp (11 paires)	3.79
		Moyenne	**12.910**

Figure: Rangs des bonnes réponses

Plan

Conclusion et Perspectives

Résultats :

- Nouveau aspect : plongement des phrases
- Bonne performance

Conclusion et Perspectives

Résultats :

- Nouveau aspect : plongement des phrases
- Bonne performance

Perspectives :

- Travailler sur un gros corpus
- Tester d'autres langues (Anglais principalement)
- Tester d'autres méthodes d'évaluation