Réunion 24/01

Compte-rendu de la Réunion sur le Multi-Aspect Topic Modeling

Introduction

Lors de notre réunion, nous avons discuté des différentes approches de **Topic Modeling multi-aspect** pour analyser les *artist-run spaces*. Nous avons défini une méthodologie d'expérimentation en testant **BERTopic, LDA et SpaCy**, ainsi que les moyens d'interaction avec Sabrina pour affiner les résultats. Ce document détaille les décisions prises, les prochaines étapes et l'organisation des réunions à venir.

1. Objectif Général

L'objectif principal est de tester différentes méthodes de **modélisation de sujets** afin d'identifier les plus pertinentes pour notre projet sur les *artist-run spaces*.

- Expérimenter plusieurs approches : **LDA, BERTopic et SpaCy**.
- Comparer les résultats et affiner l'analyse avec l'aide de Sabrina.
- Tester si la **traduction en anglais** avant modélisation influence les résultats.
- Développer une approche plus interactive, voire semi-supervisée, pour améliorer la pertinence des topics.

2. Étapes et Méthodologie

a) Expérimentation des Modèles

Nous allons tester plusieurs techniques pour évaluer leurs performances et leur capacité à extraire des thématiques pertinentes.

- **Tester BERTopic** sur un jeu de données génériques avant de l'appliquer à notre corpus.

- **Appliquer BERTopic** aux *artist-run spaces* pour obtenir une première liste de topics.
- **Comparer les résultats** entre **LDA, SpaCy et BERTopic** pour observer les différences significatives.
- **Analyser l'impact de la traduction en anglais** sur la qualité des représentations thématiques.

b) Interaction avec Sabrina

Sabrina jouera un rôle clé dans l'évaluation et le filtrage des résultats pour affiner l'analyse.

- **Évaluer la pertinence des topics obtenus** et supprimer ceux qui sont peu exploitables.
- Expérimenter un **Topic Modeling semi-supervisé**, en supprimant certains topics pour voir si le modèle ajuste mieux les thématiques restantes.
- Tester une **approche inspirée du reinforcement learning**, où Sabrina attribue des scores aux résultats pour guider les itérations suivantes.

c) Stockage et Partage des Résultats

Une bonne gestion du suivi des résultats est essentielle pour optimiser notre collaboration.

- **Utilisation de Discord** : Partage rapide d'extraits textuels et de visuels des résultats obtenus.
- **Stockage sur GitHub** : Versionner le code et conserver un historique des analyses.

d) Perspectives d'Amélioration

Nous avons également exploré les possibilités d'amélioration à long terme.

- L'**augmentation des données textuelles** avec des modèles comme **BERT ou ChatGPT** pourrait être envisagée.
- Cependant, la spécificité des textes sur les *artist-run spaces* rend cette approche plus délicate que pour d'autres types de données (images, textes génériques).

Prochaines Actions

Nous avons défini un plan d'action clair pour les jours à venir :

1. **Stocker et partager les résultats** sur Discord et GitHub pour faciliter la collaboration.
2. **Faire valider les premiers résultats par Sabrina** afin d'affiner la méthodologie et d'améliorer le modèle.
3. **Planifier un point de suivi**

FAQ

1. Pourquoi tester plusieurs modèles de Topic Modeling ?

Chaque modèle a ses propres forces et faiblesses. LDA est plus classique, tandis que BERTopic utilise des techniques avancées de clustering et d'embeddings. Comparer plusieurs approches permet d'identifier celle qui s'adapte le mieux à notre corpus.

2. Quel est le rôle de Sabrina dans cette analyse ?

Sabrina aide à **filtrer les résultats** en identifiant les topics pertinents et en supprimant ceux qui sont moins exploitables. Elle pourrait aussi influencer l'algorithme via une approche **semi-supervisée ou inspirée du reinforcement learning**.

3. Pourquoi envisager une traduction des textes avant modélisation ?

Certains modèles NLP sont plus performants en anglais qu'en français. Traduire nos textes avant analyse pourrait **modifier la répartition des topics** et offrir une perspective différente sur les thématiques.

4. Comment seront partagés les résultats ?

Nous utiliserons **Discord pour des échanges rapides** et **GitHub pour le stockage et la versioning du code**, garantissant ainsi une bonne traçabilité des analyses.

5. L'augmentation de données avec BERT ou ChatGPT est-elle pertinente ?

Cela pourrait être intéressant, mais notre corpus étant très spécifique, cette approche n'est pas aussi efficace que pour des textes plus génériques. L'expérimentation est nécessaire pour mesurer son impact réel.