Convolutional Wine Sentiment Analysis

Pierre Mével
Centrale Paris
Pierre.mevel@student.ecp.fr

Benjamin Terris
Centrale Paris
benjamin.terris@student.ecp.fr

Abstract

L'une des choses les plus importantes pour une entreprise est de prendre soin de sa clientèle. Pour cela, il faut pouvoir être à l'écoute sur de multiples canaux de communication, qui reçoivent d'énormes flux de données tous les jours. L'analyse de sentiments est un sous-domaine du NLP (Natural Language Processing) qui cherche à automatiser cette tâche. Dans ce projet, nous tentons d'exploiter les solutions existantes dans la description des vins, un domaine complexe avec un vocabulaire riche. À partir d'une review d'un vin, nous arrivons à prédire ses caractéristiques, sa provenance, son cépage, ainsi que la personne qui a écrit la review.

1. Introduction

Retrouver les caractéristiques d'un vin en ne se basant que sur les reviews que nous avons de ce vin est un problème complexe, car subjectif. Une review peut être trop concise, trop personnelle, ou erronée. Heureusement pour nous, ce dataset ne contient que des reviews écrites par des professionnelles, ce qui nous épargne au moins les erreurs dans les reviews.
Le plus souvent, les reviews ne contiennent que des éléments purement œnologiques, sur lesquels nous pouvons aisément travailler. En voici un exemple:

« Easy-going in spite of full body, this warm and generous wine is full of ripe fruit flavors and has a smooth texture. The fresh cherry and raspberry flavors are tasty and satisfying, while very light tannins keep the mouth-feel soft. »

Parfois, la review contient des informations précises sur les données que nous tentons de prédire, comme le cépage, et s'en sert pour marquer son propos. Enfin, il arrive que la review soit trompeuse, avec une faute d'orthographe sur un lieu, un cépage ou un mot important, ou qu'elle propose un cépage comme élément de comparaison :

« A wonderful wine, one of the best Petite Verdots in California, although the truth is there aren't all that many. It's sturdy in tannins and richly textured, with delicious blackberry flavors that are a little earthier and drier than Cabernet Sauvignon. »
On note ici la comparaison avec Cabernet Sauvignon ainsi que la faute sur Petite Verdot.

2. Etat de l'art

Si l'on peut trouver quelques travaux sur le vin en particulier, il s'agit généralement du travail de particuliers. Le Machine Learning n'est pas encore capable d'apprécier un bon vin ou de le trouver, ou un bon morceau de musique.

Pour remplacer la suggestion de bon vin ou de bon morceau de musique, on utilise le principe de comparaison : « vous avez écouté ce morceau, notre base de données utilisateur dit que les gens qui font de même ont aimé tels autres morceaux », pour les suggestions.

Nous savions donc que nous ne pourrions pas donner une indication de qualité à partir du dataset que nous décrirons en partie 3.

Nous avons donc choisi de travailler sur la reconnaissance d'un vin à partir de sa description, un problème pour lequel nous n'avons pas réellement trouvé d'état de l'art.

Cependant, d'un point de vue strictement technique, après avoir travaillé la donnée, il s'agit d'un problème de classification, pour lequel de nombreux algorithmes existent déjà.

3. Jeu de données

Le jeu de données provient d'un scrapping de WineEnthusiast, et peut être trouvé sur *Kaggle*. Il contient 130 000 entrées de reviews de vins, avec par exemple pour colonnes :

- Le pays d'origine
- Le cépage
- L'auteur de la review
- La province d'origine
- La varieté du vin
- Le domaine d'origine
- Le prix du vin

Avant de pouvoir prédire une caractéristique du vin à partir de la review, il nous faut travailler notre jeu de données.

Nous commençons par retirer les reviews pour lesquelles

la caractéristique en question n'a pas été spécifiée par le goûteur.

Ensuite, nous retirons toutes les classes qui n'ont pas assez de représentants. Par exemple, si nous n'avons que cinq vins du Portugal, il semble difficile de prédire qu'un vin provient de ce pays : nous n'avons pas suffisamment d'éléments de comparaison.

Une série de tests nous a permis de placer une limite qui semble donner des résultats cohérents à 150 éléments minimum pour définir une classe. Nous retirons du jeu de données les classes qui ne satisfont pas à cette règle pour une caractéristique donnée.

Cela représente environ 1/1000^{e} du jeu de données.

Un autre problème du jeu de données est la présence de virgules dans les reviews, couplée à l'utilisation de ce même symbole comme séparateur dans un csv. Nous avons donc fait attention à utiliser une librairie de chargement de fichiers adaptée à cette contrainte.

4. Convolutional Network for Sentiment Analysis

Les réseaux de convolution sont généralement utilisés pour l'analyse d'images, avec une ou plusieurs couches de convolution suivies de couches linéaires. Les couches de convolution utilisent souvent, comme vu en cours, des filtres, ou noyaux, avec sa forme (e.g, 3x3, ou autre), et produisent de nouvelles versions de l'image, processées.

Les poids associés sont ajustés par back propagation.

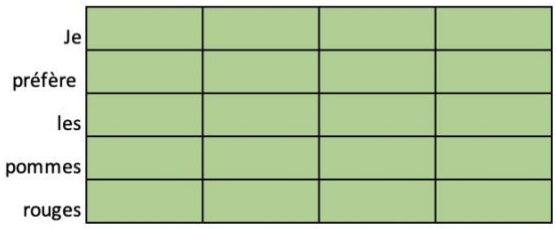
Cet ajustement permet d'apprendre quelles sont les parties les plus importantes d'une image, comme la présence d'une truffe, d'une langue rose, pour la détection d'un chien, par exemple.

L'idée dans ce projet était tout d'abord d'utiliser un filtre de taille 1*n, avec n un certain nombre de mots, pour tenter de dégager des morceaux de phrases importants, plutôt que simplement analyser le vocabulaire.

Malheureusement, le passage en une seule dimension pour du texte ne donne pas les résultats attendus.

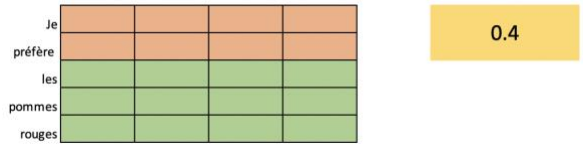
Nous allons utiliser l'approche de *John Rupert Firth* pour la représentation des mots. Au lieu d'avoir des vecteurs binaires, de type [0 0 1 0 … 0] et de taille, la taille du dictionnaire, avec le 1 signifiant le mot que le vecteur représente, nous allons créer des vecteurs de type [0.1 0.2432 ……. 0.3465334] dans une représentation où deux mots *similaires* vont idéalement avoir des vecteurs proches l'un de l'autre.

On utilise donc des filtres de taille n*embedding_dimension, avec n le nombre de mots dans notre fenêtre glissante, le filtre que nous faisons voyager par-dessus le texte, et embedding_dimension la dimension du vecteur représentatif d'un mot.

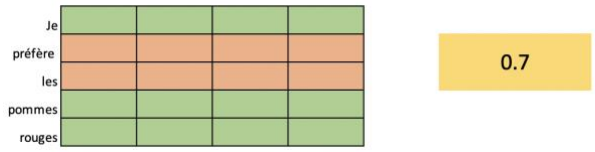


Ici on a une phrase de 5 mots, et une embedding_dimension de 4.

On peut donc faire passer un vecteur de dimension 2*4 par-dessus et obtenir un scalaire :



Et parcourir l'intégralité du texte pour obtenir ces scalaires.



Enfin, on utilise une fonction de max pooling pour prendre la valeur maximale sur une dimension (0.7 sur les deux itérations montrées si dessus). L'idée est ici de donner une importance maximale à un ensemble de deux mots consécutifs. Nous n'avons même pas à savoir de quel ensemble il s'agissait grâce à la back propagation, qui va changer les poids des filtres lorsqu'un fort indicateur de sentiments apparait. Si c'est la valeur maximale, elle va alors passer dans la fonction de max pooling et représenter le résultat final.

On travaille dans ce projet avec 100 filtres de taille 2, 3 et 4, soit 300 filtres au total, ou 300 ensembles de mots consécutifs que notre modèle pense importants.

Enfin, on les concatène en un simple vecteur, qui passe à travers la couche linéaire, pour prendre une décision finale dans la classification.

5. Résultats

Puisque nous travaillons avec une classification multi-classe, nous définissons notre précision par le pourcentage d'éléments correctement classés en utilisant le label obtenant la probabilité maximale.

En utilisant le CNN, nous arrivons, à partir d'une revue, à prédire avec 85% de précision de quel pays provient un vin. Les résultats ne changent pas en supprimant les pays avec moins de 150 vins revus, probablement car cela ne représente pas une partie assez importante du jeu de données.

Nous pouvons voir par l'exemple que rien dans les reviews n'indique cependant le pays d'origine d'un vin, ce qui rend les résultats assez impressionnants.

De nombreux cépages n'ont pas beaucoup de représentants. Sans suppression de ces classes, on obtient 60% de précision. Avec la limite des 150 représentants, on passe à 65%, et ce malgré les problèmes évoqués dans le paragraphe précédent.

6. Travail auxiliaire

Parce que c'est réputé pour ses résultats sur l'analyse de texte, nous avons également implémenté un RNN multi classes. Cependant, il n'a pas été aussi performant que notre CNN, et a été bien plus long à entrainer. Nous l'avons cependant laissé dans le notebook si un lecteur souhaite l'utiliser pour comparer.

```
Still showing its tannins, this wine is developing well. It is relatively light in texture, the sweet berry fruits
balanced with a layer of acidity.
Actual: France, predicted: France

A unique blend of fermenting orange, aging white flowers, dried apples and a musky sandalwood show on the nose of t
his wine. The palate is simultaneously rich, sour and creamy, with tangerine and banana flavors.
Actual: US, predicted: US

Sweet tobacco and overripe cherry open the nose of this thick, jammy Amarone. The wine exhibits a syrupy, bold mout
hfeel with lingering tones of smoke, beef jerky and spice on the finish.
Actual: Italy, predicted: Italy

Mineral aromas of gravel, graphite and crushed slate show on the nose of this bottling, leading into baked black pl
um and oak notes. It's a refreshing example from an appellation that tends toward richer, jammy styles. The palate
offers raspberry and dried thyme flavors, with a touch of eucalyptus.
Actual: US, predicted: US
```

Pour ce qui est de la prédiction de la province d'origine d'un vin, nous avons fait face à plusieurs difficultés. Par exemple, la province *Bordeaux* et la province *Burgundy* sont en fait une seule et même province.
On a de plus 385 classes, ce qui représente un problème de classification autrement plus difficile qu'avec 41 pays. Nous obtenons 60% de précision, et montons à 70% en appliquant le filtrage de données décrit en partie 2, pour supprimer les classes avec peu de représentants.

De manière assez inattendue, nous arrivons à prédire avec 95% de précision quel expert a écrit la review proposée. Cela peut être dû au fait que seuls 18 experts ont écrit les reviews de ce jeu de données.
Peut-être le style d'écriture finit par avoir une influence sur la prédiction

Le cépage représente un challenge plus important. Pour commencer, de nombreux cépages apparaissent sous plusieurs noms : *pinot blanc* et *pinot bianco*, des recoupements existent : *red style Bordeaux, red style Rhone,* etc…

References

[1] Word2Vec project, Google, 2013

[2] Gabriel Mordecki, how to transform text into numbers, 2017

[3] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[4] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1746–1751.

[5] Santos, C. N. dos, & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In COLING-2014 (pp. 69–78).