Data 4 Good
Off Season
Grenoble & Paris

Perspectiva

Utiliser les outils LLM pour analyser et fournir un retour plus rapide et fidèle aux contributeurs d’une consultation publique

Enjeux sur la qualité du traitement de synthèse

Cas d’usage pour la consultation citoyenne

Objectif :

Analyser et comprendre ce qu’un large groupe de personnes pense en s'appuyant sur leurs propres mots, grâce aux outils intelligents du traitement du langage.

Exemple d’application : question 163 du GRAND DEBAT.

”Que faudrait-il faire pour rendre la fiscalité plus juste et plus efficace ?”

- 52 000 contributions répondant à une question ouverte
- Traité par une société privé Qwam (interview de Christian Langevin par Maxence Fabrion, journaliste chez Decode Media)
    - 2 mois de travail
    - 5 personnes à temps complet
    - plusieurs millions €
    - outils IA (2019, NER, Sentence détection, similarité sémantique)

Exemple de donnée et analyse actuelle

contribution_id int64 0–5.37k 10.1%	content string · lengths 1–2.33k 99%
0	Repartir les richesses. suppression de la taxe d habitation pour tous les français et de la csg reindexé les retraites . Ne plus diviser les français (les patrons ont besoin des plus modestes et vice versa. Les patrons ne réussisses que si les plus modestes sont aussi avec eux) . Les français ne veulent plus de l assistanat mais veulent vivre dignement de leur salaire
1	Les droits soient automatiques, comme nos devoirs de payer les impots
2	réduire drastiquement la fraude fiscale. Imposer les grands groupes (GAFA) qui ne le sont pas suffisamment Renforcer la taxe sur les transactions …
3	diminuer le taux de prelevement pour les retraités percevant mins de 2550 euros
4	TOUT FRANÇAIS DEVRA PAYER L'IMPÔT QU'IL SOIT DOMICILIÉ OU NON EN FRANCE LES FRANÇAIS NE VOULANT PLUS PAYER L'IMPOT EN FRANCE DEVRONT CHANGER DE…
5	Rétablir l'ISF sans délai avec incorporation dans son assiette de tous les éléments de fortune (objets d'art inclus + yachts etc) sans aucune autre…

Réviser la fiscalité des entreprises	**12,0%**
Plus d'équité entre TPE/PME et grandes entreprises	6,2%
Imposer les GAFA	3,0%
Soumettre davantage à l'impôt les très grandes entreprises du CAC 40	2,0%
Revoir la fiscalité des TPE pour améliorer leur compétitivité	1,1%
Rendre l'impôt sur les sociétés plus juste	0,8%
Contrôler davantage les très grandes entreprises (grands groupes, CAC40…)	0,8%
Diminuer, voire supprimer les charges sociales pour les TPE, les indépendants	0,8%
Lutter contre les fraudes	**11,3%**
Lutter plus fermement contre la fraude fiscale	9,6%
Lutter plus fermement contre la fraude à la Sécurité sociale	0,8%
S'appuyer davanatge sur les rapports de la Cour des Comptes	0,8%
Mettre à l'amende les fraudeurs fiscaux	0,8%
Un système fiscal plus équitable et plus transparent (sans précision)	**9,9%**
Plus de justice fiscale et sociale	5,8%
La généralisation du principe de transparence	1,6%
Construire un système fiscal plus équitable	1,2%
Les mesures gouvernementanles ont été injustes pour les retraités (CSG…)	0,9%
Privilégier les impôts directs, plus justes	0,7%
Faire contribuer chaque citoyen à la hauteur de ses moyens	0,4%
Réformer, simplifier la fiscalité (sans précision, grands principes)	**28%**
Réformer l'impôt sur le revenu	15,8%
Réformer l'impôt	6,4%
Généraliser une taxation proportionnelle aux revenus	2,6%
Imposer la flat tax	1,9%
Encourager davantage la solidarité fiscale	1,7%
Revoir le calcul de la taxe foncière	1,4%
Réduire les crédits d'impôts	1,3%
Intégrer les minima sociaux (APL, prime d'activité, aides sociales …) au calcul de l'impôt sur le revenu	1,1%
Élargir le principe du prélèvement à la source	1,0%
Réformer le système des cotisations sociales	0,8%
Modifier le calcul du quotient familial	0,8%
Réformer la nature des revenus imposables	0,7%
Supprimer ou fusionner la CRDS avec la CSG	0,7%
Simplifier le système fiscal français	0,6%
Revoir le calcul de l'impôt	0,6%
Fiscaliser les individus sur la base de la nationalité française	0,3%
Elargir l'assiette fiscale, l'impôt sur le revenu pour tous	**34,5%**
Faire en sorte que tout le monde paie l'impôt	32,6%
La généralisation de l'impôt sur le revenu afin de financer les services publics	1,9%
Les entreprises présentes en France et les citoyens français résidant à l'étranger doivent payer leurs impôts en France	1,6%
Soumettre les faibles revenus à un impôt symbolique	1,0%

Problématiques :

- Qualité de l’extraction : quelle métrique ?
- Gérer les aspects positifs / négatifs syntaxiques ou sémantiques dans la formation des clusters d’idées

Métrique sur la qualité d’extraction

Métrique sur la qualité d’extraction :

Problématique :

“ En tant que responsable de l’extraction, je veux avoir une métrique qui m’assure de la qualité de l’extraction. Cette métrique peut me permettre de détecter qu’une ou un ensemble d’extraction n’est pas représentative, et donc peut mener à une interprétation erronée.”

Enjeu :

- Elaborer une métrique
- Valider cette métrique sur un jeu de donnée

Aspect négatif / positif dans les clusters :

Problématique :

“ En tant que modélisateur, lors de la formation des clusters d’idées à partir des extractions des idées principales des contributeurs, je n’arrive pas à distinguer efficacement les idées positives des idées négatives d’une même concept.

Exemple:

- *je souhaite plus d’impôts.*
- *Je ne veux plus d’impôt. “*

Enjeu :

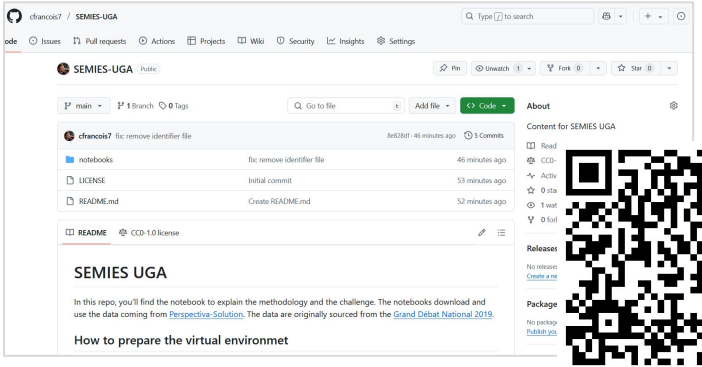
- Définir la négativité pertinente à gérer dans le traitement : est-ce syntaxique ou sémantique ?
- Pertinence de ce traitement dans la formation des idées principales et l’interprétabilité de la synthèse.

Jeu de données

Données pour réaliser le défi

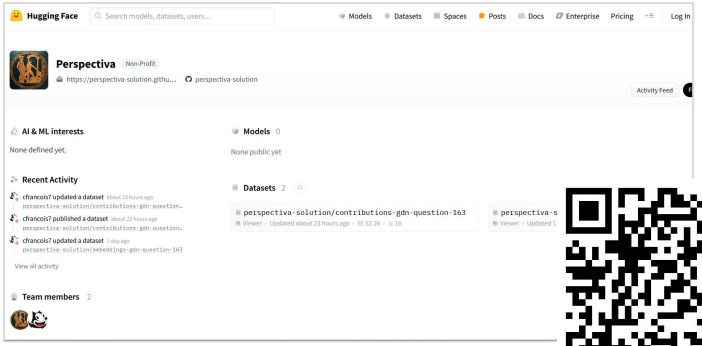
Notebook et présentation succincte de la méthodologie pour générer le dataset.

Accessible sur Github : https://github.com/cfrancois7/SEMIES-UGA



Datasets disponibles sur Hub Huggingface de la question n°163 du Grand Débat National 2019.

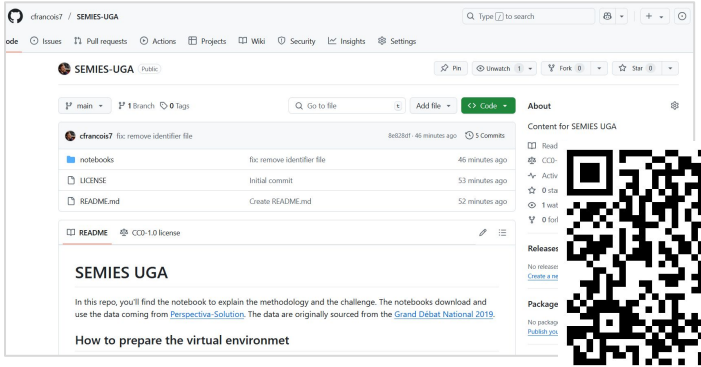
Accessible : https://huggingface.co/perspectiva-solution



Roadmap

Intégration des travaux réalisés lors du SEMIES UGA par les doctorants.

Accessible sur Github (voir Pull Request):
https://github.com/cfrancois7/SEMIES-UGA



Travaux à venir :

- **Création d'un script pipeline de traitement pour orchestrer les traitements.**
- **Développement de la compatibilité avec les outils Talk-to-the-City (T3C)**
- **Intégration du script via API**
- **Réflexion sur un scénario utilisateur et design front.**

Preuve de concept développée en 2024

Usage for public consultation.

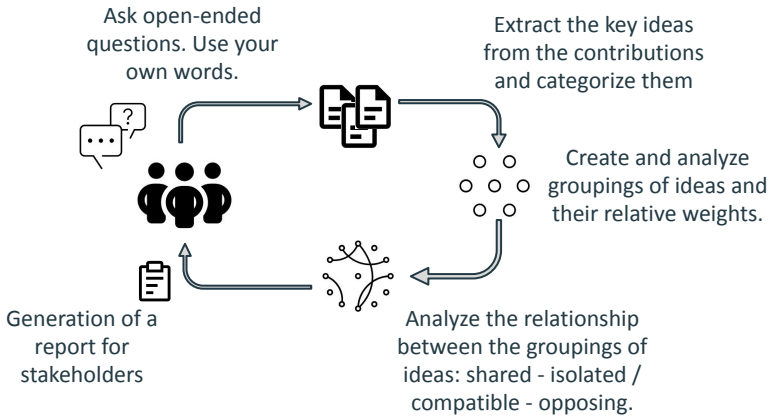
Objectives:

Analyze and understand what a large group of people thinks by relying on their own words, using intelligent language processing tools.

Principles :

- Open-ended questions to allow contributors to express themselves more freely in their own words.
- Transparency and reproducibility in the processing of proposals.
- No predefined categories, reducing errors in the analysis grid. Groupings are made based on the entirety of contributions.
- Automatic generation of simple graphs for communication.
- Empowering the collective to better understand itself.
- Maintaining dialogue with contributors through short processing cycles.
- Adaptable assessments according to needs or constraints.

Process:



Key figures :

- 24 hours for a systematic analysis of thousands of contributions.
- Technical costs (excluding human costs) comparable to those of closed-ended questions: 10 euros for processing 1,000 contributions of less than 500 characters.

Cas d'usage pour la consultation citoyenne

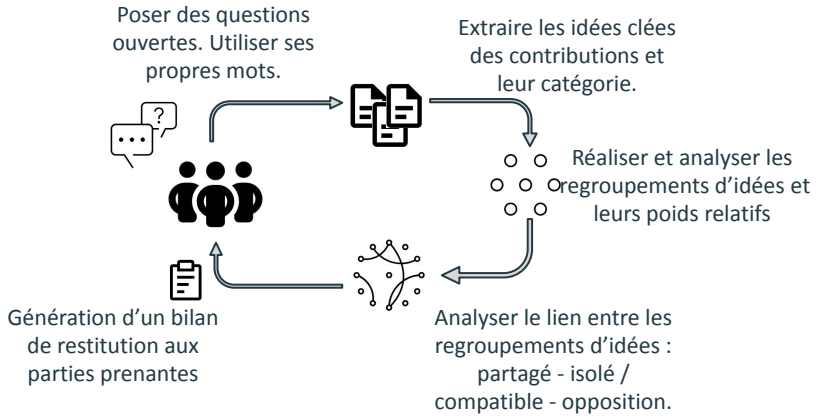
Objectif:

Analyser et comprendre ce qu'un large groupe de personnes pense en s'appuyant sur leurs propres mots, grâce aux outils intelligents du traitement du langage.

Principes :

- Des questions ouvertes pour que les contributeurs puissent s'exprimer plus librement avec leurs mots.
- Transparence et reproductibilité sur le traitement des propositions.
- Pas de catégorie à priori, donc réduction d'erreur de la grille d'analyse. Les regroupements se font sur la base de l'entièreté des contributions.
- Génération automatique de graphiques simples pour communiquer.
- Donner au collectif les moyens de se connaître
- Maintenir le dialogue avec les contributeurs par des cycles de traitement courts.
- Bilan adaptable selon besoins ou contraintes.

Etapes du traitement :



Type de données :

- Des milliers de contributions libres sur un sujet.
- Pour chaque contribution, les idées principales de cette contribution.
- Pour chaque idée, un vecteur "embeddings" est calculé.
- Les vecteurs sont regroupés par "cluster" à l'aide d'algorithme de regroupement (ex: HDBSCAN).