Arstist-run spaces : exploration et enrichissement des données textuelles

François-David Collin

Encadrant pédagogique : François-David Collin Expert domaine : Sabrina Issa

Contexte

Les artist-run spaces [1–3] ou « espaces projets » sont des espaces d'art initiés et gérés par les artistes eux-mêmes. Ces formes de structuration du milieu artistique se sont à l'origine développées en marge des structures institutionnelles et commerciales, galeries ou musées par exemple. La singularité du phénomène des artist-run spaces repose sur la variété des réponses que les artistes dans le contexte de renouvellement des politiques et de l'économie du secteur culturel de ces quarante dernières années. La plateforme artist-run-spaces.org a pour objectif de recenser ces espaces et de les intégrer dans une visualisation spatiale et temporelle. En interne, ces espaces sont tabulés avec leurs caractéristiques propres, souvent avec une description textuelle sémantiquement très riche, qui peut être fournie directement par les artistes eux-mêmes ou par une recherche bibliographique. La base de données est alimentée par recensement et contact direct avec les artistes et les structures, comporte actuellement environ 330 entrées. Un appel à contribution en cours, à visée internationale, sous forme de questions et réponses (fréquemment longues) permettra d'augmenter les données disponibles sur une cinquantaine de sites.

Nous souhaitons explorer une nouvelle méthodologie d'exploration et d'enrichissement des données textuelles en utilisant des LLM (Large Language Models, comme BERT, GPT-3, etc.) pré-entraînés et les techniques dites "few-shot learning" pour découvrir des relations sémantiques entre les artist-run spaces et des axes de représentation pour des visualisations interactives. Pour ce faire, un expert pourra valider les annotations générées ou en proposer de nouvelles sur quelques entrées. Le pipeline devra être capable de s'adapter à l'ajout de nouvelles données et de nouvelles annotations. Une attention particulière devra être portée à la visualisation des résultats et à l'interactivité de l'outil, aussi bien pour mettre en place l'entrainement que pour explorer les résultats obtenus (graphiques de type nuages de mots, graphes de relations, etc.). Pour les outils utilisés, nous proposons de tirer profit de la robustesse

et l'adaptabilité d'un LLM comme BERT, dont la taille raisonnable le rend accessible pour ce type de projet [4].

Concrètement, l'on dispose :

- D'une table de données d'un peu plus de 300 entrées, avec métadonnées et texte descriptif plus ou moins long, en langues variées
- Des contributions auprès de certains sites (collecte en cours, une douzaine maintenant, une cinquantaine dans quelques mois), sous forme de questions/réponses (les réponses sont souvent longues)

Missions

- Mettre en place un outil exploratoire/non supervisé avec visualisation des clusters obtenus et graphiques de relations sémantiques (nuages de mots, graphes de relations, etc.), grâce à BERTopic [5].
- Proposer de nouvelles annotations (classes) à l'expert avec le même outil
- Mettre en place un pipeline type *few-shot learning* pour généraliser les annotations à l'ensemble des données, et visualiser les résultats obtenus en intégrant les nouvelles classes dans les visualisations précédentes [6, 7].

Outils/Plateforme/Langages : python, Hugging Face, D3js, etc.

Dépôts notables :

- BERT
- BERTopic
- PET

Sites web :

- Artist-run-spaces.org
- Marathon du web

1. **Vincent**, **F** 2016 L'artiste-curateur. Entre création, diffusion, dispositif et lieux. URL http://www.theses.fr/2016PA01H313/document

2. **Dettere**, **G** et **Nannucci**, **M** 2012 Artist-run spaces. *Nonprofit Collective Organizations in the 1960s and 1970s.*

3. **Rosati**, **L** et **Staniszewski**, **M A** 2012 Alternative histories: New York art spaces, 1960 to 2010. *(No Title).*

4. **Rogers**, **A**, **Kovaleva**, **O**, et **Rumshisky**, **A** 2021 A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8: 842-866.

5. **Grootendorst**, **M** 2022 BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

6. **Schick**, **T** et **Schütze**, **H** 2020 Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference. *Computing Research Repository*, arXiv:2001.07676. URL http://arxiv.org/abs/2001.07676

7. **Schick**, **T** et **Schütze**, **H** 2020 It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *Computing Research Repository*, arXiv:2009.07118. URL http://arxiv.org/abs/2009.07118