**Rapport - Innovation**

**Introduction**

Dans le cadre de l’UE “Projet/Challenge” de la formation Statistique et Informatique Décisionnelle (SID) de l’Université Paul Sabatier de Toulouse, Berger Levrault, une grande entreprise d’éditeur de logiciels international et multisectoriel, nous a confié leur projet Hélios qui est de mettre un système automatique de collecte d’agrégation et de présentation de contenus des actualités de l’innovation de leurs domaines d’activités stratégiques et ceux de leurs clients aussi.

Notre groupe s’est concentré sur la classification d’articles. L’objectif est de, parmi les articles collectées :

* Distinguer ceux qui sont innovants ou non et ceux qui font référence à la gamme gestion ou non.
* Regrouper les articles par thèmes
* Identifier les doublons

J’étais dans le sous-groupe Innovation. Mon travail était de classifier les articles par des méthodes supervisées.

Dans ce rapport je vais exposer les 3 méthodes utilisées et pour chacune, je présenterai la méthodologie et les résultats obtenus.

**Méthode 1 - V0:**

**Méthodologie**

* Données

On a dû utiliser deux (2) données. Un dataframe Data.json contenant l'ensemble des données brut nettoyées.

format de sortie

* 'art\_id': id de l'article choisit de manière arbitraire permet de se repérer entre les différentes représentations
* 'art\_content': contenu nettoyé de l'article (il y a toujours la ponctuation)
* 'art\_content\_html': contenu non nettoyé de l'article avec les balises html
* 'art\_extract\_datetime': date de publication de l'article peut avoir des formats différents selon le site
* 'art\_lang': langue, fr
* 'art\_title': titre de l'article
* 'art\_url': lien vers l'article
* 'src\_name': nom de la source par exemple Direction générale des Finances publiques
* 'src\_type': xpath\_source ou rss\_source
* 'src\_url': racine du site d'ou provient l'article
* 'src\_img': lien vers l'image de présentation de l'article (pas toujours présent)
* 'art\_auth' : nom(s) de l'auteur de l'article (rarement présent)
* 'art\_tag': tags de l'articel (pas toujours présent)

On a eu aussi une liste de mots clés faisant référence à la notion d’innovation.

* Modèles

Les modèles utilisées pour l’évaluation de la classification sont la régression logistique, les forêts aléatoires et le k-plus proche voisin

* Démarche, méthode

La démarche est la suivante, d’abord il y a une phase pré-processing qui consiste à nettoyer ‘art\_content’, ‘art\_title’ et aussi la liste des mots clés, c'est -à -dire enlever les mots vides, les ponctuations et lemmatiser les mots restants; ajouter des nouvelles features :

* ‘count\_words\_content’ et ‘count\_words\_tilte’: le nombre de mots que contiennent respectivement ‘art\_content’ et ‘art\_title’ nettoyés
* ’count\_keywords\_content’ et ‘count\_keywords\_title’ : respectivement le nombre de mots clés que contiennent ‘art\_content’ et ‘art\_title’ nettoyé.
* ‘count\_sentence’ : le nombre de phrase de ‘art\_content’
* ‘ratio\_keywords\_content’ et ‘ratio\_keywords\_title’ : respectivement le rapport ‘’count\_keywords\_content’ sur ‘count\_words\_content’ et ‘count\_keywords\_title’ sur ‘count\_words\_tilte’
* ‘average\_keywords\_sentence’ : le nombre de mots clés moyen que contient une phrase.

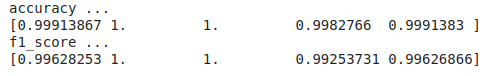
Sommer les cinq (5) features (’count\_keywords\_content’, ‘count\_keywords\_title’, ‘ratio\_keywords\_content’, ‘count\_keywords\_title’, ‘average\_keywords\_sentence’) après normalisation.

Ensuite une phase de seuillage, on a essayé de trouver un seuil où on considère toutes les articles qui sont inférieurs ou égal à ce seuil sont non innovants et les autres innovants. Pour trouver ce seuil optimum, pour chaque modèle fallait faire varié le seuil sur un intervalle et calculer le f1 score correspond puis on garde le seuil qui maximise le f1 score et le modèle associé.

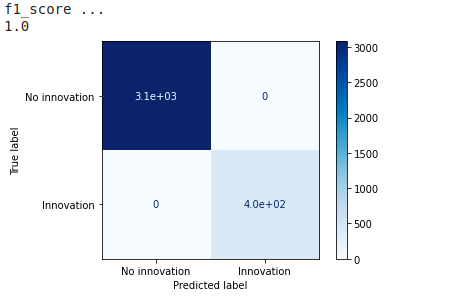
Puis une phase d’attribution de poids plus importants aux cinq (5) features citées ci dessus. Pour cela, on supprime une feature puis, on calcule le f1 score, s’il y a une importante baisse du score, on saura que cette feature est indispensable. On récupère ses features puis on les multiplie par différents coefficients et sur différents seuils on cherche le coefficient et seuil qui maximise le f1 score.

Et enfin une phase d’évaluation, par validation croisée, on évalue notre classification.

**Résultats**

****

*Les résultats par validations croisées*

****

*La matrice de confusion*

On constate que le f1 score est égal ou sensiblement égal à 1. Donc on a un bon classifieur. Par cette méthode, on a obtenu 1344 articles innovants et 10264 non innovants.

**Méthode 2 - V1:**

**Méthodologie**

* Données

Les mêmes données que V0

* Modèles

Les memes modeles que V0

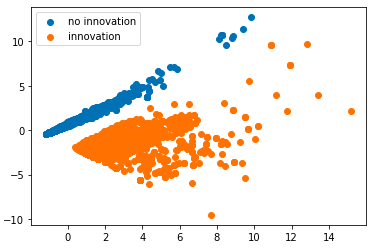
* Démarche, méthode

La démarche ici, d’abord pour la phase pré-processing est la même que V0. Ensuite on a une phase de réduction de dimension avec l’analyse des composantes principales sur ses cinq (5) features (’count\_keywords\_content’, ‘count\_keywords\_title’, ‘ratio\_keywords\_content’, ‘count\_keywords\_title’, ‘average\_keywords\_sentence’) après normalisation. Sur les trois (3) premières axes où on a eu une inertie expliquée de 90%. Sur les deux (2) premières axes, on remarque deux groupes compacts et quelques points dispersés. Étant sûr que les articles dont leurs titres et contenus ne contenaient aucun mots clés, étaient des articles non innovations. On a trouvé 100% de ses articles appartenaient à un des groupes. Par là, on en a déduit l’autre groupe.

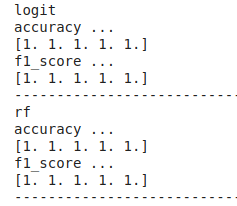
Puis une phase prédiction avec les points dispersés. Pour cela, on a utilisé les modèles de régression logistique et de random forest. Chacun des ses deux modèles ont prédit les mêmes valeurs. A partir de là, il fallait s’assurer que la classification était bien faite.

Enfin la phase évaluation, on fait une validation croisée sur l’ensemble des données avec les deux mêmes modèles.

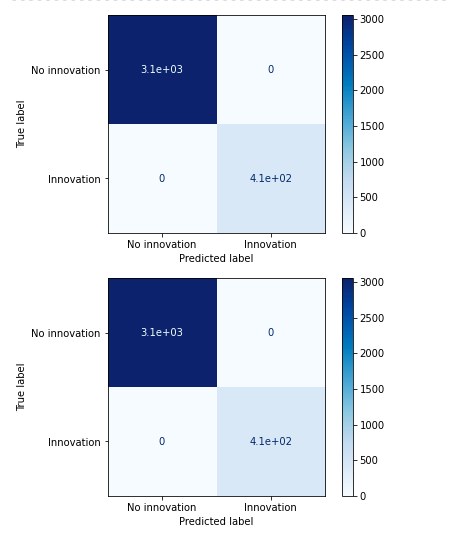
**Résultats**

****

*Nuage de point des articles sur les 2 premières axes principales*



*Résultats par validation croisées rf( random forest), logit (regression logistic)*

**

*La matrice de confusion pour chaque model*

Chacun des modèles et pour chaque partition, cinq au total, a renvoyé un f1 score égal à un (= 1) ce qui prouve la pertinence de la classification voire même parfaite.

Par cette méthode, on a obtenu 1474 articles innovants et 10134 non innovants.

Les labels ainsi obtenus peuvent servir de benchmark pour les nouvelles prédictions d’articles innovants ou non.

**Méthode 3 - V2:**

**Méthodologie**

* Données

Les mêmes données que V0

* Modèles

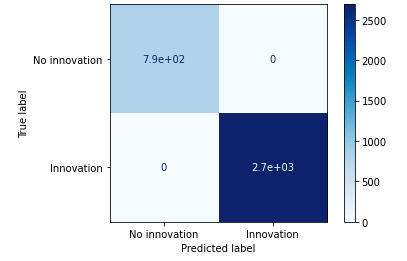
Les mêmes modèles que V0.

* Démarche, méthode

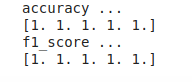
La même démarche que V0 a été utilisé à la différence que la somme de 5 features dite à la présentation du V0 a été étendue sur l’intervalle 0 et 1.

**Résultats**

Cette méthode a amélioré le seuillage et le f1 score du modèle est égal à un (=1) donc la labellisation est correcte.



*La matrice de confusion*



*Résultats par validation croisées par random forest*

Par cette méthode, on a obtenu 9004 articles innovants et 2604 non innovants. Cette méthode semble plus juste du fait que les articles collectées doivent avoir plus d’article innovant importante si on tient compte du processus du scrapping.

Les labels ainsi obtenus peuvent servir de benchmark pour les nouvelles prédictions d’articles innovants ou non.

**Conclusion**

Nous constatons que les trois (3) méthodes ne prédisent pas les mêmes nombres d’ articles innovants ou non. Aussi les deux (2) dernières ont un f1 score égal à un (=1) et la première sensiblement égal à un (~1). Si on tient compte les méthodes de collectes des données. La V2 est plus correcte.

Pour conclure, j’ai effectué deux semaines de projets très intéressant. J’ai pu approfondir mes connaissances et me suis confronté aux difficultés que cela engendre ( travail distanciel par exemple ).

Les principaux problèmes que nous avons rencontrés, d’abord les biais : le biais lié au peu données, le biais lié aux thématiques des articles collectées; ensuite les features utilisées pour l’apprentissage sont fortement corrélées et positivement d’ailleurs, l’hypothèse selon laquelle les variables doivent être indépendants et identiquement distribués n’étant pas respectés, ceux qui peut donner l'illusion qu’on a de très bon classifieur.

La perspective qui en découle est de tenter des domaines d’application moins standards que les textes: les images par exemple. Normalement les articles en contiennent.