Introduction:

Les données traitées dans ce rapport sont obtenues en se basant sur le choix et préférences alimentaires des étudiants. Cet ensemble de données comprend des informations sur les choix alimentaires, la nutrition et les préférences des enfants et d'autres informations provenant des étudiants de l'Université de Mercyhurst à Pennsylvania. Il y a 126 réponses d'étudiants. Les données sont brutes et non nettoyées.

Questions à traitées :

Quelle est l'importance de l'information nutritionnelle pour les étudiants d'aujourd'hui ? Leur goût en matière d'alimentation est-il défini par leurs préférences alimentaires lorsqu'ils étaient enfants ? Les enfants de parents qui cuisinent sont-ils plus susceptibles de faire de meilleurs choix alimentaires que les autres ? Ces enfants sont-ils susceptibles d'avoir un goût différent de celui des autres ?

Cet ensemble de données comprend un certain nombre de questions ouvertes, telles que : Quel est votre aliment de réconfort préféré ? Quelle est votre cuisine préférée ? Qui pourrait bien fonctionner pour le traitement du langage naturel (NLP).

* Top 10 cuisine ?

Outils et Méthodologie :

L'objectif principal de ce travail est de mettre en pratique tout / autant que possible les connaissances acquises lors du cours de M. Vigon, en l’occurrence, traitement des données. Ainsi, la méthodologie consiste à revoir, mettre en œuvre ou réutiliser et appliquer les méthodes vues en cours à ce cas d’étude. Pour cette raison, vous risquez de voir de nombreuses comparaisons de méthodes donnant pourtant le même résultat. En outre, certaines adaptations de codes existants en langage R disponibles sur site web ‘Kaggle’ sont incluses pour consolider l'analyse de l'ensemble des données et les conclusions qui en découlent.

NLP :

* <https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide>
* <https://www.youtube.com/watch?v=xvqsFTUsOmc>
* <https://github.com/adashofdata/nlp-in-python-tutorial/blob/master/1-Data-Cleaning.ipynb>

Plan :

* Commencer par une question
* Nettoyer les données
* L'exploration des données
* Appliquer des techniques
* Résultat et conclusion

Bases de pandas : Nettoyer les données

* Supprimer la ponctuation
* Lettres minuscules
* supprimer les numéros

Introduction :

Ce projet passe par une étape nécessaire à tout projet de traitement des données : le nettoyage des données. Le nettoyage des données est une tâche longue et fastidieuse, mais très importante. Gardez à l'esprit que "des déchets entrants sont des déchets qui sortent". En introduisant des données incomplètes dans un modèle, nous obtiendrons des résultats qui n'ont aucun sens.

Plus précisément, nous allons traverser :

* Obtenir les données - dans ce cas, nous allons extraire les données d'un site web
* Nettoyage des données - nous allons passer en revue les techniques de prétraitement de texte les plus courantes
* Organiser les données - nous organiserons les données nettoyées de manière à ce qu'elles puissent être facilement introduites dans d'autres algorithmes

Le résultat de ce carnet sera des données propres et organisées dans deux formats de texte standard :

* Corpus - un recueil de textes
* Matrice des termes du document - comptage des mots sous forme de matrice

Déclaration du problème :

Pour rappel, notre objectif est de répondre à un certain nombre de questions ouvertes, telles que : Quel est votre aliment de réconfort préféré ? Quelle est votre cuisine préférée ? Qui pourrait bien fonctionner pour le traitement du langage naturel (NLP).

Obtenir les données :

‘kaggle’

Nettoyage des données :

Lorsqu'il s'agit de données numériques, le nettoyage des données consiste souvent à supprimer les valeurs nulles et les données en double, à traiter les valeurs aberrantes, etc. Pour les données textuelles, il existe des techniques courantes de nettoyage des données, également appelées techniques de prétraitement du texte.

Avec les données textuelles, ce processus de nettoyage peut se poursuivre indéfiniment. Il y a toujours une exception à chaque étape du nettoyage. Nous allons donc suivre l'approche MVP (minimum viable product). Commencer simple et itérer. Voici un ensemble de mesures que vous pouvez prendre pour nettoyer vos données. Nous allons seulement exécuter les étapes de nettoyage communes ici et le reste pourra être fait plus tard pour améliorer nos résultats.

Étapes communes de nettoyage des données sur tout le texte :

* Mettre le texte en minuscules
* Supprimer la ponctuation
* Supprimer les valeurs numériques
* Supprimer le texte qui n’a pas de sens
* Tokenize text
* Supprimer les mots vides

Plus d'étapes de nettoyage des données après la tokenisation :

* Stemming / lemmatization
* Parties du marquage des discours
* Créer des bi-grammes ou tri-grammes
* Traiter les fautes de frappe
* Et plus encore...

Statiques descriptive :

Principe du Machine Learning :

Scikitlearn :

Régression :

Classification binaire :

Decision tree :

Ensemble learning random forest :

Recurent neural networks :