# Evaluating the Robustness of Retrieval Pipelines

Balkis Dirahoui, Miray Senyuz

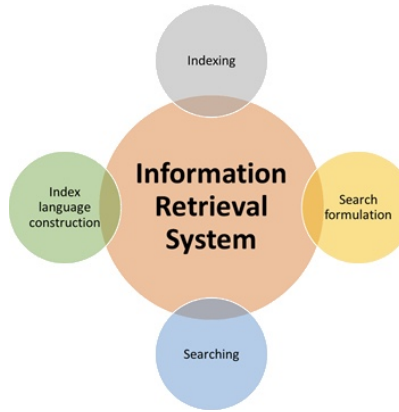
Sorbonne Université

17 septembre 2022

# Introduction

Les modèles de benchmark se sont révélés remarquablement efficaces pour les tâches de recherche d'information (RI).Néanmoins..

- Un utilisateur peut formuler **plusieurs** requêtes différentes...
- face à ce genre de requêtes,les modèles **perdent** notablement de leur performances



# Problématique et apport de l'article : Objectif

- **Objectif de [1]**
  - étudier la **robustesse**
  - **variations** de requêtes

# Problématique et apport de l'article : Variations

- ***Misspelling*** : fautes d'orthographe. e.g., day→d**y**a
- ***Naturality*** : supprimer les **stopwords**. e.g.,
  " ~~what are the~~ best restaurants near me " →"best restaurants near me"

# Problématique et apport de l'article : Variations

- ***Ordering*** : changer l'ordre des mots. e.g.,
  " **what** are the **best** restaurants near me " →"**best** are the **what** restaurants near me"
- ***Paraphrasing*** : reformuler. e.g.,
  " what are the best restaurants **near me** " →"what are the best restaurants **close to me**"

# Problématique et apport de l'article : Variations

| Category | Method Name | *M*('***what is durable medical equipment consist of***') |
|---|---|---|
| *Misspelling* | `NeighbCharSwap` | *what is durable* ***mdeical*** *equipment consist of* |
| | `RandomCharSub` | *what is durable* ***medycal*** *equipment consist of* |
| | `QWERTYCharSub` | *what is durable medical equipment* ***xonsist*** *of* |
| *Naturality* | `RemoveStopWords` | ~~***what is***~~ *durable medical equipment consist* ~~***of***~~ |
| | `T5DescToTitle` | ~~***what is***~~ *durable medical equipment* ~~**consist of**~~ |
| *Ordering* | `RandomOrderSwap` | ***medical*** *is durable* ***what*** *equipment consist of* |
| *Paraphrasing* | `BackTranslation` | *what is* ***sustainable*** *medical equipment* ~~consist of~~ |
| | `T5QQP` | *what is durable medical equipment* ~~***consist of***~~ |
| | `WordEmbedSynSwap` | *what is durable* ***medicinal*** *equipment consist of* |
| | `WordNetSynSwap` | *what is* ***long lasting*** *medical equipment consist of* |

# Protocole d'évaluation

**Dataset :**

- ***Antique*** [1].
- ***TREC 2019*** [2].

**Modèles évalués :**

- ***Modéles classiques*** : BM25,RM3.
- ***Modèles neuronaux*** : KNRM, CKNRM,
- ***Modèles transofrmer-based*** : BERT,T5,EPIC.

1. https://ir-datasets.com/antique.html
2. https://microsoft.github.io/msmarco/TREC-Deep-Learning-2019.html

## Protocole d'évaluation

**Métrique d'évaluation :** Nous utiliserons la métrique NDCG@10 :

$$NDCG@10 = \frac{DCG@10}{IDCG@10} \tag{1}$$

**Test d'hypothèse :** T-test de Student [2] avec une confiance de 95%.

# Résultats

Performances (nDCG@10) des différentes méthodes pour TREC face à différentes variations de requêtes, les flèches  démontrent un réel changement statistique selon un t-test.

| Catégorie | Variation | BM25 | RM3 | CKNRM | KNRM | EPIC | BERT | T5 |
|---|---|---|---|---|---|---|---|---|
| - | requête originale | 0.4795 | 0.5151 | **0.4795** | **0.4931** | 0.6231 | 0.4795 | 0.6995 |
| ***Mispelling*** | *NeighbCharSwap* | 0.2738 | 0.2771 | 0.3080 | 0.3034 | 0.3893 | 0.2738 | 0.4944 |
| | *RandomCharSub* | 0.2738 | 0.2338 | 0.2262 | 0.2398 | 0.2950 | 0.2314 | 0.3963 |
| | *QWERTYCharSub* | 0.2437 | 0.2509 | 0.2965 | 0.2573 | 0.3496 | 0.2437 | 0.4459 |
| ***Naturality*** | *RemoveStopWords* | $0.4778^{\downarrow}$ | 0.5104 | 0.4756 | 0.4782 | $0.6214^{\downarrow}$ | $0.4778^{\downarrow}$ | 0.6862 |
| | *T5DescToTitle* | 0.4217 | 0.4349 | 0.3927 | 0.3818 | 0.5061 | 0.4217 | 0.5717 |
| ***Ordering*** | *RandomOrderSwap* | 0.4795 | 0.5152 | 0.4707 | 0.4868 | 0.6227 | 0.4796 | $0.6970^{\downarrow}$ |
| ***Paraphrasing*** | *BackTranslation* | 0.3965 | 0.4227 | 0.3605 | 0.3909 | 0.5301 | 0.3965 | 0.6058 |
| | *T5QQP* | 0.4723 | 0.5046 | 0.4609 | 0.4423 | 0.6040 | 0.4723 | $0.7045^{\downarrow}$ |
| | *WordEmbedSynSwap* | 0.3488 | 0.3636 | 0.3605 | 0.3710 | 0.4490 | 0.3488 | 0.5457 |
| | *WordNetSynSwap* | 0.3520 | 0.3554 | 0.3680 | 0.3819 | 0.4749 | 0.3520 | 0.5602 |

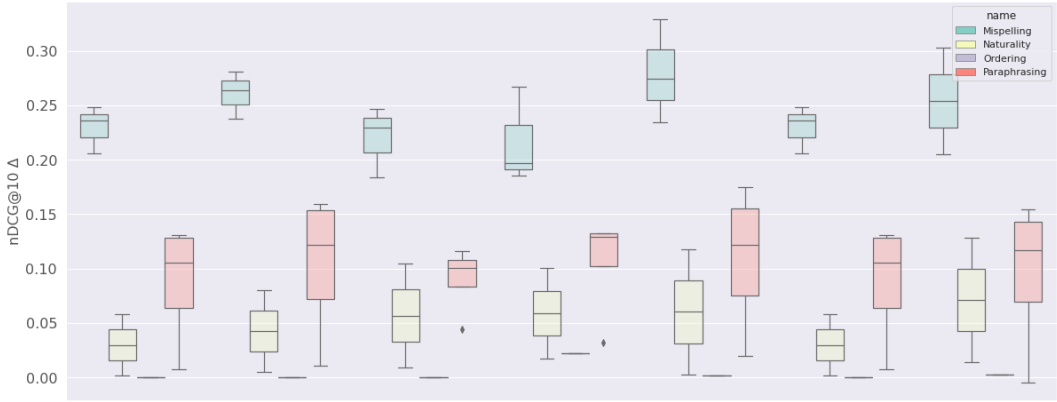
# Résultats

Performances (nDCG@10) des différentes méthodes pour ANTIQUE face à différentes variations de requêtes, les flèches démontrent un réel changement statistique selon un t-test.

| Catégorie | Variation | BM25 | RM3 | KNRM | CKNRM | EPIC | BERT | T5 |
|---|---|---|---|---|---|---|---|---|
| - | requête originale | 0.2325 | 0.2186 | 0.2175 | 0.2024 | 0.2663 | **0.3549** | **0.3350** |
| ***Mispelling*** | *NeighbCharSwap* | 0.1617 | 0.1491 | 0.1271 | 0.1442 | 0.1827 | 0.2406 | 0.2524 |
| | *RandomCharSub* | 0.1663 | 0.1581 | 0.1268 | 0.1475 | 0.1895 | 0.2355 | 0.2476 |
| | *QWERTYCharSub* | 0.1648 | 0.1528 | 0.1390 | 0.1562 | 0.1928 | 0.2465 | 0.2691 |
| ***Naturality*** | *RemoveStopWords* | 0.2312 | 0.2187$^{\downarrow}$ | 0.2123 | 0.2139 | 0.2688$^{\downarrow}$ | 0.3043 | 0.3210 |
| | *T5DescToTitle* | 0.1717 | 0.1668 | 0.1641 | 0.1680 | 0.1997 | 0.2164 | 0.2431 |
| ***Ordering*** | *RandomOrderSwap* | 0.2327 | 0.2186 | 0.1764 | 0.1959 | 0.2660$^{\downarrow}$ | 0.3284 | 0.3257 |
| ***Paraphrasing*** | *BackTranslation* | 0.1630 | 0.1536 | 0.1259 | 0.1385 | 0.2025 | 0.2522 | 0.2614 |
| | *T5QQP* | 0.2258 | 0.2138 | 0.1676 | 0.1917 | 0.2602 | 0.3170 | 0.3226 |
| | *WordEmbedSynSwap* | 0.1780 | 0.2138 | 0.1457 | 0.1675 | 0.2128 | 0.2591 | 0.2827 |
| | *WordNetSynSwap* | 0.1852 | 0.1772 | 0.1522 | 0.1729 | 0.2104 | 0.2715 | 0.2758 |

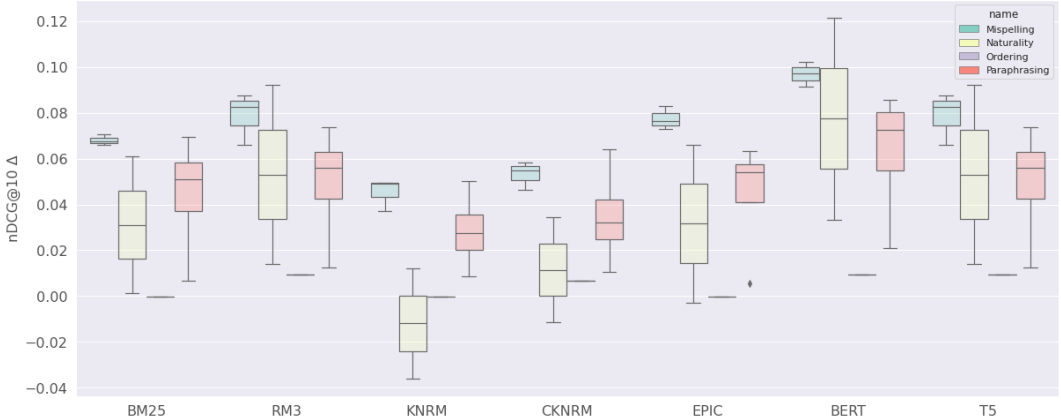
# Résultats

**Dataset TREC** : Distribution du *nDCG*@10Δ ( remplacement de la requête originale par les méthodes de chaque catégorie)



# Résultats

**Dataset Antique** : Distribution du *nDCG*@10Δ ( remplacement de la requête originale par les méthodes de chaque catégorie)



# Analyse des résultats

- Bien que certaines requêtes ont un effets positif dans les 2 datasets, les modèles perdent considérablement leur performances.
- Les **meilleurs** scores sont obtenus par la variation ordering.
- Les **pires** scores sont obtenus par les variations misspelling et naturality.
- la catégorie de variation ordeing a le **moins** d'effet sur les résultats.

# Conclusion

En conclusion, ce travail met en avant le besoin de dataset contenant des variations de requêtes, pour ensuite améliorer les résultats des modèles des modèles.

- G. Penha, A. Câmara, et C. Hauff, « Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators »
- P. Mishra, U. Singh, C. Pandey, P. Mishra, et G. Pandey, « Application of student's t-test, analysis of variance, and covariance », Ann Card Anaesth, vol. 22, n 4, p. 407, 2019, doi : 10.4103/aca.ACA_94_19