



HACKATAL 2019

Toulouse - 1 et 2 juillet 2019 avec PFIA / TALN

<https://hackatal.github.io/2019/>



GRAND DÉBAT & LEGAL TECHS



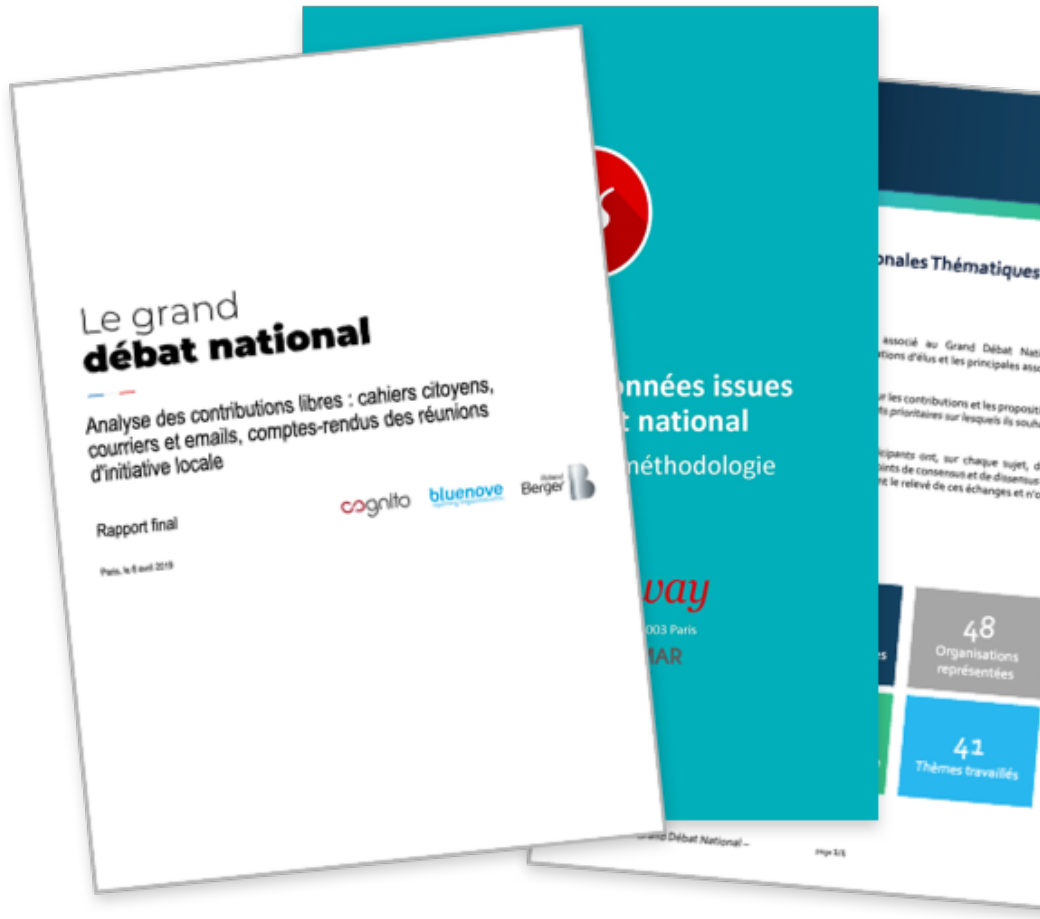
DÉFI GRAND DÉBAT

CONTEXTE

- Le Grand Débat National a eu lieu du 21 Janvier au 15 Mars permettant aux français de contribuer en ligne ou non
- Les contributions en ligne pouvaient se faire soit en répondant à des questionnaires, soit via des contributions libres
- 4 Thèmes
 - La fiscalité et les dépenses publiques
 - La Transition écologique
 - La démocratie et la citoyenneté
 - L'organisation de l'état et des pouvoirs publics
- 1,9 millions de contributions (70% de réponses aux questionnaires, 30% de propositions)
- Beaucoup d'initiatives périphériques (le Vrai Débat, La Grande Annotation, ...)

OBJECTIFS

.....



- Fournir une analyse automatique des propositions libres (334.000 enregistrements) du Grand Débat
- Deux angles d'analyse (complémentaires), qui correspondent à deux tâches
 - Cartographie des thèmes abordés
 - Extraction des arguments sur un thème donné

TÂCHE #1: ANALYSE THÉMATIQUE

- Effectuer une analyse thématique de l'ensemble du corpus de contributions
 - L'analyse n'est pas contrainte aux « thèmes » de la consultation
 - Ni à la structure des contributions (il y a plusieurs questions par contribution)
- Restitution : libre, pourvu qu'elle soit compréhensible
 - De la simple liste avec effectifs... à l'infographie 3D interactive
 - Une génération de descriptif pour chacun des clusters serait bienvenue
- Evaluation : fournir une métrique de cohérence du clustering

TÂCHE #1: LES DONNÉES UTILISABLES

- Le site « La Grande Annotation » a mis en ligne un site collaboratif d'annotations thématiques
 - Mais toutes les questions ne sont pas annotées
 - ... et les thèmes utilisés pour le tagging sont fournis par les organisateurs
 - Les données sont disponibles en ligne (<https://grandeannotation.fr/data>)
- Le site « Le Vrai Débat » a collecté des contributions (non-limitées par les thématiques du Grand Débat) et fourni des analyses sur ces données (<https://www.le-vrai-debat.fr/syntheses/>)

TÂCHE #2: EXTRACTION D'ARGUMENTS

- Ciblée sur une thématique précise: le RIC
 - Mais pas limitatif
- Un jeu de données de référence a été extrait sur le sujet du RIC (10 lots de 100 contributions)
- Vous êtes vivement engagés à annoter manuellement ces contributions
 - Pour entraîner/évaluer votre système
 - Pour gagner des points (un leaderboard est mis en place)
- Restitution
 - Les annotations manuelles sur l'échantillon
 - Les annotations automatiques sur l'échantillon et sur votre thème relatif au RIC (issu de la tâche #1)

A PROPOS DE L'ANNOTATION MANUELLE

- Préconisation d'outil : BRAT
 - Voir HOWTO et collection à annoter ici [link]
- Si vous préférez un autre outil, vérifiez que vos annotations sont bien compatibles
 - Avec le format BRAT
 - Avec la représentation des documents choisie (cf. séparateurs de documents - voir le HOWTO)
- Points bonus si deux annotateurs annotent le même lot
 - Super-bonus si vous fournissez une mesure de l'Accord Inter Annotateurs

SCHÉMA D'ANNOTATION

- On annote l'argument en `Premisse` et `Conclusion` avec éventuellement un `Connecteur` (optionnel dans le marquage, pas évalué)

« `Conclusion` [Notre gouvernement agit trop vite sans aucune réflexion.] `Conclusion` `Prémisse` [On réforme pour dire on réforme mais on ne prend pas compte les conséquences réelles sur l'emploi ou sur la population] `Prémisse` .»

« `Prémisse` [Il est anormal que l'école de la République soit incapable de répondre à sa mission pour un budget toujours plus élevé.] `Prémisse` `Connecteur` [C'est pourquoi.] `Connecteur` `Conclusion` [il faut instaurer un système scolaire unique] `Conclusion` »

- Le schéma permet de relier les prémisses aux conclusions en cas d'ambiguïté
- Voir le guide d'annotation [link] pour plus de détails

DES QUESTIONS ?

