

4DVST

TP : Analyse des discours d'investiture présidentiels américains à l'aide du NLP et de la data visualisation

Léo SOHRABI

Rapport de Visualisation de Données

Objectif

Ce projet vise à explorer un corpus de discours présidentiels américains, en appliquant des techniques de Traitement Automatique du Langage (NLP) ainsi que des visualisations efficaces à l'aide de Python.

L'objectif est de détecter les émotions exprimées, de repérer les thèmes dominants, et d'extraire les mots les plus significatifs dans le temps, tout en adoptant une démarche rigoureuse de traitement de texte et de visualisation conforme aux bonnes pratiques vues en cours.

Jeu de données utilisé :

- Source : Kaggle - Presidential Inaugural Addresses Dataset
 - Fichiers CSV utilisés : **inaug_speeches.csv**
 - Ce fichier contient les discours d'investiture des présidents américains, avec le texte intégral, le nom du président, et l'année du discours.
-

Méthodes NLP appliquées :

Ce projet applique plusieurs méthodes de traitement du langage naturel : la **tokenisation**, pour découper les discours en unités de sens ; la **lemmatisation** et le **stemming**, pour ramener les mots à leur forme de base ou racine ; l'**analyse de sentiment**, réalisée à l'aide de la bibliothèque TextBlob, qui permet de mesurer la polarité de chaque discours ; et enfin, le **TF-IDF**, qui identifie les mots les plus significatifs selon leur fréquence et leur rareté contextuelle.

🌐 Visualisations réalisées :

Plusieurs visualisations ont été utilisées pour appuyer l'analyse : un histogramme pour observer la répartition des sentiments, un nuage de mots pour faire ressortir les termes les plus fréquents, un barplot TF-IDF pour identifier les mots les plus significatifs, ainsi qu'une courbe temporelle et une analyse par président pour étudier l'évolution du ton des discours dans le temps.

1. Histogramme des sentiments

- Type : Histogramme simple
- Lib : Matplotlib
- Affiche la répartition des discours selon leur tonalité (négatif → positif)

2. Nuage de mots

- Type : WordCloud
- Lib : WordCloud
- Met en avant les mots les plus utilisés dans les discours après nettoyage

3. Barplot des mots TF-IDF

- Type : Graphique en barres verticales
- Lib : Matplotlib
- Montre les 20 mots les plus importants selon le score TF-IDF (influence contextuelle)

4. Courbe de sentiment dans le temps (bonus)

- Type : Courbe temporelle
- Affiche l'évolution du score de sentiment moyen des discours au fil des années

5. Analyse par président (bonus)

- Type : Affichage tabulaire
- Donne une moyenne du sentiment par président (si disponible dans les données)

⭐ Bonnes pratiques appliquées :

Le code a été structuré de manière modulaire avec des fonctions claires et réutilisables. Les données ont été nettoyées rigoureusement (suppression des doublons, gestion des stopwords), et les visualisations ont été soignées pour garantir une lecture fluide et esthétique des résultats.

- Code structuré et modulaire (fonctions définies en haut)
- Suppression des doublons dans les données
- Combinaison des stopwords WordCloud + NLTK pour un meilleur nettoyage
- Visualisations lisibles et cohérentes
- Titres, axes, couleurs sobres
- Taille des figures adaptée
- Analyse enrichie avec commentaires et interprétations dans le notebook

👉 Analyse des résultats et interprétations :

L'analyse des discours d'investiture révèle des tendances notables. Les discours prononcés durant des périodes de guerre ou de tensions politiques, comme ceux d'Abraham Lincoln ou de Franklin D. Roosevelt, ont tendance à être plus graves et négatifs. En revanche, les discours plus récents adoptent un ton plus positif et optimiste, mettant l'accent sur l'unité et l'espoir.

Certains mots, comme « nation », « liberté » et « peuple », apparaissent de manière récurrente dans l'ensemble des discours, reflétant des valeurs fondamentales du discours politique américain.

L'analyse TF-IDF met en évidence des termes spécifiques à certains contextes historiques. Par exemple, des mots comme « terrorisme », « reconstruction » et « esclavage » apparaissent de manière significative dans des discours liés à des événements clés de l'histoire, renforçant ainsi la pertinence de l'approche sémantique.

Conclusion personnelle :

Ce projet m'a permis de mettre en œuvre des techniques concrètes de traitement automatique du langage naturel (NLP) et de visualisation de données sur un corpus textuel riche historiquement et sémantiquement.

À travers les différentes étapes du travail, j'ai appris à préparer, transformer et analyser des textes de manière rigoureuse, tout en les représentant graphiquement de façon claire et pertinente. Cette démarche m'a également aidé à mieux interpréter les données au-delà de leur simple forme brute, et à structurer mon code de façon claire et réutilisable.

Enfin, ce projet m'a permis de développer un regard critique sur l'analyse des données textuelles et sur l'impact des choix de visualisation dans la transmission d'informations.

ANNEXE:

Distribution des sentiments des discours

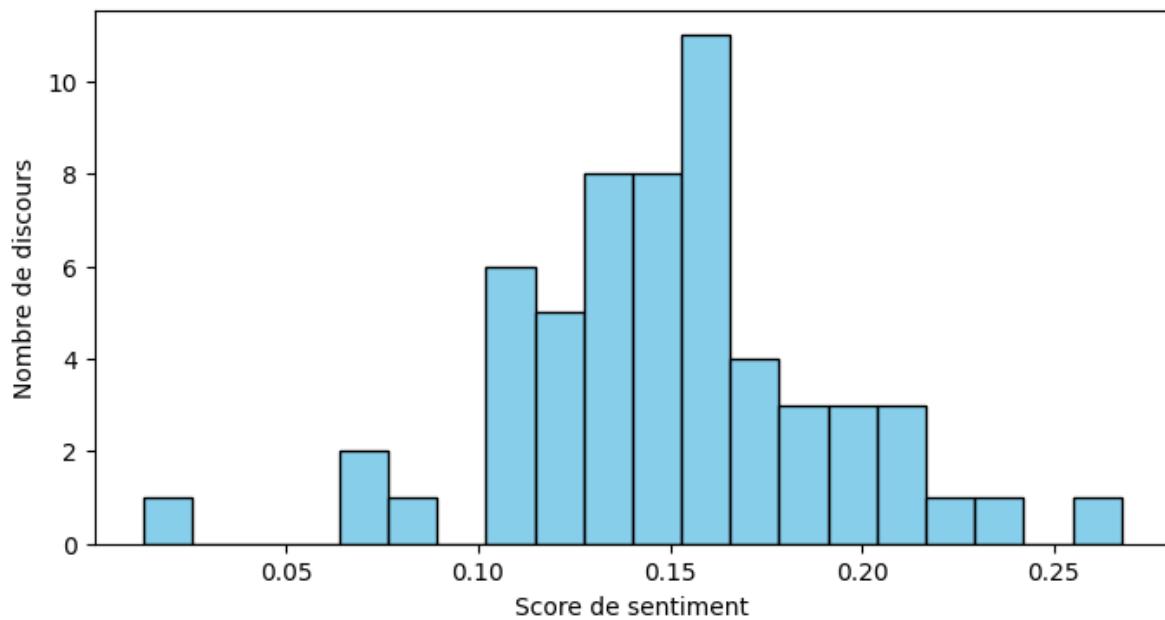


Diagramme 1

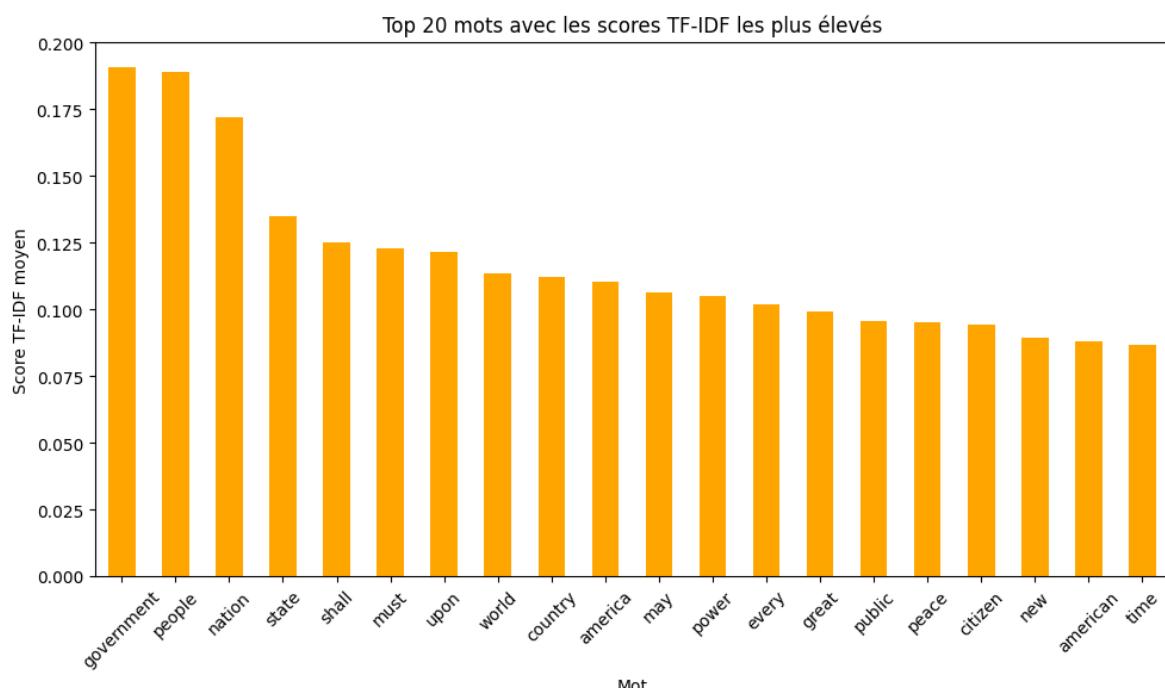


Diagramme 2

TP-presidential-speeches-nlp README.md

Mr10Wick Update README.md a7fea73 · now History

Preview Code Blame 55 lines (34 loc) · 1.96 KB

 Analyse NLP des discours d'investiture des présidents américains

Ce projet propose une analyse des discours d'investiture des présidents des États-Unis à l'aide de techniques de Traitement Automatique du Langage Naturel (NLP) et de visualisation de données.

L'objectif est d'identifier les émotions exprimées, les thèmes dominants, ainsi que les mots les plus significatifs dans le temps, à travers un corpus historique riche.

##◆ Objectif

Appliquer différentes techniques de NLP sur un corpus de discours présidentiels afin d'en extraire des tendances sémantiques, des tonalités, et des indices linguistiques pertinents, tout en illustrant les résultats via des visualisations claires.

Jeu de données

- Source : [Kaggle - Presidential Inaugural Addresses Dataset](#)
- Fichier utilisé : `inaug_speeches.csv`
- Le fichier contient le texte complet de chaque discours, ainsi que le nom du président et l'année.

Méthodes utilisées

- Tokenisation
- Lemmatisation et stemming
- Analyse de sentiment (avec TextBlob)
- Vectorisation TF-IDF
- Suppression des stopwords (personnalisée avec NLTK et WordCloud)

Visualisations produites

- Histogramme de la polarité des sentiments
- Nuage de mots des termes les plus fréquents
- Graphique en barres des scores TF-IDF les plus élevés
- Courbe de l'évolution du sentiment dans le temps
- Moyenne du sentiment par président

Bonnes pratiques appliquées

- Code Python propre, modulaire et structuré par fonctions
- Nettoyage rigoureux des données (suppression doublons, normalisation)
- Visualisations lisibles avec titres, axes et couleurs soignées
- Rapport clair avec interprétation des résultats

 Auteur

Léo – Étudiant en informatique

Fichier README