Analyse du Tourisme au Maroc via Reddit

Projet NLP avec MongoDB et Streamlit

Présenté par : Ayman Naaimi

Encadré par : Dr. Bouchra Frikh, PhD

École Nationale Supérieure d'Arts et Métiers (ENSAM), Meknès

Contents

1 Introduction **2**

2 Objectifs du Projet **2**

3 Méthodologie **2**
3.1 Scraping des données (Reddit) 2
3.1.1 Exemple de requête 2
3.2 Nettoyage des données 3
3.2.1 Statistiques du nettoyage 3
3.3 Enrichissement des données 3
3.3.1 Exemple de classification thématique 4
3.4 Analyse statistique 4
3.5 Visualisation 4
3.5.1 Exemple de visualisation Streamlit 4

4 Résultats **4**

5 Technologies utilisées **5**

6 Conclusion **5**

7 Perspectives **5**

1 Introduction

Ce projet vise à analyser les discussions touristiques sur Reddit concernant le Maroc, en utilisant des techniques de traitement du langage naturel (NLP) et des outils de visualisation modernes. L'objectif principal est d'extraire, nettoyer, enrichir et visualiser les données pour identifier les destinations populaires, les sentiments des visiteurs et les thèmes touristiques clés. Le projet s'appuie sur une combinaison de technologies incluant PRAW pour le scraping, MongoDB pour le stockage, pandas et TextBlob pour l'analyse, et Streamlit pour la visualisation interactive.

2 Objectifs du Projet

- **Extraction des données** : Collecter des posts et commentaires Reddit liés au tourisme au Maroc.
- **Nettoyage des données** : Filtrer les contenus non pertinents et spam pour obtenir un dataset de haute qualité.
- **Enrichissement des données** : Ajouter des métadonnées telles que les thèmes touristiques et les coordonnées géographiques.
- **Analyse de sentiment** : Analyser les sentiments des discussions pour comprendre les perceptions des visiteurs.
- **Visualisation** : Développer un tableau de bord interactif pour explorer les données (villes, sentiments, thèmes).
- **Analyse statistique** : Identifier les tendances à travers des analyses de mots, bigrammes et trigrammes.

3 Méthodologie

3.1 Scraping des données (Reddit)

- **Outil** : Bibliothèque PRAW pour interagir avec l'API Reddit.
- **Requêtes** : Utilisation de 10 modèles de requêtes par ville (par exemple : "visit Marrakech", "things to do in Fès").
- **Villes ciblées** : 49 villes et villages marocains, incluant Marrakech, Fès, Agadir, Chefchaouen, etc.
- **Données collectées** : Posts principaux et commentaires, stockés dans MongoDB avec un identifiant unique pour éviter les doublons.
- **Résultat** : Collecte initiale de 121 866 entrées.

3.1.1 Exemple de requête

```
query_templates = [
    "visit {}",
    "travel to {}",
    "{} tourism",
    "things to do in {}",
    "is {} safe for tourists",
    "{} travel guide",
    "{} backpacking",
```

```
    "must see in {}",
    "vacation in {}",
    "{} hotel reviews"
]
```

3.2 Nettoyage des données

- **Pré-nettoyage** : Suppression des doublons et fusion des colonnes `title` et `text` en une seule colonne `content`.
- **Filtrage anti-spam** :
  - Suppression des textes contenant des mots commerciaux (par exemple : ".com", "buy now").
  - Élimination des textes trop courts (< 30 caractères).
  - Vérification de la présence d'au moins un mot-clé touristique (par exemple : "desert", "medina").
  - Vérification du ratio de caractères alphabétiques (> 70 %).
- **Filtrage thématique** : Conservation des textes avec au moins deux mots-clés touristiques pour garantir la pertinence.
- **Résultats** :
  - Dataset initial : 121 866 entrées.
  - Après filtrage anti-spam : 38 497 entrées.
  - Après filtrage thématique : 18 562 entrées (réduction de 85 %).

3.2.1 Statistiques du nettoyage

Étape	Nombre de lignes
Dataset brut	121 866
Après filtrage anti-spam	38 497
Après filtrage thématique	18 562

Table 1: Statistiques de réduction du dataset

3.3 Enrichissement des données

- **Analyse de sentiment** : Utilisation de TextBlob pour classer les sentiments en Positif, Neutre ou Négatif (basé sur la polarité : > 0.1 pour Positif, < -0.1 pour Négatif).
- **Classification thématique** : Attribution de thèmes (Attractions Naturelles, Sites Culturels, Activités, Hébergement, Nourriture & Boissons, Transport, Sécurité) en fonction de mots-clés prédéfinis.
- **Géolocalisation** : Ajout des coordonnées (latitude, longitude) pour chaque ville à partir d'un fichier CSV.
- **Typologie** : Classification des lieux en "Ville" ou "Village" selon une liste prédéfinie.
- **Colonnes finales** : `city`, `content`, `lieu_type`, `themes`, `sentiment`, `latitude`, `longitude`.

3.3.1 Exemple de classification thématique

```
tourism_terms = {
   'Attractions Naturelles': ['desert', 'sahara', 'mountains', ...],
   'Sites Culturels': ['medina', 'kasbah', 'mosque', ...],
   ...
}
```

3.4 Analyse statistique

- **Analyse des n-grammes** :
  - **Unigrammes** : Mots les plus fréquents (par exemple : "like" : 13 905, "morocco" : 6 147).
  - **Bigrammes** : Expressions comme "day trip" (545 occurrences), "sahara desert" (190).
  - **Trigrammes** : Expressions comme "best time visit" (83), "planning trip morocco" (36).
- **Statistiques des n-grammes** :
  - Unigrammes uniques : 29 545.
  - Bigrammes uniques : 69 508.
  - Trigrammes uniques : 77 814.
- **Analyse thématique** : Fréquence des mots-clés par catégorie (par exemple : "desert" : 1 481, "medina" : 870).

3.5 Visualisation

- **Outil** : Streamlit pour un tableau de bord interactif.
- **Fonctionnalités** :
  - Filtres dynamiques : type de lieu, ville, thème, sentiment.
  - Graphiques : Top 10 villes, répartition des thèmes, répartition des sentiments.
  - Carte Folium : Visualisation géographique des mentions par ville.
  - Aperçu des avis : Affichage des commentaires avec ville, thème et sentiment.

3.5.1 Exemple de visualisation Streamlit

```
st.subheader("Nombre d'avis par ville")
top_cities = filtered_df["city"].value_counts().head(10)
fig1, ax1 = plt.subplots()
sns.barplot(x=top_cities.values, y=top_cities.index, palette="Blues_d",
   ax=ax1)
st.pyplot(fig1)
```

4 Résultats

- **Destinations populaires** : Marrakech, Fès, Agadir, et Chefchaouen dominent les discussions.
- **Thèmes principaux** : Les Attractions Naturelles (par exemple : désert, montagnes) et les Sites Culturels (par exemple : medina, souk) sont les plus mentionnés.

- **Sentiments** : Majorité de sentiments positifs, reflétant une perception favorable du tourisme au Maroc.
- **Géolocalisation** : La carte interactive montre une concentration des discussions autour des grandes villes touristiques.
- **Pertinence des données** : Le nettoyage rigoureux a permis de réduire le dataset de 85 %, garantissant des analyses fiables.

5 Technologies utilisées

- **Scraping** : PRAW (API Reddit).
- **Stockage** : MongoDB.
- **Nettoyage et analyse** : pandas, TextBlob, scikit-learn.
- **Visualisation** : Streamlit, Seaborn, Folium.

6 Conclusion

Ce projet démontre l'efficacité de l'analyse des données sociales pour comprendre les tendances touristiques. En combinant scraping, nettoyage, enrichissement et visualisation, il offre une vue détaillée des perceptions des touristes sur le Maroc. Le tableau de bord Streamlit facilite l'exploration des données, rendant les résultats accessibles à un large public.

7 Perspectives

- **Améliorations possibles** :
  - Utilisation de modèles de machine learning, deep learning ou intelligence artificielle pour améliorer lanalyse des données.
  - Intégration de modèles NLP plus avancés (par exemple : BERT pour une analyse de sentiment plus précise).
  - Extension à d'autres plateformes sociales (par exemple : Twitter via l'API de X).
  - Analyse temporelle pour détecter les tendances saisonnières.