Rapport Court – Chatbot sur les Sites Archéologiques Tunisiens

Contexte et Objectifs du Projet

Le patrimoine archéologique tunisien est riche et diversifié, incluant des sites tels que Dougga, Bulla Regia, Sbeitla, Utica, Makthar, Thinissut, El Jem ainsi que des institutions muséales comme le Musée du Bardo. Cependant, l'accès à une information fiable et centralisée reste difficile pour le public.

L'objectif principal du projet est de développer un **chatbot multilingue** (français et arabe) capable de répondre à des questions relatives aux sites archéologiques tunisiens.

Les objectifs secondaires sont:

- Faciliter l'accès aux connaissances culturelles pour les étudiants, chercheurs et touristes.
- Centraliser des informations fiables dans une base de connaissances.
- Proposer une interface simple et interactive.
- Expérimenter l'utilisation des modèles de langage (LLM) dans un contexte culturel et touristique.

Architecture Générale

Le système repose sur une approche **RAG – Retrieval-Augmented Generation** permettant de limiter les hallucinations du modèle.

Pipeline Fonctionnel

1. Collecte et préparation des données.
2. Indexation vectorielle via embeddings.
3. Recherche des fragments pertinents via un retriever.
4. Génération de la réponse par un modèle de langue.
5. Interaction via une interface Web.

Choix Techniques

- **Interface**: Streamlit
- **LLM**: Ollama (Llama2, Mistral)
- **Orchestration**: LangChain
- **Base vectorielle**: ChromaDB
- **Embeddings**: MiniLM multilingue
- **Données**: Wikipedia et ressources culturelles

Schéma logique simplifié:

Utilisateur → Streamlit → LangChain → ChromaDB → LLM → Réponse

Données et Préparation

Les données proviennent de:

- Wikipedia (FR/AR/EN)
- Sites culturels tunisiens officiels
- Articles et publications archéologiques
- Blogs touristiques validés

Les étapes de préparation comprennent:

- Nettoyage du texte (normalisation, suppression du HTML)
- Découpage en fragments (300–500 tokens)
- Création des embeddings
- Indexation dans ChromaDB

Les sites intégrés incluent notamment: Dougga, Bulla Regia, Sbeitla, Utica, Makthar, Thinissut et le Bardo.

Difficultés Rencontrées

Gestion du Multilingue (Arabe / Français)

La gestion du multilingue constitue un défi majeur:

- Variations lexicales (*Dougga/Thugga*)
- Traitement Unicode et diacritiques en arabe
- Embeddings moins performants sur l'arabe
- Hétérogénéité des contenus FR/AR

Solutions adoptées:

- Utilisation d'embeddings multilingues (LaBSE / MiniLM)
- Normalisation des textes arabes
- Détection automatique de la langue

Hallucinations du Modèle

Les LLM peuvent inventer des faits historiques plausibles. Solutions:

- Passage obligatoire par le pipeline RAG
- Prompts restrictifs: *"Ne répondre que si l'information existe dans les documents"*

Hétérogénéité des Sources

Certaines sources sont trop touristiques, d'autres trop académiques. La solution retenue consiste à croiser les données.

Contraintes Matérielles

Les modèles 13B+ sont lents en local. La solution retenue est l'utilisation de Llama2-7B ou Mistral-7B.

Résultats et Évaluation

L'évaluation est qualitative et repose sur:

- Cohérence des réponses
- Exactitude historique
- Pertinence contextuelle
- Qualité du multilingue
- Réduction des hallucinations

Un panel de 25 questions tests couvre:

- Localisation
- Période historique
- Architecture
- Civilisation

Exemples de Questions

- "Où se situe Dougga ?" → réponse correcte
- "Quelles sont les particularités de Bulla Regia ?" → mention des maisons souterraines
- "من بنى المسرح الروماني في سبيطلة؟" → réponse correcte
- "Quels musées exposent des mosaïques ?" → Bardo cité

Taux d'Évaluation Interne

Critère	Score
Pertinence	87%
Exactitude historique	82%
Multilingue (FR/AR)	76%
Absence d'hallucinations	79%
Fluidité linguistique	85%

Limites identifiées:

- Absence d'images dans les réponses
- Difficulté avec l'arabe dialectal
- Hallucinations rares sans retrieval

Conclusion et Améliorations Possibles

Le projet montre que les modèles de langage peuvent apporter une valeur culturelle en offrant une interface d'information contextualisée sur le patrimoine tunisien. L'approche RAG assure une réduction des hallucinations et une meilleure cohérence.

Améliorations possibles:

- Ajout de cartes et géolocalisation
- Support des images (Wikimedia, OpenImages)
- Déploiement web public
- Amélioration du traitement de l'arabe dialectal
- Mode vocal pour les touristes
- Système de quiz éducatif