« GRAND DÉBAT NATIONAL » ANALYSIS

IBM ADVANCED DATA SCIENCE SPECIALIZATION – CAPSTONE PROJECT

COMPARING DEEP LEARNING AND STANDARD MACHINE LEARNING ALGORITHM

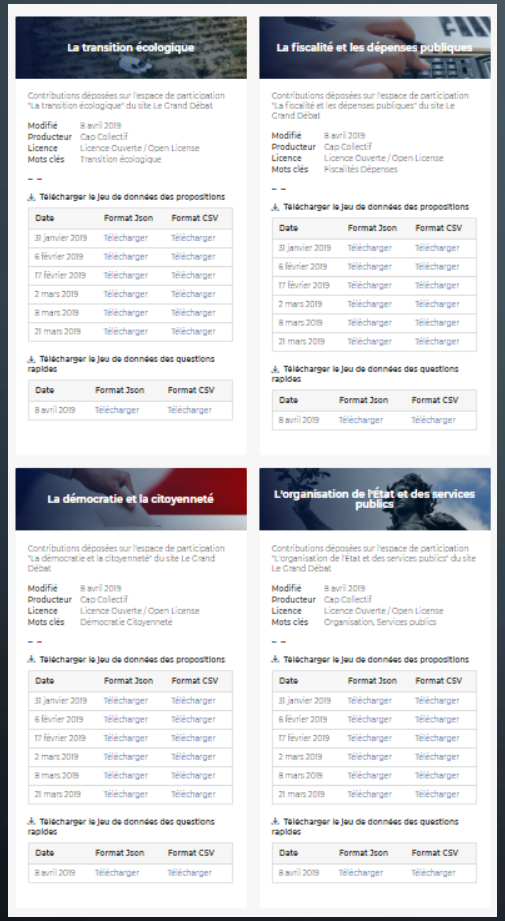
Olivier LUBET – April 2019

CONTEXT

Since October 2018, a strike called "mouvement des Gilets jaunes" growth in France. This movement starts with some unpopular decisions of the government. Due to the fact that no organization or political movement was at the roots of this movement, revendications were unclear.

Emmanuel Macron, the actual president, decided to launch the “grand débat national”, a national forum inline as well as in real life, and asked questions on four mains thematic. Answers of this debate are available as open data. As these answers, due to their volume, are not humanly readable, we are going to analyze them with data science tools, to propose classifications, and reveal trends as well as outliers.

INPUTS



- Data is published at the following address :
  - https://granddebat.fr/pages/donnees-ouvertes
- The answers of contributors are available on the 4 main topics :
  - Ecological transition
  - Fiscality and public expanses
  - Democracy
  - Public services and organization

ANALYZING ECOLOGICAL TRANSITION

THE LIGHTEST JSON FILE OF 236MB WAS USED FOR THIS ANALYSIS

MODEL DEFINITION

ANSWER

authorId	object
authorType	object
authorZipCode	int64
createdAt	object
publishedAt	object
reference	object

responses

title	object
trashed	bool
trashedStatus	object
updatedAt	object

RESPONSE

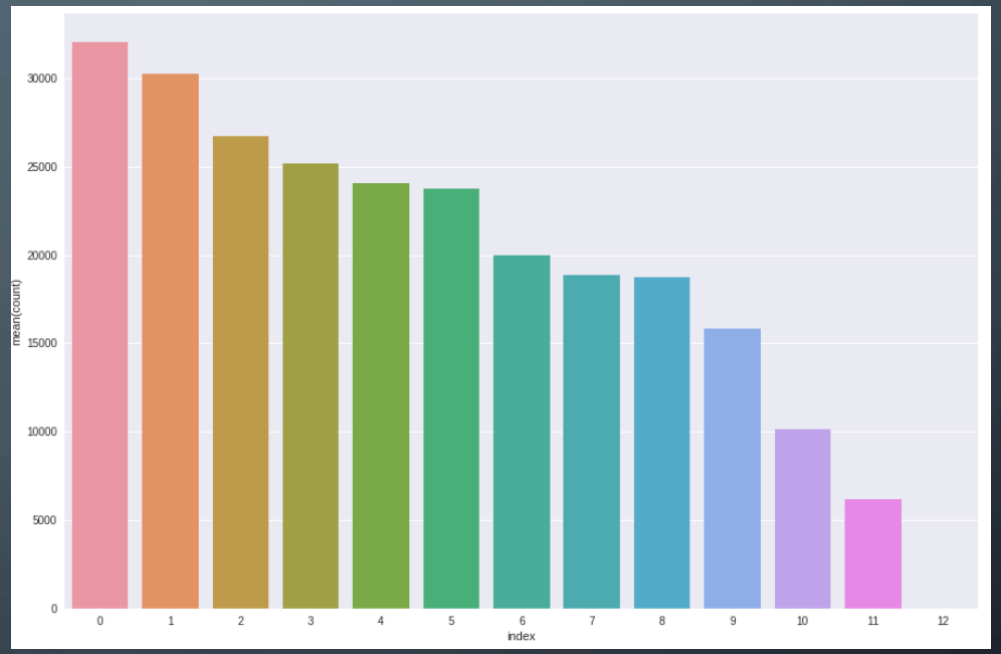
formattedValue object

questionId	object
questionTitle	object
value	object

Number of answers: 575712
Number of filtered answers: 439008
Number of filtered answers (without "Oui" or "Non"): 315721
Number of unique questions: 16

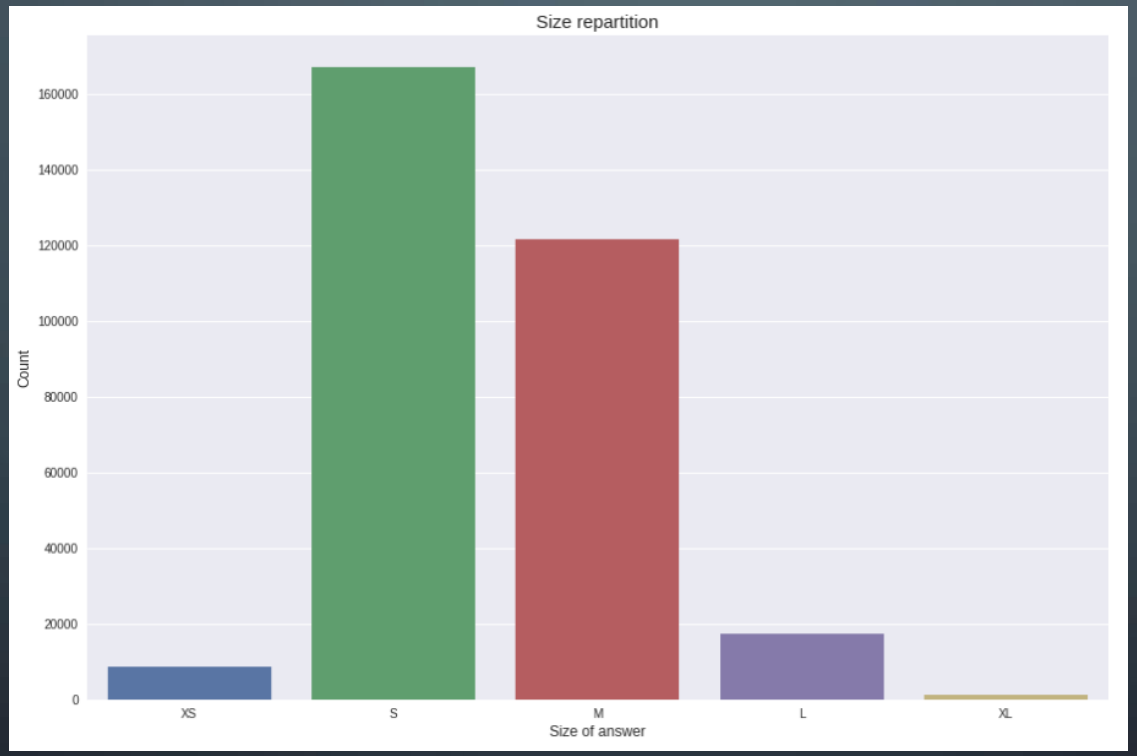
COUNT OF ANSWERS BY QUESTIONS

	question	count	index
0	Que faudrait-il faire selon vous pour apporter...	32047	0
1	Si oui, que faites-vous aujourd'hui pour proté...	30227	1
2	Qu'est-ce qui pourrait vous inciter à changer ...	26712	2
3	Que pourrait faire la France pour faire partag...	25164	3
4	Y a-t-il d'autres points sur la transition éco...	24094	4
5	Quelles seraient pour vous les solutions les p...	23765	5
6	Si oui, de quelle manière votre vie quotidienn...	19954	6
7	Si oui, que faudrait-il faire pour vous convai...	18884	7
8	Et qui doit selon vous se charger de vous prop...	18742	8
9	Si oui, que faudrait-il faire pour vous convai...	15863	9
10	Quel est aujourd'hui pour vous le problème con...	10157	10
11	Si non, quelles sont les solutions de mobilité...	6207	11
12	Avez-vous pour vos déplacements quotidiens la ...	1	12



SIZE OF ANSWERS

Size	Length of answer
XS	< 10 char
S	< 100 char
M	< 500 char
L	< 2000 char
XL	> 2000 char



WORDS COUNT BY ANSWER

mean	26
Std	70
25%	6
50%	13
75%	29
max	20162

75% answers are < 30 words, this threshold will be used

DATA PROCESSING

Data Extraction, Cleaning, Loading

Feature engineering

- BERT feature extraction
- Feature reduction
- Normalization

Clustering

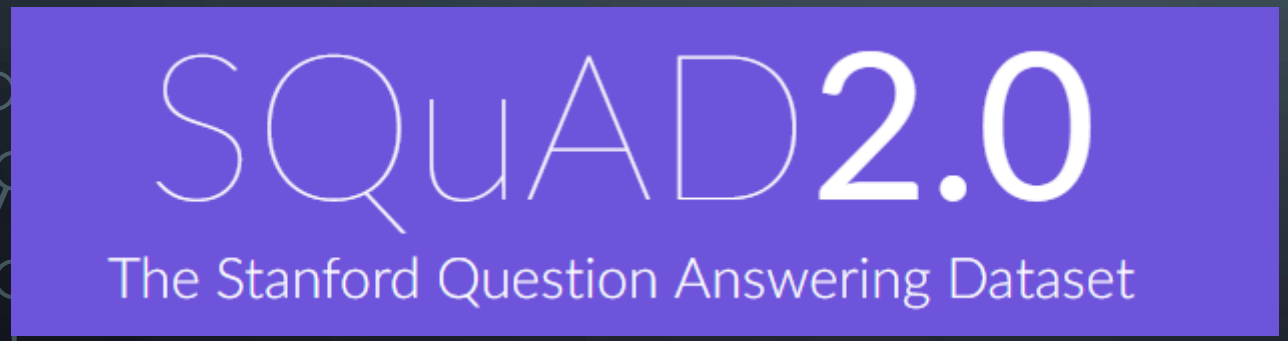
- Machine Learning Technic
- Deep Learning Technic

WHY USING BERT

BERT, or Bidirectional Encoder Representations from Transformers, is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks.

- https://arxiv.org/abs/1810.04805

BERT algorithm achieve better results than human at the SQuAD test.



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621

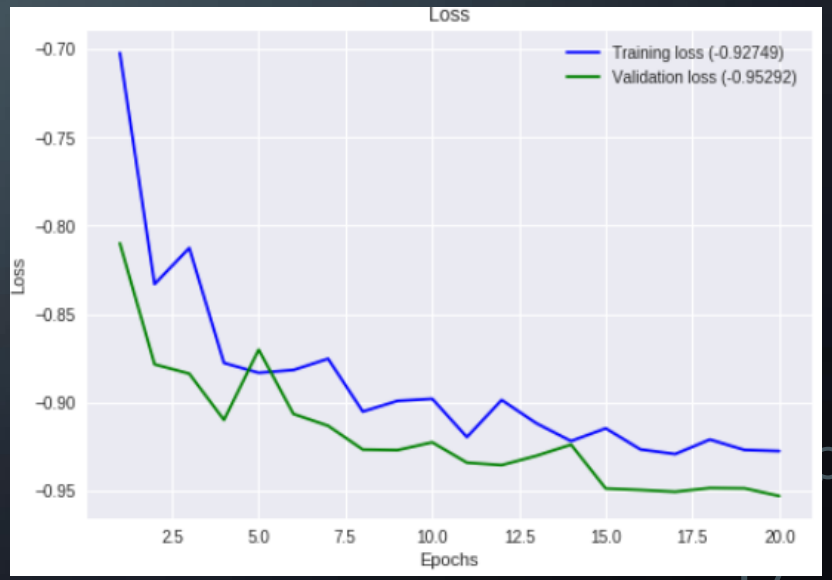
WHY USING FEATURE REDUCTION BEFORE CLUSTERING

OK, it should not be such a good idea but …

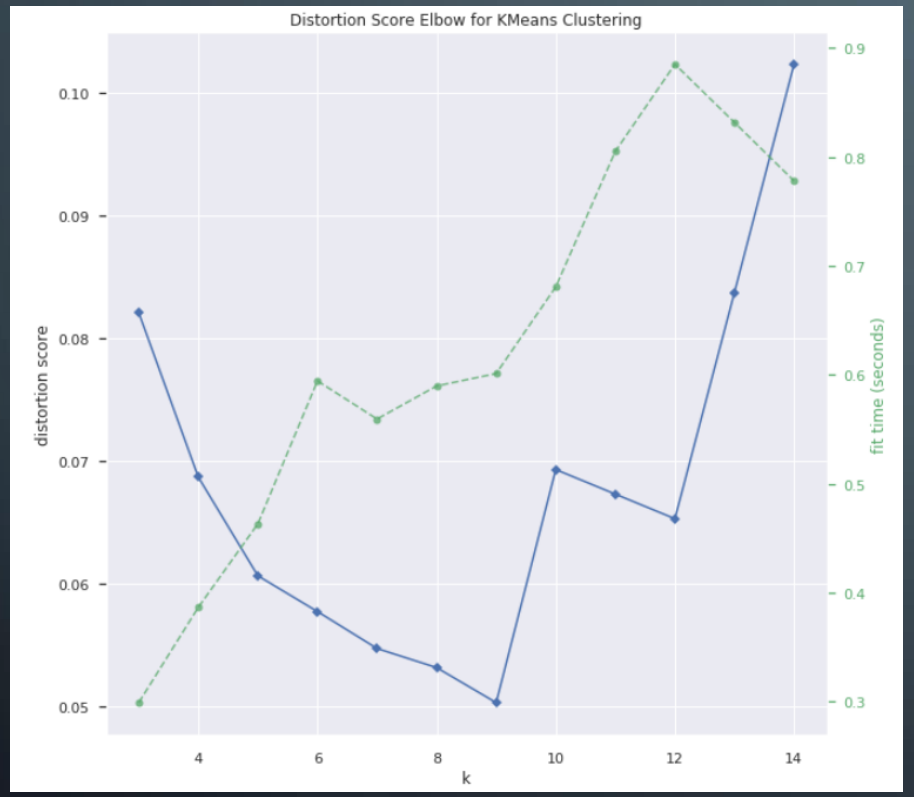
The output of BERT is huge :

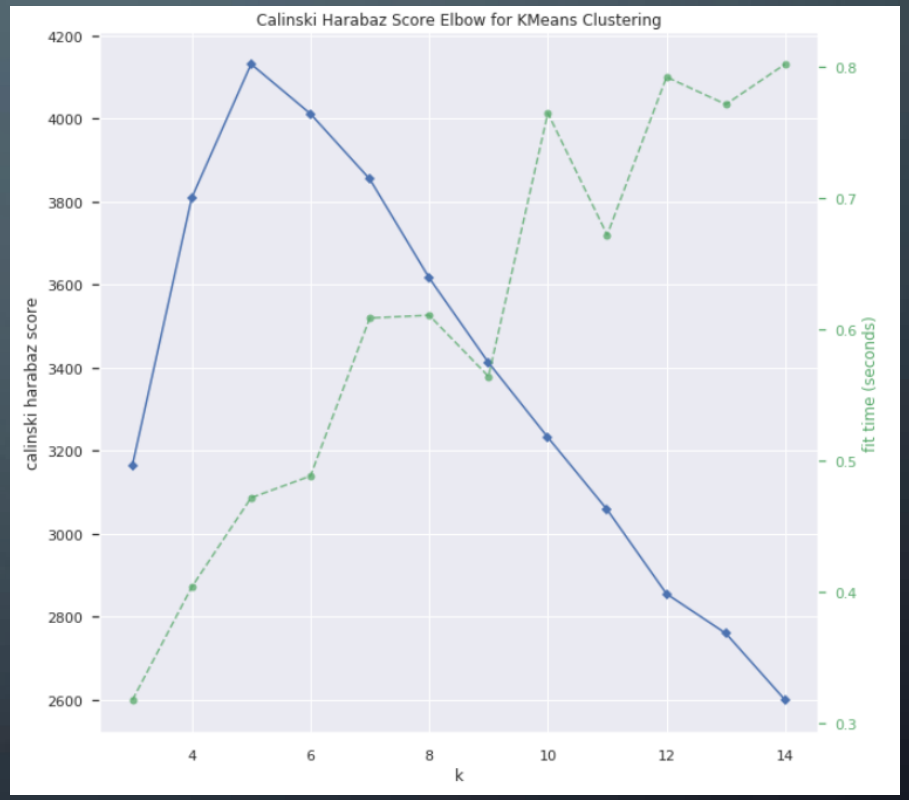
- Sentences are tokenized
- Each token is a 768 dimension vector
- For a sample of 2000 answers, BERT produced 620MB of data

AUTOENCODER with LSTM
100 epochs
> 1h with Tesla K80 GPU



CLUSTERING
DETERMINING THE NUMBER OF CLUSTERS





The elbow method shows us that an optimal K is between 5 ant 9

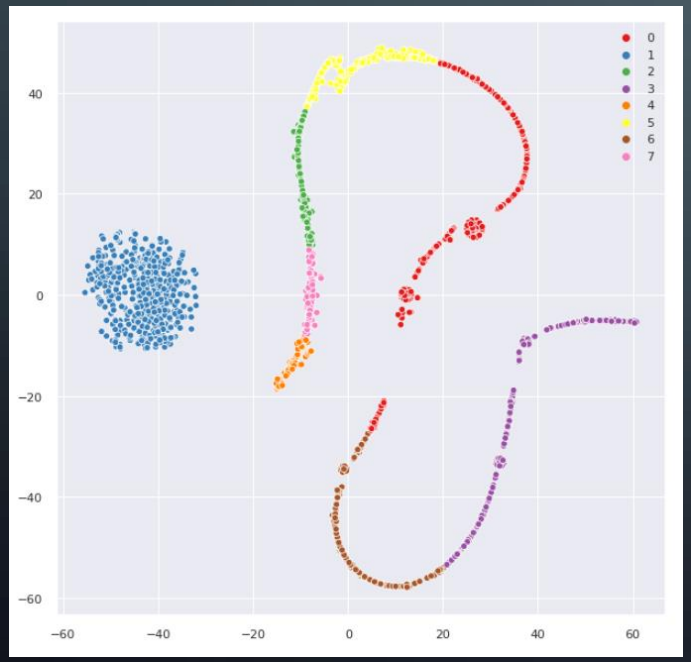
MODELS EVALUATION

FOR A 100 ITERATION COMPUTATION

NOTE : THE VISUALIZATION IN 2D IS JUST ILLUSTRATIVE AS IT DOES NOT REFLECTS THE ACCURACY OF MODELS

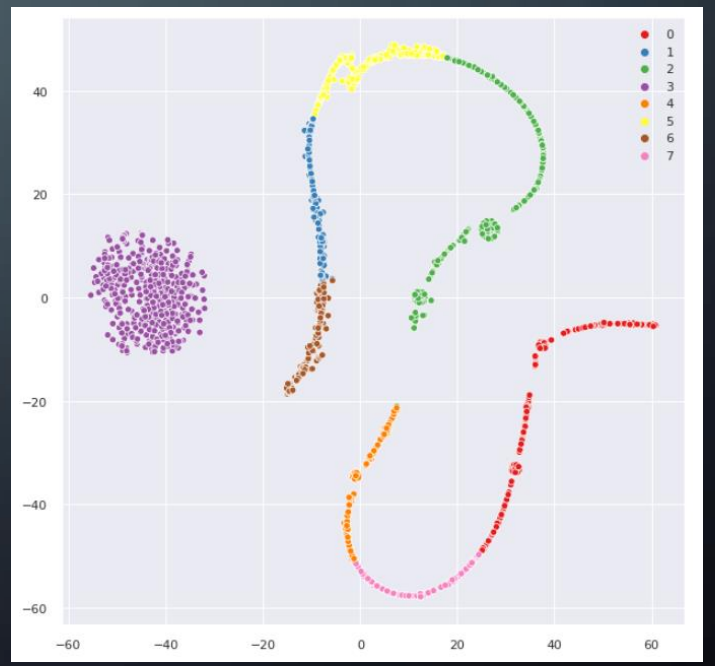
K-Means with SKLearn

Accuracy with euclidean distance : 65.27759 ★

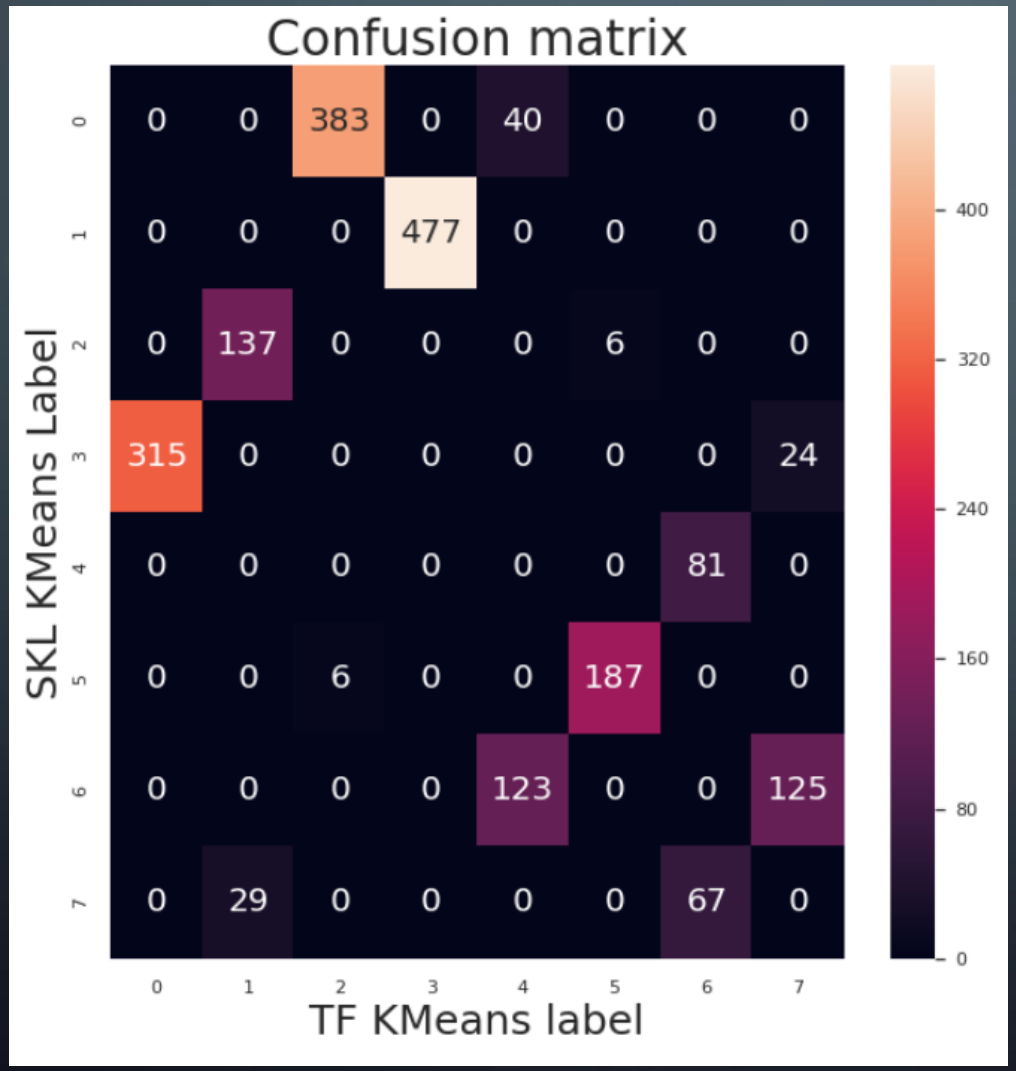


K-Means with TensorFlow

Accuracy with euclidean distance : 65.75744



CLUSTERS SIMILARITIES



The two differents models achieve to give quite similares clusters :

- Cluster 1 of SKL model reflects cluster 3 of TF Model
- Cluster 4 of SLK model is fully integrated in cluster 6 of TF model

The major difference is the SKL cluster 6, split in two clusters with TF model

TF CLUSTERS

Cluster 1 answers nb: 166

Answers near centroid
1901 A 71 ans je ne pense pas avoir besoin de vous pour me déplacer et je ne compte pas sur l'état l'assistanat ça suffit.
1290 Appliquer les mesures contraignantes à tous les usagers de son territoire, qu'ils résident en France ou dans un autre pays européen.
531 Il y a un gaspillage important d'énergie du au TV, ordinateur, tablette, GSM toujours en veille donc à maintenir branchés ou à recharger.
1322 Les catastrophes climatiques sont de plus en plus fréquentes et même si on n'est pas touché, les Assurances augmentent.
815 Rejet de fumées polluantes gigantesques par une usine : Tarnaise des panneaux alors qu'elle pourrait recycler ses fumées
Name: text, dtype: object

Answers with negatives feelings

Outliers
1172 utilisation des transports en commun , marche ,tri sélectif ,récupération de l'eau de pluie ,consommation de fruits et légumes de saison
1020 Que le coût soit équivalent à mon chauffage actuel... et que les entreprises ne se mettent pas dans la poche les bonus accordés par l'état.
1019 Diminution de ma consommation électrique de chauffage et d'éclairage et réduction de l'utilisation de mon véhicule thermique
8 je les utilise déjà mais des transports en commun seraient un gros plus. J'habite un village et il n'y a quasiment pas d'offre en transport en commun.
813 disparition des oiseaux et des insectes (abeilles) et sans doute effets sur la santé même si ce n'est pas sensible aujourd'hui
Name: text, dtype: object

More neutral answers

TF CLUSTERS

Cluster 5 answers nb: 193

Answers near centroid
1952 plus de voitures en ville : que des transports en commun, laisser les voitures à l'entrée des villes
1989 Interdire la revente des productions d'énergies renouvelables subventionnées par l'argent des Français
1124 Economie d'électricité , tri des déchets, jamais de jet de déchets dans la nature.
1122 Taxes sur le carburant. Taxes énergies (fioul, électricité). Montée des eaux
1917 Habitant en zone rurale, l'utilisation de la voiture est indispensable pour tout déplacement.
Name: text, dtype: object

Outliers
938 Tri sélectif, attention aux dates de péremption afin d'éviter le gâchis alimentaire
949 Par une attitude écoresponsable. Tri des déchets, économies d'énergie, consommation des matières premières raisonnée.
900 Faire appliquer la loi de 2005 pour les Personnes à Mobilité Réduite aux nouveaux services de mobilité active
985 diminuer les gaz d'échappement, véhicule, chauffage fioul ou gaz
995 Ne plus subventionner les enfants Taxer a partir du 3ieme enfant Délit a partir du 5ieme Crime a partir du 6ieme
Name: text, dtype: object

Imperatives formed sentences

All the 5th cluster is composed of proposition formed sentences

TF CLUSTERS

Cluster 0, Cluster 2, Cluster 4 and cluster 7 are composed of short answers, where identification of the frontiere between trends and outliers is not clear.

Cluster 0 answers nb: 315

Answers near centroid
403 Maladies / cancers de nos proches
908 déjà fait
439 La pollution de l'air
1384 régulation de chauffage efficace
1225 je fais déjà tout ça
Name: text, dtype: object

Outliers
77 Incitations fiscales
231 Montrer l'exemple.
893 La démographie
880 Aide et credit d impot
1678 Tous ces éléments vont ensemble.
Name: text, dtype: object

Cluster 2 answers nb: 389

Answers near centroid
1086 Compostage, éviction des emballages, achat d'une alimentation bio
769 tri des déchets, interdire utilisation des plastiques et emballages et pailles, etc
1222 Difficile à dire , tous les états n'ayant pas la même priorité .
1099 Soit les collectivités locales soutenues par l'état ou des initiatives citoyennes.
1127 Je n'utilise pas la voiture pour des déplacements quotidiens
Name: text, dtype: object

Outliers
1279 prime pour l'achat d'une batterie électrique pour équiper mon vélo
605 Je n'utilise pas la voiture pour des déplacements quotidiens
1494 opter pour un fournisseur d'énergie renouvelable

Cluster 4 answers nb: 163

Answers near centroid
1324 pas de voiture, ramassage des déchets
830 La biodiversité et la disparition de certaines espèces
1134 Les transports en commun, Le vélo, La trottinette
1085 La pollution de l'air ET la biodiversité
1982 Taxer le transport routier et développer le transport ferroviaire
Name: text, dtype: object

Outliers
414 permetre de créer des bus qui fonctionne netre les region
988 Les transports en commun, Le transport à la demande
1798 Valoriser les bons comportement. Des prix juste pour des choix meilleurs.
1512 SE montrer vertueuse pour changer les esprits par l'exemple

Cluster 7 answers nb: 149

Answers near centroid
1014 améliorer les pistes cyclables
506 developper les pompes a chaleur
1848 Changer immédiatement notre façon de vivre.
1586 la communauté d'agglo avec la mairie.
1001 problème global, tout se recoupe.
Name: text, dtype: object

Outliers
292 Augmentation du prix de l'énergie !
1925 Les transports en commun, Le vélo
1038 Montrer l'exemple en étant exemplaire
56 prendre le leadership en matière d'écotaxes
301 avoir les mêmes lois pour tous serait déja bien
Name: text, dtype: object

TF CLUSTERS

Cluster 3 answers nb: 477

Answers near centroid
1933 Les chasseurs prennent pour excuse qu'ils "sauvent" la nature en régulant. Sans eux, nous serions envahis d'animaux sauvages qui nous feraient toutes sortes de misères. C'est totalement faux ! Pour diminuer fortement la régulation à laquelle s'adonnent les chasseurs, je propose : - d'interdire l'élevage de gibier destiné à la chasse. […]
4 Pesée au niveau communal des déchets recyclés, puis comparaison au niveau national (par tête) pour attribuer des bonus et des malus au niveau des impôts locaux. Établir des contrats de consommation énergétique globalement avantageux pour les faibles consommations mais avec une augmentation rapide des prix après un certain seuil, ou encore globalement avantageux à l'année mais avec des malus dans les pics de consommation ou pour les consommations inutiles (lumière le jour, chauffage l'été...)
5 Baisser les prix des produits Bio et les produits vegan. Mettre en avant le Bio dans les magasins pour inciter les personnes à acheter. Pour limiter les déchets et augmenter le tri sélectif apporter une récompense (argent) à un certain seuil.
1733 Comme exprimé plus haut, la réponse va du changement individuel des mentalités à la prise de conscience mondiale. Les grandes réunions internationales accouchent de souris et ne sont pas assez contraignantes. […]
1696 L'utilisation du covoiturage nécessiterait que je me penche (avec mon hiérarchique) sur une nouvelle organisation de mes semaines de travail car je réalise des déplacements dans le cadre de celui-ci.
Name: text, dtype: object

Outliers
1517 Consommation de produits bio en priorité, tri des déchets, limitation de l'usage des emballages, achat d'un véhicule moins polluant, utilisation de panneaux solaires, travaux d'isolation de la maison, rationalisation des déplacements en voiture, diminution de ma consommation de produits polluants, de viande, réduction des achats compulsifs liés à la mode.

TF CLUSTERS

Cluster 6 answers nb: 148

Answers near centroid
1750 Je vis dans un village, les bus ne fonctionnent qu'aux horaires scolaires. L'usage raisonné et partagé d'une voiture économe est la meilleure solution pour se rendre en ville.
473 Déjà inciter les plus gros pollueurs a changer leurs comportements, ensuite si nous avons les moyens pourquoi pas, mais aujourd'hui ce n'est pas le cas, même avec vos aides.
359 Ralentir l'exploitation de la planète , l'homme préempte tout pour lui seul une espèce disparue c'est pour toujours, le gaz carbonique lui se résorbera
533 Une agriculture responsable rémunérée correctement; limiter le transfert des terres agricoles ou non agricoles au profit de l'urbanisme en enlevant au maire le droit des permis de construire n'importe quoi
310 Lorsque l'on développe de nouvelles énergies d'origine renouvelable il faudrait les substituées par des énergie fossiles afin d'avoir une véritable action sur notre bilan carbone
Name: text, dtype: object

Outliers
870 Transports en commun bien moins chère. Moins de packaging. Arrêter le tout plastique, interdire les perturbateurs endocriniens. Avoir une vrai politique de transition écologique.
1108 Mes enfants hésitent d'avoir des enfants à cause d'un avenir incertain. L'inquiétude. Surtout quand la conscience des problèmes semble être au stade molle chez beaucoup de monde.
94 taxe entreprise polluante ou imposition à se mettre aux normes (mais réduire le nombre de norme inutile). pareil pour les modes de délacement extrêmement polluant (bateau avion)
114 aider les salariés à avoir une aide kilomètrique pour les trajets où il n'y a pas de transport en commun et inciter les entreprises à aider les salariés à faible revenu
1929 tri sélectif, compost, isolation toiture et ouvertures, chauffage et eau sanitaire solaire + panneaux photovoltaïques, déplacement en transports en commun, vélo, marche à pied...
Name: text, dtype: object

GOING FURTHER

- This study was build in some weeks as an afterwork challenge, and results obtained could be criticized
  - In fact, clusters reflects more the answer size than the use of vocabulary
- I was very enthusiast to use BERT algorithm, but the fact is that :
  - It takes me a lot of time to find a good way I could use it
  - The computation time and the output of BERT is huge
  - I was not able to use technics like TF-IDF with the output of BERT as it give for a same token a unique meaning in each sentence
- This experience made me read a lots of scientific papers on NLP. It made me discover technics I should have used for a more complete approach like text summarization