Projet No 3 — Modélisation de sujets avec différentes techniques

Bazié Dureel
Coulibaly Cheick Ahmed
Sawadogo Abdel Saïd Najib

Encadrant : Dr. Rodrique KAFANDO

2 octobre 2025

Objectif & Corpus

- **Objectif** : appliquer LDA, NMF et BERTopic pour extraire et comparer thèmes d'un corpus.
- **Corpus** : 40 articles collectés manuellement sur l'Agence d'Information du Burkina (AIB).
- **Catégories représentées** : politique, économie, société, culture, sport, éducation.

Prétraitement

- Conversion en minuscules, suppression ponctuation & chiffres.
- Stopwords retirés, lemmatisation via **spaCy** (`fr_core_news_sm`).
- Conservation des tokens alphabétiques ≥ 3 caractères.
- Résultat : `articles_clean.csv` / `corpus_clean.csv`.

Méthodes utilisées

- **LDA** (scikit-learn) — CountVectorizer (n-grams 1–2).
- **NMF** — TF-IDF + factorisation NMF.
- **BERTopic** — embeddings `all-MiniLM-L6-v2`, UMAP + HDBSCAN.

Paramètres expérimentaux

- Corpus : 40 docs , seed `random_state=42`.
- **LDA** : `n_components=8`, `ngram_range=(1,2)`, `min_df=2`, `max_df=0.95`.
- **NMF** : `n_components=8`, `init='nndsvda'`, `max_iter=500`.
- **BERTopic** : `nr_topics='auto'`, `min_topic_size=2`, `top_n_words=10`.

Pipeline & reproductibilité

- Étapes : prétraitement → vectorisation → entraînement → assignation docs → évaluation (c_v) → export JSON/visualisation Streamlit.
- Sauvegardes progressives (`results/topics.json`, `results/backups/`).
- Versions consignées (`requirements.txt`).

Évaluation quantitative (cohérence c_v)

- **Scores c_v (agrégés) :**
  - **NMF : 0.65** (meilleure cohérence).
  - **LDA** : 0.41.
  - **BERTopic** : 0.32.

Exemples de topics (extraits)

- **LDA (POLITIQUE SOCIALE)** : social, régional, protection, délégation, national.
- **NMF (ÉDUCATION)** : scolaire, directeur, école, élève, enseignant.
- **BERTopic (COMMUNE / ÉDUCATION)** : commune, scolaire, initiative, élève, investissement.

Observations & limites

- **NMF** $\rightarrow$ topics plus spécifiques et interprétables.
- **LDA** $\rightarrow$ thématiques générales, parfois hétérogènes.
- **BERTopic** $\rightarrow$ capture sémantique mais sous-ajuste sur petit corpus.
- **Limites** : taille du corpus (40 docs), sensibilité aux stopwords et paramètres, BERTopic gourmand en ressources.

Comparaison synthétique

Méthode	Représentation	Cohérence c_v	Force
LDA	Counts	0.41	rapide, probabiliste
NMF	TF-IDF	**0.65**	spécifiques, interpréta
BERTopic	Embeddings	0.32	capture sémantique

Application Streamlit

- **Fonctionnalités** :
  - Choix du modèle (LDA / NMF / BERTopic).
  - Paramètres ajustables (`n_topics`, `min_df`, `min_topic_size`).
  - Visualisations : wordcloud, bar chart, liste de documents par topic.
  - Edition interactive : renommer/fusionner topics, déplacer documents, relancer cohérence.

Démonstration

Conclusions & recommandations

- Pour ce corpus (40 docs) $\rightarrow$ **NMF** recommandé (topics plus nets).
- **LDA** utile pour vue d'ensemble.
- **BERTopic** plus pertinent sur grands corpus.
- **Prochaines étapes** : élargir corpus, enrichir prétraitement (entités nommées, bigrams), automatiser cohérence.

Ressources & Annexes

- **Code, données et instructions** : `https://github.com/Saiken77/topic-modeling-pipeline-fr-app`