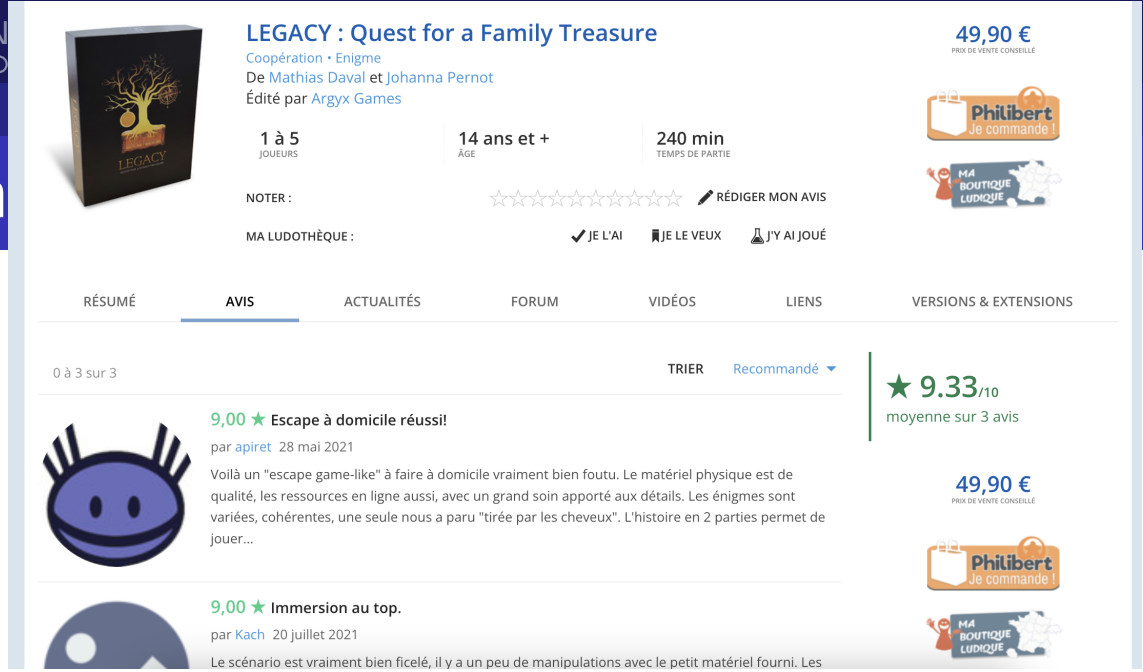
Représentations textuelles et plongements sémantiques : une application pour l'analyse de sentiment/dissertation

MARIE Clément, SAMAHA Elio, XIA Tianxiang

Encadré par Mr. Olivier Schwander

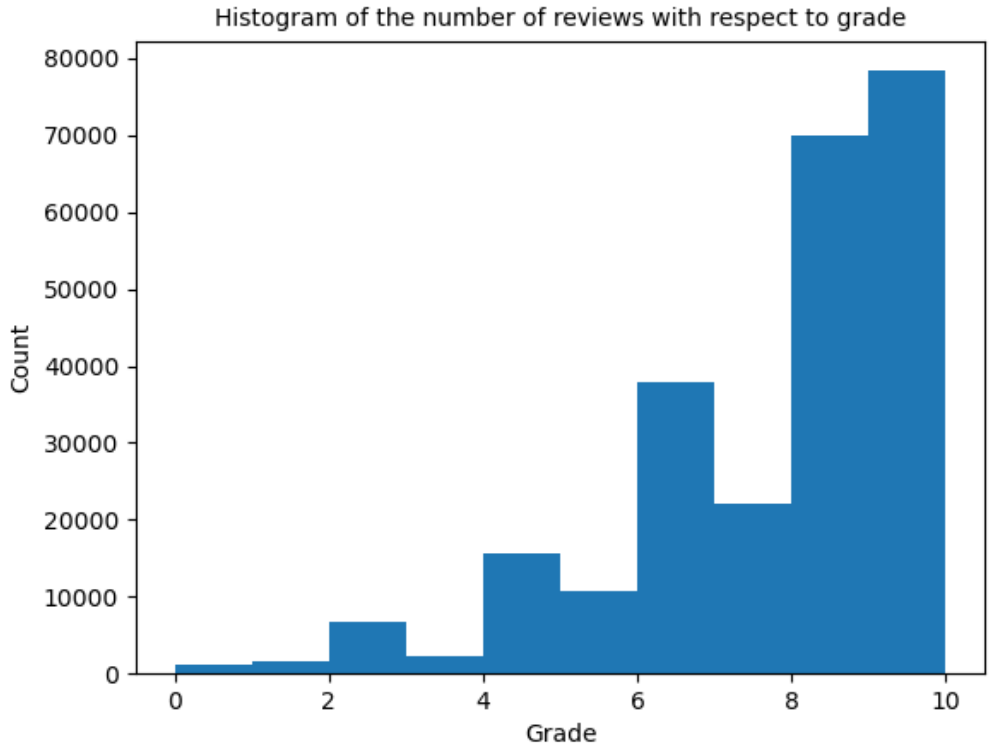
June 26, 2023

Introduction - Motivation



- 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 répresentation positive/négative
- Adapter en conséquence les métriques d'évaluation

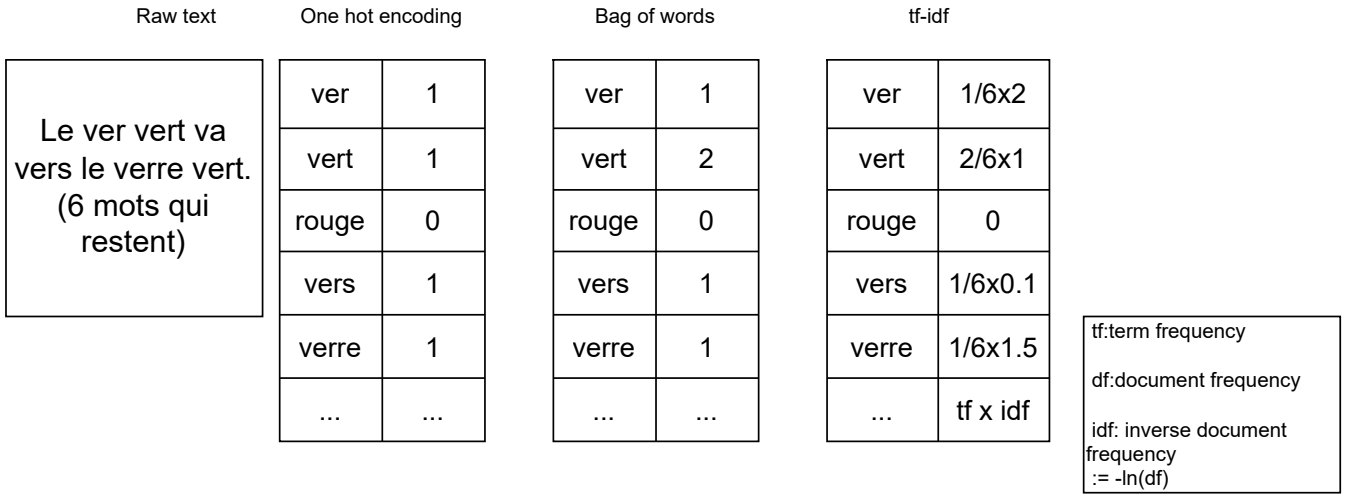
Représentation vectorielle des documents

- Besoin d'une répresentation 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'éssouflent à terme...
->
['peur', 'extens', 'éssouflent', 'term', ...]
->
[0, 0, 1.2, 1, ...]
```

Plongement lexical



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:**

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

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

  C'est la similarité cosinus entre u et v

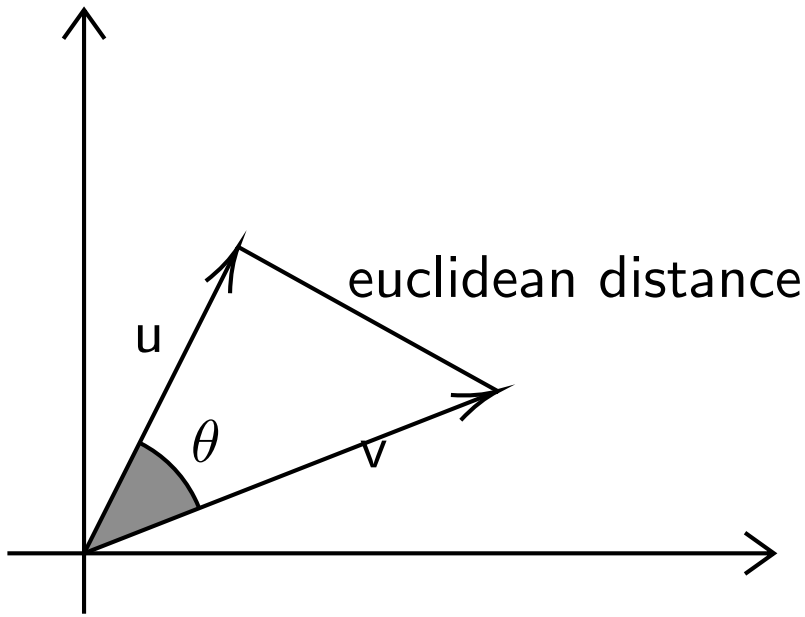


Illustration de Distance Cosinus et Euclidienne

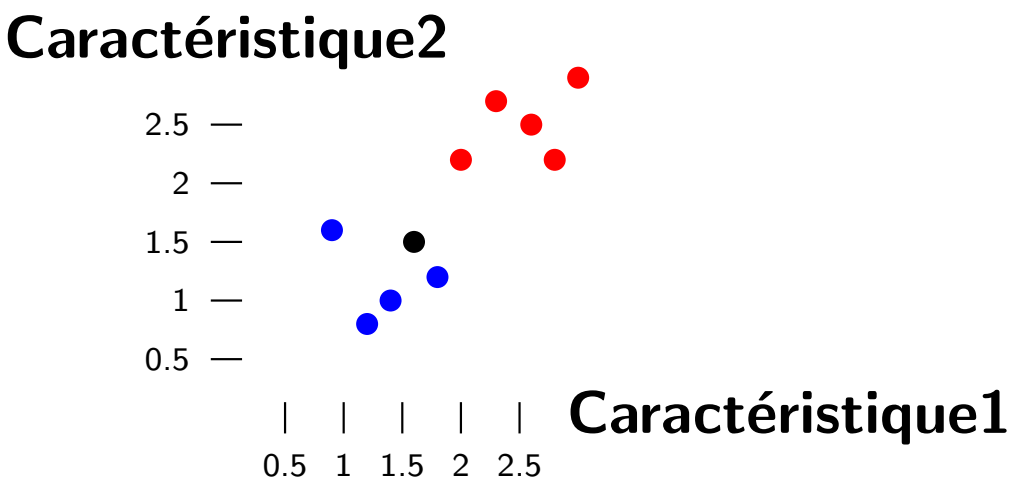
K-Plus Proches Voisins (KPP ou KNN)

Avantages :

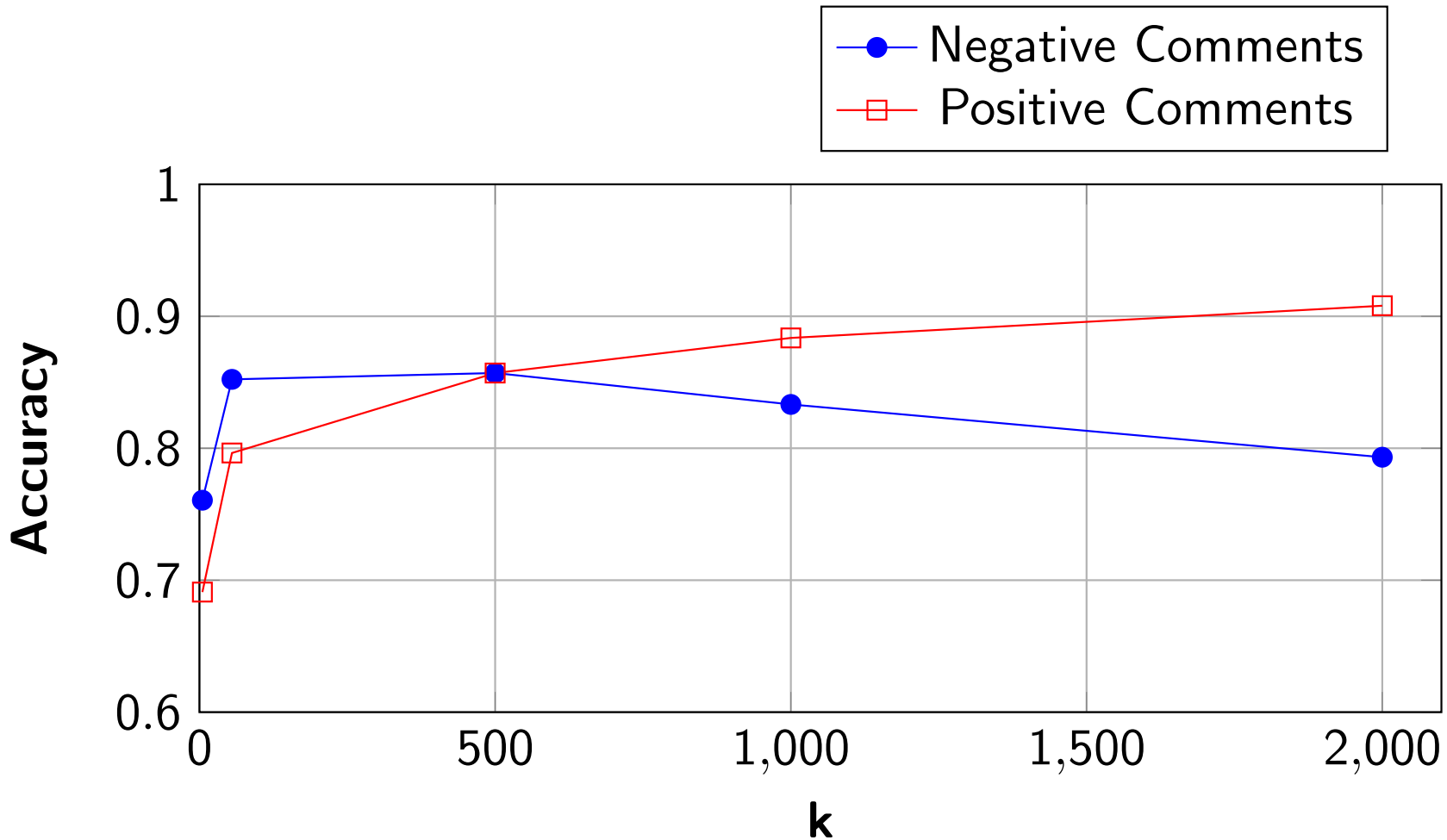
- Simple et facile à implémenter.
- Méthode non paramétrique.
- Convient à la fois pour les tâches de classification et de régression.

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.

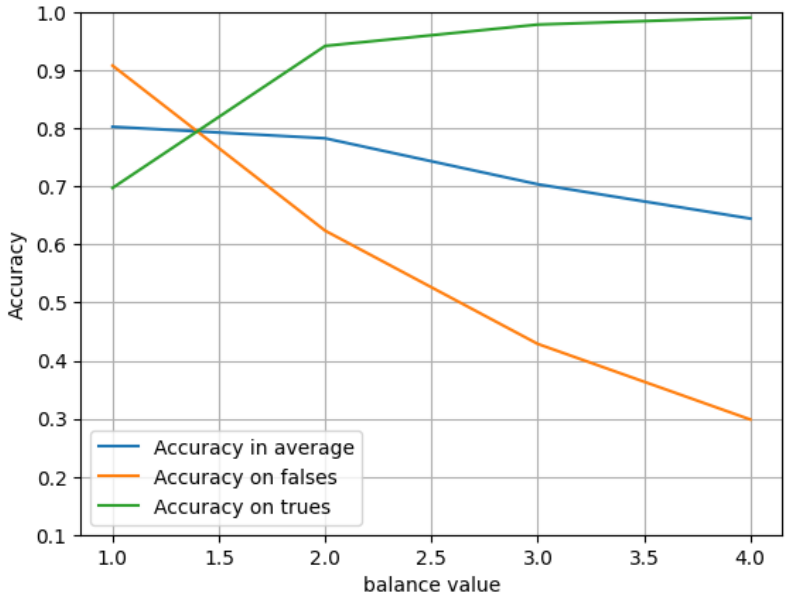


KNN Analysis (1)

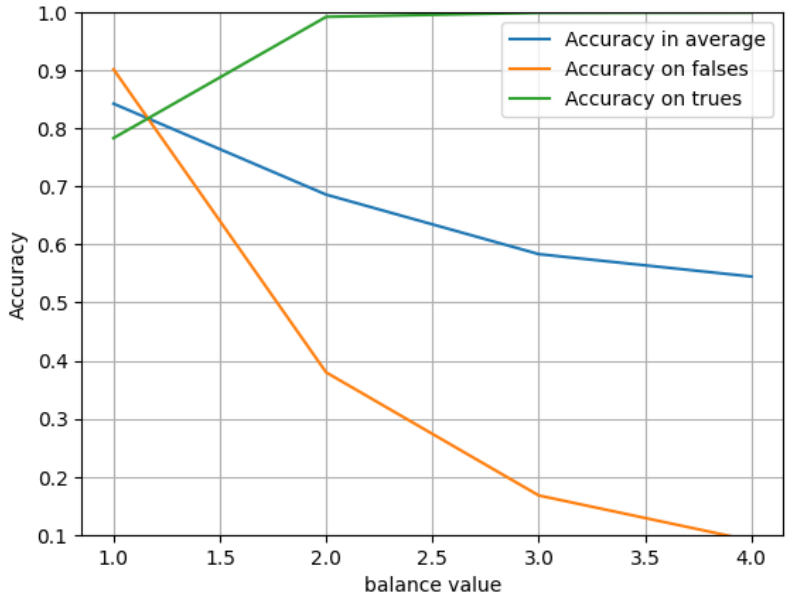


k nearest neighbor on balanced data (5000 positive/negative comments)

KNN Analysis (2)



(a) $k = 50$



(b) $k = 500$

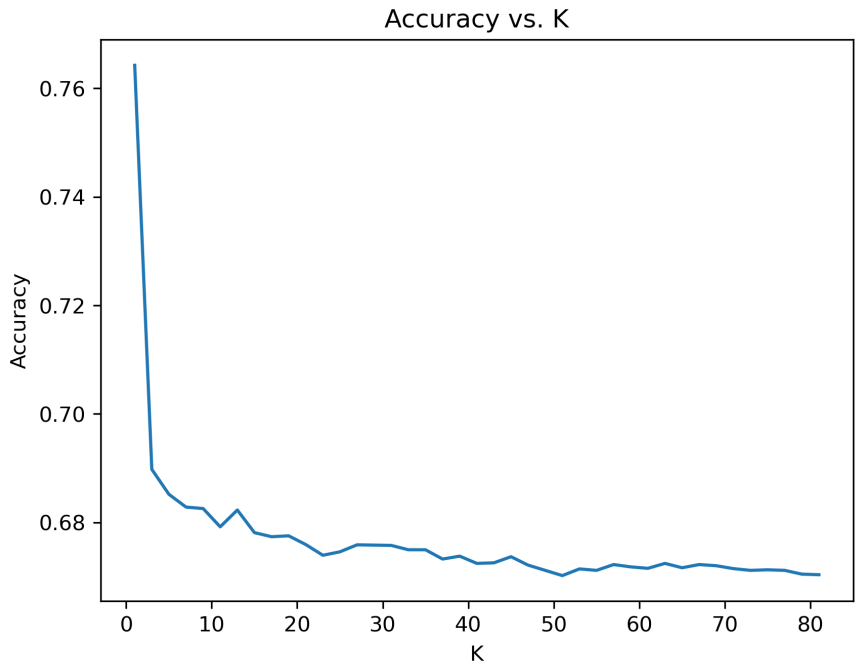
Class	Precision	Recall	F1-Score	Support
False (negative comments)	0.23	0.91	0.37	1789
True (positive comments)	0.99	0.76	0.86	22520
Accuracy = 0.77				

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)

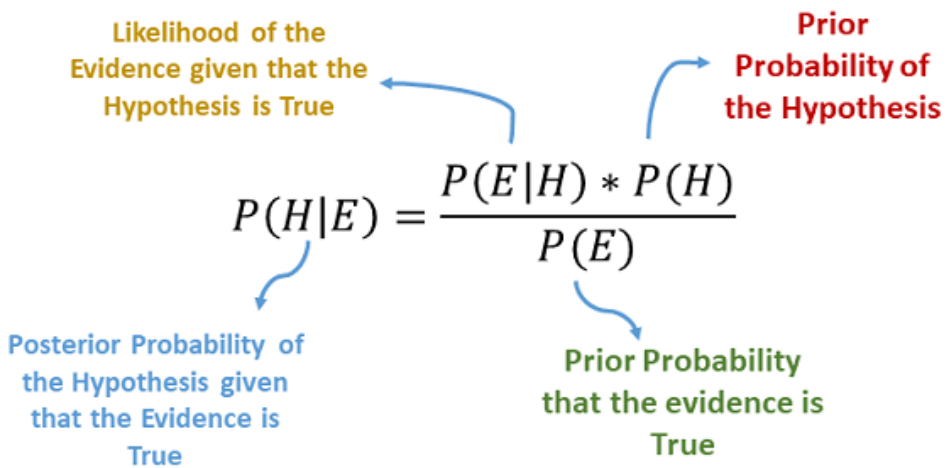
Comparative Accuracy Results, (k)



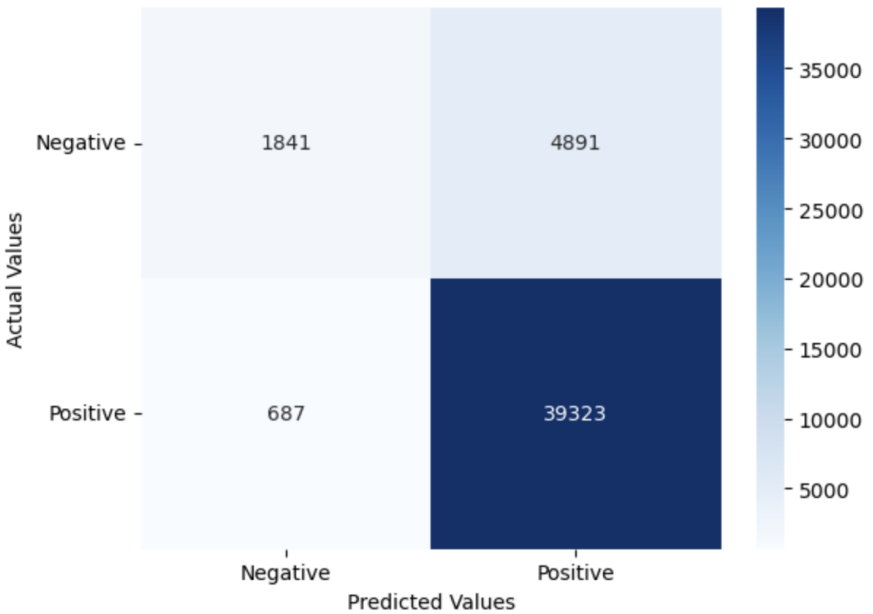
Variation of accuracy with k (KNN, euclidean distance, BoW)

Naive Bayes

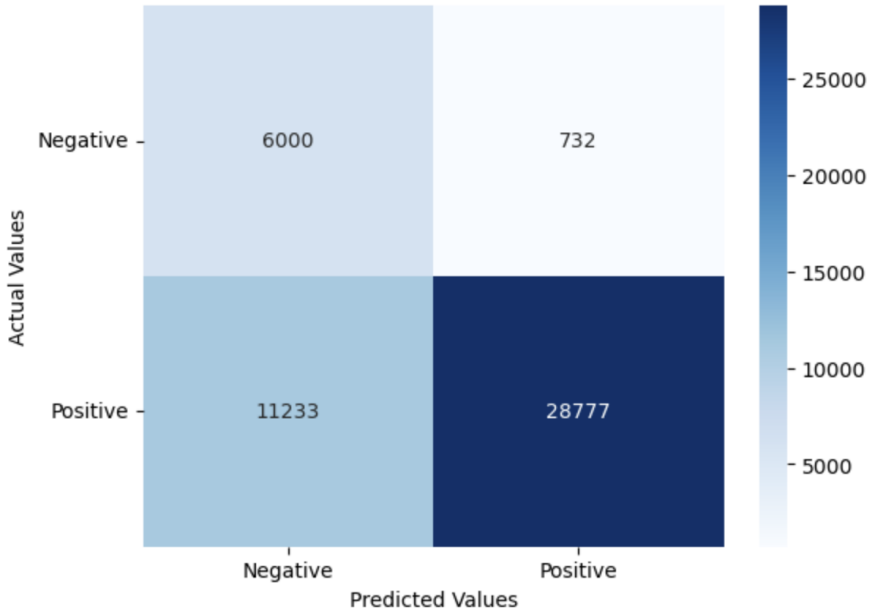
- Bag of words: décompte des occurrences de mots dans le corpus (rép. vectorielle pas nécéssaire)
- Algorithme de classification probabiliste
- Naif car suppose que les variables sont indépendantes
- Probabilités conditionnelles par classe
- Classification où la classe prédite maximise la probabilité



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

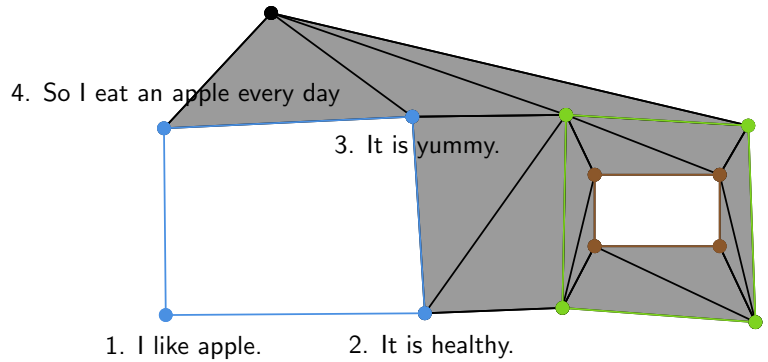
Motivation - Représentation par homologie persistente

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'entrainement. (plus robust)

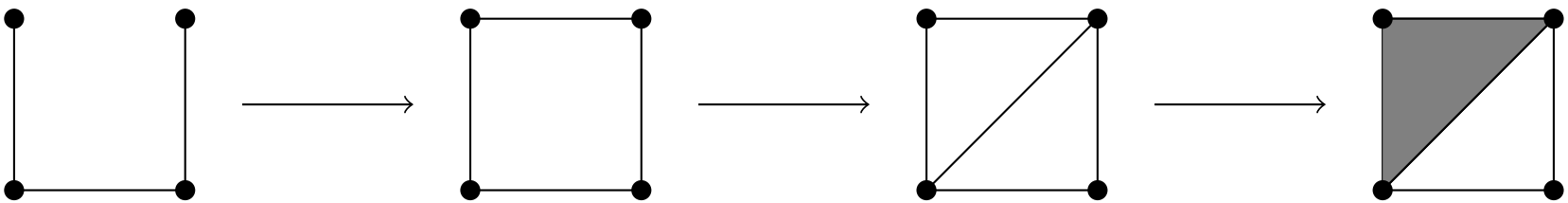
Pour la prédiction du sentiment, cela n'a pas beaucoup d'avantage par rapport à KNN ou Naive Bayes. On change à une tâche plus appropriée : analyser la richesse des structures dans une dissertation.

Introduction




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

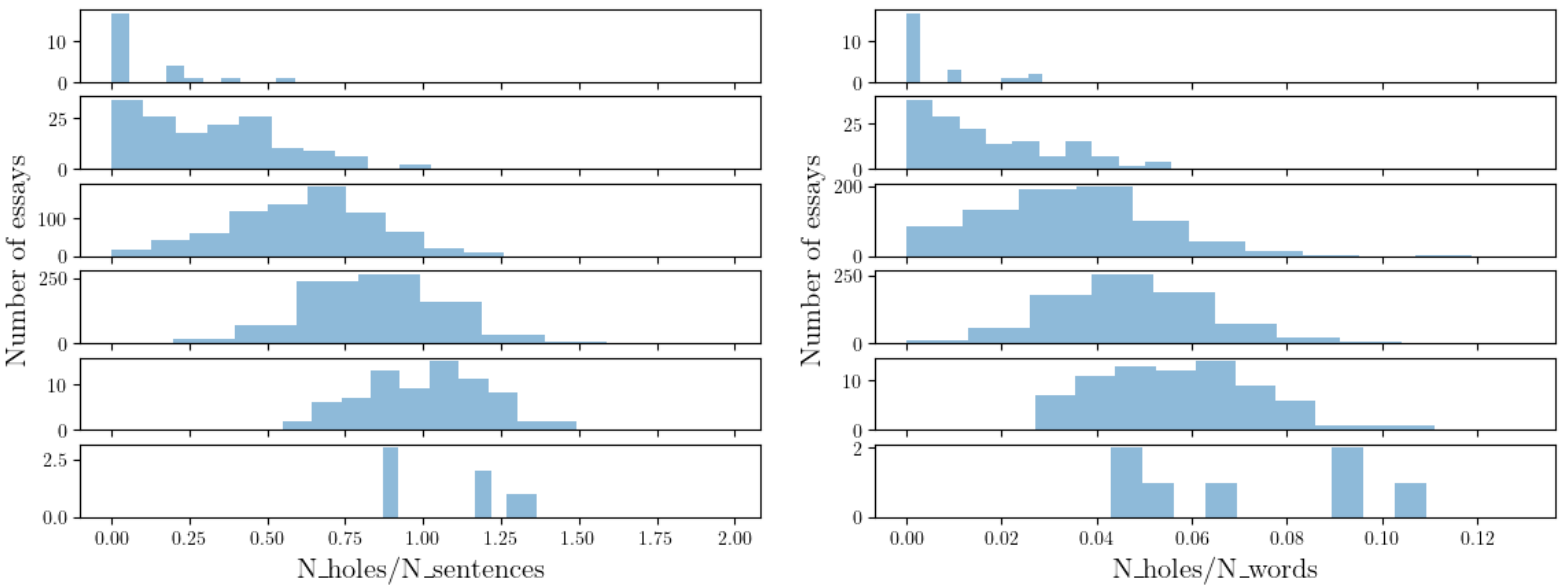
Il y a 2 trous dans ce complexe simplicial, l'un représenté par le cycle bleu et l'autre par le cycle marron.



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

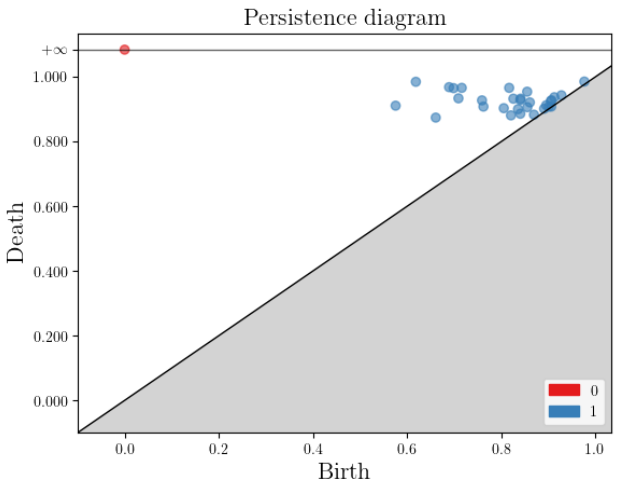
Techniques d'apprentissage

- 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

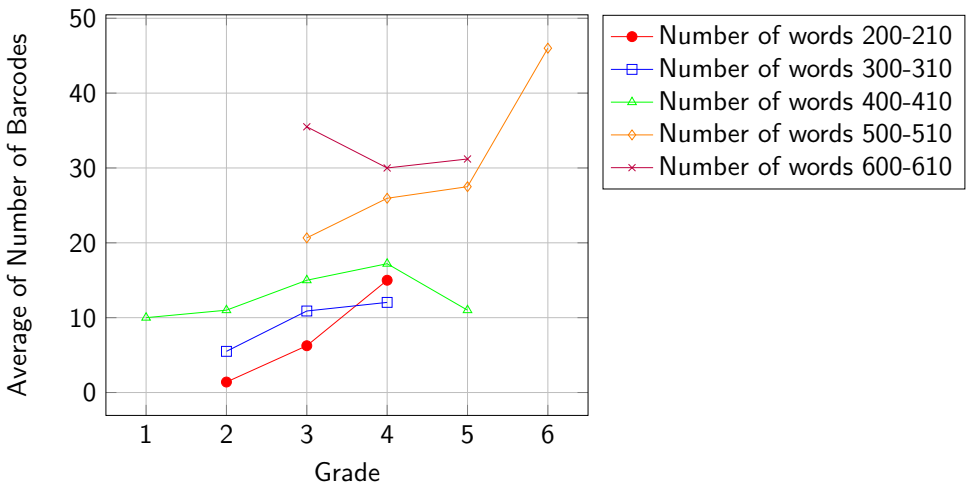
Annexe



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 censhorship 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