

PROJECT 2

Data Exploration and NLP Modeling - traduction
client reviews

Par Maud Tissot & Aurélien Pouxviel

Sommaire :

I - Explication & analyse des données

II - Traitement des avis

- Tokenization, corréction
- Fréquence des mots

III - Résumé & traduction avec T5-base

IV - Analyse des sentiments

V - Modélisation du sujet avec BERTopic

VI - Embedding de mots et d'avis

- Word2Vec & GloVe
- Visualisations TSNE & Tensorboard
- Question Answering
- Recherche sémantique

VII - Supervised learning

- TFIDF & Fasttext
- Random Forest with TFIDF and Fasttext
- Bert

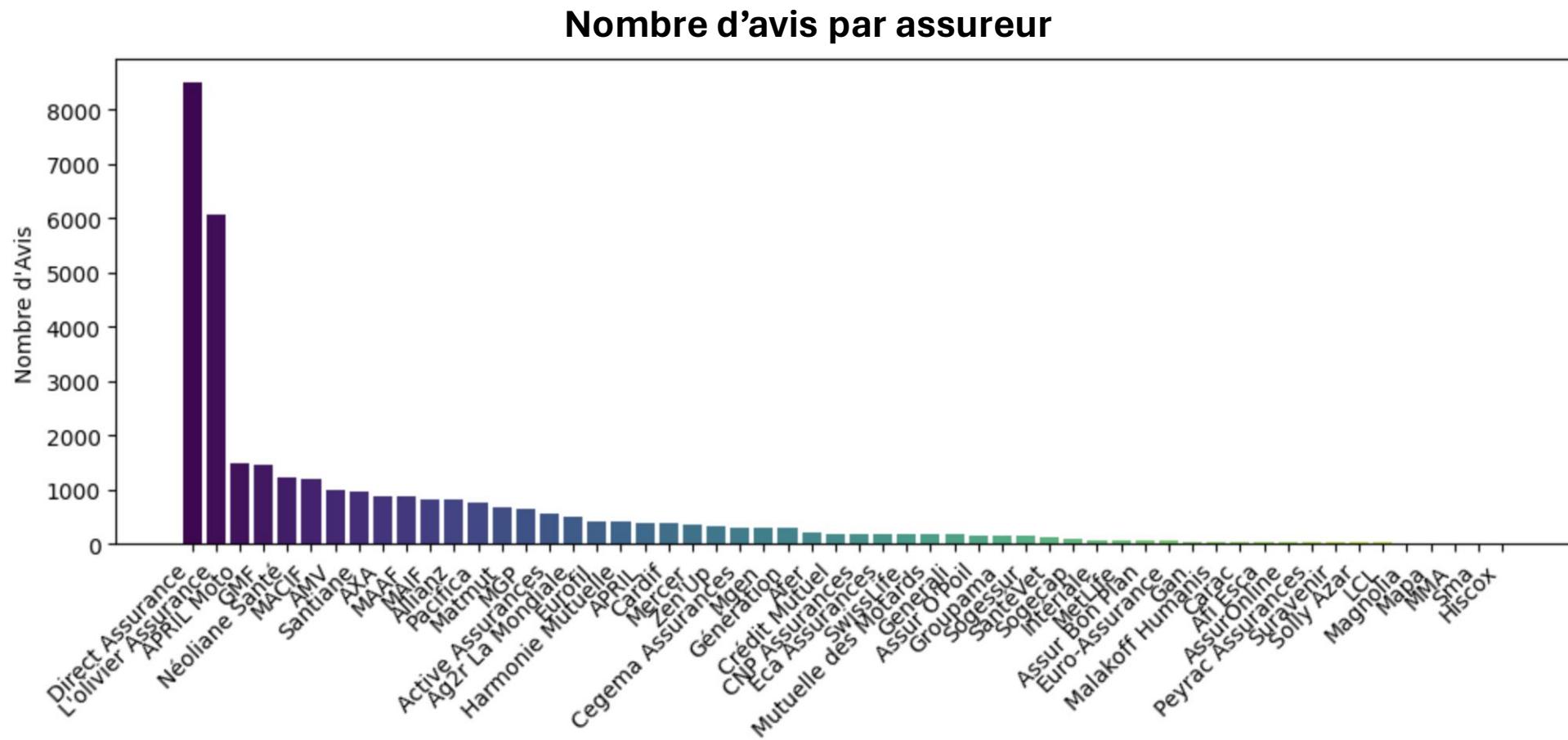
VII - Lien de l'app : [ici](#)

Github https://github.com/Aurelien7877/NLP_Project/tree/master/Project2

Explication & analyse des données

Jeu de donné choisi : Traduction avis clients.zip disponible dans l'énoncé

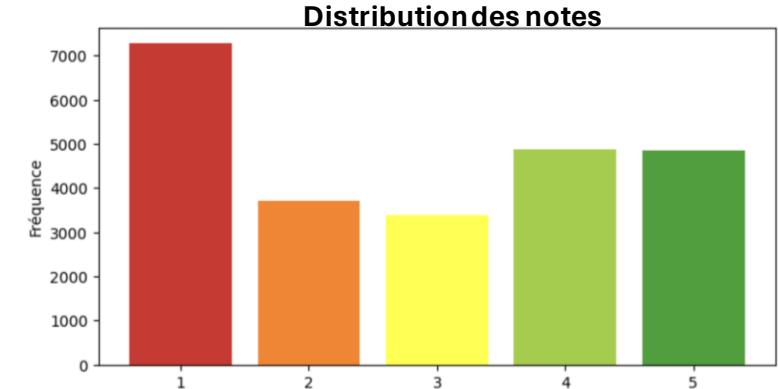
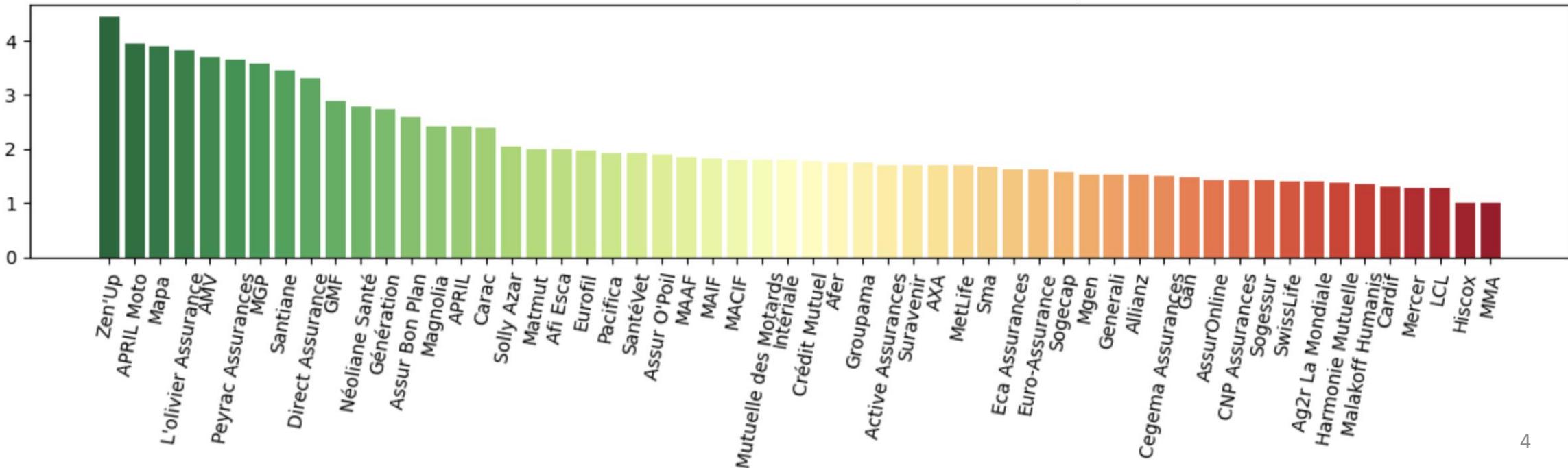
- Ensemble d'avis donnés à des assurances de la part d'usager.



Explication & analyse des données

- Ensemble d'avis donnés à des assurances de la part d'usager.
- Chaque avis peut être accompagné d'une note entre 1 et 5.

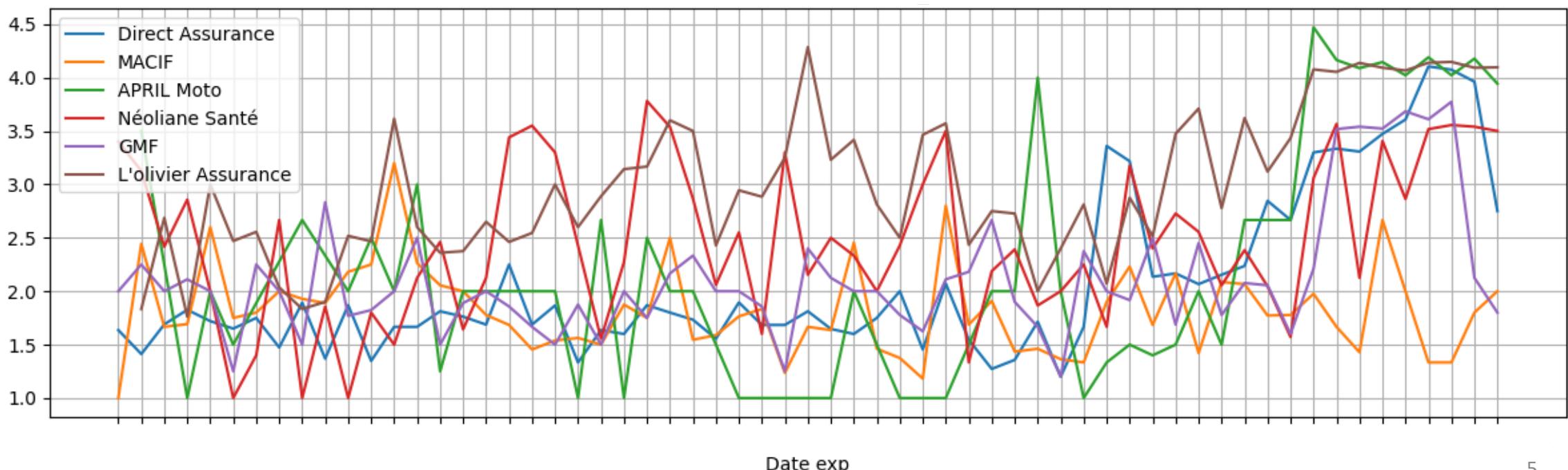
Note moyenne par assurance



Explication & analyse des données

- Ensemble d'avis donnés à des assurances de la part d'usager.
- Chaque avis peut être accompagné d'une note entre 1 et 5.
- Et d'une date, ce qui permet par exemple de suivre l'évolution de la note.

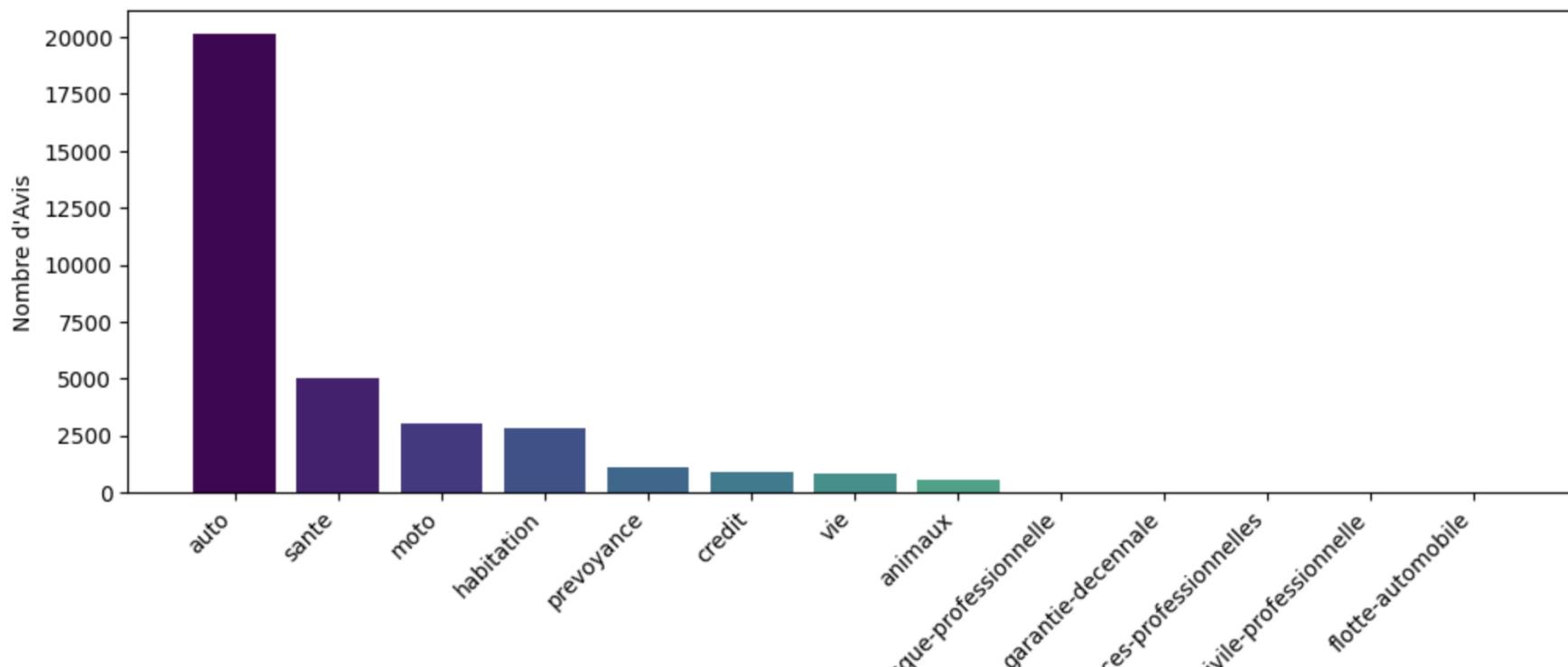
Evolution des notes par assureur



Explication & analyse des données

- Ensemble d'avis donnés à des assurances de la part d'usager.
- Chaque avis peut être accompagné d'une note entre 1 et 5.
- Et d'une date, ce qui permet par exemple de suivre l'évolution de la note.
- Les assurances sont reparties par produit.

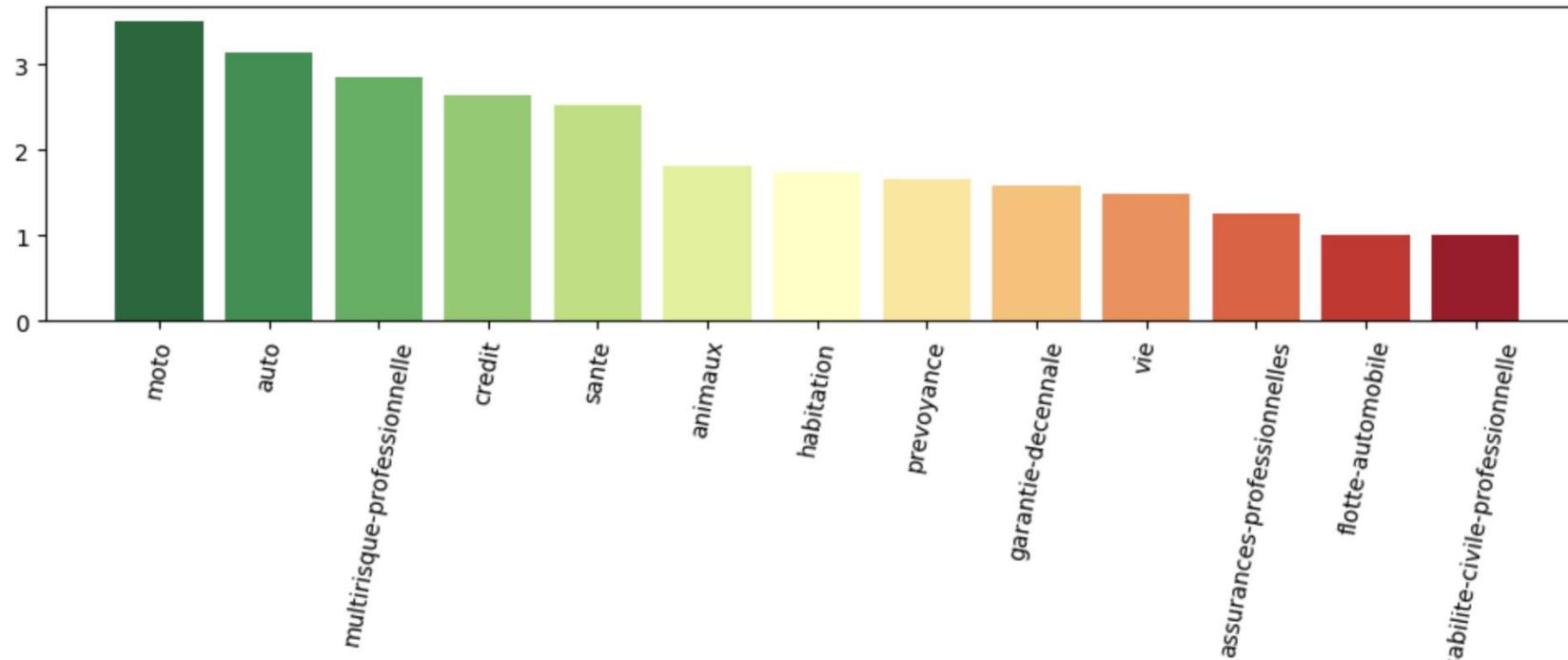
Répartition des produits dans le dataset



Explication & analyse des données

- Ensemble d'avis donnés à des assurances de la part d'usager.
- Chaque avis peut être accompagné d'une note entre 1 et 5.
- Et d'une date, ce qui permet par exemple de suivre l'évolution de la note.
- Les assurances sont reparties par produit.

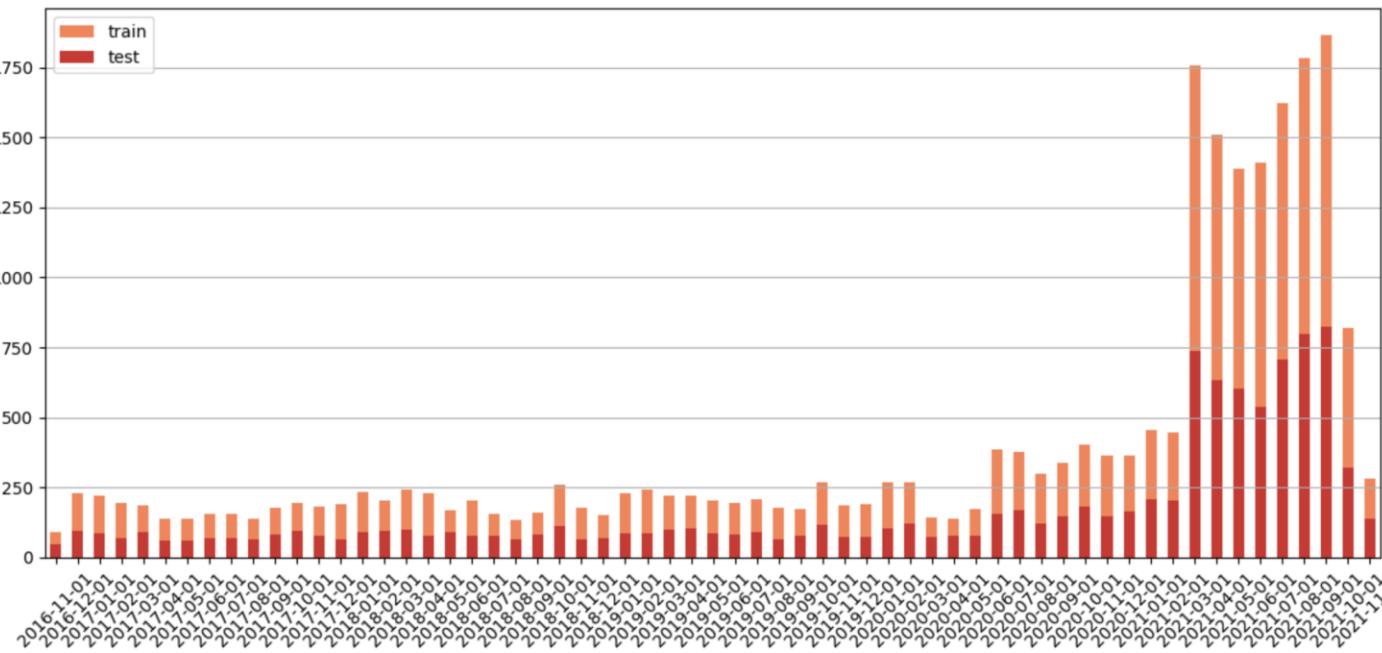
Note moyenne par produit



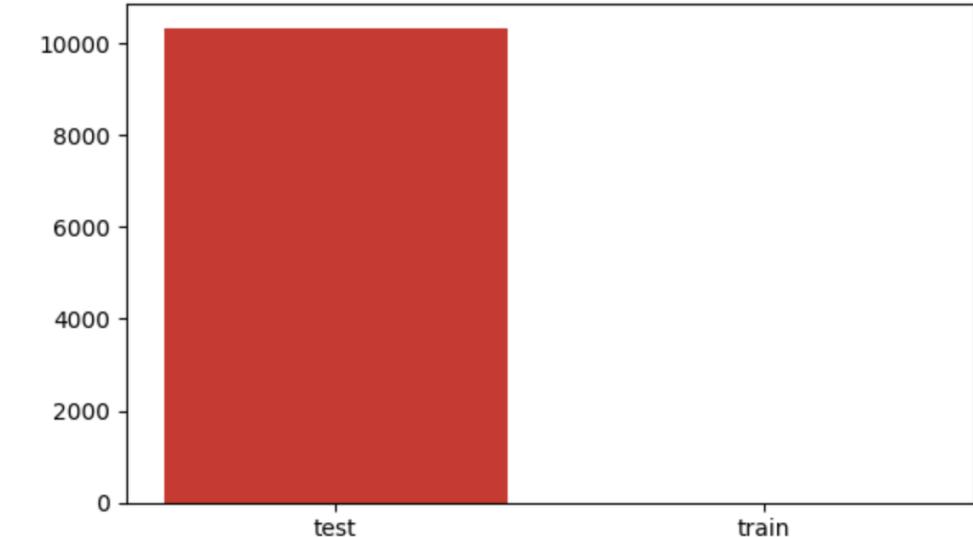
Explication & analyse des données

- Ensemble d'avis donnés à des assurances de la part d'usager.
- Chaque avis peut être accompagné d'une note entre 1 et 5.
- Et d'une date, ce qui permet par exemple de suivre l'évolution de la note.
- Les assurances sont reparties par produit.
- Les données sont reparties par type : entraînement et test. On ne connaît pas la note pour ces dernières.

Répartition du type selon la date



Répartition des valeurs nulles par type



Traitements des avis

- Determinants
 - Stopwords
 - Conjunctions,
 - Caractères spéciaux, mails, liens, code etc...

Nuage de mot avant pré-traitement

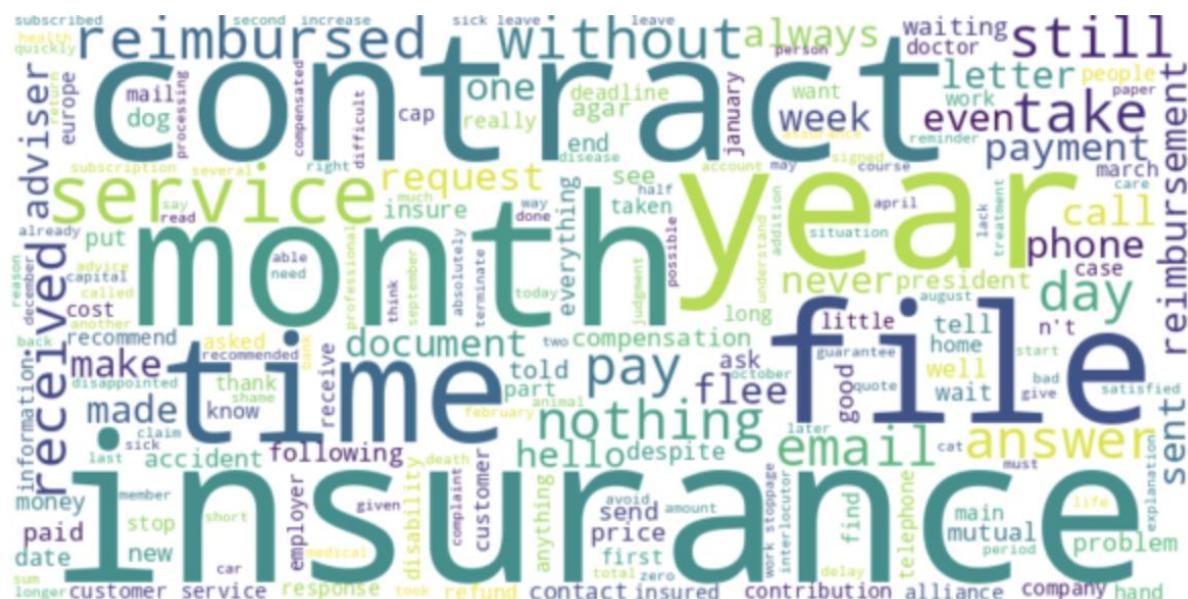


Nuage de mot après pre-traitement



Traitement des avis

NUAGE DE MOTS APRÈS CORRECTION AVEC TEXTBLOB*



NUAGE DE MOTS APRÈS LEMMATISATION

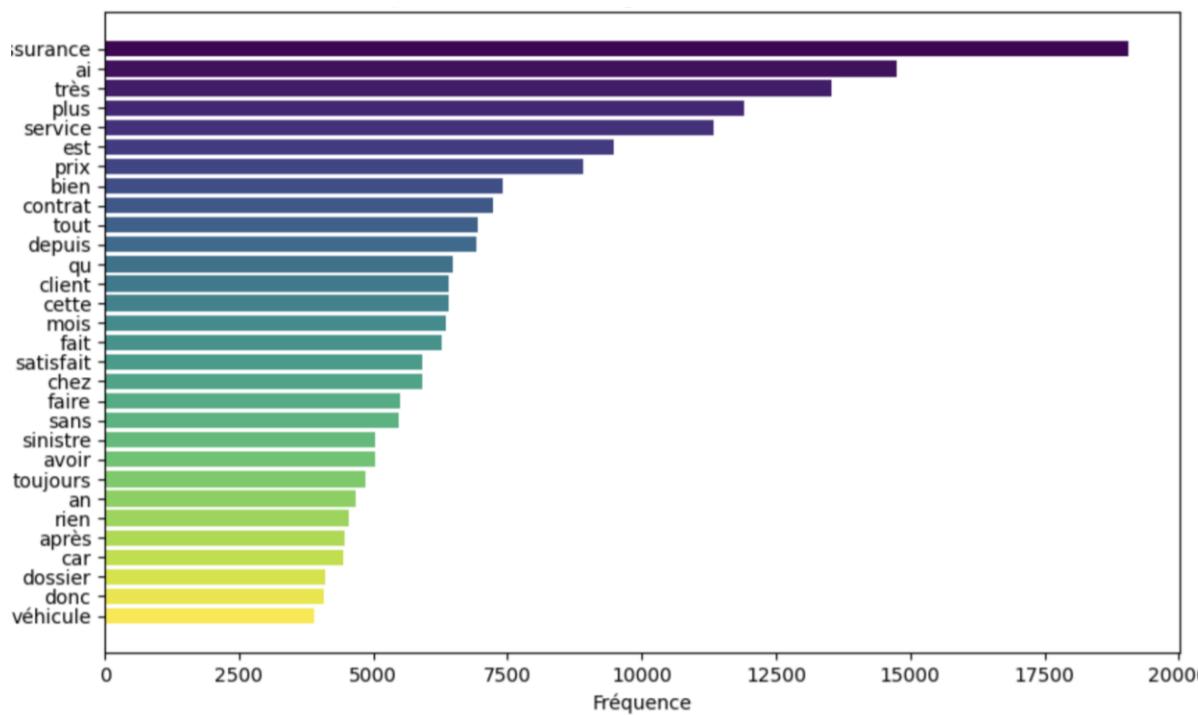


*Simplifie l'utilisation de nombreuses tâches de NLP courantes, comme l'analyse de sentiments, la classification de texte, l'extraction de phrases clés, etc. TextBlob est construite au-dessus de la bibliothèque NLTK

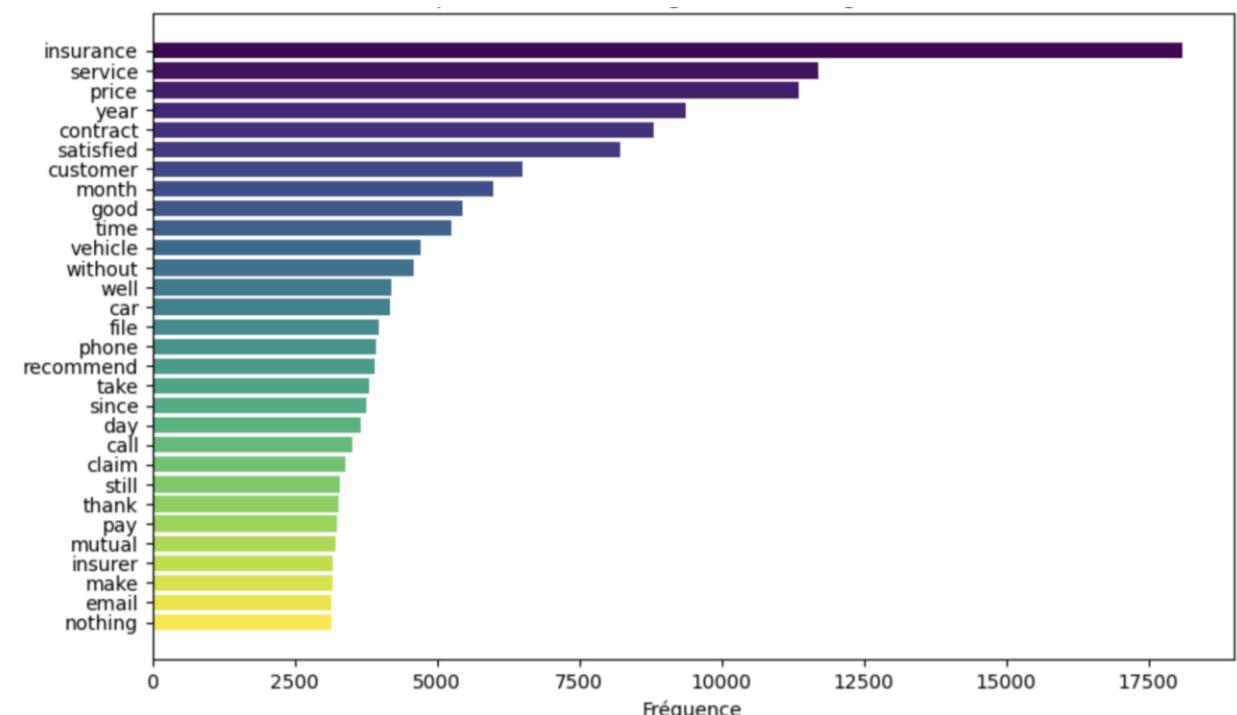
Il ne semble pas y avoir un changement visible par rapport au pré-traitement

Traitement des avis

Analyse des n-grammes dans les avis français



Analyse des n-grammes dans les avis anglais



Plus de bruit dans les avis français (déterminants, prépositions). Les librairies sont mieux adaptées à de l'anglais

Resumés et traductions avec T5-base

T5-base pour les résumés

T5, ou "Text-to-Text Transfer Transformer", est un modèle de réseau de neurones conçu par Google Research. Le modèle est basé sur les architectures suivantes: transformers, encoder et decoder avec des entrées et des sorties textuelles.

La version "base" de T5 est notamment utilisée pour générer résumé de texte. T5 reformule une entrée longue en une version courte, conservant les informations clés.

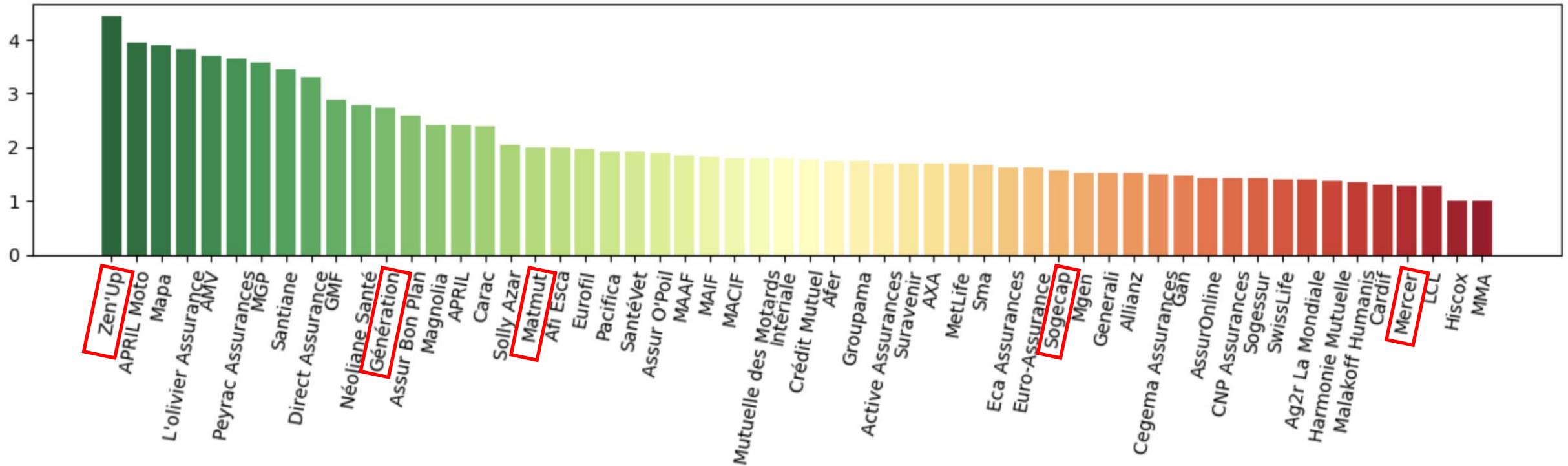
T5-base pour la traduction

T5 peut également être utilisé pour la traduction de langue.

T5 a été entraîné sur diverses paires de langues, lui permettant de comprendre et de traduire le texte efficacement en tenant compte des nuances linguistiques

Resumés et traductions avec T5-base

Note moyenne par assurance

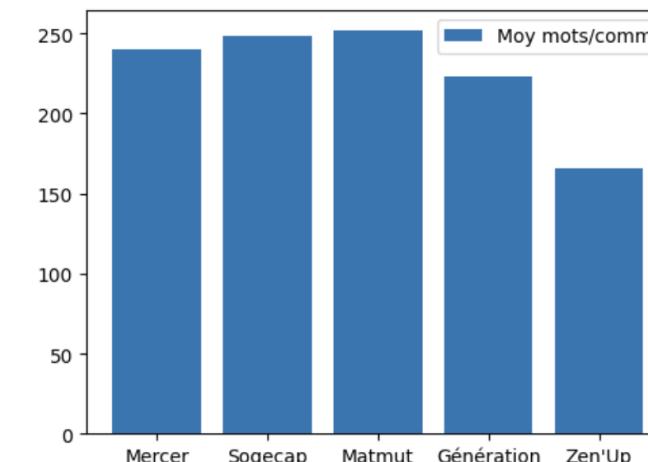
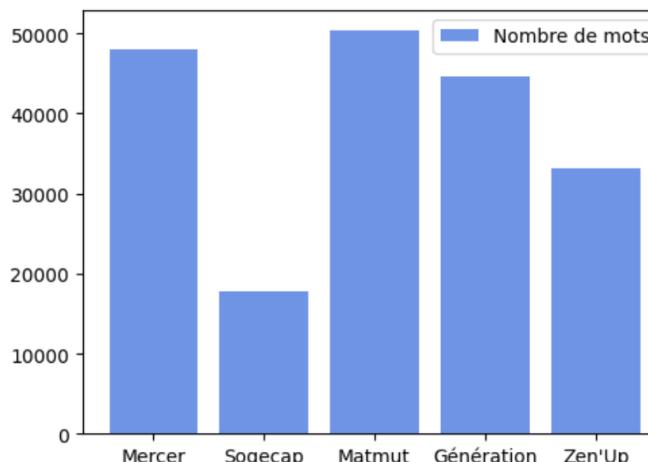
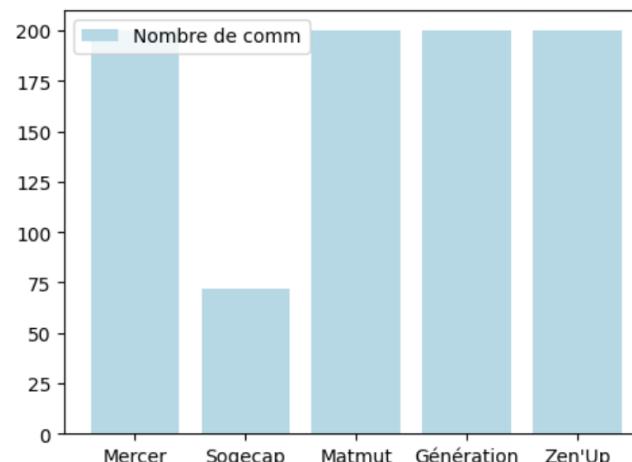


Nous allons analyser le résumé des avis pour cinq assurances différentes dont les notes sont variés. Nous souhaitons retrouver un sentiment positif ou négatifs dans les résumés. Ensuite chaque résumé sera traduis.

Resumés et traductions avec T5-base

Resumés

- Afin d'analyser le plus de commentaires, nous sélectionnons **200 avis max** par assurance.
- Pour chacune nous gardons les **300 premiers caractères**.
- Les avis ne sont pas traités avant car T5 possède son propre traitement.
- Les tenseurs ont une longueur maximum de 1000 et la taille du résumé peut osciller de 100 à 512.



Resumés et traductions avec T5-base



Mercer

Mercer (business mutual) does **not manage** the files of its customers. it's been 2 months since i **await** the validation of the portability of my mutual while the documents have been shipped more than 2 times. i have to **call at least 30 times** to perceive my reimbursement is **lamentable**. no follow -up, **no alert**, just a deposit message. the display of services that are in **problem** is hidden after 3 "actions"

Sogecap

cnn's john sutter says the company is **incompetent** and **deplorable**. sutter: "zero, zero, incompetent, deplorable service, **reimbursement delay incompatible with human problems** in the face of diseases, signature at the bottom of people who do not even exist" "i am the benefit of a life insurance contract due to my aunt who died in September 2020. sutter: "i am the benefit of a life insurance contract due to my aunt

Matmut

matmut is clearly **not a friend**, says adam mcdonald. mcdonald: "it is in adversity that we recognize his real friends" mcdonald: "the matmut, a company like no other, is clearly not a friend" mcdonald: "the matmut, a company like no other, is clearly not a friend" mcdonald: "the matmut is clearly not a friend"

Génération

the mutual is **good** in terms of price but despite having subscribed an option, i have a **lot of dependents**. despite having a robot which does not understand everything, we **quickly** have access to an **advisor who is always perfect**. the only thing that **annoyed** me is not to be able to take the formula above mine all immediately. this year sends schedule by email on december 31 for an end of contract on the same date. a contract termination within 3 months of subscription impossible.

Zen'Up

i took for the second time borrower insurance via zen'up for my new real estate credit My first was reimbursed in advance and the first insurance was closed Contact and ease of completion of the file. the questionnaire is premature since the file has just been created **I am satisfied with the zen'up service, where the agents are available. the prices announced are correct. I can gladly recommend to my entourage.** Regarding the price, the service is more **interesting** compared to the insurance offered by my bank.

Resumés et traductions avec T5-base

Traductions

- Traduction de chaque résumé.
- La longueur maximale de chaque **résumé** est de **512 caractères**.
- Les résumés ne sont pas traités avant car T5 possède son propre traitement
- Les tenseurs en entrée ont la taille du nombre de mots dans le résumé
- La longueur des traductions n'est pas contrainte

Resumés et traductions avec T5-base



Mercer

Mercer (mutuelle d'affaires) ne gère pas les dossiers de ses clients. il y a 2 mois que j'attends la validation de la transférabilité de ma mutuelle alors que les documents ont été expédiés plus de 2 fois. je dois appeler au moins 30 fois pour comprendre que mon remboursement est lamentable. pas de suivi, pas d'alerte, simplement un message de dépôt.

Sogecap

cnn's john sutter dit que l'entreprise est incompétente et déplorable. sutter : « zéro, zéro, incompétent, déplorable service, retard de remboursement incompatible avec les problèmes humains face aux maladies, signature au bas des gens qui n'existent même pas » « je bénéficie d'un contrat d'assurance-vie à cause de ma tante qui est décédée

Matmut

matmut n'est manifestement pas un ami, dit adam mcdonald. mcdonald : « C'est dans l'adversité que nous reconnaissons ses vrais amis » mcdonald : « le matmut, une entreprise comme aucune autre, n'est manifestement pas un ami » mcdonald : « le matmut, une entreprise comme aucune autre, n'est manifestement pas

Génération

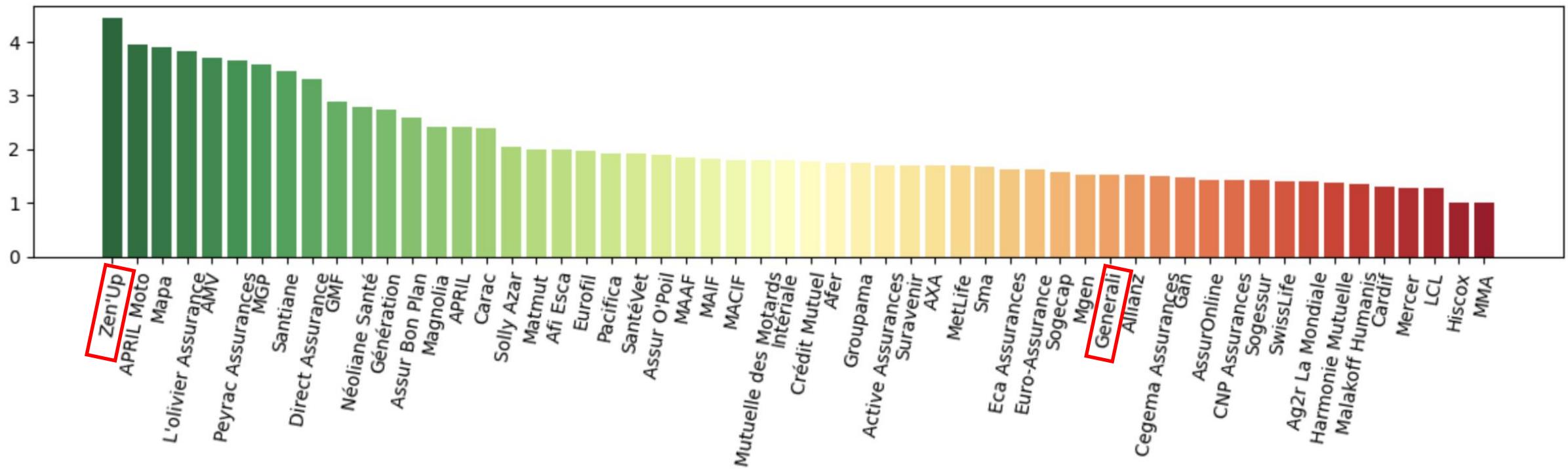
La mutuelle est bonne en termes de prix mais malgré avoir souscrit une option, j'ai beaucoup de personnes à charge. malgré avoir un robot qui ne comprend pas tout, nous avons rapidement accès à un conseiller qui est toujours parfait. la seule chose qui m'a ennuyé est de ne pas être en mesure de prendre la formule ci-dessus toutes immédiatement. cette année envoie calendrier par courriel le 31

Zen'Up

J'ai pris pour la deuxième fois l'assurance emprunteur via zen'up pour mon nouveau crédit immobilier Mon premier a été remboursé à l'avance et la première assurance a été fermée Contact et la facilité de remplir le dossier. le questionnaire est prématûré puisque le dossier vient d'être créé Je suis satisfait du service zen'up, où les agents sont disponibles. les prix annoncés sont corrects.

Analyse de sentiments

Note moyenne par assurance



Nous allons analyser le sentiment des avis pour deux assurances: Zen'Up et Generali. Au regard de la note, nous souhaitons retrouver plus d'avis positif chez Zen'up que chez Generali. Pour cela nous utilisons **reviews sentiment analysis de Hugging face**.

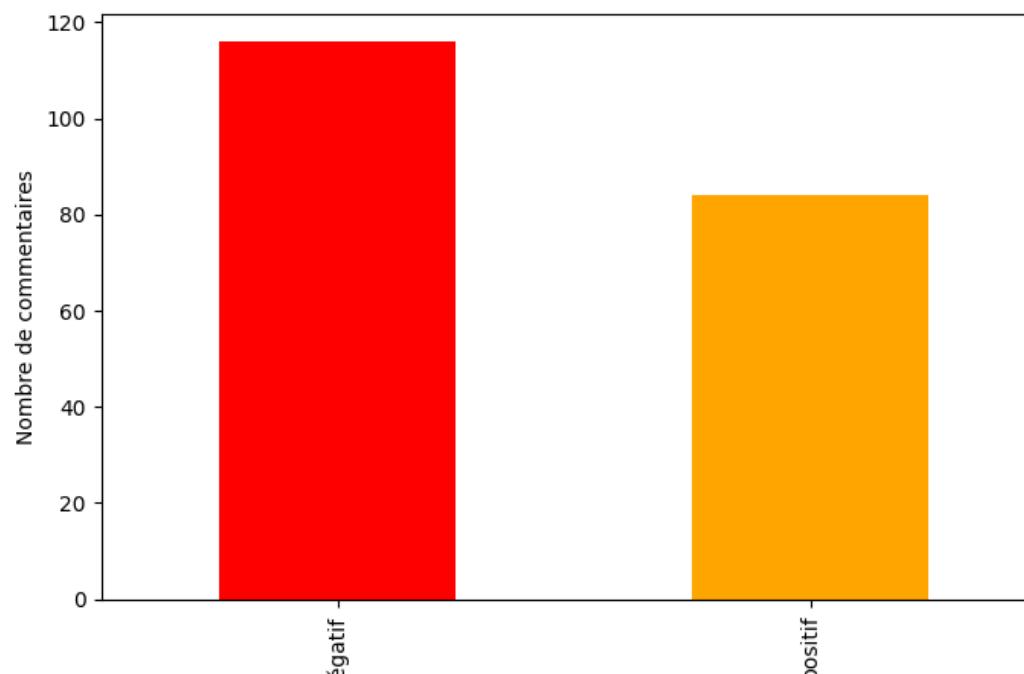
Analyse de sentiments

Utilisation d'un modèle pré-entraîné d'hugging face : juliensimon/reviews-sentiment-analysis

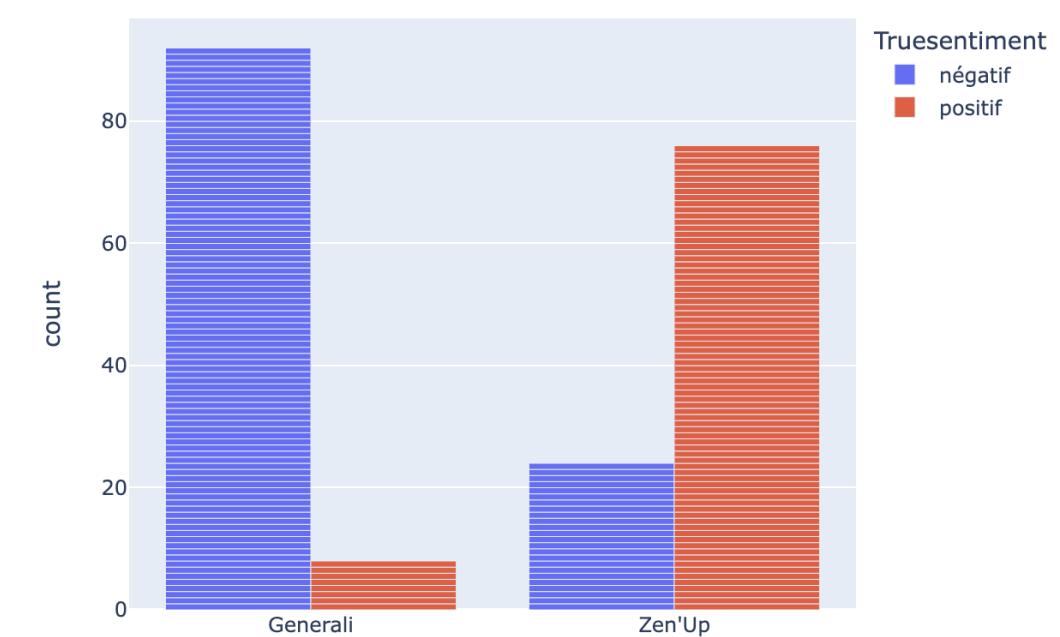
-> Distilbert modèle fine-tuné sur des avis en anglais de produits. Choisi car plus rapide et fine-tuné sur des avis déjà en anglais, ce qui ressemble fortement à des avis d'assureur.



Répartition des sentiments dans les commentaires



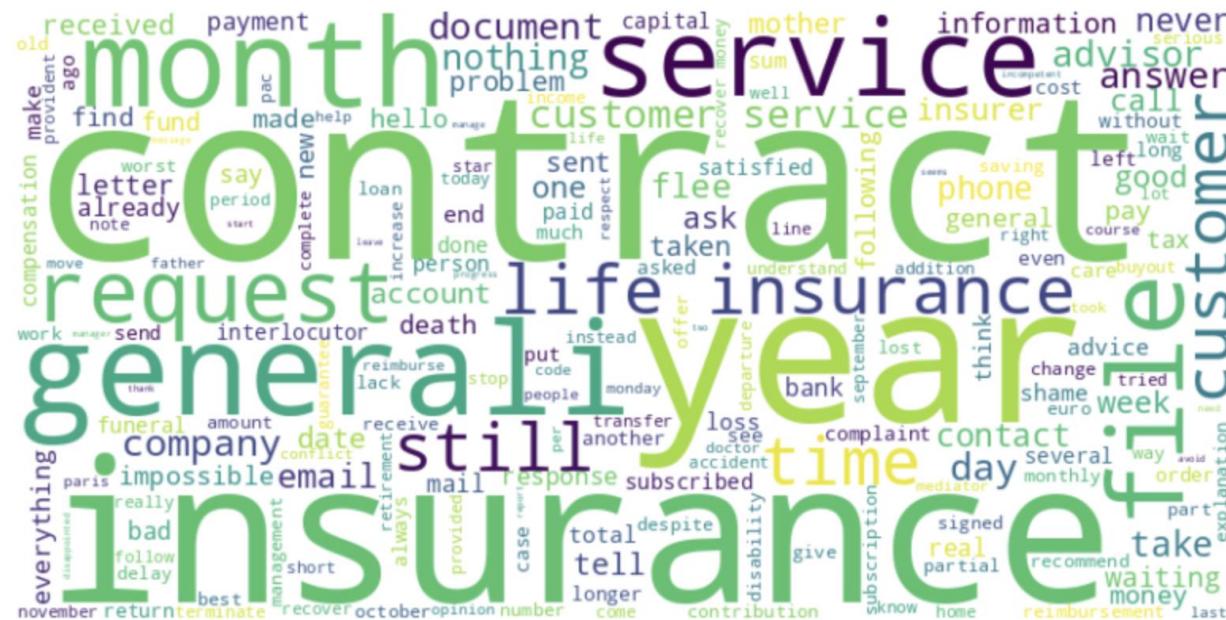
Repartition des sentiments par assurance



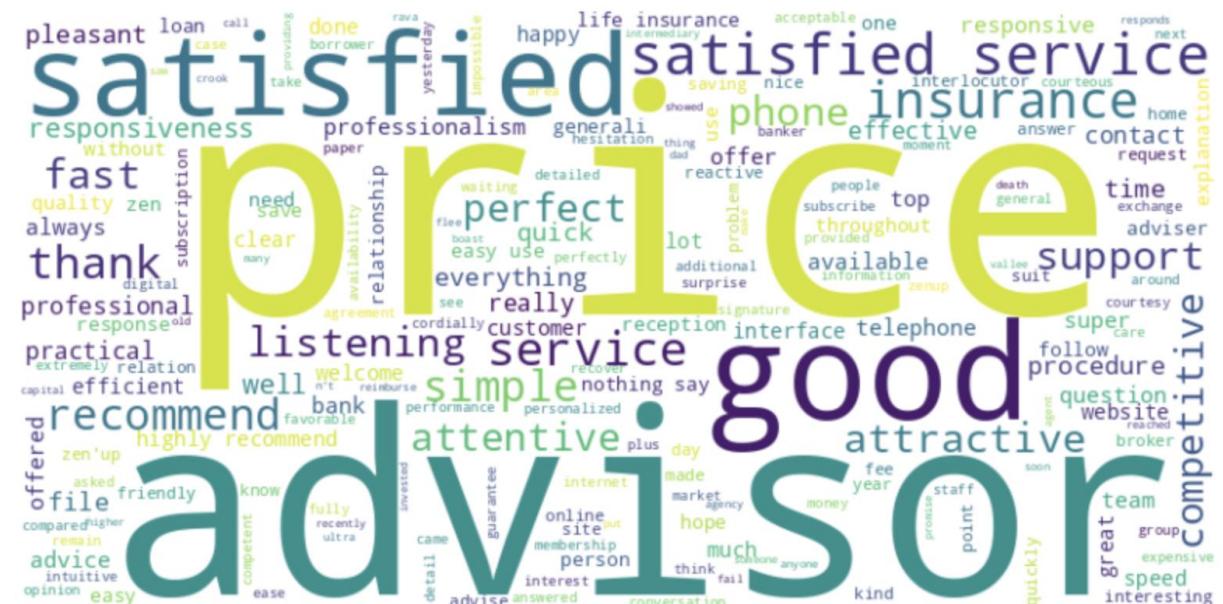
En général, les commentaires sont plus négatifs, on observe que la répartition est très inégale entre Generali et Zen'up pour laquelle les clients semblent bien plus satisfaits.

Analyse de sentiments

Nuage de mots - commentaires négatifs



NUAGE DE MOTS - COMMENTAIRES POSITIFS



On voit que pour les commentaires prédis "negatifs", des mots comme "waiting", "never", "nothing" reviennent. Pour le nuage qui prédit "positif", les mots sont beaucoup plus marqués comme "satisfied", "good", "thank".

Modélisation du sujet avec BERTopic

BERTopic

BERTopic est une technique avancée de modélisation de sujets à partir de modèles de langage contextuels comme BERT (Bidirectional Encoder Representations from Transformers). Il est basé sur l'architecture suivant : génération d'embeddings, reduction de dimension, clustering et extraction de mots clés.

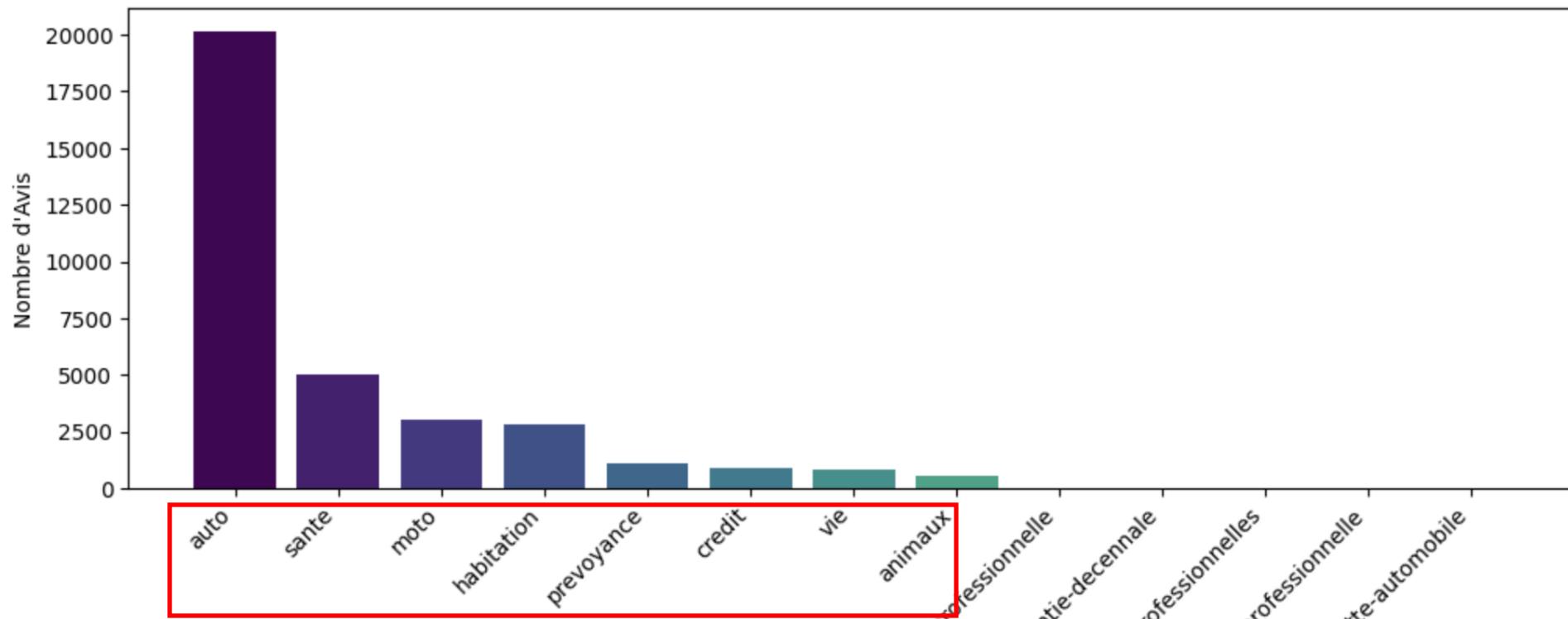
Les embeddings de phrases capturent des nuances sémantiques et des relations contextuelles entre les mots dans un texte.

Voici les champs d'utilisation de BERTopic:

- Modélisation de Sujets Contextuelle
- Clustering Dynamique
- Réduction de Dimension
- Visualisation et Interprétation

Modélisation du sujet avec BERTopic

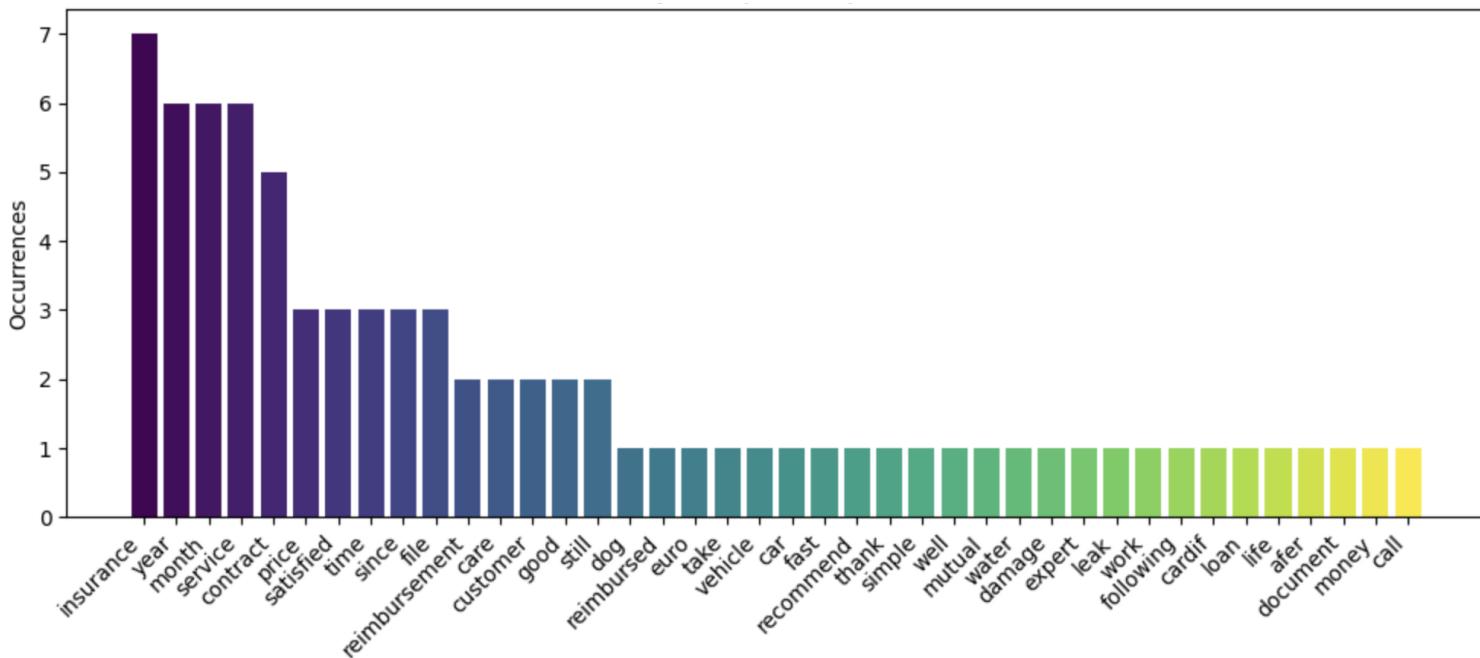
Produits les plus représentés dans le dataset



On analyse les sujets les plus fréquents pour un maximum de 200 commentaires par produit. Les produits sélectionnés sont les 8 plus fréquents dans le jeu de données

Modélisation du sujet avec BERTopic

Sujets les plus fréquents



Les sujets qu'on retrouve le plus fréquemment dans les 8 catégories sont :

- Assurance, service, contrat
- Month, year, time
- Satisfied, good, care,
- Price, reimbursement

Des sujets qui englobent toutes les catégories et qui sont plus axés sur la définition et les étapes administratives / relation clients

Modélisation du sujet avec BERTopic

Auto : vehicle, car

sante: mutual

moto: fast, recommend, thank, simple, well

habitation: water, damage, expert, leak, work, following

prevoyance: call

credit: cardif, loan

vie: life, after, document, money

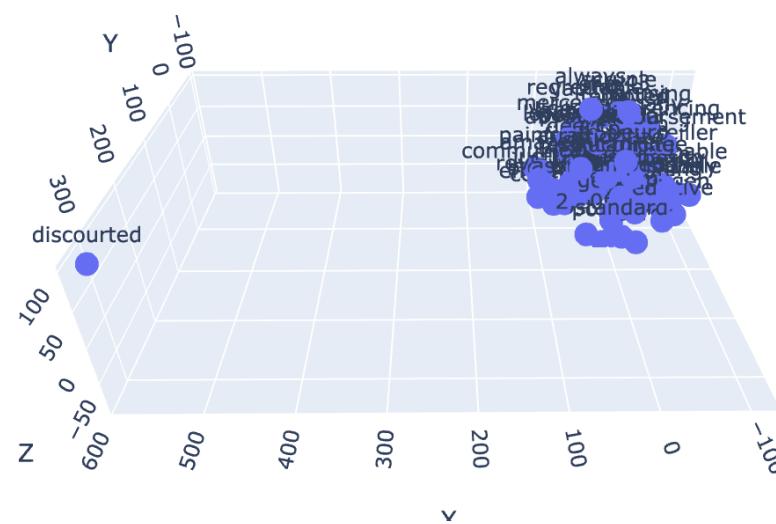
Animaux : dog, reimbursed, euro, take

Dans les sujets les plus fréquents, on retrouve aussi des sujets spécifiques à chaque assurance

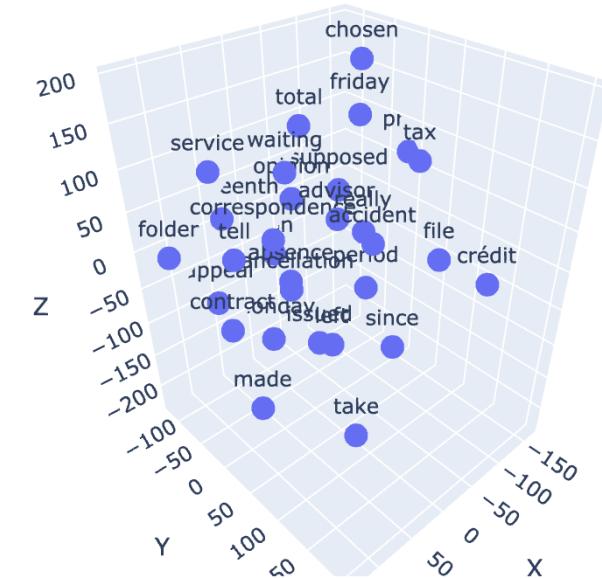
Embedding pour identifier des mots similaires

Nous avons pris un subset spécifique : les 100 premiers avis de tous les assureurs Sur ces 100 avis, nous avons visualisé avec TSNE pour une réduction en 3 dimensions les embeddings des mots, générés par Word2Vec puis Glove via spacy. La fonction prend en paramètre un assureur spécifique, ici "Generali".

**Word2Vec Embeddings on words
for Generali with t-SNE 3D**



**GloVe Embeddings on words
for Generali with t-SNE 3D**



Embedding pour identifier des mots similaires

Word2Vec

Technique de traitement du langage naturel qui convertit les mots en vecteurs numériques dans un espace vectoriel

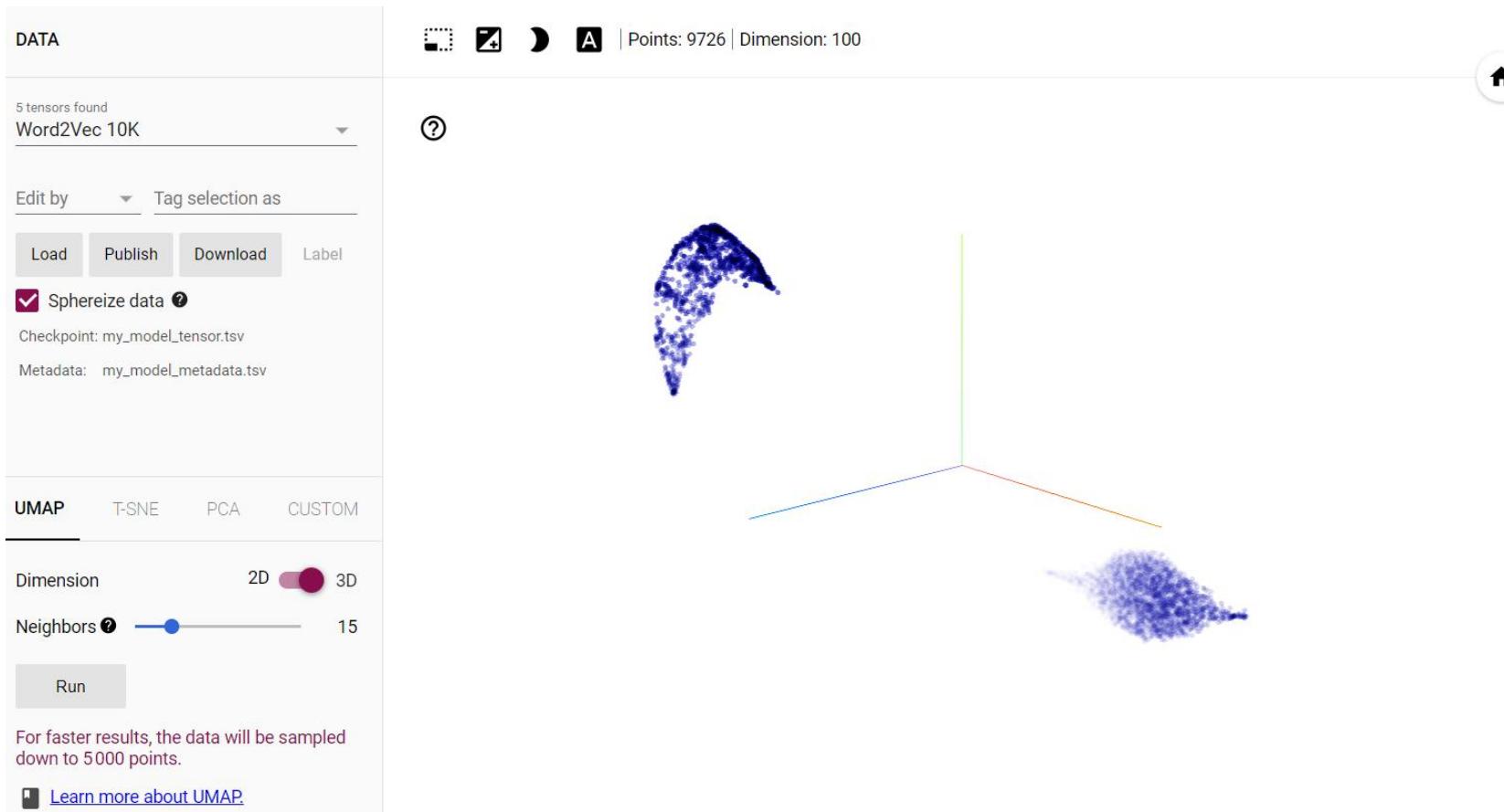
Il repose sur la sémantique distributionnelle, capturant les relations sémantiques entre les mots. Les vecteurs Word2Vec facilitent la mesure de similarité sémantique et sont souvent utilisés pour transférer des connaissances vers d'autres tâches de NLP

GloVe

Technique de représentation de mots qui vise à capturer les relations sémantiques entre les mots en analysant les co-occurrences globales dans un corpus de texte. La version de GloVe intégrée à spaCy, comme en_core_web_md, utilise des vecteurs de mots pré-entraînés de dimension moyenne (md) pour la langue anglaise. Ces vecteurs sont obtenus en analysant de vastes quantités de texte pour capturer les relations entre les mots et sont utiles pour des tâches telles que la similarité sémantique et le transfert de connaissances.

Embedding pour identifier des mots similaires

Ensuite, nous avons gardé ceux générés par Word2vec et les avons visualisés dans TensorBoard. Ici c'est pour les 100 commentaires par assureur.



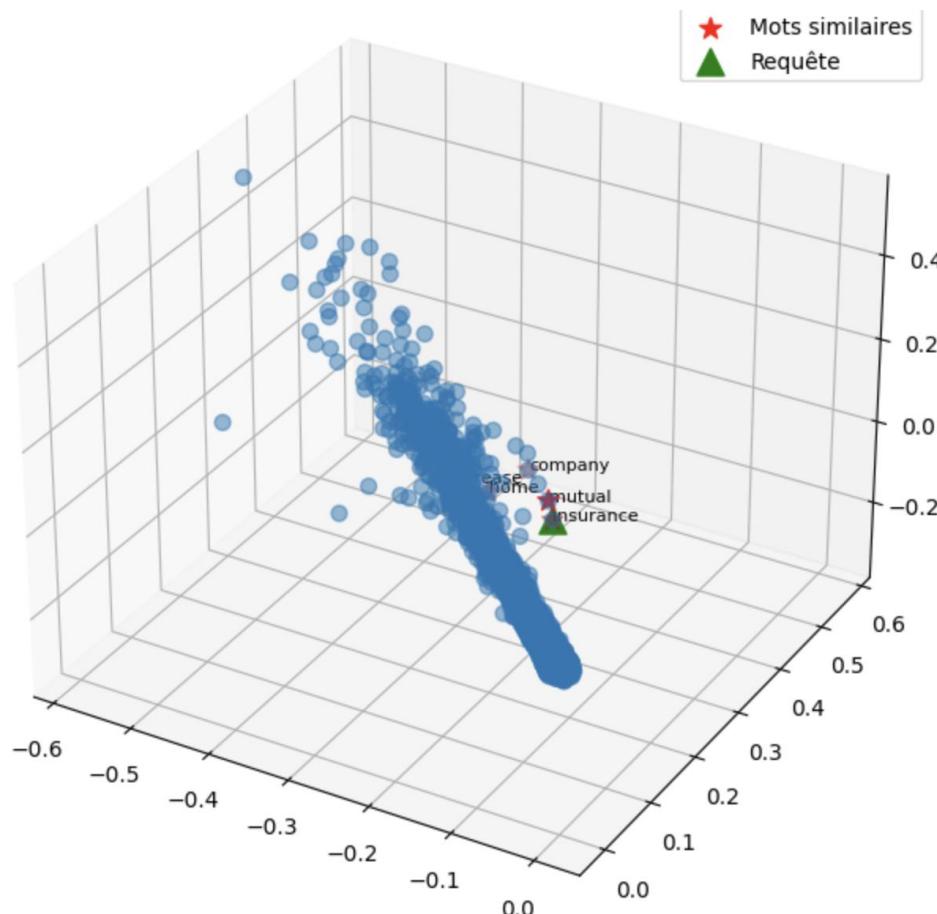
Embedding to identify similar words



Tensorboard a la possibilité de faire de la recherche sémantique. Exemple ici avec "enjo"

Embedding pour identifier des mots similaires + recherche sémantique

Recherche sémantique avec la similarité du cosinus



Ensuite, nous avons pris les 100 premiers commentaires toujours de générali, et nous avons fait des embeddings de word2vec. On tape un mot spécifique et on trace en 3D les embeddings et les 5 premiers mots les plus similaires avec notre requête, grâce à un calcul de similarité du cosinus.

```
Mot : insurance, Score : 0.9999999403953552
Mot : mutual, Score : 0.9897371530532837
Mot : home, Score : 0.9873225092887878
Mot : company, Score : 0.9852068424224854
Mot : increase, Score : 0.9837861061096191
```

Question answering

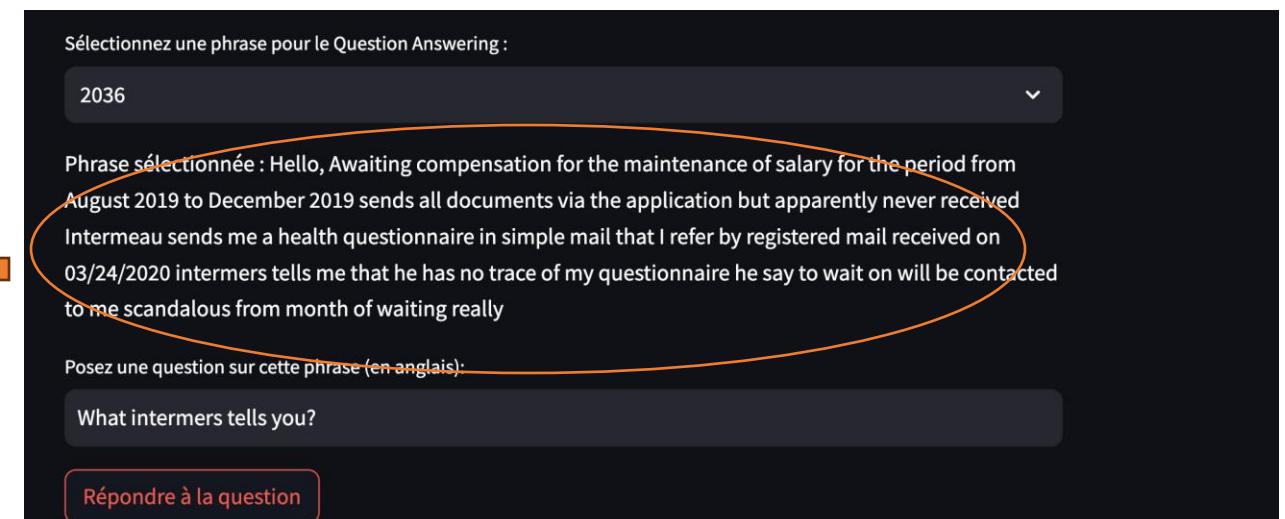
Pour faire du question answering dans notre application, nous avons utilisé le modèle **distilbert-base-cased-distilled-squad**. C'est un modèle pré entraîné qui est capable de prendre d'autres phrases en paramètre pour y répondre.

distilbert-base-cased-distilled-squad

Modèle pré-entraîné conçu pour répondre à des questions posées sur un texte (question-réponse), basé sur l'architecture DistilBERT, avec une tokenisation préservant la casse et distillé à partir du jeu de données SQuAD. Ce modèle est souvent utilisé pour des tâches spécifiques de question-réponse.

```
# Charger le modèle de question-answering
question_answering_pipeline = pipeline('question-answering', model='distilbert-base-cased-distilled-squad')

# Obtenir la réponse à la question
result = question_answering_pipeline(
    question=question_qa,
    context=df_selected_assureur.loc[selected_index_qa, 'avis_en']
)
```



Question answering

2. Question Answering ? 🎯

Modele de question answering d'Hugging face : distilbert-base-cased-distilled-squad

Sélectionnez une phrase pour le Question Answering :

2036

Phrase sélectionnée : Hello, Awaiting compensation for the maintenance of salary for the period from August 2019 to December 2019 sends all documents via the application but apparently never received Interneau sends me a health questionnaire in simple mail that I refer by registered mail received on 03/24/2020 intermers tells me that he has no trace of my questionnaire he say to wait on will be contacted to me scandalous from month of waiting really

Posez une question sur cette phrase (en anglais):

What intermers tells you?

Répondre à la question

Chargement terminé ✓

Réponse : he has no trace of my questionnaire
(Confiance : 0.1013)

Extrait dans l'application : on choisit un avis qui va nourrir le modèle, et on peut poser une question sur l'avis. Ensuite, on appelle le modèle qui répond avec un score de confiance.

[Lien de l'app](#)

Recherche sémantique de phrase

Pour cette partie, au lieu d'avoir des embeddings pour les mots, nous avons généré des embeddings pour chaque critique avec un modèle pré-entraîné "**all-MiniLM-L6-v2**" à partir de hugging face. Nous avons sélectionné 50 avis par assureur.

all-MiniLM-L6-v2

C'est un modèle de langue pré-entraîné basé sur l'architecture MiniLM avec 6 couches de transformer, développé pour des tâches de traitement du langage naturel. Ces modèles sont souvent utilisés comme points de départ pour des tâches spécifiques, tels que le question-answering, la classification de texte.

	avis_en	avis_process_en	embedding
0	My father died 4 weeks ago (05/05/2021), we im...	father died week ago immediately warned insure...	[-0.38344634, 0.29173043, 0.0046724505, -0.063...
1	I opened a general life protection contract. I...	opened general life protection contract view s...	[-0.5199887, 0.23300038, -0.03205548, -0.21843...
2	In case of conflicts it works service of the a...	case conflict work service assurance insurance...	[-0.53974533, -0.05201935, -0.29816607, -0.124...

Recherche sémantique de phrase

3. Recherche Sémantique sur l'assureur sélectionné 🔎

Embeddings fait avec Sbert : paraphrase-MiniLM-L6-v2

Puis similarité cosinus entre la query et les avis

Tapper un avis pour trouver des similaires 🏆 :

I have been waiting to be refund for month and I have no answers

Rechercher

Chargement terminé ✓

Phrases les plus similaires 🏆 :

- Reimbursement of wages which is made with weeks of delay putting the users in financial precariousness Request for documents to delay payments while contributions are collected without delay Telephone service not always reachable I plan to leave this mutual a lot once re-established and at the end of my contract (Similarité 📈: 0.4211)
- Hello, Awaiting compensation for the maintenance of salary for the period from August 2019 to December 2019 sends all documents via the application but apparently never received Interneau sends me a health questionnaire in simple mail that I refer by registered mail received on 03/24/2020 intermers tells me that he has no trace of my questionnaire he say to wait on will be contacted to me scandalous from month of waiting really (Similarité 📈: 0.4044)

Extrait de l'application. L'utilisateur écrit sa phrase, nous calculons son embedding avec le modèle pré entraîné miniLmL6, puis nous calculons avec la similarité du cosinus les 3 avis les plus similaires.

[Lien de l'app](#)

Supervised learning

TF-idf

"Term Frequency-Inverse Document Frequency" est une technique statistique utilisée pour évaluer l'importance d'un mot dans un document par rapport à un ensemble de documents.

Elle favorise les mots qui sont fréquents dans un document mais rares dans l'ensemble du corpus.

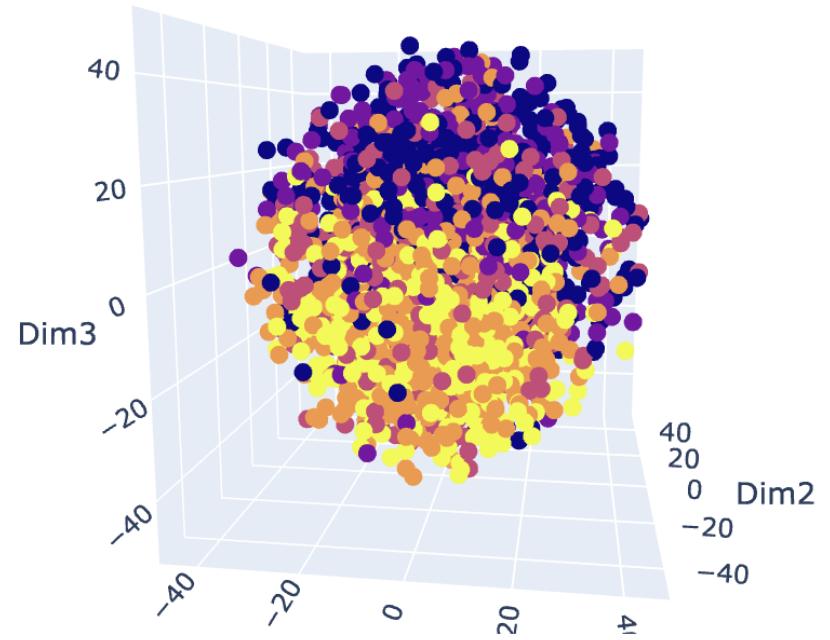
FastText

Développé par Facebook, permet de générer des embeddings de mots et de phrases.

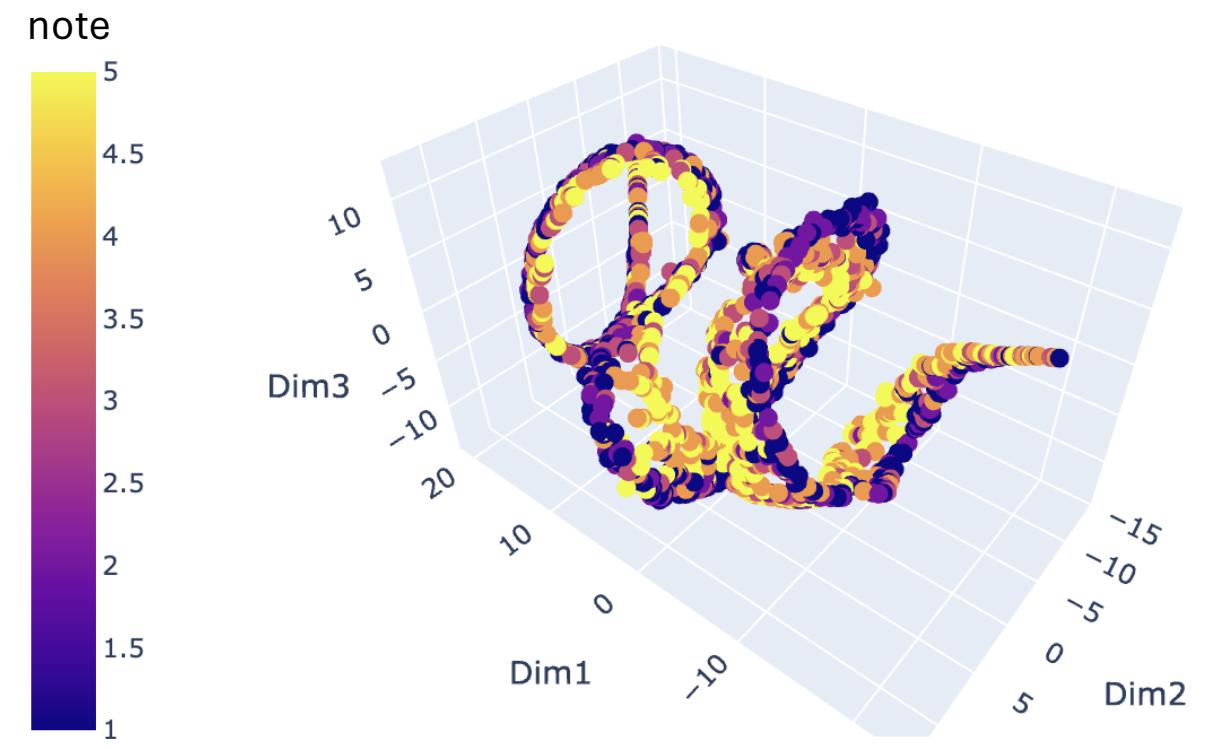
Il prend en compte la structure sous-jacente des mots (subwords), ce qui le rend efficace pour traiter des mots mal orthographiés.

Supervised learning

**TFidf embeddings sur les avis
TSNE visualisation**



**Fasttext embeddings sur les avis
TSNE visualisation**



Supervised learning

Dans cette partie, nous avons prédit la note d'un commentaire à partir des embeddings des avis.

Modèle 1

Modèle : Random forest
Classifier

Target : Note de 1 à 5
(quantitatives)

Variables : Embeddings Tfifd des
avis

Accuracy: 0.5069838196653298				
Classification Report:				
	precision	recall	f1-score	support
1.0	0.54	0.95	0.69	2220
2.0	0.50	0.02	0.03	1129
3.0	0.38	0.05	0.08	983
4.0	0.43	0.42	0.42	1459
5.0	0.51	0.61	0.55	1440
accuracy			0.51	7231
macro avg	0.47	0.41	0.36	7231
weighted avg	0.48	0.51	0.42	7231

Modèle 2

Modèle : Random forest
Classifier

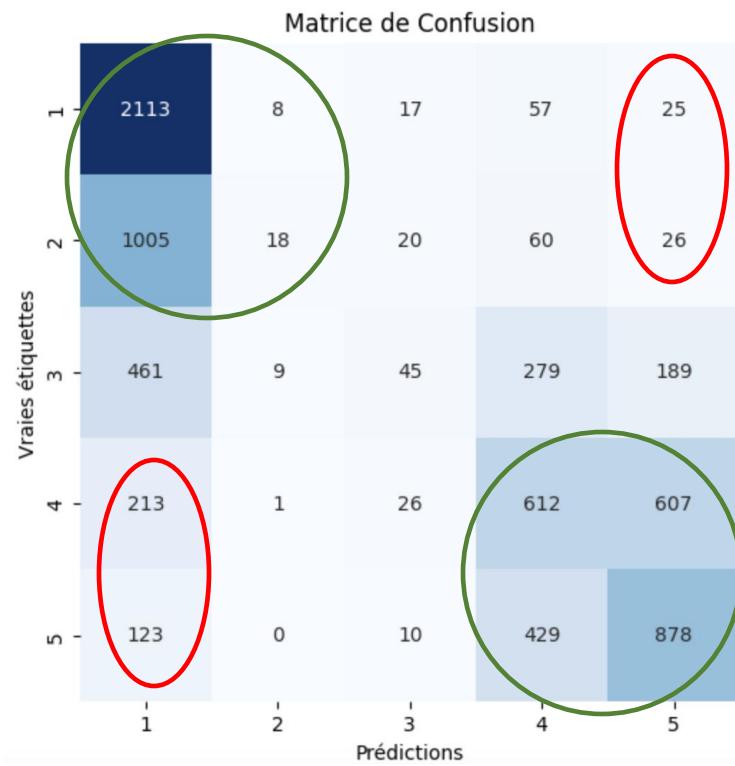
Target : Note de 1 à 5
(quantitatives)

Variables :
- Embeddings Tfifd des
avis

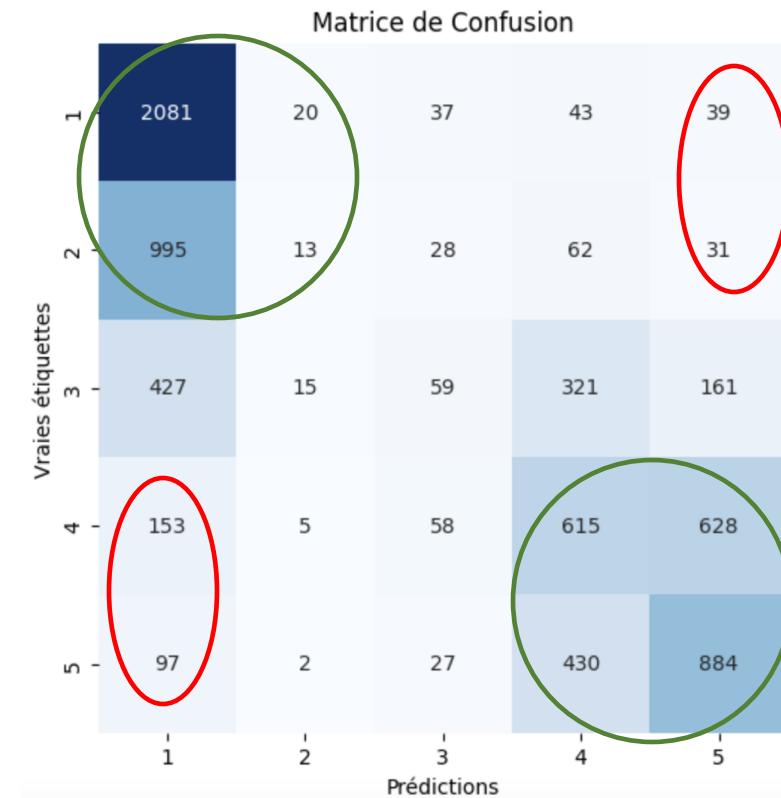
Accuracy: 0.5050477112432582				
Classification Report:				
	precision	recall	f1-score	support
1.0	0.55	0.94	0.70	2220
2.0	0.24	0.01	0.02	1129
3.0	0.28	0.06	0.10	983
4.0	0.42	0.42	0.42	1459
5.0	0.51	0.61	0.56	1440
accuracy			0.51	7231
macro avg	0.40	0.41	0.36	7231
weighted avg	0.43	0.51	0.43	7231

Supervised learning

Modèle 1



Modèle 2



Les deux modèles ont à peu près la même accuracy de 0,5. Bien que les deux modèles aient une précision générale faible, ils réussissent néanmoins à nuancer le sentiment exprimé dans les critiques sans générer d'erreurs trop extrêmes. Ex : ils peinent à distinguer des notes proches, comme 4 et 5, mais arrivent à les classer dans le bon cluster (bonnes évaluations). Les valeurs extrêmement mal prédites

Supervised learning

Modèle 2

Modèle : Random forest Classifier

Target : Note de 1 à 5 (quantitatives)

Variables :

- Embeddings **Tfidf** des avis
- Produits (qualitatives)
- Assureurs (qualitatives)

Accuracy: 0.5050477112432582				
Classification Report:				
	precision	recall	f1-score	support
1.0	0.55	0.94	0.70	2220
2.0	0.24	0.01	0.02	1129
3.0	0.28	0.06	0.10	983
4.0	0.42	0.42	0.42	1459
5.0	0.51	0.61	0.56	1440
accuracy			0.51	7231
macro avg	0.40	0.41	0.36	7231
weighted avg	0.43	0.51	0.43	7231

Modèle 3

Modèle : Random forest Classifier

Target : Note de 1 à 5 (quantitatives)

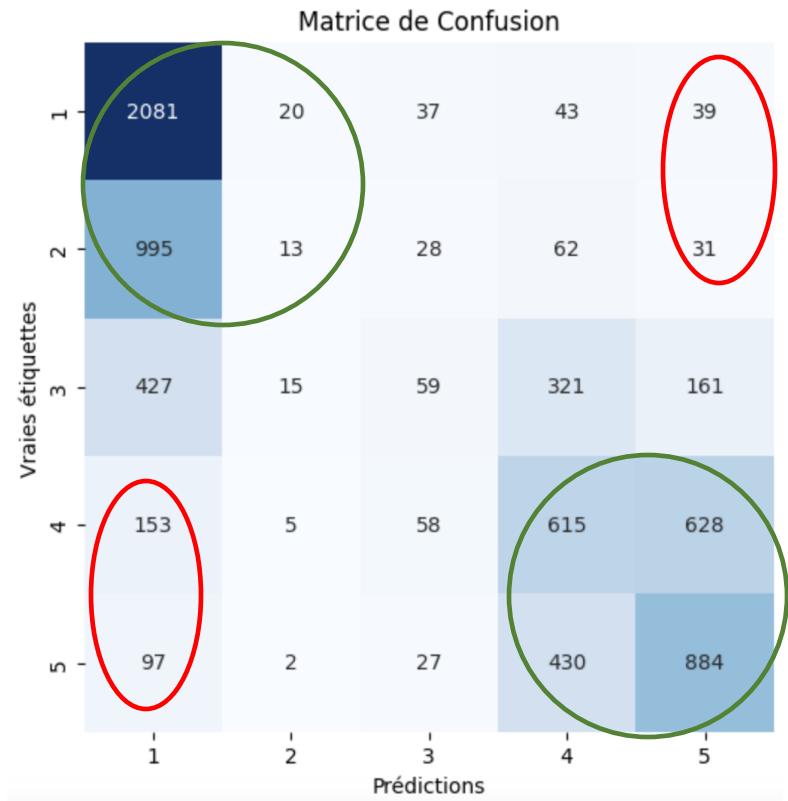
Variables :

- Embeddings **FastText** des avis
- Produits (qualitatives)
- Assureurs (qualitatives)

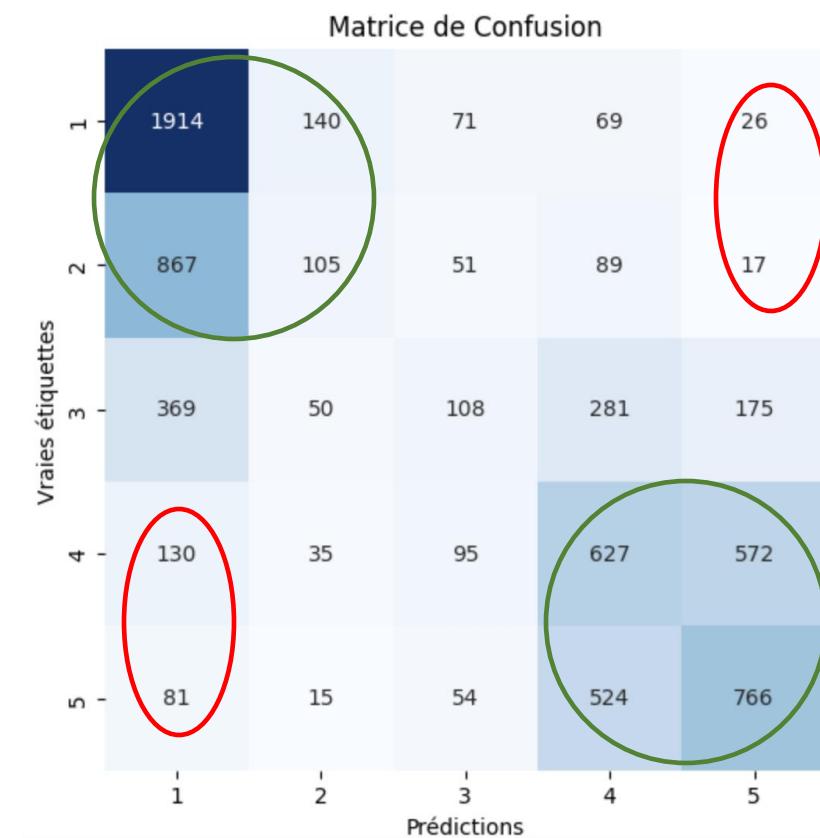
Accuracy: 0.48679297469229704				
Classification Report:				
	precision	recall	f1-score	support
1.0	0.57	0.86	0.69	2220
2.0	0.30	0.09	0.14	1129
3.0	0.28	0.11	0.16	983
4.0	0.39	0.43	0.41	1459
5.0	0.49	0.53	0.51	1440
accuracy			0.49	7231
macro avg	0.41	0.41	0.38	7231
weighted avg	0.44	0.49	0.44	7231

Supervised learning

Modèle 2



Modèle 3



Le modèle 3 présente une précision inférieure, mais il offre l'avantage de générer moins de faux positifs extrêmes pour les évaluations, même si la précision globale reste modeste. De nouveau, les deux modèles semblent être capables de regrouper correctement les avis positifs et négatifs, sans toutefois parvenir à affiner la prédiction pour atteindre la note exacte.

Supervised learning

AutoTokenizer

Ce tokenizer est conçu pour travailler avec BERT-mini. Il convertit le texte en tokens

Bert For Sequence Classification :

Ce modèle est une instance de BERT ("Bidirectional Encoder Representations from Transformers") pour la classification de séquences. "*prajjwal1/bert-mini*" est une version plus petite adapté pour les environnements avec des ressources limitées comme google collab.

L'architecture du modèle est la suivante : Transformer, réduction du nombre de couches pour bert mini, Classification des séquences

Supervised learning

Modèle 4

Modèle : Sbert

Target : Note de 1 à 5 (qualitative)

Variables :

- Tokens des avis (max = 128)

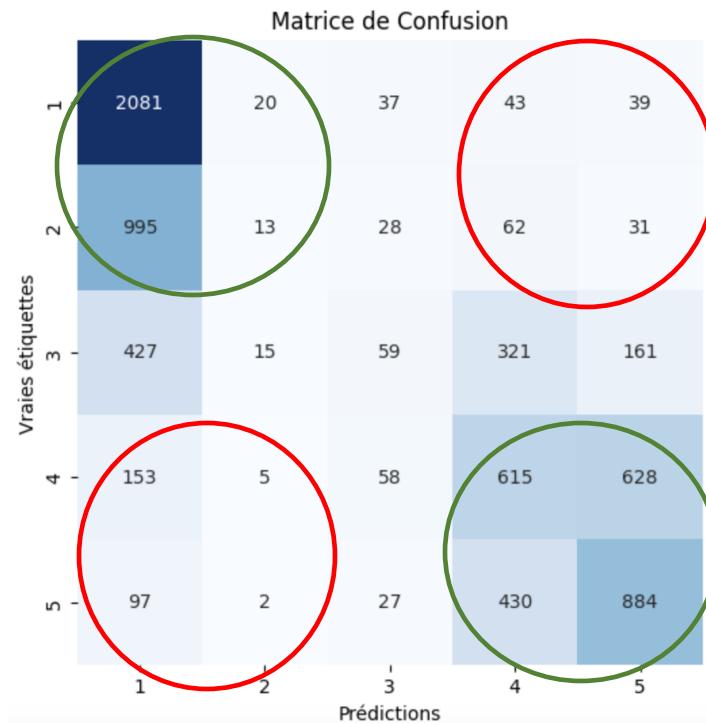
Accuracy on test set:
0.4964717741935484

Hyper-paramètres :

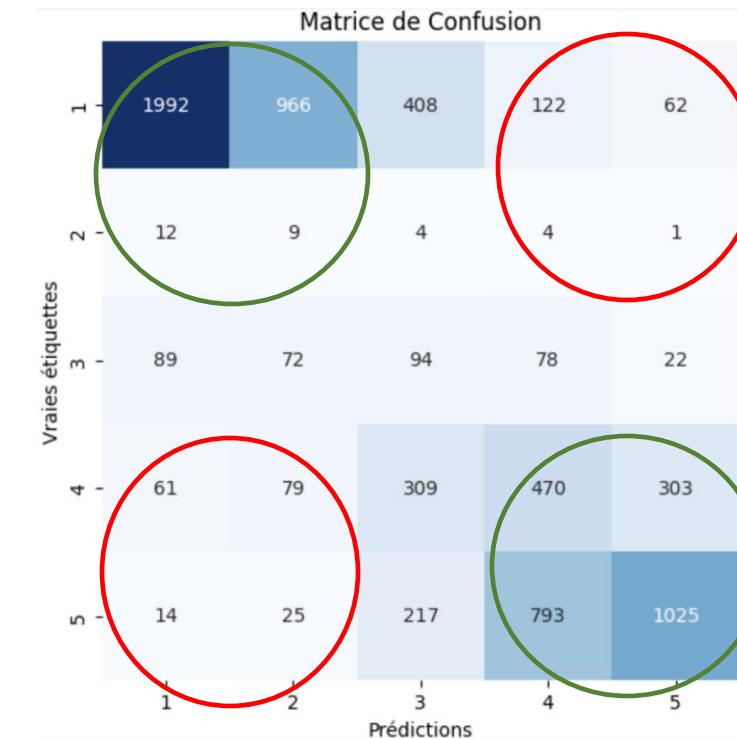
- Nombre de classes : 5
- Epoch : 3
- Batch size : 32
- optimizer : Adam
- Learning rate : 2e-5
- criterion : CrossEntropyLoss()

Supervised learning

Modèle 2



Modèle 4



Nous aboutissons aux mêmes conclusions que précédemment. Plus, Bert-mini parvient avec plus de succès à prédire et à nuancer les notes élevées par rapport aux faibles. Notre modèle présente des marges d'amélioration notables. Il est surprenant que la précision soit inférieure à celle d'un random forest.

Notre modèle pourrait être amélioré sur plusieurs aspects, notamment en considérant les classes comme quantitatives plutôt que qualitatives, en optimisant le nombre d'époques (epochs), les caractéristiques (features), ainsi que la gestion de la fonction de perte (loss). Et en s'appuyant sur des cas similaires qu'on pourrait trouver en ligne pour choisir les hyper-paramètres

Notre application : lien

Visualisation d'Embeddings en 3D et Fonctionnalités NLP - Par Maud Tissot et Aurélien Pouxviel

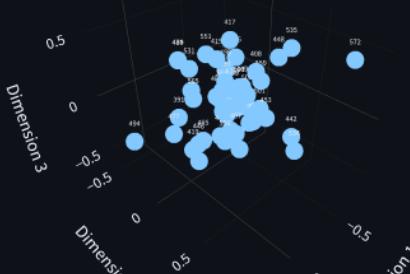
Sélectionnez un type d'assureur 🔥 :

Direct Assurance

1. Visualisation d'Embedding en 3D pour Direct Assurance

Embeddings fait avec Sbert : paraphrase-MiniLM-L6-v2

Visualisation d'Embedding en 3D pour Direct Assurance



2. Question Answering ? 🎯

Modèle de question answering d'Hugging face : distilbert-base-cased-distilled-squad

Sélectionnez une phrase pour le Question Answering :

477

Phrase sélectionnée : I am satisfied with the rates that are very attractive, have can choose several options at least. I'm happy. To see now in the long term and the real guaranteed

Posez une question sur cette phrase (en anglais) :

what is attractive ?

Répondre à la question

3. Recherche Sémantique sur l'assureur sélectionné



Embeddings fait avec Sbert : paraphrase-MiniLM-L6-v2

Puis similitude cosinus entre la query et les avis

Taper un avis pour trouver des similaires 🔎 :

im very happy with this assurance. Lots of option

Rechercher

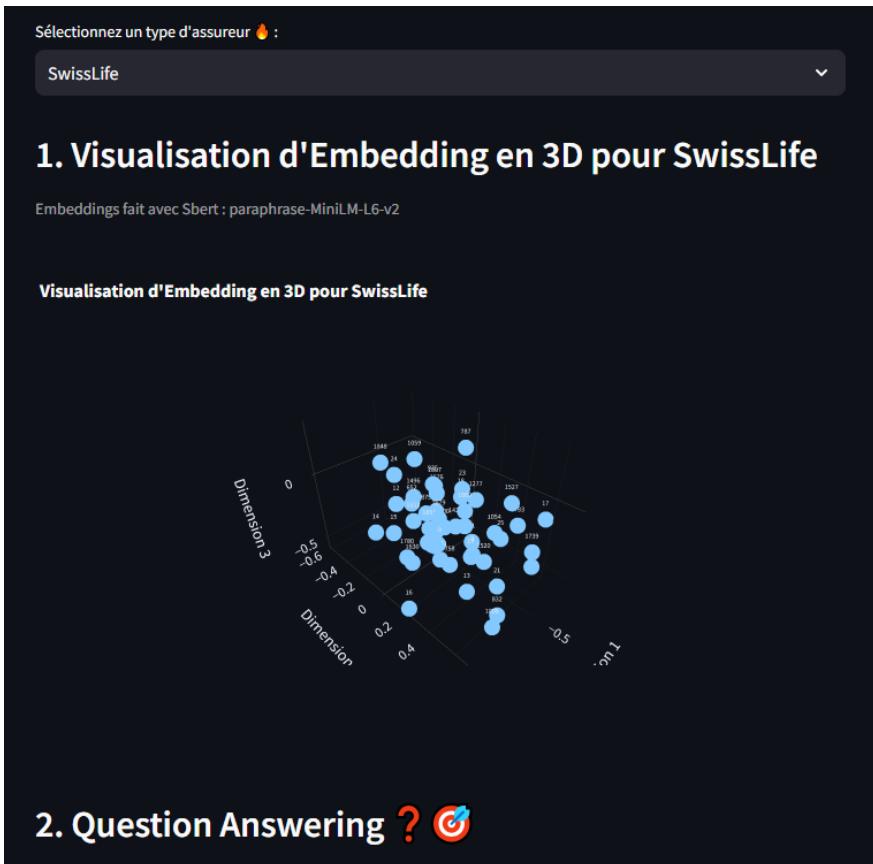
Chargement terminé ✓

Phrases les plus similaires 🏆 :

- I am satisfied with the rates that are very attractive, have can choose several options at least. I'm

Notre application : lien

1. On sélectionne un assureur et on visualise en 3D les embeddings des avis



2. On choisit un avis et on peut poser une question dessus (Question Answering)

Sélectionnez une phrase pour le Question Answering :

8

Phrase sélectionnée : My wife has been death since October 05, 2021, I made the declaration and sent all the documents that I was asked on November 07, 2021, but today November 9, 2021, I am still waiting for my transfer that 'I am promised then 3 weeks. Thank you Swisslife

Posez une question sur cette phrase (en anglais):

when did your wife died

Répondre à la question

Chargement terminé ✓

Notre application : lien

3. On peut faire de la recherche sémantique basée sur le cosinus pour un avis écrit

3. Recherche Sémantique sur l'assureur sélectionné



Embeddings fait avec Sbert : paraphrase-MiniLM-L6-v2

Puis similarité cosinus entre la query et les avis

Tapper un avis pour trouver des similaires :

Im still waiting for documents

Rechercher

Tapper un avis pour trouver des similaires :

Im still waiting for documents

Rechercher

Chargement terminé ✓

Phrases les plus similaires 🏆 :

- Hello Essential protection contract It's been 2 x that I send requests for daily allowances. 1 response outside time after 3 months of waiting 2 response, still nothing received after 1 month and a half of request. Wait in waiting .., registered since April 2010. Jamais sick. End of the year I terminate. In my advice Take another insurance. (Similarité ✅ : 0.3470)
- An incompetence, an abuse in the request of documents, a lack of respect for their customers. For more than 3 months I have been trying to buy my life insurance. At first I was faced with the incompetence of employees who certified me that no internal document was necessary. Just a buy-back letter, identity pcs, home confirmation, Having no news after 3 weeks, I contacted the management service which asked me for an additional sheet. Then two weeks later still nothing, so I recalled customer service which informed me that the 5000 document was necessary in my case. (Foreign residents). This document was very complicated to obtain in my situation. So I contacted the tax center in France which certified me that this document was not necessary, and told me about the abusive insurance/banks practices. (Know that I closed 2 other life insurance MACSF, Valvi de la Bred, and that none asked me for this document). Insurance not wanting to give up on it, I had to take a lawyer to get it the famous document 5000 which did not take up any other information already