

Analyse des données textuelles d'articles de presse

Présentation Stage Première Année

Adib HABBOU

Haut-Commissariat au Plan du Maroc

17 octobre 2022

Haut-Commissariat au Plan



- **Structure ministérielle** marocaine créée en septembre 2003
- Principal producteur de l'**information statistique** au Maroc
- Locaux dans les quartiers d'*Agdal* et de *Hay Riad* à Rabat
- Admet une **Direction des Systèmes d'Information Statistique**

Direction des Systèmes d'Information Statistique



- **Collecte des données** à travers différentes sources
- **Nettoyage, classification** et **tri** des données
- **Stockage** des données dans les **Bases de Données Statistiques**
- **Analyse des données** pour extraire des **indicateurs**
- Publication de **rapports** (*trimestriel, semestriel et annuel*)

Contexte



LE MATIN.ma



- Récupération et analyse d'articles de **presse en ligne marocaine**
 - ① Morocco World News → Anglais
 - ② Le Matin → Français
 - ③ Hespress → Arabe

Problématique

Besoin de données sur la réalité économique, sociale et culturelle du pays

- Études ministérielles et gouvernementales
- Études de marché pour le secteur privé
- Études académiques et universitaires

Problématique

Besoin de données sur la réalité économique, sociale et culturelle du pays

- Études ministérielles et gouvernementales
- Études de marché pour le secteur privé
- Études académiques et universitaires

Existant

- Observatoire des conditions de vie des ménages
- Centre d'études et de recherches démographiques
- Recensement générale de la population et de l'habitat

Problématique

Besoin de données sur la réalité économique, sociale et culturelle du pays

- Études ministérielles et gouvernementales
- Études de marché pour le secteur privé
- Études académiques et universitaires

Existant

- Observatoire des conditions de vie des ménages
- Centre d'études et de recherches démographiques
- Recensement générale de la population et de l'habitat

Solutions

- Web Scraping
- Machine Learning

Web Scraping



Technique de récupération de données à partir d'un site Web dynamique

Web Scraping



Technique de récupération de données à partir d'un site Web dynamique

- Interaction avec les navigateurs via un driver
- Navigation sur une page web dynamique (formulaires, boutons...)
- Localisation des éléments (id, name, xpath, class name...)

Textual Data Visualization



Représentation visuelle du contenu d'un document textuel

Textual Data Visualization



Représentation visuelle du contenu d'un document textuel

- Répartition des catégories, auteurs, dates avec **Plotly**, **Matplotlib**
- Synthèse des mots clés d'un texte avec **WordCloud**

Topic Modeling



Modèle non supervisé permettant d'extraire les sujets de documents

Topic Modeling



Modèle non supervisé permettant d'extraire les sujets de documents

- Preprocessing des données avec **NLTK**
- Application du modèle avec **Gensim**

Text Classification



Modèle supervisé nécessitant un entraînement sur des données labélisées

Text Classification



Modèle supervisé nécessitant un entraînement sur des données labélisées

- Preprocessing des données avec **NLTK**
- Application de différents modèles avec **Scikit-Learn**
- Évaluation des modèles de Machine Learning avec **Scikit-Learn**

Dashboard



Application Web pour présenter visuellement les résultats obtenus

Dashboard



Application Web pour présenter visuellement les résultats obtenus

- Transformation de script Python en application Web avec **Streamlit**
- Intégration de code HTML généré par **pyLDAvis**
- Intégration de graphiques crée avec **Plotly**

Web Scraping

Méthodologie adoptée

- ➊ Parcours des catégories : <https://www.site.com/category>
- ➋ Parcours des pages : https://www.site.com/category/page_number
- ➌ Extraction des liens de chaque article
- ➍ Parcours des liens d'articles et récupération des données

Résultat Web Scraping

	title	lead	author	date	content
0	Spanish FM: Spain Works on 'Constructive, Firm...	Algeria decided to suspend its friendship trea...	Safaa Kasraoui	June 09, 2022 10:53 a.m.	Rabat - Spain regrets Algeria's decision to su...
1	Afro-Atlantic Treaty: Morocco Says Only Unity,...	Morocco's Foreign Affairs Minister insists tha...	Aya Benazizi	June 08, 2022 5:08 p.m.	Casablanca - With the Moroccan capital of Raba...
2	Cape Verde to Open Consulate General in Morocc...	Togo also announced the forthcoming opening of...	Safaa Kasraoui	June 08, 2022 4:40 p.m.	Rabat - Cape Verde's Foreign Affairs Minister ...
3	Mexican Delegation To Visit Morocco, Keen To C...	Migration management, the fight against climat...	Souad Anouar	June 08, 2022 4:20 p.m.	Rabat - A Mexican parliamentary delegation is ...
4	Spanish PM Renews Support for Morocco's Autono...	A number of marginal opposition political part...	Safaa Kasraoui	June 08, 2022 2:02 p.m.	Rabat - Spain's Prime Minister Pedro Sanchez h...

Figure – Data Frame de Morocco World News

Résultat Web Scraping

	category	content
0	politics	Rabat - A confidential report from NATO has ex...
1	politics	Rabat - Top security officials from Morocco an...
2	politics	Rabat - Indie-rock band Big Thief has announce...
3	politics	Rabat - The European Union has called on Alger...
4	politics	Rabat - Spain regrets Algeria's decision to su...
5	politics	Casablanca - With the Moroccan capital of Raba...
6	politics	Rabat - Cape Verde's Foreign Affairs Minister ...
7	politics	Rabat - A Mexican parliamentary delegation is ...
8	politics	Rabat - Spain's Prime Minister Pedro Sanchez h...
9	politics	Rabat - The Permanent Representative of Morocc...

Figure – Data Frame de Morocco World News

Topic Modeling - LDA

Latent Dirichlet Allocation

- Chaque document noté M est un mélange θ d'un petit nombre de sujets α
- Attribue topic à chaque mot selon une distribution de Dirichlet :

$$\theta_i \sim \text{Dir}(\alpha) \text{ pour } 1 \leq i \leq M \text{ avec } \alpha < 1$$

- Mise à jour du topic lié à chaque mot en fonction de la probabilité :

$$\mathcal{P}(t|d) \times \mathcal{P}(w|t)$$

Topic Modeling - LDA

Latent Dirichlet Allocation

- Chaque document noté M est un mélange θ d'un petit nombre de sujets α
- Attribue topic à chaque mot selon une distribution de Dirichlet :

$$\theta_i \sim \text{Dir}(\alpha) \text{ pour } 1 \leq i \leq M \text{ avec } \alpha < 1$$

- Mise à jour du topic lié à chaque mot en fonction de la probabilité :

$$\mathcal{P}(t|d) \times \mathcal{P}(w|t)$$

Calcul de probabilité

- $\mathcal{P}(t|d)$: la probabilité que le document d soit assigné au topic t
- $\mathcal{P}(w|t)$: la probabilité que le topic t soit assigné au mot w

Topic Modeling - LDA

Preprocessing

- Tokenisation : tout en minuscule, suppression ponctuation
- Lemmatisation : tout au présent et 1ère personne
- Racinisation : réduction à la forme radicale
- Bag of Words : dictionnaires (mots/nombre d'occurrences)

Topic Modeling - LDA

Preprocessing

- Tokenisation : tout en minuscule, suppression ponctuation
- Lemmatisation : tout au présent et 1ère personne
- Racinisation : réduction à la forme radicale
- Bag of Words : dictionnaires (mots/nombre d'occurrences)

Visualisation **pyLDAvis**

- Extraction des informations du topic model
- Réalisation d'une visualisation Web interactive

Résultat Topic Modeling

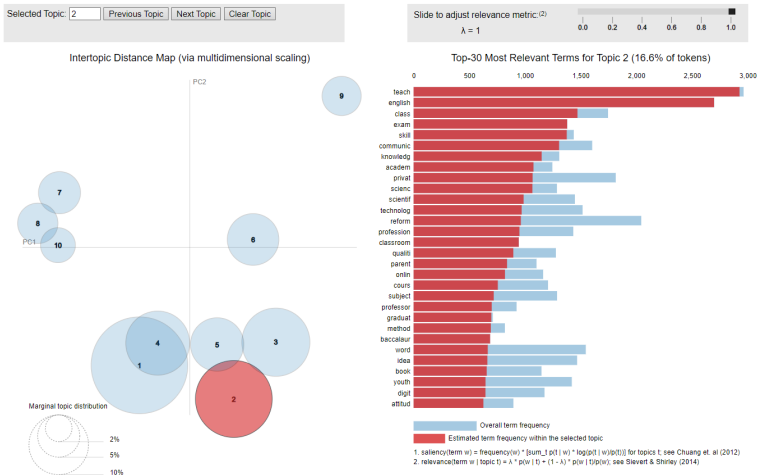


Figure – Capture d'écran résultat pyLDAvis

Text Classification

Modèles de Machine Learning

- K Nearest Neighbour
- Logistic Regression
- Random Forest
- Support Vector Machine
- Stochastic Gradient Descent
- Multi-Layer Perceptron

Text Classification

Modèles de Machine Learning

- K Nearest Neighbour
- Logistic Regression
- Random Forest
- Support Vector Machine
- Stochastic Gradient Descent
- Multi-Layer Perceptron

Évaluation des modèles

- Précision : taux de prédictions positives correctes
- Recall : taux de positifs correctement prédits
- F1-score : capacité d'un modèle à bien prédire les individus positifs
- Matrice Confusion : ligne catégorie réelle, colonne catégorie estimée

Rapport de classification

	precision	recall	f1-score	support
covid-19	0.92	0.89	0.90	277
economy	0.75	0.86	0.80	130
education	0.85	0.92	0.88	279
maghreb-news	0.85	0.90	0.88	294
opinion	0.87	0.91	0.89	280
politics	0.78	0.45	0.57	84
society	0.79	0.31	0.45	48
accuracy			0.85	1392
macro avg	0.83	0.75	0.77	1392
weighted avg	0.85	0.85	0.85	1392

Figure – Rapport de classification pour le SVM

Matrice de confusion

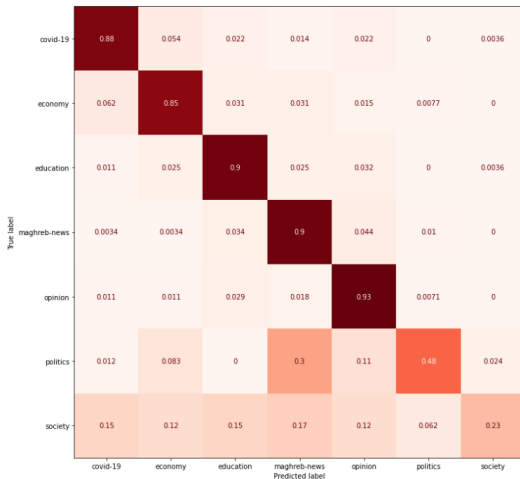


Figure – Matrice de confusion pour le SVM

Résultat Dashboard



Data Analysis of Moroccan Newspapers

Morocco World News Data Set Sample

	title	lead	author	date
0	Spanish PM: Spain Works on 'Constructive, Firm' Response to Algeria's Decision	Algeria decided to suspend its friendship treaty with Spain to further protest the country's endorsement	Safaa Kasraoui	June 09, 2022 10:53 a.
1	Afro-Atlantic Treaty: Morocco Says Only Unity, Solidarity Can Save Africa	Morocco's Foreign Affairs Minister insists that only continental solidarity can save Africa from being a p	Aya Benazid	June 08, 2022 5:08 p.m
2	Cape Verde to Open Consulate General in Morocco's Dakhla	Togo also announced the forthcoming opening of a consulate in Dakhla.	Safaa Kasraoui	June 08, 2022 4:40 p.m
3	Mexican Delegation To Visit Morocco, Keen To Consolidate Relations	Migration management, the fight against climate change, and the Western Sahara dispute are expecte	Soudad Anouar	June 08, 2022 4:20 p.m
4	Spanish PM Renews Support for Morocco's Autonomy Plan Amid Hostile Campaign	A number of marginal opposition political parties in Spain are taking issue with the Sanchez governme	Safaa Kasraoui	June 08, 2022 2:02 p.m
5	Moroccan Ambassador: Polisario's Sultana Khaya is Not a Human Rights Defender	While presenting herself as a pacifist champion of human rights, Sultana Khaya has been recorded an	Safaa Kasraoui	June 08, 2022 11:10 a.
6	Minister: Spain Wants to Preserve Ties with 'Reliable, Fraternal' Morocco	Since endorsing Morocco's Autonomy Plan for Western Sahara, Madrid has repeatedly renewed its reac	Safaa Kasraoui	June 07, 2022 5:02 p.m
7	Saudi Arabia Engages in 'Serious Talks' With Israel For Potential Normalization	Many have suggested normalization of ties between Saudi Arabia and Israel is just a matter of time.	Safaa Kasraoui	June 07, 2022 2:14 p.m

Figure – Capture d'écran du Dashboard

Résultat Dashboard

Morocco World News Categories

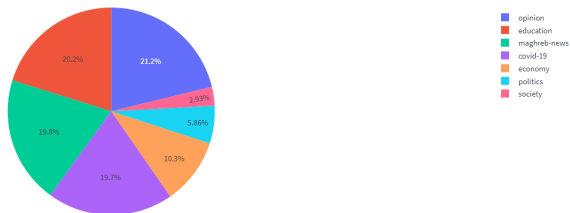


Figure – Capture d'écran du Dashboard

Résultat Dashboard

Morocco World News Authors

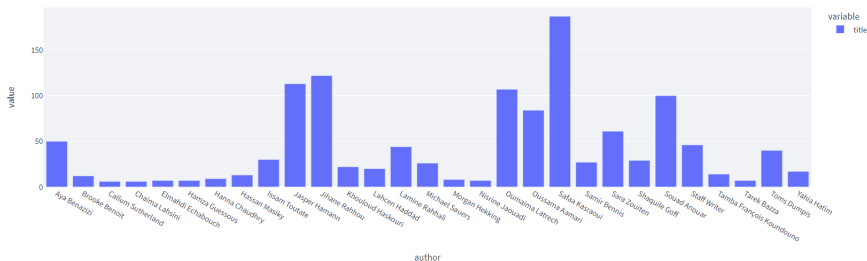


Figure – Capture d'écran du Dashboard

Conclusion

Apport

- Étudier la faisabilité de l'ensemble du processus
- Identifier l'ensemble des technologies nécessaires
- Sélectionner les modèles de *Machine Learning* les plus performants

Conclusion

Apport

- Étudier la faisabilité de l'ensemble du processus
- Identifier l'ensemble des technologies nécessaires
- Sélectionner les modèles de *Machine Learning* les plus performants

Perspective

- Infrastructure d'extraction et de nettoyage des données
- Pipeline d'application du Topic Modeling et de la Text Classification
- Développement d'une application Web de monitoring des résultats