

Représentations textuelles et plongements sémantiques : une application pour l'analyse de sentiment/dissertation

MARIE Clément, SAMAHA Elio, XIA Tianxiang

Encadré par Mr. Olivier Schwander

June 26, 2023

1 Introduction

2 K-Plus Proches Voisins

3 Naive Bayes

4 Homologie persistente

5 Outro

6 Annexe

Introduction - Motivation

The screenshot shows the product page for 'LEGACY: Quest for a Family Treasure' on the Tric Trac website. The page includes a product image, a box set, and a detailed description. The product is a cooperative board game for 1 to 5 players, suitable for ages 14 and up, with a playtime of 240 minutes. It is published by Argyx Games. The page also features a section for user reviews, showing two reviews with ratings of 9.00 stars. The first review is by 'apret' from May 28, 2021, and the second is by 'Kach' from July 20, 2021. The product is priced at 49.90 €. The page layout includes a top navigation bar, a main content area with tabs for 'RÉSUMÉ', 'AVIS', 'ACTUALITÉS', 'FORUM', 'VIDÉOS', 'LIENS', and 'VERSIONS & EXTENSIONS', and a right sidebar with a 'Philibert le commandant' logo and a '49,90 €' price tag.

LEGACY : Quest for a Family Treasure
Coopération • Enigme
De Mathias Daval et Johanna Pernot
Édité par Argyx Games

1 à 5 JOUEURS 14 ans et + ÂGE 240 min TEMPS DE PARTIE

NOTER : ★★★★★ RÉDIGER MON AVIS

MA BIBLIOTHÈQUE : ✓ JE L'AI JE LE VEUX J'Y AI JOUÉ

RÉSUMÉ **AVIS** ACTUALITÉS FORUM VIDÉOS LIENS VERSIONS & EXTENSIONS

0 à 3 sur 3 TRIER Recommandé ▼

9,00 ★ **Escape à domicile réussit**
par **apret** 28 mai 2021
Voilà un "escape game-like" à faire à domicile vraiment bien foutu. Le matériel physique est de qualité, les ressources en ligne aussi, avec un grand soin apporté aux détails. Les énigmes sont variées, cohérentes, une seule nous a paru "briée par les cheveux". L'histoire en 2 parties permet de jouer...

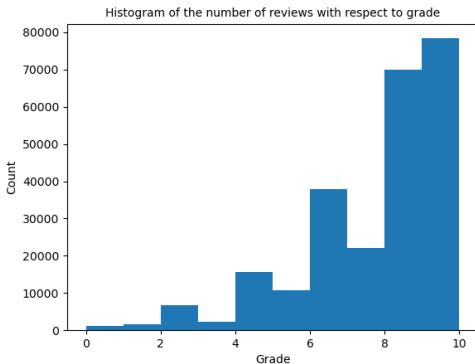
9,00 ★ **Immersion au top.**
par **Kach** 20 juillet 2021
Le scénario est vraiment bien ficelé, il y a un peu de manipulations avec le petit matériel fourni. Les

49,90 €
prix de vente conseillé

Philibert le commandant
MA BIBLIOTHÈQUE

- Site web Tric Trac: site communautaire autour des jeux de société
- Collecte d'avis et de notes des utilisateurs sur les jeux
- Traitement de données par représentation vectorielle de mots: analyse multidimensionnelle
- Apprentissage statistique (avis en paramètre et la classe comme variable prédite), algorithmes de classification
- Enjeux pour l'entreprise

Analyse descriptive des notes: anticipation du problème



- Tendance positive des notes: déséquilibre de classes
- Solution: harmoniser la représentation positive/négative
- Adapter en conséquence les métriques d'évaluation

Représentation vectorielle des documents

- Besoin d'une représentation numérique: structurée et quantifiable
- Prétraitement (ponctuation, stemming, stopwords, etc.)
- Plongement lexical (One hot encoding ou bag of words ou tf-idf, etc.)

Exemple:

J'avais peur que les extensions s'essoufflent à terme...

->

['peur', 'extens', 'essoufflent', 'term', ...]

->

[0, 0, 1.2, 1, ...]

Plongement lexical

Raw text

Le ver vert va

vers le verre vert.

(6 mots qui restent)

One hot encoding

ver	1
vert	1
rouge	0
vers	1
verre	1
...	...

Bag of words

ver	1
vert	2
rouge	0
vers	1
verre	1
...	...

tf-idf

ver	1/6x2
vert	2/6x1
rouge	0
vers	1/6x0.1
verre	1/6x1.5
...	tf x idf

tf:term frequency

df:document frequency

idf: inverse document frequency

:= -ln(df)

Avantages:

- Préserve la signification sémantique des mots dans le document.
- Considère moins les mots non-important. (tf-idf)

Inconvénients:

- Ignore la grammaire et l'ordre des mots, entraînant une perte d'informations contextuelles.
- Augmente la dimensionnalité et la parcimonie de la représentation vectorielle.

Distances: Cosinus et Euclidienne

■ Distance Euclidienne:

$$D(u, v) := \sqrt{\sum_{i=1}^n (u_i - v_i)^2}$$

■ Distance Cosinus:

$$D(u, v) := 1 - S(u, v)$$

$$\text{Où } S(u, v) := \cos(\theta) = \frac{u \cdot v}{\|u\| \|v\|}$$

C'est la similarité cosinus entre u et v

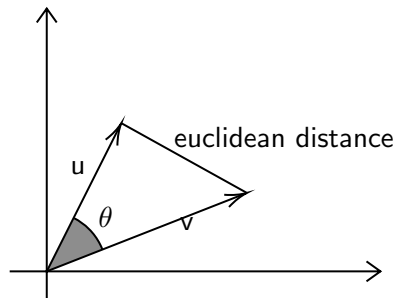


Illustration de Distance
Cosinus et Euclidienne

K-Plus Proches Voisins (KPP ou KNN)

Avantages :

- Simple et facile à implémenter.
- Méthode non paramétrique.
- Convient à la fois pour les tâches de classification et de régression.

Inconvénients :

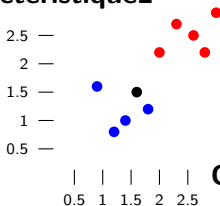
- Coûteux en calcul pour les grands ensembles de données.
- Sensible au choix de k et de la métrique de distance.
- Performances médiocres avec des données de grande dimension.

Caractéristique2

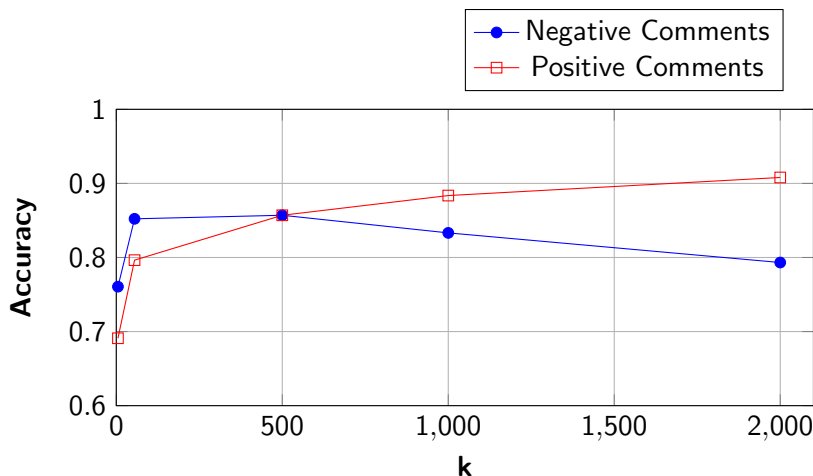
2.5 —
2 —
1.5 —
1 —
0.5 —

| | | | |
0.5 1 1.5 2 2.5

Caractéristique1

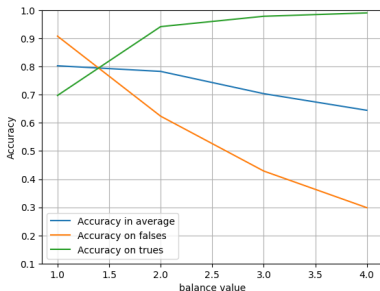
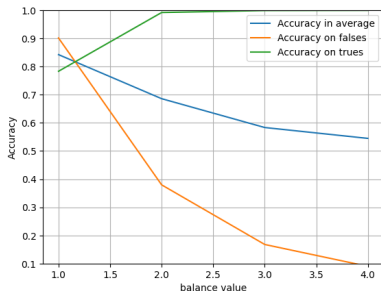


KNN Analysis (1)



k nearest neighbor on balanced data (5000 positive/negative comments)

KNN Analysis (2)

(a) $k = 50$ (b) $k = 500$

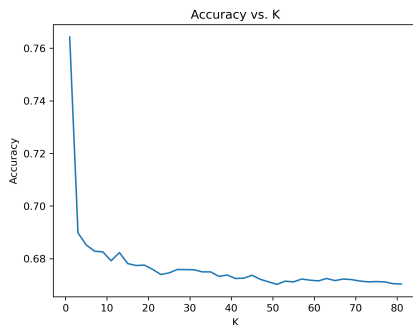
Class	Precision	Recall	F1-Score	Support
False (negative comments)	0.23	0.91	0.37	1789
True (positive comments)	0.99	0.76	0.86	22520
Accuracy = 0.77				

Classification report with undersampling (balance = 1, $k = 500$)

KNN Analysis with Different Options

	BoW	tf-idf	One-hot
Euclidean	0.7642 (1)	0.7425 (113)	0.6864 (5)
Cosine	0.7725 (1)	0.7771 (1)	0.6864 (5)

Comparative Accuracy Results, (k)



Variation of accuracy with k (KNN, euclidean distance, BoW)

Naive Bayes

- Bag of words: décompte des occurrences de mots dans le corpus (rép. vectorielle pas nécessaire)
- Algorithme de classification probabiliste
- Naïf car suppose que les variables sont indépendantes
- Probabilités conditionnelles par classe
- Classification où la classe prédite maximise la probabilité

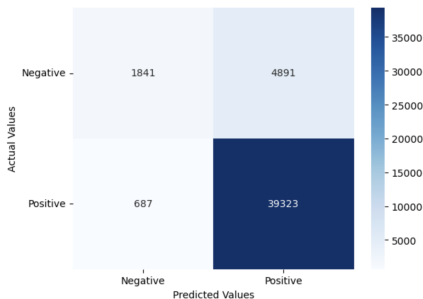
The diagram shows the Naive Bayes formula with labels for each term:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

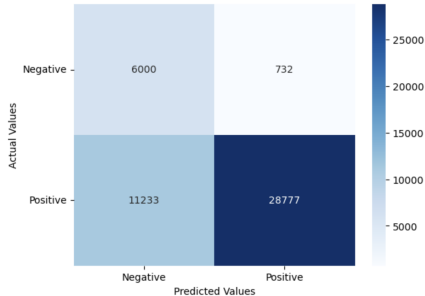
Labels and arrows:

- Likelihood of the Evidence given that the Hypothesis is True** (orange text, arrow points to $P(E|H)$)
- Prior Probability of the Hypothesis** (red text, arrow points to $P(H)$)
- Posterior Probability of the Hypothesis given that the Evidence is True** (blue text, arrow points to $P(H|E)$)
- Prior Probability that the evidence is True** (green text, arrow points to $P(E)$)

Naive Bayes - Metrics



(a) Pas d'undersampling



(b) Undersampling

Matrices de confusion et métriques associées

	F1 Pos.	F1 Neg.	Accuracy
Undersampling	0.83	0.50	0.74
No undersampling	0.93	0.40	0.88

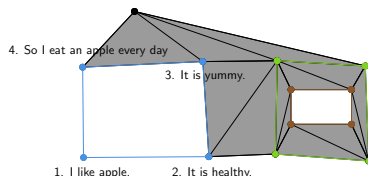
Motivation - Représentation par homologie persistente

Maintenant on veut une méthode qui

- considère la structure des phrases dans un document (qui a été ignorée);
- n'a pas de paramètre qui dépend des données d'entraînement. (plus robust)

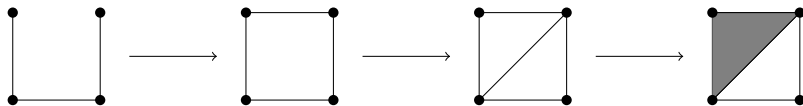
Pour la prédiction du sentiment, cela n'a pas beaucoup d'avantage par rapport à KNN ou Naive Bayes. On change à une tâche plus appropriée : analyser la richesse des structures dans une dissertation.

Introduction



...
 1. I like apple. ->
 2. It is healthy. ->
 3. It is yummy. ->
 4. So I eat an apple every day. ->
 ...

Il y a 2 trous dans ce complexe simplicial, l'un représenté par le cycle bleu et l'autre par le cycle marron.

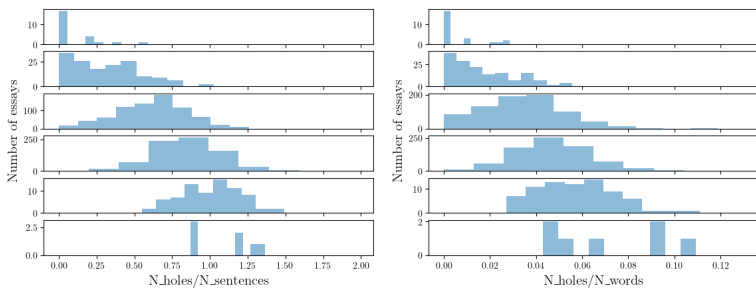


Nous continuons à relier et à remplir lorsque le seuil similarité augmente.

- On compte le nombre d'apparitions de trous comme une mesure de la richesse des structures dans une dissertation.

Analyse sur un ensemble de données réelles

source de données (plus précisément, l'ensemble d'essais 2, qui est discursif) : https://www.kaggle.com/datasets/thevirusx3/automated-essay-scoring-dataset/code?select=training_set_rel3.tsv



La moyenne des trous de chaque phrase/mot augmente avec le niveau scolaire (de 1 à 6, histogramme du haut vers le bas).

Analyse - Conclusions

Chargement des données

- Gérer un grand jeu de données
- Décisions sur le prétraitement

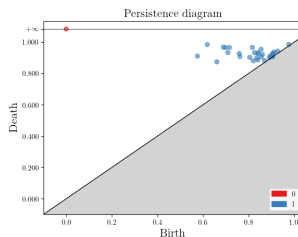
Techniques d'apprentissage

- Choix de la représentation des données
- Choix de la distance
- Choix du modèle d'apprentissage
- Importance de la gestion du déséquilibre des classes

Limitations et améliorations

- Gérer les sentiments subjectifs
- Défis de l'adaptation au domaine
- Intervention de l'homologie persistante
- Exploration de modèles de Deep Learning

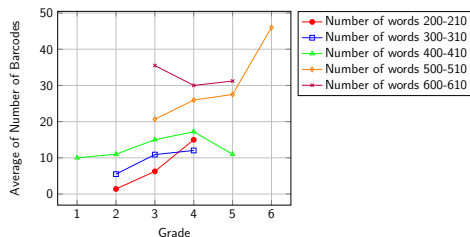
Annexe



In @DATE1's world, there are many things found offensive. Everyone has their own opinion on what is offensive and what is not. Many parents are becoming upset because they think their children are viewing things that they should not. Other people are upset because they think the libraries are offending their culture or way of life. This is even taken to the extreme where people want censorship on libraries to avoid this, which is wrong. Some people are becoming concerned about the materials in libraries...(~450 words)

Le diagramme de persistance d'un essai de niveau 4/6 : Un point bleu (x, y) dans ce diagramme signifie qu'un trou apparaît à $d = x$ et disparaît à $d = y$. qu'il y a un trou qui apparaît à $d = x$ et disparaît à $d = y$. Un point rouge signifie qu'il n'y a qu'une seule composante connectée tout le temps, parce que nous lions toutes les phrases par ordre dans la dissertation.

Annexe



Notes et moyennes
des nombres de trous
dans des plages de
nombres de mots fixes