Machine Learning for NLP : Insurance reviews

ESILV A5 DIA1 –

Romain GIRODET and Annabel MERCERON

Sommaire

1. Objectifs et problématique
2. Nettoyage du jeu de données
3. Visualisation

4. Vectorisation des avis
5. Apprentissage supervisé: estimer la notation
6. Apprentissage non-supervisé : création de cluster
7. Interprétation du jeu de données

Objectifs et problématique

Le jeu de données contient des informations sur les **avis laissés par les clients sur un produit d'assurance**. Chaque ligne comprend une date, une note sur l'avis, l'auteur de l'avis, le texte de l'avis, l'assureur et le type de produit.

Problématique : Quelle est la raison lorsque la note est basse ?

Objectifs :

- Prédire les notes à l'ensemble de données « test »
- Trouver les raisons pour lesquelles les notes de certaines compagnies d'assurance sont si basses

Nettoyage du jeu de données

- Problème : la colonne date est inexploitable

```
06 septembre 2021 suite Ã\xa0 une expÃ©rience en septembre 2021
03 mai 2021 suite Ã\xa0 une expÃ©rience en mai 2021
21 mars 2021 suite Ã\xa0 une expÃ©rience en mars 2021
```

- Résultats : encodage et formatage puis création de 3 colonnes utilisables

day_date	day_mounth	day_year
06	septembre	2021
03	mai	2021
21	mars	2021

Nettoyage du jeu de données

- Problème : la colonne avis est inutilisable

```
"je suis globalement satisfait , sauf que vous avez un problÃ¨me avec votre site internet ,impossible de dÃ©clarer un sinistre en ligne aprÃ¨s plusieurs tentatives dÃ©claration faite par tÃ©lÃ©phone ou tout c'est trÃ¨s bien passÃ© , interlocutrice compÃ©tente et trÃ¨s agrÃ©able "
```

```
"Prix tres abordable plusieurs options s'offrent a nous comme le boitier connecter Ã\xa0 la voiture, l'option tranquiliter et zero franchise ce qui est tout a fait plaisant"
```

- Résultats : encodage puis nettoyage et traduction en anglais

```
'I am generally satisfied except that you have a problem with your website impossible to declare a claim online after several attempts declaration made by telephone or everything went very well competent and very pleasant interlocutor'
```

```
'very affordable price, several options are available to us such as the box to connect to the car, the option of tranquility and zero deductible, which is quite pleasant'
```

Nettoyage du jeu de données

- Problème : les valeurs catégoriques de produit et assureur n'étaient pas au même format et pouvait ne pas correspondre.

```
['NÃ©oliane SantÃ©',
 'SantÃ©Vet',
 'IntÃ©riale',
 'GÃ©nÃ©ration',
 'CrÃ©dit Mutuel',
 'Hiscox']
```

Assureur dans le train

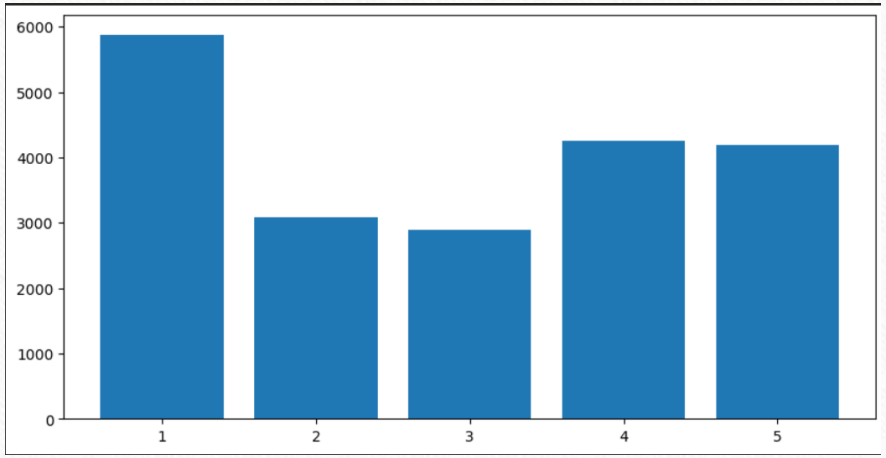
```
['Néoliane Santé', 'SantéVet', 'Intériale', 'Génération', 'Crédit Mutuel']
```

Assureur dans le test

- Résultats : l'ensemble de ces variables a été encodé, écrit en minuscule.

Visualisation des données

Pour la visualisation, nous avons choisi plusieurs types de graphiques. Nous vous en présentons ici 3 mais vous en trouverez bien plus dans notre notebook.



Nombre de commentaires par note donnée

- Nombre de commentaires par note

Grâce au graphique suivant, on voit par exemple que les avis concernent majoritairement des avis très négatifs (1 sur 5), puis des avis très positifs (4 et 5/5). Peu de gens mettent des commentaires si leur avis est mitigé.

Visualisation des données

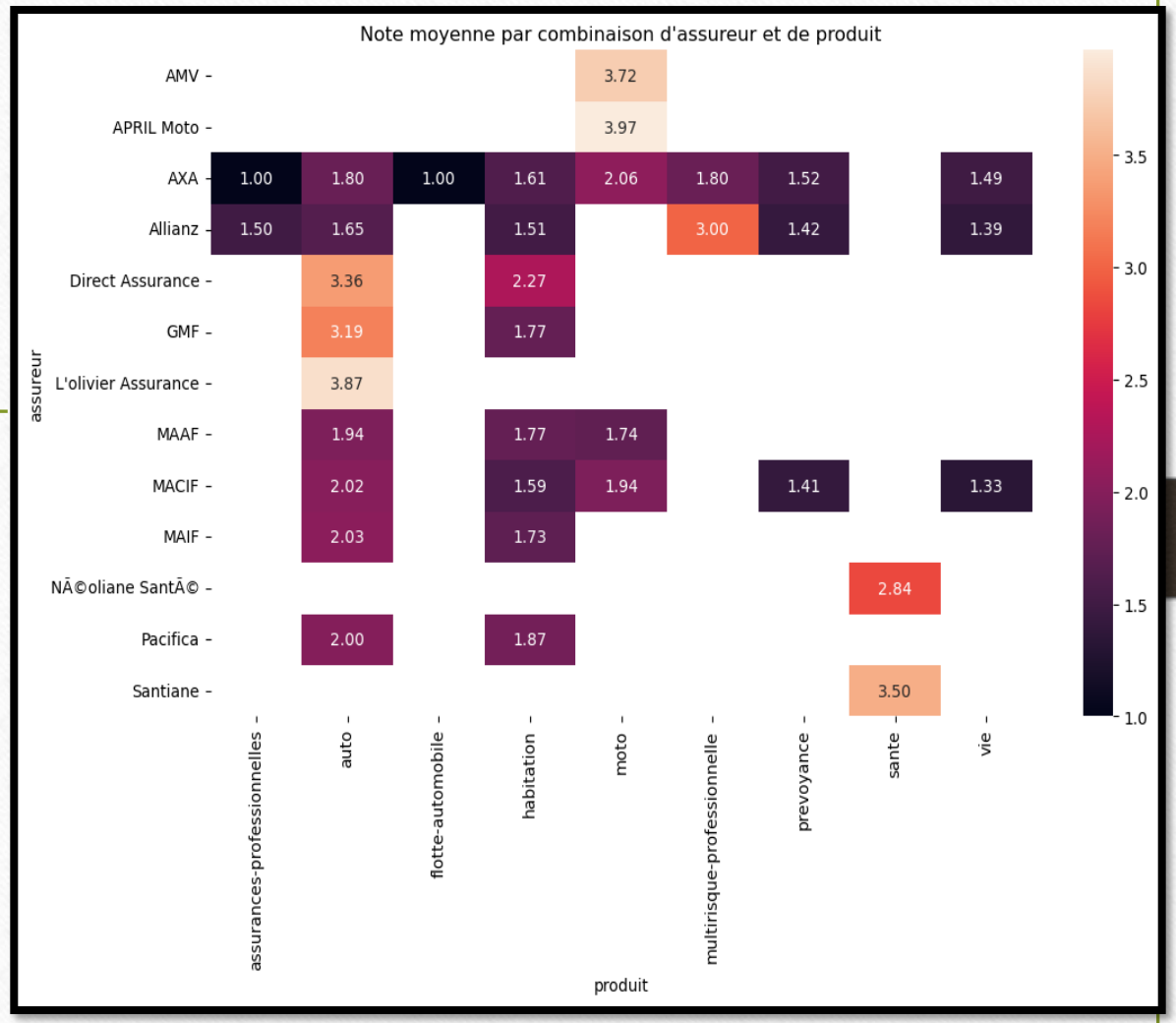
- Wordcloud pour visualiser la fréquence des termes utilisés dans les avis

Certains mots sont neutres (comme assurance, snaté, réponse…) et ne permettent pas de conclure. Mais d'autres qui reviennent fréquemment ne sont pas anodins comme "fraude", "attente" ou "problème" qui sont des mots négativement connotés.
On remarque également qu'il n'y a pas de mot "positivement" connoté.

Visualisation des données

La fonction suivante affiche une carte thermique indiquant le nombre de notations pour chaque combinaison d'assureur et de produit. Ici, seuls les assureurs ayant plus de 300 avis sont affichés.

On remarque que certains assureurs sont plus ou moins bien noté que les autres. Par exemple, si l'on vient à étudier les avis négatifs, **AXA sera l'assureur le plus concerné par les avis négatifs du dataset.**



Vectorisation des avis

note	token_avis_en	vecteur_moyen	vecteur_0	vecteur_1	vecteur_2	vecteur_3	vecteur_4	vecteur_5	vecteur_6	vecteur_7	vecteur_8	vecteur_9
5	[insurance, speed, company, attractive, qualit...	[-0.024839262, 0.015172017, 0.0071390495, -0.0...	-0.024839	0.015172	0.007139	-0.017712	-0.016423	-0.027853	0.047951	-0.023569	-0.019401	0.005957
4	[satisfied, service, very, quick, response, se...	[-0.022542713, -0.019362811, -0.007623002, 0.0...	-0.022543	-0.019363	-0.007623	0.004122	-0.019122	-0.017050	0.015299	-0.016616	0.024054	0.015240
1	[generali, life, insurance, patrimony, very, b...	[0.008720321, -0.018020166, 0.011874797, 0.001...	0.008720	-0.018020	0.011875	0.001687	-0.013880	0.006640	0.017586	0.003373	0.002346	0.026025

Pour chaque avis, les mots de l'avis sont transformés en vecteur afin de calculer le vecteur moyen.

Ce vecteur moyen de taille 10 est ensuite décomposé en 10 colonnes.

Cette décomposition permet de à nos algorithmes de prendre en compte l'avis utilisateur qui était inutilisable sous forme de string.

Apprentissage supervisé: notre RMSE

- Utilisation de **Pycaret** une librairie d'entrainement et de comparaison des performances d' algorithmes de régression.

- **Nous avons sélectionné les 5 meilleures algorithmes** de **régressions** selon la métrique RMSE pour construire nos modèles :
  1. Light Gradien Boosting Machine
  2. Random forest Regressor
  3. Extra Trees regressor
  4. Gradient boosting regressor
  5. Linear regression

→ **Notre meilleure RMSE est 1.0950 avec le modèle Light Gradient Boosting**

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
gbr	Gradient Boosting Regressor	0.8797	1.2197	1.1037	0.4217	0.3043	0.4259	0.1820
lr	Linear Regression	0.8812	1.2333	1.1099	0.4153	0.3059	0.4139	0.4360
ridge	Ridge Regression	0.8839	1.2357	1.1109	0.4141	0.3054	0.4167	0.0140
br	Bayesian Ridge	0.8859	1.2397	1.1127	0.4123	0.3058	0.4179	0.0220
rf	Random Forest Regressor	0.8850	1.2525	1.1182	0.4061	0.3067	0.4237	0.4640
lightgbm	Light Gradient Boosting Machine	0.8812	1.2577	1.1204	0.4039	0.3072	0.4177	0.0370
huber	Huber Regressor	0.8855	1.3005	1.1393	0.3835	0.3146	0.4176	0.1070
et	Extra Trees Regressor	0.8941	1.3397	1.1565	0.3645	0.3156	0.4146	0.2870
omp	Orthogonal Matching Pursuit	0.9540	1.3879	1.1776	0.3418	0.3222	0.4589	0.2080
ada	AdaBoost Regressor	1.0115	1.4806	1.2164	0.2978	0.3352	0.4962	0.0470
knn	K Neighbors Regressor	1.0330	1.6948	1.3011	0.1961	0.3600	0.5408	0.0310
lasso	Lasso Regression	1.2705	2.1137	1.4537	-0.0025	0.4016	0.6603	0.2120
en	Elastic Net	1.2705	2.1137	1.4537	-0.0025	0.4016	0.6603	0.0150
llar	Lasso Least Angle Regression	1.2705	2.1137	1.4537	-0.0025	0.4016	0.6603	0.4150
dummy	Dummy Regressor	1.2705	2.1137	1.4537	-0.0025	0.4016	0.6603	0.0140

Apprentissage supervisé: estimer la notation

Résultat RMSE sur un validation set du train set.

GradientBoostingRegressor : 1.1037
LinearRegression : 1.1099
Ridge : 1.1109
BayesianRidge :1.1127
RandomForestRegressor : 1.1182

Résultat RMSE sur test set

```
{'GradientBoostingRegressor': 1.1169071346326371,
 'LinearRegression': 1.1161385470224867,
 'Ridge': 1.0976210594221891,
 'BayesianRidge': 1.0976868217132079,
 'RandomForestRegressor': 1.1292005712880544}
```

Apprentissage non-supervisé : création de cluster

Nous avons créé un modèle non supervisé pour mieux comprendre les avis, et créé des segmentations que nous avons pu interpréter. En effet, nous pourrions **expliquer quels sujets sont présents dans l'ensemble de données,** former un modèle d'intégration de mots et analyser des mots similaires et des analogies de mots.

- 1ère étape : création d'un dataset OneHot

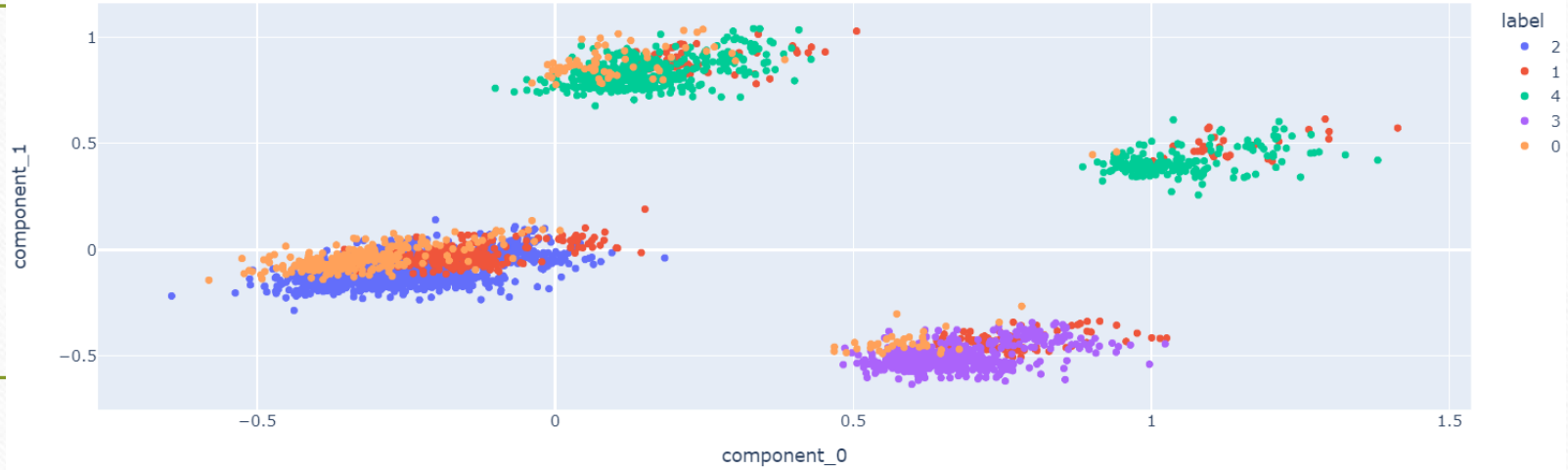
note	avis_en	abandon	abber	aberr	abil	abl	abnorm	abomin	abreast	...	wrote	yamaha	yassin	year	yearsi	yearspo	yesterday	youdriv	youness	young
5	insurance speed company attractive quality spe...	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Grâce à ce dataset, nous pouvons faire une PCA et créer des cluster en 2D.

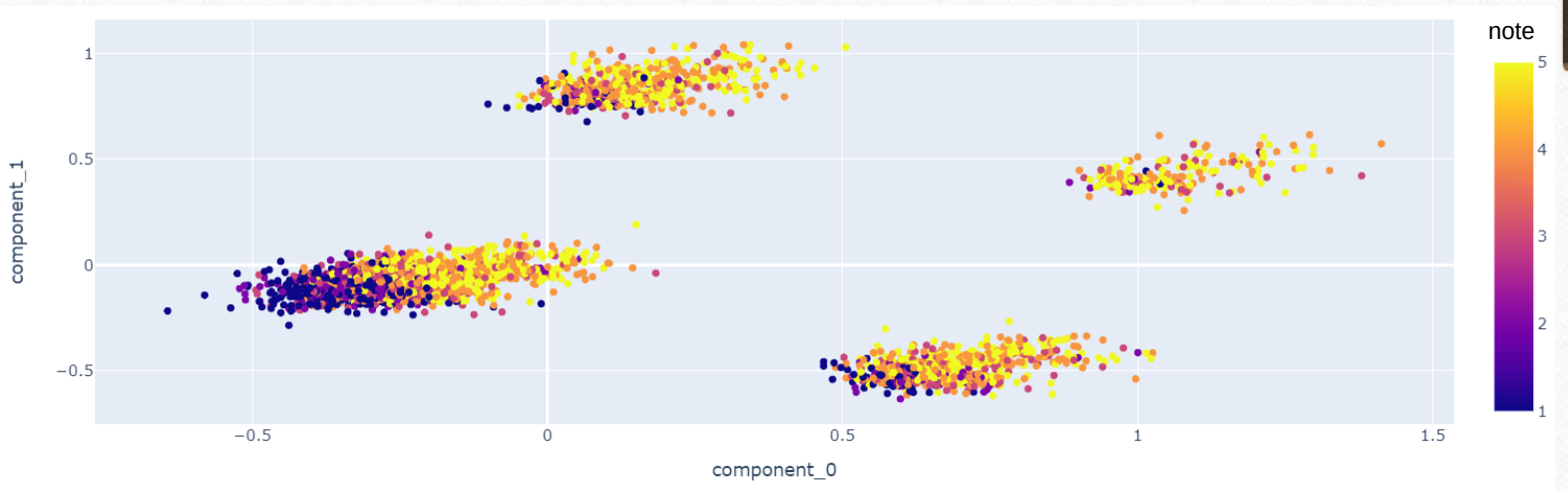
Apprentissage non-supervisé : création de cluster

Nous avons utilisé la function Kmeans pour créer les clusters.

Afin de situer où sont les notes, le 2ème graphe est colorié en function des notes. On remarque que les mauvaises notes sont majoritairement regroupées au même endroit.



Clusters crées à l'aide de la function KMeans

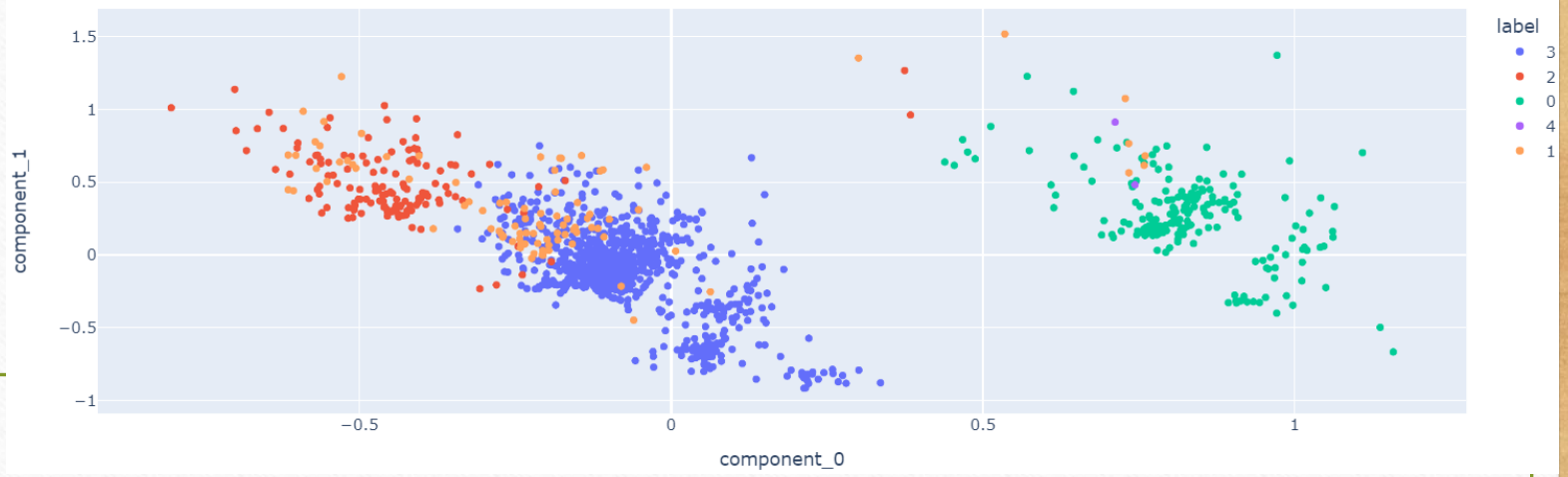


Clusters crées à l'aide des notes des commentaires

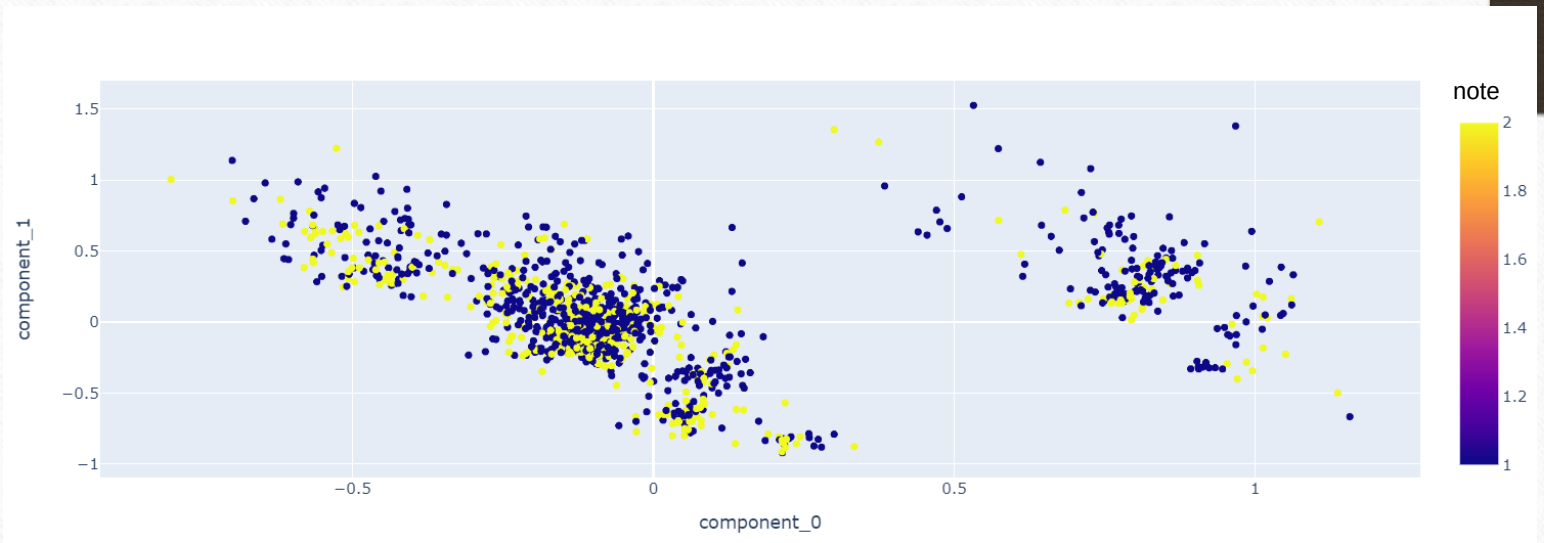
Apprentissage non-supervisé : les mauvais avis

Dans le premier graphe, on voit clairement deux clusters se detacher mais en coloriant en function des notes, les avis 1/5 et 2/5 sont mélangés.

On peut donc considérer que les personnes ayant mis des avis 1/5 et 2/5 sont liés aux mêmes plaintes et on peut donc les assimiler.



Clusters crées à l'aide de la function Kmeans avec les notes 1 et 2/5



Clusters crées à l'aide des notes des commentaires 1/5 et 2/5

Apprentissage non-supervisé : analyse des mauvais avis

Notre problématique est axée sur les avis négatifs.

- Mots revenant le plus dans les commentaires notés 1/5 et 2/5

Cluster 0: [('nan', 153), ('not', 32), ('insurance', 24), ('service', 17), ('very', 16), ('customer', 15), ('year', 9), ('increase', 9), ('price', 9), ('was', 9)]

Cluster 1: [('nan', 74), ('not', 7), ('service', 6), ('price', 4), ('customer', 3), ('impossible', 3), ('phone', 3), ('never', 3), ('still', 3), ('contract', 3)]

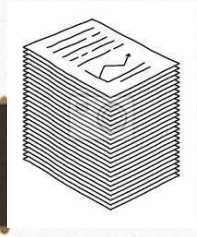
Cluster 2: [('nan', 92), ('not', 32), ('very', 16), ('service', 11), ('contract', 11), ('insurance', 10), ('termination', 8), ('customer', 7), ('was', 7), ('increase', 7)]

Cluster 3: [('nan', 865), ('not', 192), ('insurance', 145), ('very', 63), ('contract', 60), ('customer', 57), ('service', 53), ('flee', 44), ('be', 42), ('still', 38)]

Etudions le vocabulaire mis dans les commentaires 1/5 et 2/5 :

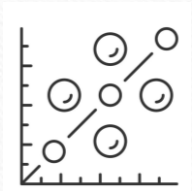
- Grande majorité de mauvais notes sans commentaires (“nan”) : l’analyse des commentaires négatifs est assez inexacte
- Beaucoup de termes négatifs “not”, “impossible”, never”
- Répétition des mêmes mots : “increase” et “price”. **On comprend que la raison principale des mauvais commentaires sont liés à l’augmentation des prix des assurance au fil des années.**

Interprétation du jeu de données



- Analyse Supervised learning :

Pour vectoriser nous avons entrainer un modèle Gensim, ce dernier ne pouvait vectoriser les mots qu'il ne connaissait pas. Nous avons dû supprimer beaucoup de mots présents dans le test et absent du train. Cette suppression de données nécessaire pour faire tourner l'algorithme a entrainer une perte d'information conséquente. Malgré cette perte nous avons conservé de très bons résultats.



- Analyse Unsupervised learning :

En étudiant les clusters formés par le vocabulaire utilisé dans les mauvais avis, on se rend compte que le principal reproche fait envers les assurances est **l'augmentation des prix de l'assurance**. Il est bien de précisé que la majorité de ces mauvais avis concernent majoritairement **l'assureur AXA**.