COMMUNIQUÉ DE PRESSE
Paris, le 30 mars 2022

Méthodologie

La Data Intelligence au service de l'analyse des discours des principaux candidats à l'élection présidentielle

ou comment l'IA peut nous aider à mieux comprendre
les propositions et les valeurs des principaux candidats ?

selon **Acadys – Beyond Data Sciences**

Date : 30 mars 2022

Objectif : décrire la méthodologie utilisée pour disséquer les discours et programmes politiques des candidats à l'élection présidentielle française 2022 afin d'identifier les thématiques communes mais aussi leurs spécificités.

Candidats : les 12 candidats validés définitivement par le conseil constitutionnel.

Identification des sources de données

Seules les sources émanant directement des candidats ont été considérées :

- Sites web officiels des candidats : description de leurs programmes officiels soit directement en ligne soit au travers de documents pdf téléchargeables.
- Vidéos enregistrées de leurs meetings officiels soit mises à disposition sur leurs sites de campagne soit relayées sur YouTube.
- Déclarations officielles à la presse (articles écrits ou déclarations filmées).
- Publications sur Twitter sur les comptes personnels des candidats.

Les interviews pouvant être biaisées par les types de questions posées par les journalistes ou les informations reprises par les médias n'ont pas été considérées dans cette analyse.

Cette identification des sources suppose une veille constante pour être à l'affût des mises à jour des informations relayées par les sites web des candidats, notamment la diffusion des vidéos de leurs meetings et de leurs posts sur les réseaux sociaux.

COMMUNIQUÉ DE PRESSE
Paris, le 30 mars 2022

Constitution des corpus et techniques utilisées

Les sources de données de chaque candidat sont constituées d'un ensemble de fichiers non structurés qu'il faut analyser pour en extraire des informations verbales (pdf, vidéos, images) déversées dans un data lake.

Voici rapidement les étapes qui vont transformer ces données source en textes bruts :

- Sites web de campagne des candidats : scrapping des pages web concernées (extraction de données textuelles).
- Documents pdf : téléchargement et extraction de texte grâce à des algorithmes spécialisés.
- Images avec texte non extractible directement : Pour certains documents mis à disposition sous formes de visuels, des algorithmes de type OCR ont été appliqués et quelquefois des algorithmes de du Deep Learning dans le cas d'images de mauvaise résolution.
- Vidéos YouTube : téléchargement des vidéos, extraction du contenu audio, reconnaissance vocale et transcription de l'audio en texte.
- Tweets : extraction de tweets via une API.

Au terme des étapes précédentes, le corpus de chaque candidat d'un ensemble de fichiers de type texte brut.

Traitement des données des corpus

Dans un premier temps, chaque corpus subit un pre-processing ou « nettoyage » afin de le formater correctement en éliminant les caractères indésirables ainsi que les mots qui n'ont pas d'intérêt dans l'analyse envisagée (les « stopwords » ou mots vides en français : ce sont les mots qui, dans une langue, n'ajoutent pas beaucoup de sens à une phrase. Ils peuvent être ignorés sans danger sans sacrifier le sens de la phrase : articles, mots de liaison…).

Dans un second temps, ces corpus « nettoyés » vont être analysés par un ensemble d'algorithmes de Data Science et d'IA. En particulier, des techniques de Machine Learning permettent d'entraîner certains algorithmes afin d'améliorer leurs performances au cours du temps (meilleure reconnaissance des voix, meilleure extraction de mots-clés).

Ces algorithmes sont principalement des algorithmes qui entrent dans la catégorie NLP (Natural Language Processing) qu'on nomme également TAL en français (Traitement Automatisé du Langage). Le Natural Language Processing (NLP), ou traitement du langage naturel, est une branche de l'intelligence artificielle qui s'attache à comprendre le langage humain tel qu'il est écrit et/ou parlé. Les algorithmes NLP utilisés sont spécifiques à la langue française.

Nous utilisons également des algorithmes de visualisation (nuages de mots, diagrammes en barres, diagrammes en bulles) ainsi que des algorithmes mathématiques pour calculer la proximité entre des ensembles de mots.