UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Analyse de la performance d'un système de recommandation de vins pour un mets

Auteur·e(s) : Alexia De Poorter et Marie Hartman
Promoteur·rice(s) : Corentin Vande Kerckhove
Année académique 2021-2022
Travail de fin d'études (TFE) en vue d'obtenir le titre de
Master (60) en Sciences de Gestion
Horaire de jour

Abstract

Ce mémoire analyse les performances d'un système de recommandation qui utilise des données textuelles exclusivement trouvées sur Internet pour suggérer des vins pour des plats spécifiques aux utilisateurs mais sans prendre en compte leurs préférences.

Ce système de recommandation, fondé sur le travail de Roald Schuring (Schuring, 2019a), est un système hybride basé d'une part sur le contenu en exploitant des données textuelles sur les vins et d'autre part sur le contexte en considérant les mets comme éléments contextuels. Un des avantages de ce système est l'utilisation de données non-structurées qui peuvent être collectées facilement et abondamment.

Notre contribution dans ce mémoire est triple : premièrement, nous conceptualisons et fournissons une explication détaillée de ce système de recommandation. Ensuite, nous implémentons une méthode d'évaluation du processus d'embedding du vin/des mets et montrons que celui-ci donne des résultats plus performants qu'un processus aléatoire. Finalement, nous implémentons deux méthodes d'évaluation des performances du système de recommandation et démontrons que ce système de recommandation de vins pour des mets spécifiques fournit de meilleures associations qu'un système aléatoire.

Remerciements

Tout d'abord, nous tenons à remercier notre promoteur Corentin Vande Kerckhvove pour son aide précieuse et sa disponibilité tout au long de cette année. Il nous a permis d'améliorer et d'approfondir notre travail en nous prodiguant des conseils utiles et en nous donnant de nombreuses pistes et idées intéressantes. Son intérêt et son implication dans notre sujet ont été très motivants et nous ont aidées à produire ce rendu final de notre travail.

Ensuite, nous sommes reconnaissants envers Sébastien Colla pour sa compréhension et son soutien constant tout au long de la réalisation de notre mémoire.

Enfin, nous tenons particulièrement à remercier Carole Lens et Martine Otjacqes pour leur relecture assidue de notre mémoire.

Table des matières

1 Introduction

La croissance rapide d'Internet a engendré, entre autres, une augmentation considérable de la quantité d'informations générées et partagées par des organisations dans presque toutes les industries et tous les secteurs (Blumberg et Atre, 2003). Nous vivons dans un monde où tout dans notre vie est enregistré numériquement et les données disponibles augmentent de jour en jour : par exemple, les données des médias sociaux, les données des smartphones ou les données de préférences en matière de comportement humain (Sarker, 2021). Dans ce contexte de l'analyse des données et de l'informatique, le Machine Learning et notamment les algorithmes de recommandation ont connu une croissance rapide ces dernières années (Sarker, 2021). Ces algorithmes suggèrent aux utilisateurs différents types de produits, tel que le vin.

Le vin est la boisson la plus consommée par l'homme pour le plaisir depuis de nombreux millénaires (Fehér et al., 2007). L'amour pour le vin n'est pas apparu du jour au lendemain : des preuves montrent que les humains consomment du vin depuis environ dix mille ans (Fehér et al., 2007). Aujourd'hui, la dégustation de vin prend de plus en plus d'importance mais la variété de vins sur le marché est énorme et chaque vin a un goût différent (Katarya et Saini, 2022). Le vin a en effet une incroyable diversité : il existe plus de 10 000 variétés différentes de raisins dans le monde, et chacune d'entre elles peut être transformée de cent mille façons différentes (Martinez et al., 2018). Les sommeliers consacrent d'ailleurs leur vie entière à l'art de la dégustation du vin. Une de leurs compétences principales est la recherche d'accords harmonieux entre les vins et repas. Ils élaborent des profils de saveurs en utilisant leur grande expérience pour fournir des évaluations nuancées d'innombrables bouteilles de vin chaque année. Cependant, la majorité n'ont ni le temps ni l'argent d'essayer de nombreuses variétés de vins pour acquérir une telle expérience, développer leur palais et déterminer quel vin associer à quel repas. (Martinez et al., 2018)

C'est dans ce cadre que les systèmes de recommandation, un domaine de recherche relativement récent en Machine Learning, peuvent s'avérer utiles. Les systèmes de recommandation suggèrent aux utilisateurs des produits qu'ils sont susceptibles d'acheter ou de consommer. Ces produits, appelés items, peuvent être de tout type : il peut s'agir de films, de nouvelles sur un site de presse, de livres ou, comme dans le cadre de ce mémoire, de vins. Le développement des systèmes de recommandation a débuté d'une observation assez simple : les individus ont tendance à s'appuyer sur les recommandations données par d'autres pour prendre des décisions courantes et quotidiennes (Resnick et al., 1994 ;

Shardanand et Maes, 1995) : par exemple, le choix d'un film est influencé par l'opinion des pairs et les critiques de cinéma ou les employeurs se basent sur les lettres de recommandation dans leur décision de recrutement. Les systèmes de recommandation peuvent donc s'avérer particulièrement utiles aux utilisateurs dans le choix du vin associé à un repas spécifique au vu de l'énorme quantité de variétés de vins disponibles.

1.1 Question de recherche

Dans le cadre de ce mémoire, nous souhaitons évaluer une méthode de recommandation qui utilise exclusivement des données en ligne pour suggérer des vins pour des plats spécifiques aux utilisateurs mais sans prendre en compte leurs préférences. A notre connaissance, il n'existe pas de tels systèmes de recommandation démontrés performants dans la littérature scientifique mais nous avons étudié une méthode de recommandation avec ces caractéristiques dont les performances n'ont pas encore été évaluées. L'objectif de ce mémoire va donc être d'analyser les performances de ce système de recommandation et ainsi, de répondre à la question de recherche suivante :

Est-il possible de recommander de manière performante des vins pour des plats spécifiques en exploitant des données exclusivement trouvées sur Internet ?

1.2 Motivations managériales

Les fonctions des systèmes de recommandation peuvent être multiples. Il convient d'abord de distinguer le rôle de ces systèmes pour les fournisseurs de service d'une part, et pour les utilisateurs d'autre part.

Pour les fournisseurs de service en ligne, il existe de nombreuses raisons pour lesquelles ils peuvent vouloir exploiter cette technologie (Ricci et al., 2015) :

- *Augmenter le nombre d'items vendus :* puisque les systèmes de recommandation suggèrent aux utilisateurs des items qu'ils sont susceptibles d'apprécier, ces systèmes permettent de vendre plus d'items par rapport à ceux habituellement vendus sans recommandation.
- *Vendre des articles plus diversifiés :* les systèmes de recommandation permettent aux utilisateurs de choisir des items auxquels ils n'auraient pas songé sans recommandation précise.
- *Augmenter la satisfaction de l'utilisateur :* si le système de recommandation suggère des recommandations efficaces et précises, l'utilisateur trouvera les recommandations intéressantes et pertinentes. Si en plus l'interface du système est agréable à utiliser, il prendra plaisir à l'utiliser.
- *Augmenter la fidélité de l'utilisateur :* si un client est satisfait, il aura plus de chance d'être fidèle. Or, les systèmes de recommandation exploitent les informations acquises auprès de l'utilisateur lors d'interactions précédentes. Par conséquent, plus l'utilisateur interagit avec le site, plus le modèle s'affine et la représentation des préférences de l'utilisateur se précise. Ainsi, si un utilisateur est fidèle à un site Web, il va revisiter et interagir davantage avec ce site, ce qui va améliorer la qualité des recommandations et par conséquent sa satisfaction, ce qui aura pour effet d'augmenter sa fidélité : il s'agit donc d'un cercle vertueux.

— *Meilleure compréhension de ce que veut l'utilisateur :* la description des préférences de l'utilisateur recueillies soit explicitement, soit par prédiction par le système peut être utilisée pour de nombreuses autres applications telles que l'amélioration de la gestion du stock, la production de l'article ou la diffusion d'un type de message promotionnel.

Toutes ces motivations s'appliquent à un système de recommandation de vins.

Les utilisateurs peuvent également trouver des avantages à l'utilisation de systèmes de recommandation si ceux-ci soutiennent efficacement leurs tâches et leurs objectifs. Un système de recommandation doit donc trouver un équilibre entre les besoins de ces deux acteurs et offrir un service qui leur soit utile à tous : notamment, un système de recommandation de vin doit prendre en compte les motivations du commerçant de vins mais également être utile au consommateur de vins.

1.3 Structure du travail

Ce mémoire est structuré de la manière suivante. Le Chapitre 2 présente une revue de la littérature des systèmes de recommandation de manière générale ainsi que de différents travaux dans le domaine du vin et des mets, puis présente une description des différents types de goûts et leurs caractéristiques. Dans le Chapitre 3, la méthode de recommandation étudiée dans ce mémoire est expliquée, puis critiquée. Ensuite, le Chapitre 4 décrit la méthode d'évaluation de l'embedding du vin et de mets utilisée dans le système de recommandation étudié et analyse les résultats pour ensuite suggérer des pistes pour améliorer cette méthode d'évaluation. Le Chapitre 5 décrit la méthode d'évaluation des performances de ce système de recommandation, analyse les résultats et détaille des pistes d'améliorations du système de recommandation. Finalement, le Chapitre 6 conclut ce travail.

2 État de l'art

Rappelons que l'objectif de ce mémoire est d'évaluer les performances d'un algorithme de systèmes de recommandation de vins pour des aliments spécifiques. Ce chapitre a donc pour but de fournir un préambule sur les concepts nécessaires des systèmes de recommandation et de présenter différents travaux dans le domaine du vin et de la nourriture. Ensuite, pour mieux comprendre la manière dont les spécialistes accordent les vins aux repas, une section sur la description des différents types de goûts et de leurs caractéristiques est présentée.

2.1 Systèmes de recommandation

Les Systèmes de Recommandation (SRs) sont une sous-classe des systèmes de filtrage de l'information qui regroupe un ensemble d'outils de programmation et de techniques tels que les arbres de décision, machines à vecteurs de support (support vector machines en anglais), réseaux neuronaux, régression logistique, etc. Les SRs permettent d'obtenir des suggestions d'items les plus susceptibles d'intéresser un utilisateur particulier (Burke, 2007 ; Resnick et al., 1994 ; Resnick et Varian, 1997).
"Item" est un terme général qui désigne ce que le système recommande aux utilisateurs. Il peut s'agir de films, de musiques ou, comme dans ce mémoire, de vins. Les suggestions sont liées à divers processus de décision, tels que le film à regarder, la musique à écouter ou les vins à acheter.

Les systèmes de recommandation ont été considérés comme un domaine de recherche indépendant depuis le milieu des années 1990 (Balabanović et Shoham, 1997 ; Goldberg et al., 1992 ; Resnick et al., 1994 ; Resnick et Varian, 1997 ; Shardanand et Maes, 1995). Ces dernières années, l'intérêt pour les systèmes de recommandation a très fortement augmenté. Pour s'en convaincre, on peut citer le fait que les SRs sont très utilisés dans les sites Internet renommés tels que Amazon, Youtube, Netflix, Spotify, Facebook, Instagram... ou le fait que de nombreux ateliers et conférences sont dédiés à ce sujet, telles que les conférences *Association of Computing Machinery's (ACM)* et *Conference Series on Recommender Systems (RecSys)*, établies en 2007 ou encore le fait que des cours sur le domaine sont maintenant donnés dans les universités, par exemple à l'Ecole Polytechnique de Louvain à l'UCLouvain.

Pour pouvoir établir des recommandations, les SRs doivent rassembler et exploiter diffé-

rents types de données. De manière générale, les données utilisées se réfèrent à 3 types de classes : les items, les utilisateurs et les transactions entre les items et les utilisateurs (Ricci et al., 2015). Par ailleurs, les SRs utilisent des données qui peuvent être créées de différentes manières. Certaines données sont dites **structurées** : il s'agit de données avec une structure bien définie, conforme à un modèle de données suivant un ordre standard, hautement organisée et facilement accessible, et utilisée par une entité ou un programme informatique (Sarker, 2021). D'autres méthodes de recommandation utilisent des données dites **non-structurées** : il s'agit de données qui n'ont pas de format prédéfini, ce qui rend leur traitement et leur analyse plus complexes ; elles contiennent principalement du texte et du matériel multimédia (Sarker, 2021). De nombreux travaux utilisent des données non-structurées dans leurs méthodes (Balducci et Marinova, 2018 ; Blumberg et Atre, 2003 ; Pazzani et Billsus, 2007 ; Protasiewicz et al., 2016 ; Subramaniyaswamy et al., 2015 ; Tanwar et al., 2015).

Dans le cadre de ce mémoire, nous utilisons des données non-structurées puisque nous considérons des bases de données en ligne regroupant des avis provenant de sites Internet (`winemag` et `Amazon`).

2.1.1 Classes des systèmes de recommandation

Pour obtenir des suggestions sur des items à recommander aux utilisateurs, les SRs doivent prédire l'utilité de certains items ou comparer leur utilité entre eux et sur base de cela, décider quels items recommander. Plus spécifiquement, un système de recommandation essaie d'estimer la fonction d'évaluation ou de classement R :

$$R : Utilisateur \times Item \rightarrow Note\ d'évaluation\ ou\ Classement \tag{2.1}$$

pour les pairs (utilisateur, item) qui n'ont pas encore été évalués par les utilisateurs. Le système recommandera à chaque utilisateur les items avec la plus grande utilité estimée. On peut classer les SRs selon leurs sources de données, c'est-à-dire l'origine des données nécessaires pour faire des recommandations. Les sources de données peuvent par exemple être les préférences des autres utilisateurs ou des données déductives sur le domaine (Burke, 2007). Burke distingue 6 classes de techniques de recommandation selon leurs sources de données, celles-ci sont représentées sur la figure 2.1 :

— **Filtrage collaboratif :** le système fait des recommandations à l'utilisateur actif en se basant sur les articles que d'autres utilisateurs ayant des goûts similaires ont aimés dans le passé (Ricci et al., 2015). C'est la technique en SR la plus populaire et la plus largement implémentée. Ce système utilise seulement des informations sur les profils d'évaluation des différents utilisateurs.

— **Système basé sur le contenu :** le système recommande des items qui sont similaires à ceux que l'utilisateur a aimés dans le passé (Ricci et al., 2015). Ce système utilise 2 sources de données : les caractéristiques associées aux produits et les évaluations qu'un utilisateur leur a attribuées.

— **Système démographique :** le système recommande des items en se basant sur le profil démographique des utilisateurs : les items sont recommandés pour différentes niches démographiques en combinant les évaluations des utilisateurs de ces niches (Burke, 2007 ; Ricci et al., 2015).

— **Système basé sur les connaissances :** le système recommande des items en se basant sur la façon dont certaines caractéristiques de l'item répondent aux besoins et aux préférences des utilisateurs et sur l'utilité de l'item pour les utilisateurs (Burke, 2007 ; Ricci et al., 2015). Parmi les SRs basés sur la connaissance, on retrouve

  — **Système basé sur le cas :** dans ces systèmes, une fonction de similarité estime dans quelle mesure les besoins de l'utilisateur (description du problème) correspondent aux recommandations (solutions du problème).

  — **Système basé sur la contrainte :** les systèmes exploitent principalement des bases de connaissances prédéfinies qui contiennent des règles explicites sur la manière de relier les besoins des clients aux caractéristiques des articles. Parmi ces systèmes, on retrouve les **systèmes basés sur la communauté**. Il s'agit de systèmes qui recommandent des items sur base des préférences des amis de l'utilisateur. Ce système nécessite des informations sur les relations sociales des utilisateurs et les préférences des amis de chaque utilisateur.

— **Système basé sur le contexte :** le contexte, tels que le temps, le lieu ou la compagnie d'autres personnes, peut permettre de mieux personnaliser la sortie du système. Pour les systèmes de recommandation basés sur le contexte, on suppose que les informations relatives au contexte sont connues et définies par un ensemble d'attributs contextuels et que ceux-ci influent sur les évaluations. La fonction d'évaluation ou de classement R (2.1) à estimer est alors donnée par

$$R : Utilisateur \times Item \times Contexte \rightarrow Note\ d'évaluation\ ou\ Classement$$

où *Contexte* spécifie les informations contextuelles connues. (Ricci et al., 2015)
Ricci et al., 2015 présentent trois paradigmes algorithmiques populaires pour incorporer le contexte dans le processus de recommandation :

  — **Système basé sur la réduction (préfiltrage) :** seules les informations qui correspondent au contexte d'utilisation actuel, par exemple les notes des items évalués dans le même contexte, sont utilisées pour calculer les recommandations.

  — **Post-filtrage contextuel :** le système ignore le contexte et la sortie de l'algorithme est filtrée pour inclure seulement les recommandations pertinentes pour le contexte considéré.

  — **Modélisation du contexte :** dans ce système, les données de contexte sont explicitement utilisées dans le modèle de prédiction. C'est l'approche la plus sophistiquée des trois.

— **Système hybride :** Ces systèmes sont basés sur une combinaison de deux ou plusieurs des techniques mentionnées ci-dessus pour améliorer les performances de recommandation. On retrouve les

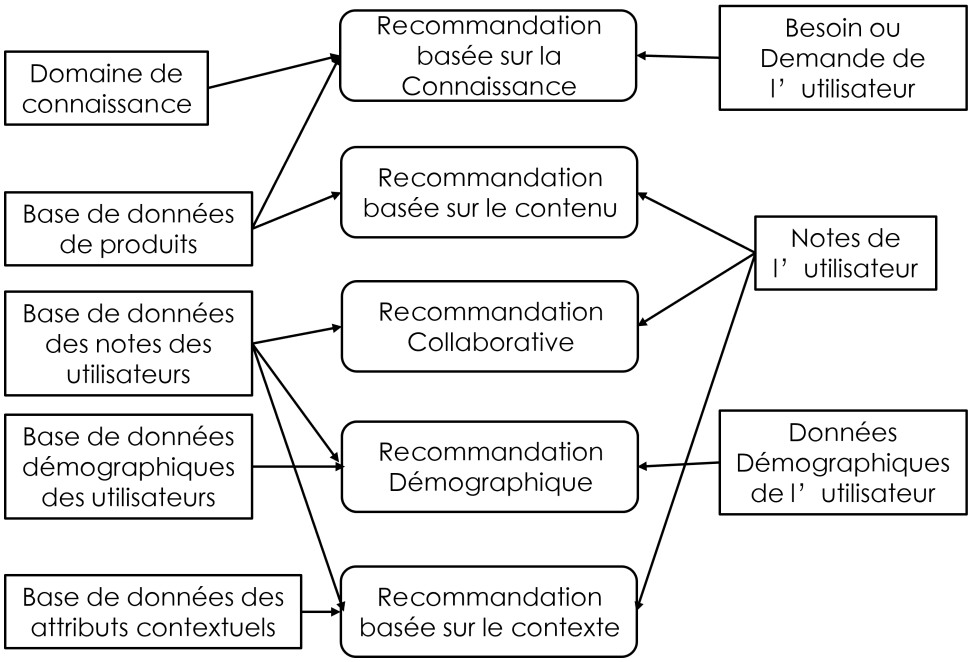


FIGURE 2.1 – Techniques de recommandation et leurs sources de données. Image traduite et inspirée de (BURKE, 2007 ; RICCI et al., 2015)

Dans le cadre de ce mémoire, nous étudions un algorithme de recommandation hybride puisque notre système est basé sur le contenu mais également sur le contexte.
D'une part, le système de recommandation étudié est basé sur le contenu puisque des bases de données d'avis sur le vin et sur la nourriture sont utilisées dans l'algorithme étudié pour établir les recommandations : ces avis sont uniquement textuels et ne donnent donc pas de notes d'évaluation.
D'autre part, on s'intéresse dans ce mémoire à l'association de vin avec un repas. On considère donc seulement l'item et le contexte, c'est-à-dire le vin et le repas mais pas l'utilisateur. Il faut donc estimer la fonction d'évaluation ou de classement R sans prendre en compte les préférences de l'utilisateur :

$$R : Item \times Contexte \rightarrow Note\ d'évaluation\ ou\ Classement$$

Notons que cette fonctionnalité pourrait être ajoutée et implémentée par la suite mais cela complexifie fortement le problème. En effet, ajouter une dimension à la fonction R peut faire apparaître certains défis : un problème de sparsité, la détermination des préférences contextuelles des utilisateurs, le manque de données disponibles publiquement, l'évaluation, etc. (YUJIE et LICAI, 2010).

2.1.2 Méthode d'évaluation des systèmes de recommandation

Un aspect très important des systèmes de recommandation est la nécessité d'évaluer leurs qualités et leurs valeurs. Il faut évaluer le système à différentes étapes et pour différentes raisons.
Au moment de la conception du système de recommandation, l'évaluation permet de vérifier que le choix de l'approche de recommandation est approprié : cette évaluation est implémentée **hors-ligne**. En effet, les expériences hors-ligne utilisent des ensembles de données existants et un protocole qui modélise le comportement de l'utilisateur afin

d'estimer la performance des recommandations telles que la précision des prédictions et afin de comparer les performances de plusieurs algorithmes. (Adomavicius et Tuzhilin, 2005 ; Ricci et al., 2015)

Cependant, ces évaluations ne permettent pas d'obtenir des informations quant à la satisfaction de l'utilisateur ou à l'expérience avec le système. Ainsi, les algorithmes pourraient fournir des recommandations très précises qui ne seraient pas acceptées par les utilisateurs. Il est donc nécessaire d'effectuer une évaluation centrée sur l'utilisateur. Dans (Ricci et al., 2015), les auteurs distinguent 2 types d'évaluations centrées sur l'utilisateur :

— **Expérience en-ligne :** ces expériences évaluent les performances des systèmes de recommandation sur des utilisateurs réels qui ne sont pas au courant de l'expérience menée.

— **Étude d'utilisateurs :** on demande à un petit groupe d'utilisateurs de réaliser un ensemble de tâches avec différentes versions du système de recommandation, tel que répondre à des questions à propos de leur expérience.

Dans le cadre de ce mémoire, nous effectuons uniquement des expériences hors-ligne : en effet, pour évaluer les performances de l'algorithme, nous utilisons un ensemble de données construites à partir de plusieurs sites Internet, autrement dit construites à partir de données existantes, et nous n'effectuons aucune évaluation centrée sur l'utilisateur.

2.2 Travaux relatifs au vin et à l'alimentation

Maintenant que les systèmes de recommandation ont été présentés de manière générale, on s'intéresse au domaine spécifique étudié dans ce mémoire : comment associer le vin et les aliments. Nous allons donc présenter différents travaux de la littérature relatifs au vin, à l'alimentation et à l'association de vin et d'aliments.

2.2.1 Travaux relatifs au vin

Dans la littérature, il existe de nombreux articles qui étudient le "vin" : les recherches dans le domaine sont nombreuses et couvrent des sujets variés. Pour s'en convaincre, citons une liste non-exhaustive d'exemple de travaux de recherche dans ce domaine qui abordent divers sujets. Par exemple, certains travaux s'intéressent aux déterminants de la consommation de vin (Brunner et Siegrist, 2011 ; Peršurić et al., 2018), d'autres travaux étudient les systèmes de recommandation de vins pour chaque utilisateur sans prendre en compte le contexte tel que le repas (Cortez et al., 2009 ; Katarya et Saini, 2022 ; Martinez et al., 2018), d'autres travaux s'intéressent aux préférences des utilisateurs pour certains attributs du vin (Pomarici et al., 2017 ; Risius et al., 2019) ou encore aux effets des évaluations des experts sur le prix du vin (Gokcekus et Gokcekus, 2019).

2.2.2 Travaux relatifs à l'alimentation

Il n'est pas toujours trivial de fournir des recommandations appropriées relatifs à l'alimentation car choisir un aliment est une décision complexe, avec de multiples facettes et dépend fortement du contexte : c'est pourquoi il existe de nombreux travaux sur la recommandation personnalisée d'aliments. Dans (Elsweiler et al., 2022), les auteurs détaillent les nombreuses facettes et complexités de ce problème, puis passent en revue

les technologies qui ont été proposées pour pallier ces difficultés. On peut citer une liste non-exhaustive d'exemples de travaux de recherche sur des solutions algorithmiques au problème des recommandations alimentaires. Il y a entre autres des travaux de recherche liés aux systèmes de recommandation d'aliments/de recettes qui donnent des recommandations sur base des préférences des utilisateurs ainsi que de leurs besoins nutritionnels (Elahi et al., 2014 ; El-Dosuky et al., 2012 ; Elsweiler et al., 2015 ; Freyne et al., 2011 ; Ueta et al., 2011). Il existe également plusieurs travaux sur la recommandation de produits alimentaires à acheter durant les courses (Christodoulou et al., 2017 ; Waltner et al., 2017) ou sur la recommandation de nourriture saine pour des groupes cibles (Alian et al., 2018 ; Khan et Hoffmann, 2003 ; Lee et al., 2010).

Finalement, Elsweiler et al., 2022 donnent également un aperçu des ressources disponibles pour étudier la recommandation alimentaire en tant que problème de recherche : plus précisément, ils présentent des ressources permettant d'une part, d'implémenter les systèmes de recommandation tels que les ensembles de données de recettes, de produits alimentaires, de repas, de menus et de restaurants, et d'autre part, de calculer et mesurer les valeurs nutritionnelles, l'impact sur la santé ou la durabilité des aliments. Par exemple, les ressources suivantes sont citées : (D.H. and FSA, 2016 ; EDAMAM, s. d. ; GGDOT, s. d. ; Group, s. d. ; Instacart, 2017).

2.2.3 Travaux relatifs à l'association vin et aliments

Pour aider le consommateur à accorder un vin à son repas, des livres sur les différentes approches d'accords mets et vins ont été publiés (Arnone et Simonetti-Bryan, 2013 ; Chartier, 2012 ; Hambleton, 2008 ; Puckette et Hammack, 2018 ; Simonetti-Bryan, 2010). D'autres experts reconnaissent que les préférences des consommateurs peuvent influencer la pertinence d'associations des vins et des aliments mais procèdent néanmoins à des recommandations d'appariement spécifiques (Goldstein, 2006 ; R. Harrington, 2007 ; Peynaud, 1996 ; Thomas, 2013).
Par ailleurs, dans la littérature scientifique, des chercheurs se sont également intéressés au problème de l'association du vin et de la nourriture. Par exemple, (Nygren et al., 2001 ; Nygren et al., 2002) étudient comment les saveurs du vin changent quand la nourriture, tel que le fromage, est consommée avec le vin. (Nygren et al., 2001 ; Nygren et al., 2002) explorent les correspondances préférées des experts et des consommateurs entre les vins et les fromages. (Sherwin, 2017) vérifie si les préférences en matière de vin des consommateurs novices peuvent être prédites à l'aide de leur consommation historique et actuelle d'aliments et de boissons. (R. J. Harrington et Seo, 2015) s'intéressent à l'impact de l'appréciation du vin et des aliments sur les perceptions des associations vin-aliment. (Klosse, 2011) étudie une nouvelle approche pour associer le vin et la nourriture basée sur les saveurs de la nourriture et du vin. Ces travaux étudient donc l'association entre le vin et les aliments mais ne fournissent pas d'algorithmes pour recommander automatiquement des vins et des repas, comme c'est le cas dans notre mémoire.

Néanmoins, il existe des systèmes de recommandation qui suggèrent de manière automatique des vins avec des repas : (Javanmardian, 2005) associe le vin à un repas en encapsulant les connaissances sur les accords mets et vins, et en utilisant d'autres types de connaissances pour réduire les options alimentaires mais ils ne considèrent que 7 possibilités de vins. (Michaelis et al., 2008) regroupent les recommandations des internautes sur les accords mets-vins, en utilisant les technologies de pointe du web sémantique : il

s'agit donc d'un système de recommandation basé sur le contenu et le contexte mais qui utilise des données structurées.

Il existe donc une large documentation à propos de l'association entre le vin et la nourriture mais à notre connaissance, aucun algorithme de la littérature ne possède les mêmes caractéristiques que notre algorithme, c'est-à-dire un système de recommandation qui utilise des **données non-structurées** d'Internet et qui est basé sur le **contenu** et le **contexte**. L'utilisation de données non-structurées offre l'avantage de ne pas nécessiter de données prédéfinies, elles sont donc plus nombreuses et peuvent être collectées plus facilement et rapidement.

2.3 Description des différents types de goûts

Dans le domaine de l'association de vins et repas, les spécialistes ne s'accordent pas toujours sur ce qui est le plus important dans le choix des vins à servir avec un repas. Est-ce la texture de l'aliment et du vin ? Est-ce que ce sont les saveurs ? Ou encore les cinq composantes sensorielles primaires ? (R. Harrington, 2007)
En effet, le vin présente l'un des profils gustatifs et olfactifs les plus complexes avec une immense diversité de goûts et d'arômes. Par exemple, l'acidité des aliments est un facteur important, mais seulement lorsqu'elle dépasse un certain seuil. De même, un arôme peut impacter significativement les associations mais uniquement lorsque son intensité est supérieure à la norme. Ces différents éléments à prendre en compte démontrent la difficulté de déterminer les facteurs clés de choix d'associations d'un vin avec un repas (R. Harrington, 2007 ; Van Niekerk, 2012).

Une étape indispensable consiste donc à séparer ces éléments en trois catégories générales et à comprendre exactement comment celles-ci diffèrent les unes des autres lors de la dégustation de vins et d'aliments. Les trois catégories sont les suivantes :

1. Le goût principal
2. La texture
3. Les saveurs ou arômes

Bien que ces trois catégories ne s'excluent pas mutuellement, le fait de les séparer permet de mieux distinguer les facteurs clés des associations repas et vins possibles.

Dans le cadre de ce mémoire, nous distinguons deux catégories : les **non-arômes** comprenant le goût principal (1) et la texture (2), ainsi que les **arômes** comprenant les saveurs (3). Les non-arômes permettront de définir des règles d'association tandis que les arômes permettront d'établir un classement des vins en fonction du mets à associer.

Caractéristiques du goût principal

Le goût principal est la façon dont nous décrivons la perception sensorielle de base avec la langue. Nous avons des papilles gustatives qui nous donnent la capacité intrinsèque d'enregistrer cinq sensations gustatives primaires dans la bouche : **le sucré, l'acidité, le salé, l'amertume et l'umami** (R. Harrington, 2007 ; IQWiG, 2006 ; Pediaopolis, 1947 ; Van Niekerk, 2012 ; Vilela et al., 2016 ; Wikipedia, 2022b).
Des expériences scientifiques ont démontré que ces cinq goûts existent et sont distincts

les uns des autres : chacun d'entre eux est présent à un certain degré dans tous les mets (Wikipedia, 2022b) : par exemple, une salade de chou frisée (amer), lardons (salé), pommes (sucré), jus de citron (acide) et parmesan (umami) est composée des 5 goûts principaux.

A l'heure actuelle, les cinq goûts spécifiques de base ne sont pas communs à tous les spécialistes et continuent d'évoluer pour diverses raisons :

— Au début du vingtième siècle, les spécialistes ne considéraient pas l'umami comme cinquième goût, mais aujourd'hui, un grand nombre d'autorités le reconnaissent comme tel (Pediaopolis, 1947).

— Dans les pays asiatiques, principalement dans les cultures chinoise et indienne, le **piquant** était traditionnellement considéré comme un sixième goût(Pediaopolis, 1947).

— En 2015, des chercheurs ont suggéré un nouveau goût appelé **gras**. Le débat actuel consiste à savoir s'il existe un récepteur qui réside dans nos papilles gustatives pour détecter les graisses (Keast et Costanzo, 2015 ; Van Niekerk, 2012).

Dans le cadre de ce mémoire, l'algorithme étudié considère le gras et le piquant comme goûts principaux mais pas l'umami : ainsi, les six goûts principaux considérés pour décrire un mets sont : **sucré, acidité, salé, amertume, piquant et gras**. Afin de les distinguer des arômes, nous les qualifions de non-arômes.

Notons que le salé, l'amertume, le piquant et le gras sont rarement perçus dans les vins, mais il est important d'en tenir compte dans l'association des repas et vins. Puisque les sensations gustatives peuvent déformer et influencer le goût du vin, il est donc nécessaire de définir quelques règles basiques d'associations de vins et repas. Cependant, ces règles ne sont pas universelles et varient selon les auteurs (Arnone et Simonetti-Bryan, 2013 ; Chartier, 2012 ; Hambleton, 2008 ; Puckette et Hammack, 2018 ; Simonetti-Bryan, 2010 ; Van Niekerk, 2012).

Dans le cadre de notre mémoire, nous considérons les règles d'associations suivantes (Puckette et Hammack, 2018 ; Van Niekerk, 2012) :

1. Le vin doit être plus sucré que les mets.
2. Le vin doit être plus acide que les mets.
3. Les vins amers ne se marient pas bien avec les mets amers.
4. L'amer et le piquant ne s'accordent pas ensemble.
5. L'amertume et l'acidité ne s'accordent pas ensemble.
6. L'acide et le piquant ne s'accordent pas ensemble.

Caractéristiques de la texture

Une deuxième catégorie clé est la texture en bouche liée au corps, à la puissance, au poids et à la structure du plat et du vin à associer. La texture est la caractéristique d'un aliment ou d'un vin qui crée une sensation buccale spécifique dans chaque coin de la bouche, plutôt qu'une saveur perceptible dans des parties spécifiques de la langue (R. Harrington, 2007).

La texture peut être décrite de diverses manières. Dans le vin, elle peut être décrite par la tannicité des vins rouges, la finesse des vins blancs ou encore la sensation crémeuse

liée au vieillissement du vin. Dans les aliments, elle peut être décrite comme granuleuse, croquante, huileuse ou rugueuse. La température peut également servir de contraste de texture.
Même si les combinaisons d'aliments et de vins peuvent être similaires ou contrastées, il est conseillé d'associer des mets et des vins ayant une texture similaire (R. Harrington, 2007 ; Van Niekerk, 2012). Les contrastes peuvent néanmoins être efficaces, par exemple des repas chauds servis avec du vin froid peuvent constituer un contraste rafraîchissant et satisfaisant. (R. Harrington, 2007)

La représentation de la texture la plus importante dans l'association d'un vin et repas est **le poids**, allant du léger au lourd. Par exemple, un poisson est considéré comme léger tandis qu'un steak est considéré comme lourd. En général, les vins légers ont tendance à être légèrement fruités, avec un faible taux d'alcool tandis que les vins lourds ou corsés seront profondément fruités, avec beaucoup d'alcool (>13%). (Van Niekerk, 2012)

Il est indispensable de chercher à équilibrer le poids des aliments et celui du vin, de manière à ce qu'aucun ne domine l'autre (R. Harrington, 2007 ; Van Niekerk, 2012). Reprenons l'exemple du steak, celui-ci est généralement accompagné d'une sauce crémeuse au poivre ou aux champignons, ce qui apporte de la lourdeur au plat. L'onctuosité de la sauce exige un vin tout aussi crémeux, ce qui nous oriente immédiatement vers un vin blanc, par exemple un Chardonnay qui apporte à la fois le poids dont le steak a besoin et la texture crémeuse que la sauce exige. (Van Niekerk, 2012)

Dans le cadre de ce mémoire, nous considérons **le poids** comme le premier non-arôme puisque c'est le non-arôme le plus important à considérer dans le cadre de l'association vins et plats (Van Niekerk, 2012). Les non-arômes sont donc au nombre de sept : **poids, sucré, acidité, salé, amertume, piquant et gras**.

Caractéristiques des arômes

Une troisième catégorie d'éléments dans les aliments et le vin est celle des arômes. Le goût et les arômes sont souvent confondus, mais alors que le goût principal est perçu par la langue, les arômes sont liés aux perceptions olfactives. Dans la dégustation du vin, les spécialistes sentent le vin avant d'en prendre une gorgée afin d'identifier certaines saveurs qui pourraient être présentes : ces saveurs représentent les arômes.

Les arômes sont plus variés que les goûts : la langue humaine est limitée aux goûts primaires perçus par les récepteurs gustatifs de la langue - acide, amer, salé, sucré et umami. A contrario, les récepteurs olfactifs perçoivent un large éventail de notes aromatiques tels que les saveurs fruitées, terreuses, cuirées, florales, herbacées, minérales et boisées présentes dans le vin et les mets (R. Harrington, 2007 ; Wikipedia, 2022a).

En oenologie, au vu de la diversité des notes aromatiques, le Dr. Ann Noble, professeur émérite de l'Université de Californie à Davis, et une équipe de vignerons ont mis en place un outil : la roue des arômes du vin de l'UC Davis montre la correspondance entre des termes spécifiques et standardisés pour caractériser les saveurs du vin (Fallis, 2014). Celle-ci est visible à la figure 2.2.
Au centre de la roue se trouvent les termes d'arômes les plus généraux, par exemple l'arôme fruité. Ensuite, le dégustateur essaie de distinguer le type de fruit que présente l'arôme du vin : agrumes, baies, fruits tropicaux ou fruits cuits/frais. Parfois, lorsqu'un arôme est très intense, les dégustateurs réussissent à distinguer le type de baies qu'ils

perçoivent : par exemple, fraise, mûre, framboise ou cassis. La roue à vin permet ainsi aux dégustateurs d'apprendre à reconnaître les arômes de vin en décrivant aussi précisément que possible ce qu'ils perçoivent.

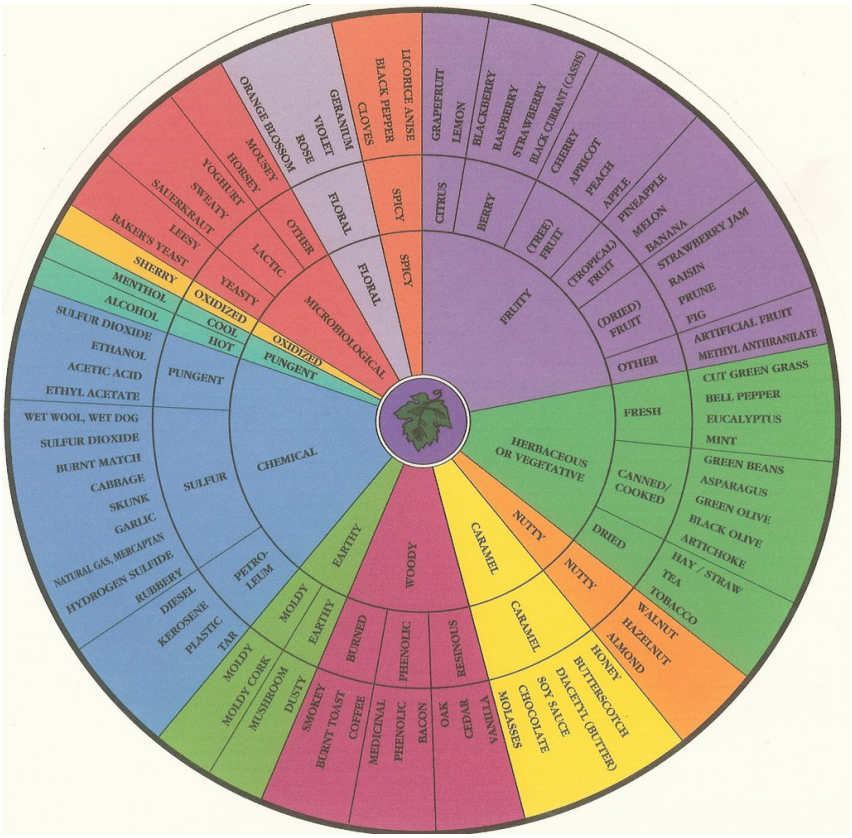


FIGURE 2.2 – La roue des arômes du vin de l'UC Davis (DAVIS, 1980).

Pour conclure, ce chapitre présente les caractéristiques principales de l'algorithme de recommandation hybride étudié : d'une part, ce système est basé sur le contenu puisque des bases de données textuelles sur les vins sont utilisées pour établir les recommandations : ces bases de données sont non-structurées. D'autre part, ce système est basé sur le contexte puisque nous considérons le repas comme élément contextuel.

Puis, il existe de nombreux travaux relatifs aux vins et à l'alimentation mais à notre connaissance, aucun algorithme ne possède toutes ces caractéristiques. En effet, l'utilisation de données non-structurées donnent un réel avantage à notre système de recommandation puisque celles-ci ne devant pas être prédéfinies, elles peuvent être collectées plus facilement et sont également plus nombreuses.

Finalement, la description des trois types de goûts principaux, i.e. les non-arômes comprenant le goût principal et la texture ainsi que les arômes, permet de comprendre les choix des spécialistes lors de l'association des vins et repas. Ces descriptions permettent de comprendre les étapes réalisées dans le système de recommandation étudié.

3 Système de recommandation

Ce chapitre a pour but de fournir une description du système de recommandation de vins fondé sur le travail de Roald Schuring (SCHURING, 2019a) en fournissant une explication détaillée des différentes étapes du code PYTHON dans le github `wine_recommender` (SCHURING, 2021a) que nous avons légèrement modifié et adapté. Celles-ci sont séparées en trois grandes parties : le preprocessing réalisé sur l'ensemble des données, les méthodes d'embedding utilisées pour caractériser les vins/ingrédients et finalement la méthode d'association d'un vin avec un plat à l'aide de plusieurs règles empiriques.

3.1 Démarche générale

Afin de recommander le vin idéal à un plat spécifique, deux étapes sont nécessaires : la première est la quantification des caractéristiques aromatiques et non aromatiques des mets et des vins. Cela permet de réaliser la deuxième étape : l'association des vins et mets selon les mêmes dimensions dans l'espace.

Pour ce faire, nous utilisons une base de données correspondant à un corpus de critiques de vins et de produits alimentaires suffisamment large pour nous permettre de caractériser de manière précise chaque produit/vin. La base de données des vins (SCHURING, 2021b) est constituée d'environ 150 000 critiques professionnelles extraites du site *winemag* (STRUM, 2022) : ces critiques couvrent environ 20 ans, des dizaines de pays et des centaines de cépages et chaque vin est associé à une critique. La base de données des aliments *Amazon Fine Foods* (STANFORD, 2017) est constituée d'environ 500 000 avis sur un large panel de produits alimentaires mais les avis ne se rapportent pas à des aliments spécifiques.

Dans un premier temps, un préprocessing est appliqué sur l'ensemble des données ce qui permettra de les traiter plus facilement par la suite. Les différentes étapes de préprocessing sont issues de l'article *Wine Embeddings and a Wine Recommender* de Roald Schuring (SCHURING, 2019b).

Ensuite, les données prétraitées seront entraînées afin d'obtenir un modèle `Word2Vec` qui transforme les termes du corpus de texte en vecteurs de nombres réels, caractérisant son profil sensoriel.

Finalement et à l'aide de règles d'associations, il sera possible d'établir un système de recommandation de vins pour une grande variété de plats.

3.2 Preprocessing

Pour chaque corpus de texte, i.e. critiques de vins et critiques de plats, les quatre étapes suivantes vont être appliquées séparément. Un exemple de l'application du preprocessing sur une phrase d'un avis dans le corpus de vin est donné à la TABLE 3.1.

Étape 1	That umami kick meets with black-plum sauce, white pepper and roasted meats.
Étape 2	'umami', 'kick', 'meet', 'blackplum', 'sauc', 'white', 'pepper', 'roast', 'meat'
Étape 3	'umami', 'kick', 'meet', 'blackplum', 'sauc', 'white_pepper', 'roast_meat'
Étape 4	'**umami**', 'kick', 'meet', '**plum**', 'sauc', '**white_pepper**', '**roasted_meat**'

TABLE 3.1 – Exemple d'application du preprocessing sur une phrase d'un avis

3.2.1 Tokenisation

La première étape du préprocessing est de tokeniser les termes du corpus de texte, c'est-à-dire séparer un morceau de texte en unités plus petites appelées tokens. Les tokens peuvent être des groupes de mots, des mots ou des caractères. Dans ce cas-ci, la tokenisation permet de séparer chaque avis en une liste de phrases. Dans la TABLE 3.1, l'étape 1 montre une phrase obtenue après avoir appliqué cette étape.

3.2.2 Normalisation des mots

Chaque phrase du corpus de critiques est normalisée en supprimant la ponctuation du texte brut, en remplaçant les mots par leur forme de base (à l'infinitif pour un verbe et au singulier pour un nom) et en supprimant les stopwords, c'est-à-dire l'ensemble des mots couramment utilisés dans la langue tels que "the", "a", "are" en anglais. Ainsi, dans la TABLE 3.1, on peut voir que de l'étape 1 à l'étape 2, la ponctuation est supprimée, le verbe "meets" est remplacé par son infinitif "meet" et les mots "That", "with", et "and" sont supprimés.

3.2.3 Enrichissement des mots (modèles bi- et tri-grams)

Les modèles bi- et tri-grams permettent de tenir compte de la possibilité que certains des termes que nous voulons extraire des descriptions soient en fait des combinaisons de mots ou de phrases. En effet, les termes qui nous intéressent ne sont pas forcément des mots isolés mais il peut s'agir de termes composés de deux mots ou plus. Pour ce faire, le package *Phrases* de `gensim` est utilisé pour d'abord appliquer un modèle bigram à chaque corpus de texte, i.e. celui du vin et celui de la nourriture. On applique ensuite un modèle trigram à chacun des deux nouveaux corpus de texte. Par exemple, en appliquant successivement les modèles bi et tri-grams sur les mots de la phrase obtenus à l'étape 2, les mots "white" et "pepper" , souvent associés ensemble, deviendront "white_pepper" (voir TABLE 3.4).

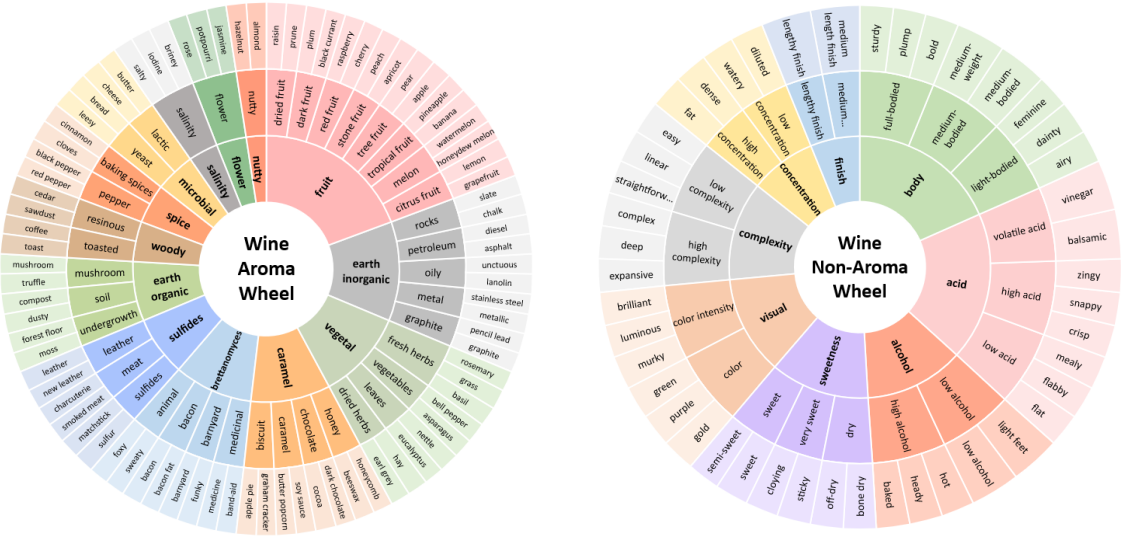
3.2.4 Normalisation des descripteurs de vins et de plats

Une des étapes les plus importantes est la normalisation des descripteurs de vin. En effet, les critiques de vins sont souvent créatives et utilisent parfois des mots différents

pour décrire des choses qui sont apparemment identiques. De plus, il existe un langage particulier dans le monde de la dégustation du vin avec des termes tels que "suave" ou encore "robuste" qui ont des significations spécifiques.

Afin d'uniformiser ce langage, comme mentionné dans la Section 2.3, l'UC Davis a mis au point une roue informatique du vin qui catégorise et fait correspondre divers termes relatifs au vin (Davis, 1980). A l'aide des recherches de Bernard Chen (Chen et al., 2014), ainsi que d'autres contributions (Puckette, 2014 ; Davis, 1980), deux roues à vin RoboSomm ont été créées. Pour construire ces deux roues, les 5 000 termes les plus fréquents relatifs au vin dans le corpus de texte ont été examinés pour, d'une part, déterminer s'ils peuvent être obtenus lors d'une dégustation à l'aveugle et d'une autre part, s'ils sont informatifs (des jugements comme "agréable" et "délicieux" ne sont pas considérés comme informatifs). Les quelques 1 000 descripteurs restants ont ensuite été mis en correspondance avec un descripteur normalisé, une classe déterminée par **arôme ou non-arôme** et dans le cas des non-arômes, une catégorie parmi **poids, sucré, acide, salé, piquant, gras, amer** (Schuring, 2019b, 2021a). Il n'est pas nécessaire de définir des catégories pour les arômes puisque ceux-ci sont utilisés uniquement pour classer les vins selon leur proximité avec les caractéristiques aromatiques des mets tandis que chaque non-arôme permet d'établir certaines règles d'associations : les vins qui ne respectent pas ces règles d'associations sont alors supprimés. Ces règles dépendent exclusivement des non-arômes (voir Section 3.4).

La première roue est une roue d'arômes représentée à la figure 3.1a, qui catégorise une variété de descripteurs aromatiques tandis que la deuxième, à la figure 3.1b, est une roue non-aromatique qui tient compte d'autres caractéristiques, telles que le corps, la douceur et les niveaux d'acidité.



(a) Roue des arômes de vin

(b) Roue des non-arômes de vin

Figure 3.1 – Représentation des deux roues à vin RoboSomm (Schuring, 2019b).

Les termes relatifs au vin peuvent être normalisés selon l'un des trois niveaux de la roue, ou même ne pas être normalisés. Dans notre mémoire, nous avons pris la décision de normaliser les termes du corpus de texte relatif au vin selon le niveau trois, c'est-à-dire la couche extérieure de la roue, car cela permet de caractériser les termes de manière plus spécifique et de tenir compte de plus d'informations. Notons que seuls les termes

correspondants aux 1 000 descripteurs sont normalisés. Par exemple, dans la TABLE 3.4, les descripteurs qui ont été normalisés sont mis en gras ; les autres termes sont soit non informatifs, soit ambigus dans le contexte de l'analyse du vin.

Pour le corpus de texte relatif à la nourriture, la même procédure est appliquée en normalisant seulement les descripteurs d'arôme, c'est-à-dire sans normaliser les descripteurs de non-arômes puisque nous ne disposons pas de descripteurs explicites caractérisant les non-arômes pour les ingrédients : ceux-ci seront donc décrits à partir des arômes du descriptor mapping.

3.3 Embedding des mots du corpus

Ensuite, l'ensemble des données sur le vin et sur les aliments sont combinées afin d'entraîner les embeddings. Il est important d'entraîner le modèle sur l'ensemble des données afin de s'assurer que les embeddings des aliments et du vin soient calculés dans le même espace de caractéristiques, ce qui permettra par la suite de calculer les vecteurs de similarité.

Le modèle appliqué sur le corpus de mots est le modèle `Word2Vec` qui génère, pour chaque mot du corpus combiné, un vecteur de nombres réels à N dimensions résumant son profil sensoriel. Les vecteurs de mots sont positionnés dans l'espace de telle sorte que les mots qui partagent des contextes communs dans le corpus sont situés à proximité les uns des autres dans l'espace.
Notons que le choix des hyper-paramètres optimaux du modèle Word2Vec est spécifique à la tâche réalisée, il est donc difficile d'estimer la dimension du vecteur de ces embeddings (ADEWUMI et al., 2020). Néanmoins, de nombreux articles (MIKOLOV et al., 2013 ; PENNINGTON et al., 2014 ; SARZYNSKA-WAWER et al., 2021) utilisent par défaut une dimension de taille 300, c'est pourquoi nous utiliserons également un embedding dans un espace à $N = 300$ dimensions.

A l'aide de ce modèle, pour chaque vin et chaque ingrédient, nous attribuons un vecteur représentant l'arôme ainsi qu'une valeur pour chaque non-arôme. Ces méthodes d'embedding diffèrent pour le vin et pour les ingrédients : elles sont expliquées séparément dans les parties suivantes.

3.3.1 Embedding du vin

Pour chaque type de vin, nous calculons un vecteur d'arôme et 7 scalaires de non-arômes en implémentant une procédure en quatre étapes : celles-ci sont détaillées ci-dessous et sont visibles sur la FIGURE 3.2.

Avant de réaliser les quatre étapes, il est nécessaire de restructurer notre base de données de critiques de vin. En effet, il est peu probable que les descriptions d'un seul vin contiennent suffisamment d'informations sur tous les non-arômes et les arômes pour produire des recommandations cohérentes et fiables. Par conséquent, nous considérons que les vins sont définis par deux caractéristiques principales : **le cépage** et **la localisation** [1]. Les recommandations sont alors produites pour chaque vin appartenant à un même type

1. La localisation d'un vin est définie par la sous-région, la région, la province et le pays.

défini par le couple (cépage, localisation). Par ailleurs, pour produire des recommandations cohérentes et fiables, il est nécessaire d'avoir un minimum de données pour chaque type de vin. Ainsi, seuls les types de vins qui apparaissent suffisamment fréquemment dans la base de données sont conservés : dans ce cas-ci, les vins doivent apparaître au minimum 30 fois. Parmi les 1802 variétés de vin, 1320 sont supprimées, ce qui réduit la liste à 482 types de vin et le nombre d'avis à 119 527.

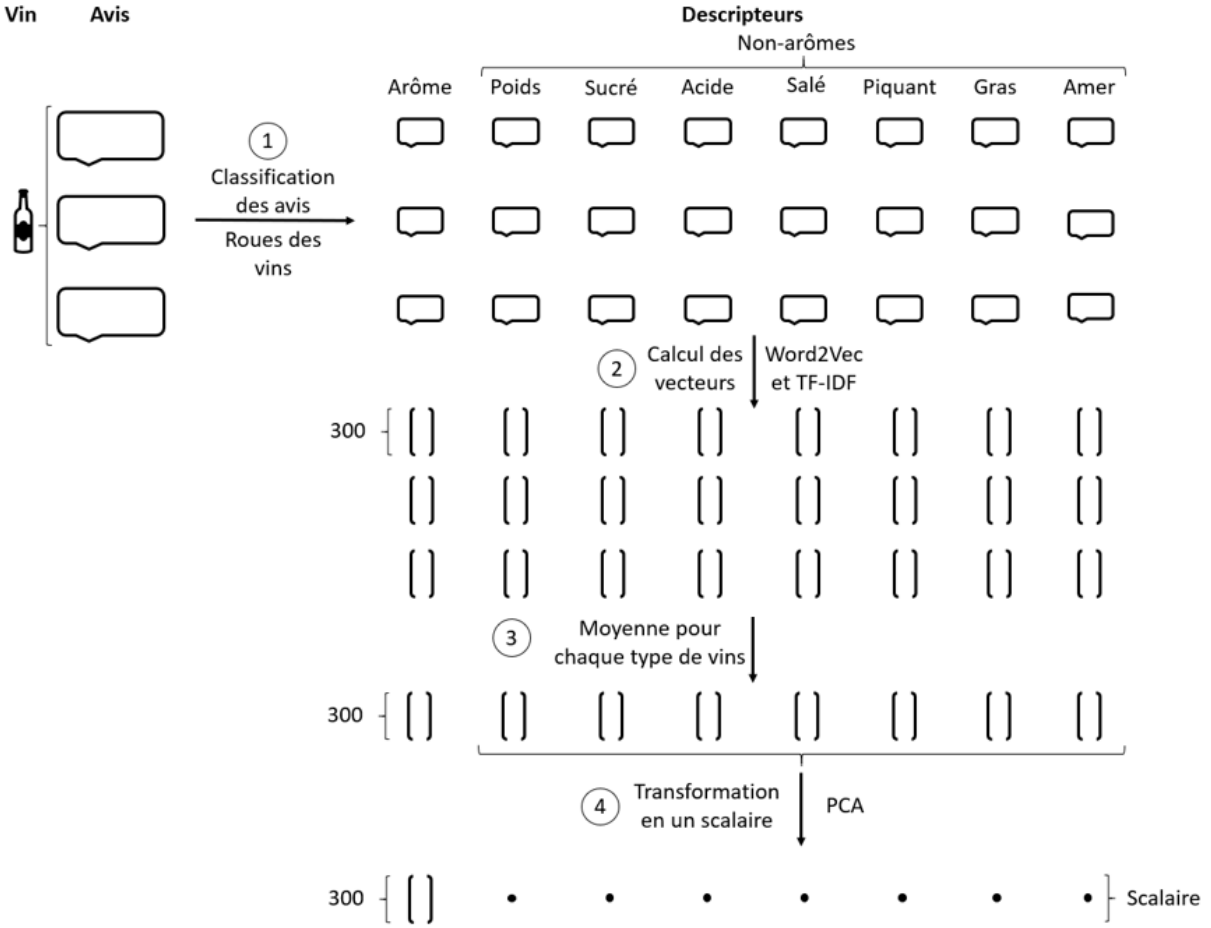


FIGURE 3.2 – Procédure de l'embedding de l'arôme et des non-arômes pour un type de vin.

Étape 1 : Classification des avis selon les arômes/non-arômes

D'abord, les mots de chaque avis sont répartis dans la classe d'arômes ou dans une des sept catégories de non-arômes à l'aide des roues à vin RoboSomm. Ensuite, pour chaque arôme/non-arôme, les termes de chaque avis vont être remplacés par les descripteurs de niveau trois correspondants, les termes ne correspondant à aucun des descripteurs sont alors ignorés. On obtient ainsi une liste de descripteurs normalisés pour chaque arôme/non-arôme pour l'ensemble des avis. Pour rappel, ces descripteurs sont visibles sur la couche extérieure des roues à vin RoboSomm de la FIGURE 5.1. Chaque mot de chaque avis est donc, si possible, mis en correspondance avec un descripteur normalisé qui lui-même appartient à une classe (arôme ou non-arôme) et dans le cas des non-arômes, une catégorie.

Etape 2 : Calcul des vecteurs d'arôme/non-arômes

Avant d'appliquer le modèle `Word2Vec`, nous calculons au préalable une pondération `TF-IDF` : pour chaque arôme/non-arôme, un modèle `TF-IDF` est obtenu à partir de l'ensemble des descripteurs de cet arôme/non-arôme. Cette méthode tient compte de deux facteurs :

— **TF** prend en considération le nombre total de descripteurs par critique. S'il y a 20 descripteurs dans un avis et 5 dans un autre, chaque descripteur individuel dans le premier avis contribue probablement moins au profil général du vin que dans le second.

— **IDF** prend en considération la fréquence à laquelle chaque descripteur apparait dans le corpus de critiques de vin et attribue plus de poids aux descripteurs rares, un mot assez commun à toutes les critiques de vin tel que "doux" aura donc moins de poids qu'un mot plus distinctif tel que "armoise".

Ensuite, à partir du modèle `Word2Vec`, nous pouvons maintenant extraire un vecteur pour chaque descripteur normalisé appartenant à une classe d'arôme/non-arômes et ce vecteur est multiplié par sa pondération `TF-IDF`. Ensuite, pour chaque classe de chaque avis, nous calculons la moyenne des vecteurs des descripteurs normalisés appartenant à cette classe. Nous obtenons donc, pour chaque avis et chaque arôme/non-arômes, un vecteur à 300 dimensions (cf. étape 2 de la Figure 3.2).

Notons que si nous ne disposons pas d'une valeur pour un arôme ou un non-arôme dans un avis, nous prenons simplement la moyenne des valeurs de ce non-arôme sur l'ensemble des avis de la base de données.

Étape 3 : Caclul de l'embedding moyen

Cette étape permet de combiner l'ensemble des vecteurs représentant un même type de vin défini par le couple (cépage, localisation). Pour obtenir l'embedding moyen de chaque type de vin, nous calculons la moyenne des vecteurs sur l'ensemble des vins d'un même type pour chaque arôme/non ârome.

Étape 4 : Transformation des non-arômes en scalaires

Pour appliquer les règles d'association qui dépendent du caractère non-aromatique des vins, nous déterminons un scalaire pour chaque non-arôme : pour cela, un modèle `PCA` (Principal Component Analysis) est construit à partir des vecteurs obtenus à l'étape 3 pour chaque non-arôme. Cette méthode permet de résumer l'information en réduisant le nombre de dimensions : dans notre cas, un vecteur de taille 300 devient un vecteur de taille 1, i.e. un scalaire. Ce modèle permet donc d'obtenir un scalaire pour chaque non-arôme de chaque type de vin.

Par la suite, pour pouvoir interpréter les scalaires de non-arômes, il faut d'abord les normaliser entre 0 et 1. Puis, pour comparer les caractéristiques de non-arômes des vins et des ingrédients, toutes les valeurs des non-arômes sont normalisées sur la même échelle pour obtenir une valeur entière comprise entre 1 (peu similaire) et 4 (très similaire) : pour cela, nous pourrions utiliser la méthode des quartiles qui sépare le nombre de valeurs en quatre groupes équivalents de sorte que, pour chaque non-arôme, chaque intervalle englobe un quart des ingrédients ou des vins.
Cependant, cette méthode part du postulat que le nombre de vins dans chaque intervalle

doit être égal, ce qui est faux : il n'existe pas, par exemple, autant de vins faiblement acides que de vins fortement acides (Van Niekerk, 2012). Une autre méthode est donc appliquée : des vins sont choisis avec une teneur faible, moyenne-faible, moyenne-élevée et élevée du non-arôme et ensuite, les limites des intervalles sont choisies manuellement en regardant où ces vins se situent sur l'échelle 0-1. Cette méthode, plus subjective que celle des quartiles, nous semble plus cohérente en terme de descriptions des vins car elle permet de tenir compte du fait que le nombre de vins dans chaque intervalle n'est pas forcément égal. La Table 3.2 représente la correspondance des valeurs des non-arômes d'un vin avec les valeurs normalisées entre 1 et 4.

	1	2	3	4
Poids	[0, 0.25]	]0.25, 0.45]	]0.45, 0.75]	]0.75, 1]
Sucré	[0, 0.25]	]0.25, 0.6]	]0.6, 0.75]	]0.75, 1]
Acide	[0, 0.05]	]0.05, 0.25]	]0.25, 0.5]	]0.5, 1]
Salé	[0, 0.15]	]0.15, 0.25]	]0.25, 0.7]	]0.7, 1]
Piquant	[0, 0.15]	]0.15, 0.3]	]0.3, 0.6]	]0.6, 1]
Gras	[0, 0.25]	]0.25, 0.5]	]0.5, 0.7]	]0.7, 1]
Amer	[0, 0.2]	]0.2, 0.37]	]0.37, 0.6]	]0.6, 1]

Table 3.2 – Correspondance des valeurs de chaque non-arôme d'un vin avec les entiers de 1 à 4 (Schuring, 2021a).

3.3.2 Embedding d'un plat

Pour calculer l'embedding d'un plat, il est nécessaire de le décomposer en ses différents ingrédients et pour chaque ingrédient, nous calculons un vecteur d'arôme et 7 scalaires de non-arômes à l'aide du modèle `Word2Vec`. Cependant, nous devons adopter une approche différente de celle de l'embedding du vin étant donné que la structure des bases de données est différente : pour les vins, chaque vin est associé à une critique tandis que pour les mets, la base de données regroupe un ensemble d'avis qui ne se rapporte pas à un met spécifique. Il n'est donc plus pertinent d'entraîner un modèle `TF-IDF` sur notre base de données, ni d'appliquer PCA pour transformer les non-arômes en scalaires puisque nous ne disposons pas d'un avis pour chaque ingrédient. Nous allons donc adopter la stratégie suivante.

D'une part, pour le calcul du vecteur d'arôme, chaque ingrédient sera directement représenté par un vecteur d'arôme à 300 dimensions à l'aide le modèle `Word2Vec`. On supposera que le vecteur d'arôme d'un plat complet est donné par la moyenne des vecteurs des ingrédients.

D'autre part, pour le calcul des 7 scalaires de non-arômes pour chaque ingrédient d'un plat, nous ne disposons pas de descripteurs explicites caractérisant les non-arômes pour les ingrédients. Afin de remédier à ce problème, chaque non-arôme est caractérisé par une liste d'ingrédients le représentant. Les différents aliments suivants ont été choisis pour représenter les 7 non-arômes.

— **Poids** : lourd, cassoulet, corsé, épais, lait, viande hachée, steak, gras, pizza, pâte, crémeux, pain

— **Sucré** : sucré, sucre, gâteau, mangue, stevia

— **Acide** : acide, aigre, vinaigre, yaourt, ceviche

— **Salé** : salé, parmesan, huître, pizza, bacon, salaison, saucisse, chips

— **Piquant** : épicé

— **Gras** : gras, friture, crémeux, cassoulet, foie gras, beurre, gâteau, saucisse, brie, carbonara

— **Amer** : amer, chou frisé

Ainsi, à chaque non-arôme, nous associons la moyenne des vecteurs des ingrédients le représentant : chaque non-arôme est donc représenté par un vecteur à 300 dimensions obtenu à partir du modèle `Word2Vec`.

Puis, pour chaque non-arôme, un score de similarité est calculé à l'aide de la distance en cosinus entre le vecteur d'arôme représentant le plat et chaque vecteur représentant les non-arômes, ce qui résulte en sept scores de similarité au total.
Ensuite, pour normaliser ces sept non-arômes en valeurs comprise entre 0 et 1, la distance en cosinus entre l'embedding de chaque non-arôme et les vecteurs d'arôme d'une gamme d'aliments courants est calculée. En effet, on définit la distance maximale entre chacun des embeddings des non-arômes et cette gamme d'aliments. Les aliments qui ressemblent le moins et le plus à chaque non-arôme nous permettront ainsi de créer une échelle normalisée entre 0 (très dissemblable) et 1 (très similaire). A l'aide d'un curseur MinMax, cette échelle est utilisée pour normaliser le vecteur constitué des sept non-arômes du plat.

Considérons un exemple afin de mieux comprendre comment les non-arômes d'un plat sont normalisés. Si nous considérons un hot-dog, celui-ci peut être décomposé en plusieurs ingrédients : pain, tomate, oignon, cornichon, saucisse, moutarde, ketchup. Le vecteur d'arôme du hot-dog est donné par la moyenne des vecteurs d'arôme des ingrédients. Pour obtenir les scalaires associés à chaque non-arôme, on calcule un score de similarité entre ce vecteur d'arôme et les vecteurs représentant chacun des non-arômes. Il faut ensuite normaliser chaque scalaire de non-arômes. Prenons l'exemple du non-arôme "salé" : en calculant la distance entre le non-arôme "salé" et une série d'aliments, on observe sur la FIGURE 3.3 que l'aliment le plus proche de l'embedding du non-arôme "salé" est le bacon et le plus éloigné est la framboise. On normalise alors le scalaire du non-arôme "salé" du plat par rapport aux valeurs du bacon et de la framboise. Le hot-dog est alors classé sur l'échelle de salinité avec une valeur de 0,9.

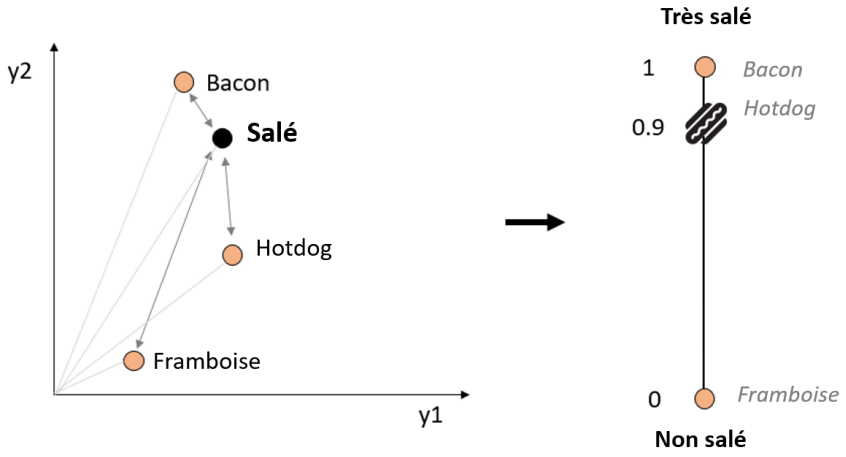


FIGURE 3.3 – Représentation en deux dimensions des vecteurs d'arômes des ingrédients *bacon* et *framboise*, du repas *hot-dog* et du non-arôme *salé* (SCHURING, 2019a).

En appliquant cette procédure pour chaque non-arôme, nous obtenons un profil de saveur approximatif représentant le hot-dog, celui-ci est observable sur la FIGURE 3.4. On remarque que le hot-dog est particulièrement acide probablement dû aux cornichons, il est également gras et épicé mais peu amer et peu sucré.

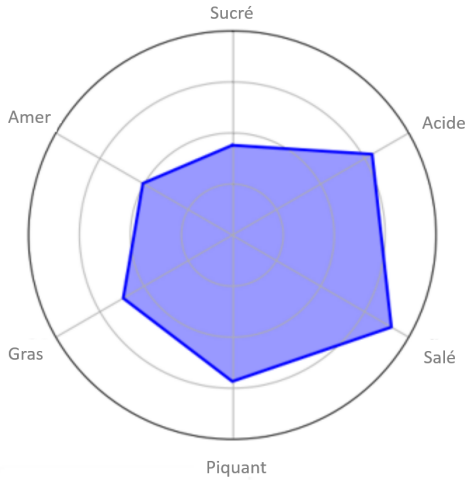


FIGURE 3.4 – Profil de saveur du repas hot-dog (SCHURING, 2019a).

De nouveau, la comparaison des caractéristiques de non-arômes des vins et des ingrédients doit se faire selon la même échelle, toutes les valeurs sont donc normalisées pour obtenir une valeur entière comprise entre 1 (peu similaire) et 4 (très similaire) : la TABLE 3.3 représente la correspondance des valeurs des non-arômes d'un plat avec les valeurs normalisées entre 1 et 4. De la même manière que pour les vins, les limites des intervalles pour chaque non-arôme ont été choisies manuellement en considérant des ingrédients avec une teneur faible, moyenne-faible, moyenne-élevée et élevée du non-arôme pour ensuite regarder où ils se situaient sur l'échelle 0-1.

	1	2	3	4
Poids	[0, 0.3]	]0.3, 0.5]	]0.5, 0.7]	]0.7, 1]
Sucré	[0, 0.45]	]0.45, 0.6]	]0.6, 0.8]	]0.8, 1]
Acide	[0, 0.4]	]0.4, 0.55]	]0.55, 0.7]	]0.7, 1]
Salé	[0, 0.3]	]0.3, 0.55]	]0.55, 0.8]	]0.8, 1]
Piquant	[0, 0.4]	]0.4, 0.6]	]0.6, 0.8]	]0.8, 1]
Gras	[0, 0.4]	]0.4, 0.5]	]0.5, 0.6]	]0.6, 1]
Amer	[0, 0.3]	]0.3, 0.5]	]0.5, 0.65]	]0.65, 1]

TABLE 3.3 – Correspondance des valeurs de chaque non-arôme d'un plat avec les entiers de 1 à 4 (SCHURING, 2021a).

3.4 Association d'un vin à un plat

Dans un premier temps, une étape de **suppression** est nécessaire afin d'éliminer des associations non-aromatiques de vins avec des repas qui n'ont pas de sens. Pour cela, des règles

d'association ont été établies, celles-ci sont inspirées des règles d'association des vins exposées des livres (R. Harrington, 2007 ; Puckette et Hammack, 2018 ; Szabo, 2013 ; Van Niekerk, 2012). Ensuite, la seconde étape est le **tri** des vins restants en fonction des caractéristiques aromatiques afin d'établir un classement des vins à recommander.

3.4.1 Suppression

Les règles d'association suivantes sont appliquées successivement :

1. Le poids du vin, noté p_v, doit correspondre approximativement au poids de l'aliment, noté p_a : $p_a - 1 \leq p_v \leq p_a$.
2. Exclure tous les vins dont les attributs non aromatiques ne s'accordent pas avec ceux de l'aliment en utilisant quelques règles empiriques :
   - Le vin doit être plus acide que le mets.
   - Le vin doit être plus sucré que le mets.
   - Les vins amers ne s'accordent pas avec les mets amers.
   - Les vins/mets amers ne s'accordent pas avec les vins/mets piquants.
   - Les vins/mets amers ne s'accordent pas avec les vins/mets acides.
   - Les vins/mets piquants ne s'accordent pas avec les vins/mets acides.

   Notons que ces règles empiriques sont appliquées uniquement si elles retiennent un nombre suffisant de vins dans la sélection, i.e. un minimum de 5 vins a été décidé dans le code Python (Schuring, 2021a).

3.4.2 Tri

Classer les vins restants en fonction de l'écart, i.e. la distance en cosinus, entre l'embedding de l'arôme du vin et l'embedding de l'arôme du repas. Les recommandations sont triées par ordre croissant de distance : la priorité est donc donnée aux vins qui partagent des caractéristiques aromatiques avec le repas.

4 Validation de l'embedding

Avant d'évaluer la performance du système de recommandation de vins pour un plat spécifique, nous estimons la performance des méthodes d'embedding utilisées et expliquées dans la Section 3.3 permettant de transformer les plats et les vins en des vecteurs (nous n'évaluons donc pas encore les recommandations mais les étapes préliminaires permettant d'obtenir les recommandations). Dans un premier temps, nous allons expliquer notre méthode de validation. Cette méthode se base sur une approche incluant un ensemble de catégories de type d'aliments/vins. Nous allons donc, pour la nourriture et pour les vins, définir ces catégories pour ensuite appliquer et analyser les résultats de notre méthode de validation. Finalement, nous présenterons quelques pistes d'améliorations de la méthode de validation.

4.1 Méthode de validation générale

Pour valider les méthodes d'embedding, nous appliquons une approche assez intuitive : nous définissons, pour la nourriture et pour les vins, un ensemble de catégories et nous considérons l'hypothèse forte selon laquelle des éléments (aliments ou vins) appartenant à une certaine catégorie sont relativement plus proches en terme de goût que des éléments de catégories différentes. Si la méthode d'embedding fonctionne, les éléments d'une même catégorie devraient donner des vecteurs d'arôme/non-arômes plus proches que des éléments de catégories différentes. Notons que cette méthode de validation repose donc largement sur la définition des différentes catégories : ces catégories regroupent des éléments ayant des caractéristiques communes et sont définies dans les Sections 4.2 et 4.3.

Concrètement, la procédure que nous avons appliquée est décrite dans l'**Algorithme** 1. Notons que, dans cet algorithme, t vaut 300 (respectivement 7) si on compare les vecteurs d'arôme (respectivement non-arômes).

Par ailleurs, afin d'évaluer la distance, nous comparons trois mesures différentes régulièrement utilisées dans la littérature scientifique, particulièrement dans le domaine du Machine Learning (Bora et Gupta, 2014 ; Hasnat et al., 2013 ; Vadivel et al., 2003). Celles-ci sont présentées ci-dessous, avec $x, y \in \mathbb{R}^t$:

1. La distance de Manhattan : $dist(x, y) = \|x - y\|_1 = |x_1 - y_1| + \cdots + |x_t - y_t|$.
2. La distance euclidienne : $dist(x, y) = \|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + \cdots + (x_t - y_t)^2}$.
3. La distance en cosinus : $dist(x, y) = \frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2} \cdot \sqrt{\sum_{i=1}^{t} y_i^2}}$.

Algorithme 1 Méthode de validation générale de l'embedding

Input:
N : nombre d'itérations de la méthode
liste_categories : liste des différentes catégories
element_categorie : liste contenant les éléments considérés et leur catégorie associée
$dist$: type de distance

Output: hit_rates_N : taux de succès de l'embedding

hit $\leftarrow 0$
for i=*1* **to** N **do**
  // Sélectionner aléatoirement 2 catégories distinctes
  $\texttt{categorie}_A, \texttt{categorie}_B \leftarrow$ Selection_Aleatoire(liste_categories,2)

  // Sélectionner aléatoirement 2 éléments distincts de la catégorie A dans la liste "element_categorie" et calculer les vecteurs d'arôme/non-arômes associés
  $x_A \leftarrow [x_1, x_2, \cdots, x_t]$ // Élément 1 de la catégorie A
  $y_A \leftarrow [y_1, y_2, \cdots, y_t]$ // Élément 2 de la catégorie A

  // Sélectionner aléatoirement 1 élément de la catégorie B dans la liste "element_categorie" et calculer le vecteur d'arôme/non-arômes associé
  $z_B \leftarrow [z_1, z_2, \cdots, z_t]$ // Élément de la catégorie B

  // Déterminer si les éléments de la même catégorie donnent des vecteurs plus proches
  if $dist(x_A, y_A) < dist(x_A, z_B)$ **then**
    $hit \leftarrow hit + 1$
  end
  $hit\_rates_i \leftarrow \frac{hit}{i}$
end
return hit_rates_N

4.2 Validation de l'embedding des plats

4.2.1 Catégories de nourriture

Pour la nourriture, puisqu'il n'existe pas de catégories de nourriture universelles, nous considérons 13 catégories à partir de catégories définies dans différents articles scientifiques (Ahn et al., 2011 ; Gannon et al., 2008 ; Moltedo et al., 2018) afin d'associer les aliments relativement proches en terme de goût à une même catégorie : plus spécifiquement, nous considérons une catégorie si elle était présente dans au moins deux des trois articles considérés. Les catégories et leurs descriptions sont données dans la Table 4.1 ci-dessous.

Ensuite, à partir d'une liste d'aliments courants, nous avons sélectionné de manière aléatoire des aliments et nous les avons associés manuellement à une catégorie spécifique jusqu'à ce que chaque catégorie contienne un minimum de 5 aliments. En appliquant cette procédure, nous obtenons alors une liste de 291 aliments et leur catégorie corres-

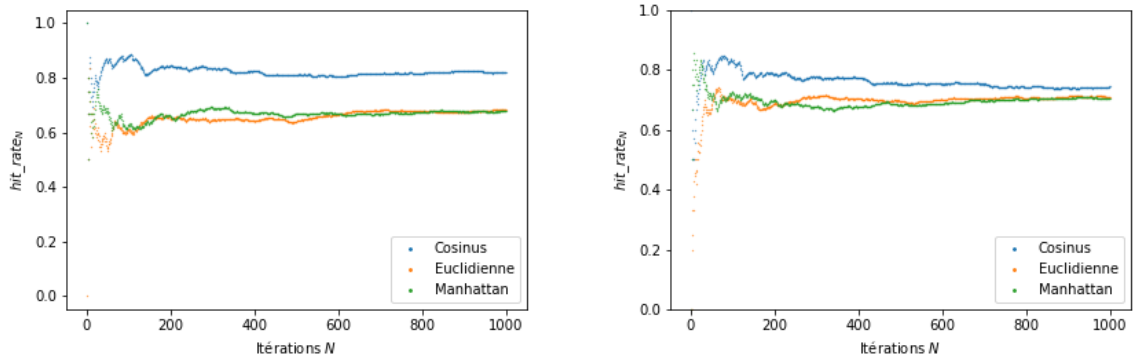
pondante. Les aliments qui semblaient n'appartenir à aucune de ces catégories, telle que la mayonnaise qui peut être à base d'oeufs ou de citron, n'ont pas été considérés.

Catégorie	Description
Fruits	Fruits frais (pommes, poires, agrumes, etc.), fruits congelés et en conserve
Noix et graines	Noix et graines grillées ou non (noix de pécan, noix de cajou, noisettes, amandes, etc.)
Poissons et fruits de mer	Poissons et crustacés frais et secs
Produits laitiers	Lait (entier, demi-écrémé, écrémé, etc.), fromage (fromage blanc, à pâte molle, à pâte dure, etc.), yaourt (y compris de chèvre, de brebis et de soja) et autres produits laitiers
Épices	Poivre, coriandre, vanille, cumin, noix de muscade, etc
Viandes	Bacon, jambon, boeuf, veau, agneau, porc, poulet, dinde, etc et autres produits majoritairement à base de viande (y compris les conserves de viande)
Herbes	Thym, origan, menthe, persil, etc
Dérivés de plantes	Infusion, thé (vert, noir, jasmin, etc.), café, huile végétale, vinaigre, moutarde, sauce soja et autres produits à base de plantes
Sucreries	Boissons gazeuses, jus et aliments sucrés tels que chocolats, bonbons, biscuits, confitures et gâteaux
Légumes	Légumes frais (carottes, tomates, brocolis, salades et pois, haricots, etc.), légumes surgelés et en conserve
Oeufs	Oeufs (de poule, de canard, de pintade, etc.) et plats à base d'oeufs
Céréales	Blé, sorgho, millet et toutes autres céréales ou aliments fabriqués à partir de ces céréales (pâtes, riz, pain, nouilles, etc.)
Pommes de terre	Pommes de terre bouillies, écrasées et frites/rôties, salades et plats de pommes de terre

TABLE 4.1 – Description des différentes catégories de nourriture.

4.2.2 Résultats

En utilisant la liste d'aliments associés avec leur catégorie, nous avons calculé, pour les embeddings du vecteur d'arôme/non-arômes, les valeurs de hit_rate_N pour un nombre d'itérations maximum de 1000 et nous obtenons les résultats présentés à la FIGURE 4.1.



(a) Embedding d'arôme (b) Embedding de non-arômes

FIGURE 4.1 – Évolution de hit_rate_N en fonction du nombre d'itérations N et selon les trois types de distance pour l'embedding d'arôme et de non-arômes des aliments.

Tout d'abord, nous observons sur les deux graphes de la FIGURE 4.1 que les valeurs de hit_rate_N se stabilisent après environ $N = 500$ itérations et convergent vers des valeurs relativement inférieures à 1 pour les trois types de distance considérés. L'erreur restante pourrait s'expliquer par le fait que des aliments d'une même catégorie n'ont pas forcément le même goût : par exemple, la courgette, le panais et la tomate appartiennent à la même catégorie mais ont des goûts variés ou encore les épices telles que la vanille et le poivre n'ont pas du tout des goûts similaires. Ainsi, par exemple, si nous considérons les aliments vanille et poivre de la catégorie "épices" et meringue de la catégorie "sucreries" et en calculant leurs vecteurs d'arôme dénotés $x_{vanille}, x_{poivre}$ et $x_{meringue}$, on obtient :

$$dist(x_{vanille}, x_{poivre}) = 0.73 > dist(x_{vanille}, x_{meringue}) = 0.63$$

Ce résultat semble cohérent puisque, intuitivement, la vanille et la meringue semblent plus proches en terme de goût que la vanille et le poivre. Cependant, dans notre méthode de validation, cela impactera négativement le taux de succès puisque la vanille et la meringue n'appartiennent pas à la même catégorie.

Cependant, pour les trois types de distance considérés, ces valeurs sont supérieures à 0.5 qui correspond à la valeur obtenue de $hit\_rate_{N\to\infty}$ si on définit de manière aléatoire lesquels des trois ingrédients sont les plus proches.

Par ailleurs, les résultats dépendent du type de distance considérée. En effet, alors que les valeurs de hit_rate_N avec la distance euclidienne et la distance de Manhattan convergent vers 0.69, la distance en cosinus donne de meilleurs résultats puisque la valeur de hit_rate_N converge vers 0.82 dans le cas de l'embedding d'arôme (FIGURE 4.1a) et vers 0.74 dans le cas de l'embedding des non-arômes (FIGURE 4.1b). Cela pourrait s'expliquer par le fait que la distance en cosinus ne considère que les angles entre les vecteurs alors que les deux

autres distances évaluent la distance entre ces vecteurs : des aliments proches pourraient avoir des angles similaires dans l'espace mais une distance plus grande. Cela justifie donc le choix de la distance en cosinus pour l'embedding des ingrédients (Schubert, 2021).

Grâce à notre méthode de validation, nous avons pu montrer que la méthode d'embedding implémentée permet de déterminer quels aliments sont les plus proches de manière plus performante que l'aléatoire. Cependant, notre méthode de validation présente quelques défauts : pour l'améliorer, différentes pistes sont données dans la sous-section suivante.

4.2.3 Pistes d'amélioration

Un des défauts de notre méthode de validation est la définition de nos catégories. En effet, pour chacun des 291 aliments, nous avons associé manuellement une des catégories définies à la sous-Section 4.2.1 : ce processus non automatisé peut impacter négativement les résultats de notre méthode.
De plus, certains aliments au sein d'une même catégorie ne sont pas toujours similaires en terme de goût, comme c'est le cas des légumes ou des épices. Il serait donc intéressant de déterminer d'autres catégories plus cohérente en terme de goût, ou de se focaliser sur une méthode de validation qui ne nécessite pas de définir des catégories. Nous avons réfléchi à plusieurs améliorations concrètes pour solutionner ce défaut :

— Nous pourrions redéfinir nos catégories en terme de goût, par exemple en fonction de la chimie du goût, c'est-à-dire les molécules perçues en bouche lors de la dégustation d'aliments. Cela est cependant très complexe et nécessite des connaissances et des études plus approfondies .

— Nous pourrions établir une méthode de validation sans catégorie prédéfinie, deux procédures pourraient être implémentées :

1. On génère trois ingrédients et détermine manuellement lesquels sont les deux plus proches en terme de goût, ensuite on applique une méthode tel que $hit = 1$ si la distance minimale entre les trois paires d'ingrédients correspond à la même paire d'ingrédients que notre décision faite au préalable, $hit = 0$ dans le cas contraire.
2. Avec la méthode d'embedding, on génère les 10 aliments les plus proches d'un aliment spécifique, on vérifie ensuite que ces aliments font sens en terme de goût. Cette méthode est subjective mais permet d'apporter une confiance vis-à-vis de la méthode d'embedding implémentée.

4.3 Validation de l'embedding des vins

4.3.1 Catégories de vin

Comme pour la nourriture, il est nécessaire de commencer par définir les catégories de vins. Pour ce faire, nous considérons deux possibilités :

— Les catégories sont définies par le cépage des vins : deux vins ayant le même cépage appartiennent à la même catégorie.

— Les catégories sont définies par le cépage des vins et leur localisation (sous-région, région, province, pays) : deux vins ayant le même cépage et exactement la même localisation appartiennent à la même catégorie.

Pour rappel, dans le système de recommandation de base, chacun des 482 types de vins [1] est caractérisé par la moyenne de l'ensemble des vecteurs d'arôme/non-arômes des avis d'un vin appartenant à un même type (cf. étape 3 de l'embedding du vin dans la sous-Section 3.3.1). Ensuite, pour chaque non-arôme, nous appliquons PCA sur les 482 vecteurs de non-arômes, ce qui les réduit en un scalaire (cf. étape 4 de l'embedding du vin dans la sous-Section 3.3.1) : chaque type de vins est finalement représenté par un vecteur d'arômes à 300 dimensions et 7 scalaires de non-arômes.

Cependant, regrouper les vins par même type ne nous permet pas d'appliquer notre méthode d'évaluation puisque la majorité des catégories ne comporte qu'un seul vin : si on considère le cépage et localisation comme catégorie de vins, nous n'aurons qu'un vin par catégorie et si on considère uniquement le cépage, nous aurons 40 catégories sur 63 qui contiendront un seul vin. Il est donc impossible de déterminer si différents vins d'une même catégorie sont proches puisqu'il n'y a plus qu'un seul vin par catégorie.

D'une part, pour évaluer l'embedding des arômes, on évalue les vecteurs d'arômes de chaque avis obtenu après l'étape 2 de l'embedding du vin (cf. sous-Section 3.3.1), c'est-à-dire avant d'avoir regrouper les avis des vins par même type. Notons que nous considérons uniquement les avis des vins qui contiennent des descripteurs aromatiques, i.e. dont le vecteur d'arôme n'est pas vide.

D'autre part, pour évaluer l'embedding des non-arômes, nous devons adapter les étapes de l'algorithme de système de recommandation : nous n'appliquons pas l'étape 3 de l'embedding des vins et passons directement de l'étape 2 à l'étape 4. Autrement dit, nous ne regroupons pas les avis des vins par même type [1] et nous appliquons PCA sur l'ensemble des vecteurs de non-arômes des vins de chaque avis, i.e. sur les 119 527 vecteurs de non-arômes, qui sont alors réduits en 119 527 scalaires. Ainsi, chaque avis correspondant à un vin est finalement représenté par un vecteur de 7 scalaires de non-arômes.

Étant donné qu'il existe plusieurs avis par vin, nous aurons dès lors un nombre suffisant de vecteurs d'arôme/non-arômes par catégorie, que ce soit pour les catégories définies par le cépage et la localisation ou celles définies uniquement par le cépage.

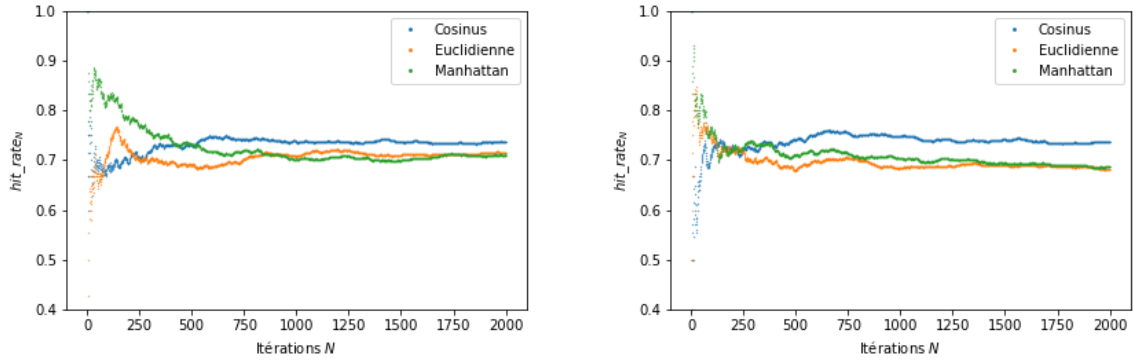
Cependant, la méthode d'évaluation de l'embedding des non-arômes de vin est impossible à implémenter en pratique : en effet, PCA nécessite le calcul de la matrice de corrélation qui aurait une taille 119 527$\times$ 119 527, ce qui requière beaucoup de mémoire. Nous allons donc présenter uniquement les résultats de l'évaluation des embeddings d'arôme des vins.

4.3.2 Résultats

Puisque les données nécessaires aux catégories sont déjà fournies dans la base de données, nous pouvons directement appliquer la procédure et calculer le hit_rate_N pour un nombre d'itérations de $N = 1$ à 2000 : cette valeur de N permet d'assurer une stabilisation du hit_rate_N pour tous les types de distance considérés. Nous obtenons les résultats visibles à la FIGURE 4.2.

On observe que les deux méthodes donnent des résultats similaires : les valeurs de hit_rate_N convergent autour de 0.7 pour les trois types de distance et selon les deux types de catégorie considéré. Comme pour la nourriture, cela est relativement inférieur à 1 mais reste supérieur à 0.5 qui correspond à la valeur obtenue de $hit\_rate_{N\to\infty}$ si on définit de ma-

1. Un type de vins est défini par un cépage et une localisation.



(a) Cépage (b) Cépage et localisation

FIGURE 4.2 – Évolution de hit_rate_N en fonction du nombre d'itérations N pour l'embedding d'arôme des vins selon les différentes catégories considérées.

nière aléatoire lesquels des trois vins sont les plus proches.
Par ailleurs, pour les deux types de catégories, la distance en cosinus donne des résultats légèrement meilleurs : en effet, pour la catégorie définie par le cépage (resp. le cépage et la localisation), hit_rate_N converge vers 0.74 avec la distance en cosinus alors qu'il converge vers 0.71 (resp. 0.68) avec les deux autres distances. Puisque la distance en cosinus donne de meilleurs résultats pour l'embedding du vin et de la nourriture, cela semble cohérent de l'utiliser pour trier les recommandations de vins selon leur proximité avec les caractéristiques aromatiques des ingrédients.
De plus, la similitude des résultats entre les deux types de catégories peut se justifier par le caractère dominant du cépage dans la caractérisation des vins (CAMPO et al., 2005 ; PALOMO et al., 2007 ; PALOMO et al., 2005).

Finalement, nous pouvons tirer la même conclusion que pour l'évaluation de l'embedding des plats : la méthode d'embedding implémentée donne des résultats plus performants que l'aléatoire mais il serait également intéressant d'évaluer l'embedding des non-arômes.

4.3.3 Piste d'amélioration

Il serait intéressant d'évaluer les vecteurs d'arômes et les scalaires de non-arômes obtenus après avoir regroupé les avis des vins par type, autrement dit après avoir appliqué l'étape 3 et 4 de l'embedding du vin (voir sous-Section 4.3) : on aurait alors un vecteur d'arôme et sept scalaires de non-arômes pour chacun des 482 types de vins. Pour cela, une possibilité est de définir d'autres catégories de manière à ce qu'il y ait suffisamment de vins par catégories. Par exemple, dans (PUCKETTE, 2015), l'auteure sépare les vins blancs et les vins rouges en 3 catégories : d'une part, vin sec, vin doux et vin riche pour les vins blancs et d'autres part, vin léger, vin moyen et vin puissant pour les rouges. Pour pouvoir appliquer la méthode de validation avec ces catégories, il faudrait donc d'abord associer chacun des 482 types de vins à une catégorie, ce qui nécessiterait une recherche au préalable ou l'aide d'un sommelier.

5 Performances du système de recommandation

Le Chapitre 4 montre que les méthodes d'embedding implémentées dans le système de recommandation sont plus performantes qu'une stratégie aléatoire. Il faut donc maintenant évaluer la performance du système de recommandation dans sa globalité et ainsi répondre à notre question de recherche principale : *est-il possible de recommander de manière performante des vins pour des plats spécifiques en exploitant des données exclusivement trouvées sur Internet ?*. Pour ce faire, nous avons commencé par établir manuellement une matrice de données remplie par des associations de vins avec des repas : ainsi, les performances du système étudié seront comparées avec cette matrice de référence. Ensuite, nous évaluons l'effet des deux étapes principales de l'association d'un vin à un plat présentées dans le Chapitre 3 : la suppression et le tri. Pour cela, nous décrivons deux méthodes d'évaluation distinctes, une basée sur un score "top N" et une basée sur un score de classement, et analysons les résultats de ces méthodes. Finalement, plusieurs pistes d'améliorations du système de recommandation étudié sont données.

5.1 Données d'association vins et repas

Afin de pouvoir évaluer les performances du système de recommandation, nous avons rempli une matrice de données $Data \in \mathbb{R}^{nv \times nr}$, où nv et nr correspondent respectivement au nombre de vins et de repas considérés, tel que

$$Data_{v,r} = \begin{cases} 1 & \text{si le vin } v \text{ s'accorde bien avec le repas } r, \\ 0 & \text{si l'information n'est pas disponible.} \end{cases}$$

Pour remplir cette matrice, nous définissons un ensemble de $nr = 40$ repas et nous nous basons sur les données trouvées sur quatre sites Internet (Beckett, s. d. ; Hammack et al., 2011 ; Jani et Simonetta, s. d. ; PlatsNetVins, s. d.) pour déterminer les associations adéquates. De plus, nous avons rempli la matrice des données $Data_{v,r}$ tel que, à chaque vin correspond au minimum un plat. Cette matrice est présentée en détail dans l'Annexe 7.1.

Ensuite, il est important de remarquer qu'une valeur de 0 dans notre matrice de données pour un certain vin et un certain plat ne signifie pas que ce vin ne s'accorde pas avec ce plat mais simplement que cette donnée n'était pas disponible sur les sites Internet considérés : ces valeurs sont donc considérées comme manquantes. Afin d'enrichir et de valider notre matrice de données, il pourrait être intéressant de prendre contact avec des sommeliers.

Par ailleurs, pour caractériser les vins, nous considérons comme unique critère le cépage puisque celui-ci est le plus déterminant (Campo et al., 2005 ; Palomo et al., 2007 ; Palomo et al., 2005). Ainsi, 2 vins appartiennent à un même type s'ils ont le même cépage, alors que dans le code fourni, les vins sont définis par leur cépage et leur localisation, i.e. 2 vins sont de même type s'ils ont le même cépage et la même localisation. Cela permet de réduire le nombre de vins nv de 482 à 65. Pour cela, les valeurs pour chaque vin (un vecteur pour l'arôme et 7 scalaires pour chaque non-arôme) ont été calculées en prenant la moyenne des données des vins avec un même cépage, quelle que soit la localisation.

Notons que plusieurs obstacles ont été rencontrés lors du remplissage de cette matrice : la liste de vins considérée à partir de la base de données n'étant pas exhaustive, il existait régulièrement des vins qui étaient conseillés sur les sites Internet mais qui ne se retrouvaient pas dans la liste de vins considérée. De plus, les vins "Red Blends" et "White Blends" étant très vagues à associer, ceux-ci ont été supprimés de la liste de vins, ce qui donne au final une liste composée d'un total de $nv = 63$ vins.

5.2 Méthode d'évaluation des performances

Une fois la matrice des données *Data* remplie, nous pouvons évaluer la performance de l'algorithme de recommandation étudié. Rappelons que cet algorithme comporte deux étapes (voir sous-Section 3.4) :

1. **Suppression** : l'algorithme commence par supprimer les vins qui ne s'accordent pas avec le repas en suivant une série de règles d'associations vins/plats.
2. **Tri** : les vins restants sont triés en fonction de la similarité entre l'embedding de l'arôme du vin et l'embedding d'arôme moyen des ingrédients du repas, par ordre décroissant, autrement dit la priorité est données aux vins qui partagent des caractéristiques aromatiques avec les ingrédients du repas.

Nous allons donc étudier l'effet de ces deux étapes séparément et de manière conjointe. Pour cela, nous considérons deux méthodes : la première est basée sur une **recommandation "top N"** et la deuxième est basée sur le **classement des vins pertinents**. Notons que les métriques utilisées dans ces méthodes sont inspirées des métriques `precision` et `recall` largement utilisées dans la littérature. Ces métriques ne peuvent cependant pas s'appliquer à notre cas puisqu'elles supposent que toutes les données sont exactes alors que dans notre matrice de données, un élément mis à 0 signifie simplement que nous n'avons pas d'information.

5.2.1 Méthode basée sur un score "top N"

5.2.1.1 Évaluation du tri des recommandations

Pour pouvoir évaluer le triage des vins avec cette méthode, on s'intéresse à la proportion des N premiers vins du tri qui sont pertinents par rapport à l'ensemble des vins pertinents : les vins pertinents correspondent aux vins qui s'accordent bien avec un repas et qui valent donc 1 dans la matrice de données *Data*. Pour évaluer la performance, nous appliquons alors la procédure présentée dans l'**Algorithme** 2.

Algorithme 2 Méthode "top N" d'évaluation du tri des recommandations

Input:

nr : nombre de repas
nv : nombre de vins
$Data_{v,r}$: Matrice d'association des vins v et repas r
P_r : l'ensemble des vins pertinents pour un repas r

Output: $score_N$: score pour un nombre N de vins sélectionnés

for *r=1* **to** nr **do**
  // Trier les vins selon leurs similarités aromatiques avec le repas r en utilisant la distance en cosinus entre les vecteurs d'arômes
  for *v=1* **to** nv **do**
    $\texttt{Similarites}_{vr} \leftarrow 1 - cosinus\_distance(vecteur\_arome_r, vecteur\_arome_v)$
  end
  $\texttt{Recommandations}_r \leftarrow Tri(\texttt{Similarites}_{vr})$

  // Considérer les N meilleures recommandations de vins données par le tri
  $\texttt{N}_r \leftarrow Top_N(\texttt{Recommandations}_r)$

  // Calculer la proportion de ces vins qui sont pertinents par rapport à l'ensemble des vins pertinents
  $score_{N,r} \leftarrow \frac{\sum_{v \in N_r \cap P_r} Data_{v,r}}{|P_r|}$
end
// Calculer la moyenne des scores sur l'ensemble des repas
$score_N \leftarrow \frac{\sum_{r=1}^{nr} score_{N,r}}{nr}$
return $score_N$

Ensuite, pour comparer ces résultats et ainsi déterminer si l'algorithme étudié exploite et assimile bien les données, nous vérifions que celui-ci est plus performant qu'une stratégie simple et naïve : une stratégie aléatoire. Pour ce faire, nous évaluons également la performance d'un algorithme qui sélectionnerait N vins de manière aléatoire : la procédure suivante est appliquée.

Algorithme 3 Méthode "top N" d'évaluation du tri aléatoire des recommandations

Input:
n_iter : nombre d'itérations de la méthode
nr : nombre de repas
$liste\_vins$: liste contentant l'ensemble des vins
$Data_{v,r}$: Matrice d'association des vins v et repas r
P_r : l'ensemble des vins pertinents pour un repas r
Output: $score\_aleatoire_N$: score pour un nombre N de vins sélectionnés

for *i=1* **to** n_iter **do**
  for *r=1* **to** nr **do**
    // Sélectionner de manière aléatoire N vins parmi l'ensemble des vins
    $A_N \leftarrow$ Selection_Aleatoire(liste_vins,N)

    // Calculer la proportion de ces vins qui sont pertinents par rapport à l'ensemble des vins pertinents
    $score\_aleatoire_{N,r,i} \leftarrow \frac{\sum_{v \in P_r \cap A_r} Data_{v,r}}{|P_r|}$
  end
  // Calculer la moyenne des scores sur l'ensemble des repas et sur l'ensemble des itérations
  $score\_aleatoire_N \leftarrow \frac{\sum_{i=1}^{n\_iter} \sum_{r=1}^{nr} score\_aleatoire_{N,r,i}}{n\_iter \cdot nr}$
end
return $score\_aleatoire_N$

5.2.1.2 Évaluation de la suppression suivi du tri des recommandations

Nous appliquons les mêmes procédures que celles décrites ci-dessus mais avec deux modifications : d'une part, dans le calcul du score de l'algorithme étudié, l'étape de suppression est appliquée avant celle de tri. D'autre part, puisque des vins sont supprimés de la liste, il se peut que le nombre de vins restants après la suppression soit inférieur à N. Le nombre de vins considérés pour l'algorithme étudié est donc donné par le minimum entre N et le nombre de vins restants après la suppression : l'algorithme aléatoire considère alors le même nombre de vins afin que la comparaison entre le système de recommandation et l'algorithme aléatoire soit cohérente. Finalement, la procédure ainsi que le calcul du score sont ensuite les mêmes.

Notons que dans le cas où N est supérieur ou égal au nombre de vins restants après la suppression pour l'ensemble des plats, cela revient à évaluer uniquement la suppression puisque tous les vins restants après la suppression sont sélectionnés et le tri n'a donc pas d'impact.

5.2.2 Méthode basée sur un score de classement

5.2.2.1 Évaluation du tri des recommandations

Pour cette méthode, on s'intéresse au classement moyen des vins pertinents pour un repas dans la liste de recommandations. Pour pouvoir calculer ce classement moyen, nous appliquons la procédure suivante :

Algorithme 4 Méthode de classement d'évaluation du tri des recommandations

Input:
nr : nombre de repas
nv : nombre de vins
P_r : l'ensemble des vins pertinents pour un repas r

Output: *classement_moy* : classement moyen des vins pertinents sur l'ensemble des repas

for *r=1* **to** nr **do**
  // Trier les vins selon leurs similarités aromatiques avec le repas r en utilisant la distance en cosinus entre les vecteurs d'arôme
  for *v=1* **to** nv **do**
    $\texttt{Similarités}_{rv} \leftarrow 1 - cosinus\_distance(vecteur\_arôme_r, vecteur\_arôme_v)$
  end
  $\texttt{Recommandations}_r \leftarrow Tri(\texttt{Similarités}_r)$

  // Calculer la moyenne des classements des vins pertinents dans la liste des recommandations
  $index_{v,r} \leftarrow Index(\texttt{Recommandations}_r, v)$ // Index du vin v dans la liste de recommandation pour le repas r
  $classement\_moy_r \leftarrow \frac{\sum_{v \in P_r} index_{v,r}}{|P_r|}$
end
// Calculer le classement moyen sur l'ensemble des repas
$classement\_moy \leftarrow \frac{\sum_{r=1}^{nr} classement\_moy_r}{nr}$
return *classement_moy*

Nous pouvons facilement comparer notre méthode à un algorithme qui classerait les vins au hasard : en effet, le classement moyen des vins pertinents dans une liste qui classe les vins de manière aléatoire est donné par

$$\frac{nv+1}{2} = \frac{63+1}{2} = 32.$$

Elle est également facile à comparer avec un algorithme qui classerait les vins de manière idéale : les vins pertinents seraient toujours classés en premiers. Dans ce cas, le classement moyen est donné par

$$\frac{\sum_{r=1}^{nr} \frac{|P_r|+1}{2}}{nr} = \frac{nr + \sum_{r=1}^{nr} |P_r|}{2 \cdot nr} = 4.13.$$

Les détails de ce calcul sont donnés dans l'Annexe 7.2.

5.2.2.2 Évaluation de la suppression suivi du tri des recommandations

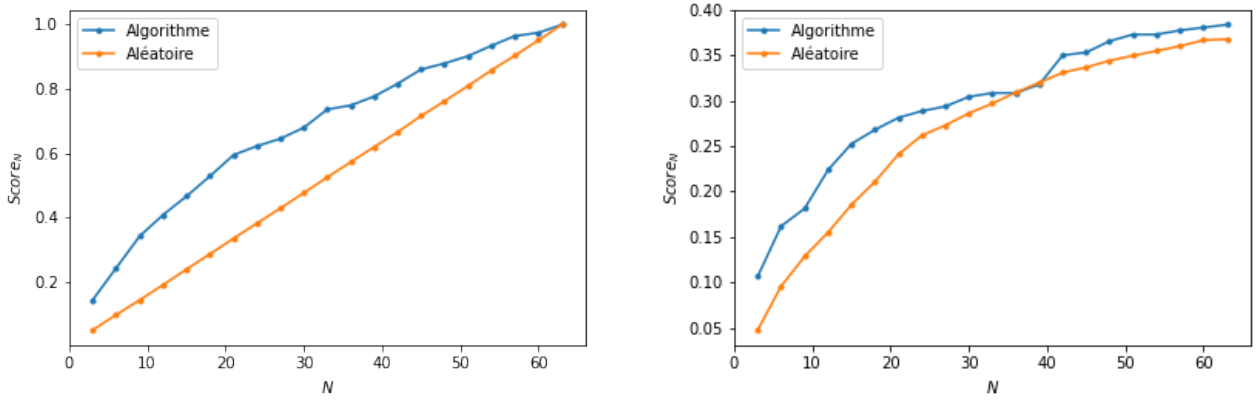
Cette méthode est également presque semblable à celle décrite ci-dessus. Il y a deux différences : d'une part, dans le calcul du classement moyen pour un repas, l'étape de suppression est appliquée avant celle de tri. D'autre part, les index des vins supprimés sont équivalents au classement moyen de l'ensemble des vins supprimés, autrement dit si nv dénote le nombre de vins et nvs le nombre de vins restants après la suppression, l'index de l'ensemble des vins supprimés est donné par $index_{v,r} = \frac{nvs+1+nv}{2}$.

5.3 Résultats

Nous avons appliqué les procédures décrites ci-dessus pour $nr = 40$ repas et pour des valeurs de N allant de 3 à $nv = 63$. Nous obtenons les résultats visibles à la FIGURE 5.1 et à la TABLE 5.1.

5.3.1 Méthode basée sur un score "top N"

Sur la FIGURE 5.1, nous observons tout d'abord que l'algorithme sans et avec l'étape de suppression sont plus performants que l'aléatoire pour la majorité des valeurs de N : cela nous indique que l'algorithme de recommandation étudié a un effet positif puisque les vins sont mieux sélectionnés avec l'algorithme que de manière aléatoire. Notons que les scores de l'algorithme aléatoire des FIGURES 5.1a et 5.1b évoluent différemment puisque l'étape suppression supprime des vins et impacte donc le score.



(a) Tri (b) Tri et suppression

FIGURE 5.1 – Comparaison entre les résultats de la méthode basée sur un score top N pour l'algorithme aléatoire et l'algorithme sans et avec l'étape de suppression.

Pour la méthode d'évaluation du tri des recommandations, nous observons sur la FIGURE 5.1a une valeur de score égale à 1 lorsque $N = 63$: ce résultat est logique puisque dans ce cas, nous sélectionnons l'ensemble des vins et donc la proportion des vins pertinents parmi cette sélection par rapport à l'ensemble des vins pertinents vaut forcément 1.

De plus, sur la FIGURE 5.1b, on peut évaluer l'effet de la suppression seule lorsque $N = 63$ puisque cela revient à sélectionner la totalité des vins. On constate que la suppression est légèrement plus performante que le hasard puisque le score pour l'algorithme vaut 0.384, ce qui est légèrement supérieur à celui donné par l'aléatoire, i.e. 0.367. La suppression permet donc d'améliorer légèrement la sélection des vins. Nous observons également une baisse de performance entre $N = 33$ et $N = 40$ mais nous ne parvenons pas à interpréter et ni à expliquer ses résultats.

Finalement, la suppression ne semble pas améliorer les résultats du tri : en effet, pour de faibles valeurs de N, on peut voir que le score de l'algorithme est meilleur sans la suppression (FIGURE 5.1a) qu'avec (FIGURE 5.1b). Pour des valeurs de N plus élevées, la comparaison est plus difficile puisque le nombre de vins sélectionnés n'est pas forcément le même et peut être fortement plus faible avec la suppression : en effet, le nombre de

vins sélectionnés avec la suppression correspond au minimum entre N et le nombre de vins restants après la suppression. Ainsi, si le nombre de vins sélectionnés est supérieur, le score sera probablement supérieur également.

5.3.2 Méthode basée sur un score de classement

Sur la Table 5.1 ci-dessous, on constate que le tri est moins performant qu'un algorithme idéal mais qu'il est tout de même plus performant que l'aléatoire puisque l'algorithme avec le tri classe en moyenne les vins pertinents en 22ème position. Par contre, l'algorithme avec le tri et la suppression est nettement moins performant qu'un algorithme idéal et seulement légèrement plus performant que l'aléatoire. Cela semble indiquer que la suppression ne permet pas d'améliorer le tri et que l'effet du tri est donc plus efficace que l'effet de la suppression et du tri conjointement.

Tri	Tri et suppression	Aléatoire	Idéal
22.23	30.12	32	4.13

Table 5.1 – Comparaison entre les classements moyens des vins sélectionnés sans et avec l'étape de suppression et de manière aléatoire.

Pour conclure, ces résultats nous permettent donc de répondre positivement à notre question de recherche : une méthode basée sur l'embedding des vins et plats à partir d'avis trouvés sur Internet peut recommander des vins pour des plats spécifiques de manière plus performante que l'aléatoire. De plus, la suppression des vins sur base de certaines règles d'associations ne semble pas améliorer les performances du triage des recommandations de vins selon leur proximité avec les caractéristiques aromatiques des ingrédients du plat. Par ailleurs, il est important de noter que les résultats des algorithmes avec ou sans la suppression peuvent être fortement impactés par le fait qu'un vin qui s'accorde très bien avec un plat pourrait être bien classé par les algorithmes mais ne pas être considéré comme un vin pertinent dans nos données. En effet, un vin non-pertinent, c'est-à-dire mis à 0 dans la matrice de données, ne signifie pas que le vin ne s'associe pas bien avec le repas mais seulement que l'information n'est pas disponible. Cela pourrait donc impacter les résultats obtenus avec nos deux méthodes d'évaluation.
Finalement, même si nous avons montré que ce système de recommandation est prometteur, il peut être amélioré : des pistes d'amélioration sont présentées dans la section suivante.

5.4 Pistes d'amélioration

Lors de l'analyse de l'algorithme de recommandation, plusieurs choix et données peuvent selon nous être critiqués et remis en question. Dans cette section, nous allons aborder ces différentes critiques et y apporter des pistes d'amélioration.

5.4.1 Qualité du corpus de critiques alimentaires

Critique :
Nous utilisons une base de données correspondant à un corpus de critiques d'aliments et repas provenant d'Amazon (Stanford, 2017) : celles-ci peuvent correspondre à des avis

d'aliments pour animaux ou encore des avis de type "Ce sont les mêmes produits que vous pouvez acheter dans les grandes surfaces.". Ce genre de critiques n'apporte pas de réelles informations en terme de goût, de texture et de saveurs de l'ingrédient d'un repas. Or, notre algorithme de recommandation se base sur le goût, la texture et les saveurs, i.e. les non-arômes et les arômes, des ingrédients du repas pour effectuer les recommandations de vins.

Améliorations :
Nous pourrions utiliser d'autres bases de données afin d'enrichir notre base de données initiale : par exemple, la base de données (Alvin, 2021) comprenant plus de 500 000 recettes avec une liste des ingrédients pourrait être utilisée.

5.4.2 Élargissement à d'autres variétés de vins

Critiques :
Premièrement, nous avons remarqué que la liste des vins (Schuring, 2021b) extraite des 150 000 critiques professionnelles n'inclut pas certaines variétés de vins. Celle-ci ne considère ni les vins rosés, ni les vins pétillants : ces vins sont pourtant des incontournables à recommander. Par exemple, le Moscato d'Asti est un vin blanc pétillant d'Italie qui, de par sa douceur et sa faible teneur en alcool, se marie particulièrement bien avec les desserts. De plus, de nombreux vins populaires, tels que le Chablis, le Beaujolais ou encore le Gewürztraminer, ne figurent pas dans la base de données.

Ensuite, lors de la restructuration de la base de données, les vins sont supprimés s'ils apparaissent moins de 30 fois : parmi les 1802 variétés de vins, 1320 sont supprimées, ce qui réduit considérablement la liste à 482 vins. La valeur choisie du nombre minimal d'apparition, i.e. 30, est une décision fortement subjective et qui a un impact important sur le nombre de vins sélectionnés.

De plus, le cépage étant le facteur le plus influant pour caractériser un vin (Campo et al., 2005 ; Palomo et al., 2007 ; Palomo et al., 2005), nous avons pris la décision de ne considérer que celui-ci pour évaluer les performances du système de recommandation, sans la localisation : la liste de vins est donc réduite à 63 vins de cépages différents. Cela modifie fortement les recommandations faites par l'algorithme et peut donc empirer les résultats.

Améliorations :
Premièrement, nous pourrions utiliser une autre base de données, plus exhaustive, qui considère les vins rosés et pétillants, ou enrichir la base de données déjà existante, avec plus de critiques : pour cela, les bases de données (McGuire, 2022 ; Thoutt, 2018) comprenant des avis sur différentes variétés de vins peuvent par exemple être utilisées.

Ensuite, il serait intéressant d'analyser l'impact du choix du nombre minimal d'apparition de vins sur la performance du système de recommandation : le choix optimal peut être trouvé en testant l'impact du nombre minimal d'apparition sur les scores de performance. Il s'agit donc d'un compromis entre précision et diversification : d'une part, on peut restreindre notre base de données en imposant un nombre minimal d'apparition élevé mais garder une représentation précise et significative des caractéristiques du vin puisque le nombre d'avis pour chaque vin est élevé. D'autre part, on peut garder une base de données diversifiée en considérant tous les vins mais perdre en précision.

De plus, nous pourrions décider de considérer la localisation, en plus du cépage, dans la

caractérisation des vins : la liste s'agrandirait de 63 vins à 482 vins. Notre matrice de données serait plus complexe à construire car elle nécessiterait beaucoup plus de données pour être remplie et les informations tirées d'Internet ne sont pas toujours suffisamment précises. En effet, la plupart des sites Internet conseillent des vins uniquement en fonction de leur cépage. De nouveau, nous sommes face à un compromis : est-il plus pertinent de considérer une plus large liste de vins à défaut d'une matrice de données moins dense ou restreindre la liste de vins avec l'avantage d'une plus grande facilité pour remplir la matrice de données.

5.4.3 Règles d'associations

Critique :
Les règles d'associations de vins et repas, utilisées dans la sous-Section 3.4 pour supprimer les associations qui ne font pas sens, ne sont pas exhaustives et universelles. D'autres règles d'associations existent (Arnone et Simonetti-Bryan, 2013 ; Chartier, 2012 ; Hambleton, 2008 ; Puckette et Hammack, 2018 ; Simonetti-Bryan, 2010) mais déterminer quelles règles appliquer pour l'association des vins et repas est complexe.

Amélioration :
Nous pourrions définir un ensemble de règles d'associations de vins, en considérant toutes les règles trouvées dans les articles et livres. Ensuite, nous pourrions mesurer l'impact de chaque règle d'association : pour cela, il faudrait, par exemple, ne pas considérer chacune des règles tour à tour et analyser l'impact engendré sur les performances, ce qui permettrait de pouvoir calibrer les règles aux données.

5.4.4 Choix des goûts principaux

Critique :
Le choix des 6 goûts principaux, i.e. **sucré, acide, salé, amer, piquant et gras**, peut être remis en question. Dans la littérature, la plupart des écrits comme (R. Harrington, 2007 ; IQWiG, 2006 ; Pediaopolis, 1947 ; Van Niekerk, 2012 ; Vilela et al., 2016 ; Wikipedia, 2022b) définissent les 5 goûts principaux suivants : **sucré, acide, salé, amer et umami**.

Amélioration :
Le choix de considérer gras et piquant comme des goûts principaux peut se justifier par les différents débats actuels expliqués dans la Section 2.3 mais nous pourrions intégrer la saveur umami comme 7è goût principal. On pourrait analyser l'impact des choix des différents goûts (dont les règles d'associations considérées dépendent directement) sur les performances et considérer uniquement ceux avec un impact significatif.

5.4.5 Choix des descripteurs non-aromatiques des aliments

Critique :
A la sous-Section 4.2, différents aliments ont été choisis pour décrire les 7 non-arômes. Bien que nous comprenons l'utilité de ces descripteurs, ceux-ci sont choisis de manière trop subjective.

Améliorations :
Afin d'avoir des descripteurs pour les non-arômes définis de manière plus subjective, une stratégie serait de générer des aliments et de les associer, si une correspondance existe

selon nous, avec un des non-arômes. L'idée serait alors que des personnes distinctes élaborent leurs propres associations et ensuite, de considérer un aliment comme descripteur d'un non-arôme uniquement si toutes les personnes ont associé cet aliment au même non-arôme. Il suffirait alors de générer des aliments jusqu'à ce que chaque non-arôme possède un minimum d'aliments le caractérisant. Notons donc que plus le nombre de personnes qui participent à ce processus est élevé, moins la méthode est subjective.
Une autre idée serait de trouver des sources fiables dans la littérature scientifique permettant d'associer des aliments aux non-arômes.
Il serait ensuite intéressant d'étudier l'impact du choix des descripteurs de non-arômes sur les scores de performance.

5.4.6 Agrégation des embeddings d'un plat

Critique :
Lors du calcul d'embedding d'un plat, chaque ingrédient est représenté par un vecteur d'arôme attribué par le modèle `Word2Vec`. Ensuite, l'embedding du plat complet est donné par la moyenne des embeddings des ingrédients : tous les ingrédients d'un plat ont donc le même poids dans le calcul de la moyenne. Par exemple, les ingrédients "chou-fleur" et "oignon" dans une soupe de chou-fleur auront la même importance dans les décisions d'associations, ce qui ne semble pas très cohérent.

Améliorations :
Pour adapter la méthode, on pourrait utiliser des poids représentant l'importance des ingrédients du repas : l'embedding du plat correspondrait alors à la moyenne pondérée des vecteurs d'arôme des ingrédients.
On considère deux stratégies pour le calcul des poids : la première est d'utiliser un poids pour chaque ingrédient défini par l'utilisateur lui-même. Ces poids seront ensuite normalisés et utilisés pour le calcul de la moyenne pondérée. Cependant, si le plat nécessite beaucoup d'ingrédients, cela risque d'être contraignant pour l'utilisateur de définir des poids pour chacun des ingrédients. La deuxième stratégie serait donc de définir uniquement deux catégories : les ingrédients principaux et les ingrédients secondaires. Nous appliquerions donc un poids différent en fonction de la catégorie dans lequel l'ingrédient appartient.

5.4.7 Dimension des vecteurs dans le modèle Word2Vec

Critique :
Lorsque le modèle Word2Vec est construit, la dimension des vecteurs est fixée à 300. Même si c'est une valeur qui est utilisée par défaut dans de nombreux articles (Mikolov et al., 2013 ; Pennington et al., 2014 ; Sarzynska-Wawer et al., 2021), elle n'est pas forcément idéale.

Amélioration :
Il serait intéressant de considérer d'autres valeurs de dimension et d'analyser l'impact du choix de la dimension sur les performances de l'algorithme.

Pour conclure, il existe de nombreuses pistes pour améliorer les performances du systèmes de recommandation. Elles ne doivent pas toutes être implémentées à la fois mais de manière générale, il serait très intéressant de comparer l'impact d'une ou plusieurs modifications sur les performances du système de recommandations.

6 Conclusion

Dans ce mémoire, nous étudions les systèmes de recommandation dans le domaine de l'association du vin avec la nourriture. Plus particulièrement, nous analysons le système de recommandation de vins implémenté par Roald Schuring et décrit dans l'article *Food and Wine Pairing* (Schuring, 2019a) que nous avons adapté : il s'agit d'un système de recommandation hybride basé sur le contexte et le contenu construit à partir de données textuelles non-structurées.
L'objectif de ce mémoire est donc de déterminer si ce système recommande de manière plus performante qu'une recommandation aléatoire des vins pour des plats spécifiques exploitant des données textuelles exclusivement trouvées sur internet en évaluant ses performances avec des expériences hors-ligne.

Dans un premier temps, nous avons mis en contexte le système de recommandation étudié en présentant les systèmes de recommandation de manière générale, les travaux existants dans le domaine du vin et de la nourriture ainsi que les caractéristiques des différents types de goûts. Nous avons ensuite fourni une explication détaillée du fonctionnement de ce système de recommandation.

Puis, nous évaluons les méthodes d'embedding d'arôme/non-arômes du vin et de la nourriture permettant de transformer les vins et les aliments en des vecteurs/scalaires, à l'exception de l'évaluation de l'embedding des non-arômes du vin qui requiert trop de mémoire. A l'aide d'une méthode de validation basée sur la définition de catégories, nous montrons que ces méthodes d'embedding sont plus performantes que l'aléatoire puisqu'elles sélectionnent mieux les vins/aliments appartenant à la même catégorie. Cependant, la définition des catégories de vins/aliments peut être améliorée : des pistes d'améliorations concrètes sont proposées.

Enfin, nous répondons à la question de recherche principale de ce mémoire : il est effectivement possible de montrer qu'un système de recommandation de vins pour des plats spécifiques qui utilise des données textuelles d'Internet est plus performant que l'aléatoire, en implémentant une méthode d'évaluation hors-ligne. Cependant, le système de recommandation présente plusieurs défauts et peut donc largement être amélioré. Des pistes d'amélioration concrètes sont finalement présentées dans ce mémoire.

Pour conclure, dans ce mémoire, nous nous sommes focalisées sur un système de recommandation dans le domaine de l'association entre les vins et la nourriture sans prendre en compte les préférences des utilisateurs. Dans des travaux de recherche futurs, il pourrait être intéressant d'intégrer cette dimension dans le système de recommandation mais cela augmentera considérablement la complexité du problème de recommandation. Finalement, selon nous, ce système de recommandation s'avère prometteur et mériterait que l'on y accorde de l'attention.

Bibliographie

Adewumi, T. P., Liwicki, F. & Liwicki, M. (2020). Word2vec : optimal hyper-parameters and their impact on nlp downstream tasks. *arXiv preprint arXiv :2003.11645*.

Adomavicius, G. & Tuzhilin, A. (2005). Personalization technologies : a process-oriented perspective. *Communications of the ACM*, *48*, 83-90. https://doi.org/10.1145/1089107.1089109

Ahn, Y.-Y., Ahnert, S. E., Bagrow, J. P. & Barabási, A.-L. (2011). Flavor network and the principles of food pairing. *Scientific reports*, *1*(1), 1-7.

Alian, S., Li, J. & Pandey, V. (2018). A Personalized Recommendation System to Support Diabetes Self-Management for American Indians. *IEEE Access*, *6*, 73041-73051. https://doi.org/10.1109/ACCESS.2018.2882138

Alvin. (2021). *Food.com - Recipes and Reviews*. https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews?select=reviews.csv (accessed : 17.05.2022)

Arnone, K. & Simonetti-Bryan, J. (2013). *Pairing with the Masters : A Definitive Guide to Food Wine*. Delmar Cengage Learning.

Balabanović, M. & Shoham, Y. (1997). Fab : content-based, collaborative recommendation. *Communications of the ACM*, *40*(3), 66-72. https://doi.org/10.1145/245108.245124

Balducci, B. & Marinova, D. (2018). Unstructured data in marketing. *Academy of Marketing Science*, *46*, 557-590. https://doi.org/10.1007/s11747-018-0581-x

Beckett, F. (s. d.). *The Ultimate Resource for Food and Wine Pairing*. https://www.matchingfoodandwine.com/ (accessed : 15.05.2022)

Blumberg, R. & Atre, S. (2003). The Problem with Unstructured Data. *DM Review*, *13*, 42-46.

Bora, D. J. & Gupta, A. K. (2014). Effect of Different Distance Measures on the Performance of K-Means Algorithm : An Experimental Study in Matlab. *CoRR*. http://arxiv.org/abs/1405.7471

Brunner, T. A. & Siegrist, M. (2011). Lifestyle determinants of wine consumption and spending on wine. *International Journal of Wine Business Research*, *23*(3), 210-220. https://doi.org/10.1108/17511061111163041

Burke, R. (2007). Hybrid web recommender systems. In P. Brusilovsky, A. Kobsa & W. Nejdl (Éd.), *The Adaptive Web : Methods and Strategies of Web Persona-*

lization (p. 377-408). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_12

Campo, E., Ferreira, V., Escudero, A. & Cacho, J. (2005). Prediction of the wine sensory properties related to grape variety from dynamic-headspace gas chromatography-olfactometry data. *Journal of Agricultural and Food Chemistry*, *53*(14), 5682-5690.

Chartier, F. (2012). *Taste Buds and Molecules : The Art and Science of Food, Wine, and Flavor*. John Wiley ; Sons.

Chen, B., Rhodes, C., Crawford, A. & Hambuchen, L. (2014). Wineinformatics : Applying data mining on wine sensory reviews processed by the computational wine wheel. *2014 IEEE International Conference on Data Mining Workshop*, 142-149.

Christodoulou, P., Christodoulou, K. & Andreou, A. S. (2017). A real-Time targeted recommender system for supermarkets. *19th International Conference on Enterprise Information Systems*, 703-712. https://doi.org/10.5220/0006309907030712

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547-553. https://doi.org/https://doi.org/10.1016/j.dss.2009.05.016

Davis, U. (1980). *A.C. Noble wine aroma wheel*. https://www.winearomawheel.com (accessed : 08.05.2022)

D.H. and FSA. (2016). *Guide to creating a front of pack (FoP) nutrition label for pre-packed products sold through retail outlets*. https://www.food.gov.uk/sites/default/files/media/document/fop-guidance_0.pdf (accessed : 18.05.2022)

EDAMAM. (s. d.). *EDAMAM - Leading provider of nutrition data and analytics*. https://www.edamam.com/ (accessed : 18.05.2022)

Elahi, M., Ge, M., Ricci, F. & Massimo, S., D.and Berkovsky. (2014). Interactive Food Recommendation for Groups. *Recsys posters*.

El-Dosuky, M. A., Rashad, M. Z., Hamza, T. T. & EL-Bassiouny, A. H. (2012). Food recommendation using ontology and heuristics. *Advanced Machine Learning Technologies and Applications*, 423-429.

Elsweiler, D., Harvey, M., Ludwig, B. & Said, A. (2015). Bringing the" healthy" into Food Recommenders. *Doctoral Midwifery Research Society*, 33-36.

Elsweiler, D., Hauptmann, H. & Trattner, C. (2022). Food Recommender Food recommenderSystems. In F. Ricci, L. Rokach & B. Shapira (Éd.), *Recommender Systems Handbook* (p. 871-925). Springer US. https://doi.org/10.1007/978-1-0716-2197-4_23

Fallis, C. (2014). *Origin of the Wine Aroma Wheel, Dr. Ann Noble*. https://www.planetgrape.com/origin_of_the_wine_aroma_wheel (accessed : 08.05.2022)

Fehér, J., Lengyel, G. & Lugasi, A. (2007). The cultural history of wine - theoretical background to wine therapy. *Central European Journal of Medicine volume*, *2*, 379-391. https://doi.org/10.2478/s11536-007-0048-9

Freyne, J., Berkovsky, S. & Smith, G. (2011). Recipe Recommendation : Accuracy and Reasoning. *User Modeling, Adaption and Personalization*, 99-110.

Gannon, R. H., Millward, D. J., Brown, J. E., Macdonald, H. M., Lovell, D. P., Frassetto, L. A., Remer, T. & Lanham-New, S. A. (2008). Estimates of daily net endogenous acid production in the elderly UK population : analysis of the National Diet and Nutrition Survey (NDNS) of British adults aged 65 years and over. *British journal of nutrition*, *100*(3), 615-623.

GGDOT. (s. d.). *Greenhouse Gas and Dietary choices Open source Toolkit*. https://www.ggdot.org/ (accessed : 18.05.2022)

Gokcekus, O. & Gokcekus, S. (2019). Empirical evidence of lumping and splitting : Expert ratings' effect on wine prices. *Wine Economics and Policy*, *8*(2), 171-179. https://doi.org/https://doi.org/10.1016/j.wep.2019.09.003

Goldberg, D., Nichols, D., Oki, B. M. & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61-70. https://doi.org/10.1145/138859.138867

Goldstein, E. (2006). *Perfect pairings : A Master Sommelier's Practical Advice for Partnering Wine with Food*. University of California Press.

Group, A. F. (s. d.). *EatingWell*. https://www.eatingwell.com/ (accessed : 18.05.2022)

Hambleton, C. (2008). *The Wine Planner : Selecting the Right Wines to Complement Your Favorite Foods*. Adams Media.

Hammack, J., Puckette, M. & Wasser, C. (2011). *World Wide Wines*. winefolly.com (accessed : 15.05.2022)

Harrington, R. (2007). *Food and Wine Pairing : A Sensory Experience*. John Wiley ; Sons.

Harrington, R. J. & Seo, H.-S. (2015). The Impact of Liking of Wine and Food Items on Perceptions of Wine–Food Pairing. *Journal of Foodservice Business Research*, *18*(5), 489-501. https://doi.org/10.1080/15378020.2015.1093455

Hasnat, A., Halder, S., Bhattacharjee, D., Nasipuri, M. & Basu, D. (2013). Comparative study of distance metrics for finding skin color similarity of two color facial images. *ACER : New Taipei City, Taiwan*, 99-108.

Instacart. (2017). *Instacart Market Basket Analysis*. https://www.kaggle.com/c/instacart-market-basket-analysis (accessed : 18.05.2022)

IQWiG. (2006). *How does our sense of taste work ?* https://www.ncbi.nlm.nih.gov/books/NBK279408/ (accessed : 08.05.2022)

Jani & Simonetta. (s. d.). *World Wide Wines*. https://www.cardsofwine.com/ (accessed : 15.05.2022)

Javanmardian, K. (2005). *A Wine Pairing Recommendation System* [Working paper].

Katarya, R. & Saini, R. (2022). Enhancing the wine tasting experience using greedy clustering wine recommender system. *Multimed Tools Appl*, *81*, 807-840. https://doi.org/10.1007/s11042-021-11300-5

Keast, R. S. & Costanzo, A. (2015). Is fat the sixth taste primary ? Evidence and implications. *Flavour*, *4*(1), 1-7.

Khan, A. S. & Hoffmann, A. (2003). Building a case-based diet recommendation system without a knowledge engineer. *Artificial Intelligence in Medicine*, *27*(2), 155-179. https://doi.org/https://doi.org/10.1016/S0933-3657(02)00113-6

Klosse, P. (2011). Food and wine pairing : A new approach. *Research in Hospitality Management*, *1*(1), 5-8. https://doi.org/10.1080/22243534.2011.11828269

Lee, C.-S., Wang, M.-H. & Hagras, H. (2010). A Type-2 Fuzzy Ontology and Its Application to Personal Diabetic-Diet Recommendation. *IEEE Transactions on Fuzzy Systems*, *18*(2), 374-395. https://doi.org/10.1109/TFUZZ.2010.2042454

Martinez, R. D., Angus, G. & Mahdavian, R. (2018). Grapevine : A Wine Prediction Algorithm Using Multi-dimensional Clustering Methods. *CoRR*, *abs/1807.00692*(3).

McGuire, S. (2022). *Wine Reviews Data - Dataset from scrape wine reviews.* https : / / www . kaggle . com / datasets / samuelmcguire / wine - reviews - data (accessed : 17.05.2022)

Michaelis, J. R., Ding, L. & McGuinnes, D. L. (2008). The TW Wine Agent : A Social Semantic Web Demo. *International Semantic Web Conference (Posters Demos).*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26.*

Moltedo, A., Sánchez, C. Á., Troubat, N. & Cafiero, C. (2018). Optimizing the use of ADePT-Food Security Module for Nutrient Analysis.

Nygren, L., Gustafson, I.-B., Haglund, Å., Johansson, L. & Noble, A. (2001). Flavor changes produced by wine and food interactions : Chardonnay wine and Hollandaise sauce. *Journal of Sensory Studies*, *16*(5), 461-470. https://doi.org/10.1111/j.1745-459X.2001.tb00313.x

Nygren, L., Gustafson, I.-B. & Johansson, L. (2002). Perceived flavor changes in white wine after tasting blue mold cheese. *Food Service Technology*, *2*(4), 163-171. https://doi.org/10.1046/j.1471-5740.2002.00048.x

Palomo, E. S., Diaz-Maroto, M., Viñas, M. G., Soriano-Pérez, A. & Pérez-Coello, M. (2007). Aroma profile of wines from Albillo and Muscat grape varieties at different stages of ripening. *Food control*, *18*(5), 398-403.

Palomo, E. S., Hidalgo, M. D.-M., Gonzalez-Vinas, M. & Pérez-Coello, M. (2005). Aroma enhancement in wines from different grape varieties using exogenous glycosidases. *Food chemistry*, *92*(4), 627-635.

Pazzani, M. J. & Billsus, D. (2007). Content-Based Recommendation Systems. In P. Brusilovsky, A. Kobsa & W. Nejdl (Éd.), *The Adaptive Web : Methods and Strategies of Web Personalization* (p. 325-341). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_10

Pediaopolis, D. W. K. (1947). *The Five Senses.* http://udel.edu/~addieg/project_2cc/taste.html (accessed : 08.05.2022)

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove : Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Peršurić, A. S. I., Damijanić, A. T. & Kerma, S. (2018). The relationship between autochthonous wine attributes and wine consumption motives. *Economics of Agriculture*, *65*(4), 1337-1357. https://doi.org/10.5937/ekoPolj1804337I

Peynaud, E. (1996). *The Taste of Wine : The Art and Science of Wine Appreciation, 2nd ed.* John Wiley ; Sons.

PlatsNetVins. (s. d.). *Recherche des accords entre mets et vins*. https://www.platsnetvins.com/ (accessed : 15.05.2022)

Pomarici, E., Lerro, M., Chrysochou, P., Vecchio, R. & Krystallis, A. (2017). One size does (obviously not) fit all : Using product attributes for wine market segmentation. *Wine Economics and Policy*, *6*(2), 98-106. https://doi.org/https://doi.org/10.1016/j.wep.2017.09.001

Protasiewicz, J., Pedrycz, W., Kozłowski, M., Dadas, S., Stanisławek, T., Kopacz, A. & Gałężewska, M. (2016). A recommender system of reviewers and experts in reviewing problems. *Knowledge-Based Systems*, *106*, 164-178. https://doi.org/https://doi.org/10.1016/j.knosys.2016.05.041

Puckette, M. (2014). *Updated Wine Flavor Wheel with 100+ Flavors*. https://winefolly.com/tutorial/wine-aroma-wheel-100-flavors/ (accessed : 08.05.2022)

Puckette, M. (2015). *Wine folly : The essential guide to wine*. Penguin.

Puckette, M. & Hammack, J. (2018). *Wine Folly : Magnum Edition : The Master Guide*. Avery.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J. (1994). *GroupLens : an open architecture for collaborative filtering of netnews*. https://doi.org/10.1145/192844.192905

Resnick, P. & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56-58. https://doi.org/10.1145/245108.245121

Ricci, F., Rokach, L. & Shapira, B. (2015). *Recommender Systems Handbook - Second Edition*. Springer.

Risius, A., Klann, B.-O. & Meyerding, S. G. (2019). Choosing a lifestyle ? Reflection of consumer extrinsic product preferences and views on important wine characteristics in Germany. *Wine Economics and Policy*, *8*(2), 141-154. https://doi.org/https://doi.org/10.1016/j.wep.2019.09.001

Sarker, I. (2021). Machine Learning : Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(160). https://doi.org/10.1007/s42979-021-00592-x

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M. & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135.

Schubert, E. (2021). A triangle inequality for cosine similarity. *International Conference on Similarity Search and Applications*, 32-44.

Schuring, R. (2019a). *Food and Wine Pairing*. https://medium.com/towards-data-science/robosomm-chapter-5-food-and-wine-pairing-7a4a4bb08e9e (accessed : 15.02.2022)

Schuring, R. (2019b). *Wine Embeddings and a Wine Recommender*. https://towardsdatascience.com/robosomm-chapter-3-wine-embeddings-and-a-wine-recommender-9fc678f1041e (accessed : 15.03.2022)

Schuring, R. (2021a). *Wine food pairing*. https://github.com/RoaldSchuring/wine_food_pairing (accessed : 15.02.2022)

Schuring, R. (2021b). *Wine Reviews*. https://www.kaggle.com/datasets/roaldschuring/wine-reviews (accessed : 25.02.2022)

Shardanand, U. & Maes, P. (1995). *Social information filtering : algorithms for automating "word of mouth"*. https://doi.org/10.1145/223904.223931

Sherwin, C. P. B. A. L. (2017). Predicting wine preference : testing the premises of the vinotype theory. *International Journal of Wine Business Research*, *29*(3), 251-268. https://doi.org/10.1108/IJWBR-08-2016-0027

Simonetti-Bryan, J. (2010). *The Everyday Guide to Wine. The Great Courses*. The Teaching Company.

Stanford. (2017). *Amazon Fine Food Reviews*. https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews (accessed : 08.05.2022)

Strum, J. (2022). *WineEnthusiast*. https://www.winemag.com/ (accessed : 08.05.2022)

Subramaniyaswamy, V., Vijayakumar, V., Logesh, R. & Indragandhi, V. (2015). Unstructured Data Analysis on Big Data Using Map Reduce. *Procedia Computer Science*, *50*, 456-465. https://doi.org/https://doi.org/10.1016/j.procs.2015.04.015

Szabo, J. (2013). *Pairing Food and Wine For Dummies*. John Wiley & Sons.

Tanwar, M., Duggal, R. & Khatri, S. K. (2015). Unravelling unstructured data : A wealth of information in big data. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 1-6. https://doi.org/10.1109/ICRITO.2015.7359270

Thomas, C. (2013). *Practical Wine Talk : A Physician-Winemaker Examines Wine*. Authorhouse.

Thoutt, Z. (2018). *Wine Reviews - 130k wine reviews with variety, location, winery, price, and description*. https://www.kaggle.com/datasets/zynicide/wine-reviews (accessed : 17.05.2022)

Ueta, T., Iwakami, M. & Ito, T. (2011). A Recipe Recommendation System Based on Automatic Nutrition Information Extraction. In H. Xiong & W. B. Lee (Éd.), *Knowledge Science, Engineering and Management* (p. 79-90). Springer Berlin Heidelberg.

Vadivel, A., Majumdar, A. & Sural, S. (2003). Performance comparison of distance metrics in content-based image retrieval applications. *International Conference on Information Technology (CIT), Bhubaneswar, India*, 159-164.

Van Niekerk, K. (2012). *The Food & Wine Pairing Guide*. Penguin Random House South Africa.

Vilela, A., Inês, A., Cosme, F. & Desk, S. (2016). Is wine savory ? Umami taste in wine. *SDRP Journal of Food Science & Technology*, *1*(3).

Waltner, G., Schwarz, M., Ladstätter, S., Weber, A., Luley, P., Lindschinger, M., Schmid, I., Scheitz, W., Bischof, H. & Paletta, L. (2017). Personalized Dietary Self-Management Using Mobile Vision-Based Assistance. In S. Battiato, G. M. Farinella, M. Leo & G. Gallo (Éd.), *New Trends in Image Analysis and Processing – ICIAP 2017* (p. 385-393). Springer International Publishing. https://doi.org/10.1007/978-3-319-70742-6_36

Wikipedia. (2022a). *Aroma of wine*. https://en.wikipedia.org/wiki/Aroma_of_wine (accessed : 08.05.2022)

Wikipedia. (2022b). *Taste*. https://en.wikipedia.org/wiki/Taste#Basic_tastes (accessed : 08.05.2022)

Yujie, Z. & Licai, W. (2010). Some challenges for context-aware recommender systems. *2010 5th International Conference on Computer Science and Education (ICCSE)*, 362-365. https://doi.org/10.1109/ICCSE.2010.5593612

7 Annexes

7.1 Matrice de données

La matrice de données d'association vins et repas est fournie ci-dessous : chaque ligne représente un vin v, avec un nombre total de vins $nv = 63$ et chaque colonne représente un repas r, avec un nombre total de repas $nr = 40$. La dernière ligne donne le nombre de vins pertinents pour le repas r : $|P_r|$.

v \ r	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1											1	1								1
2		1			1						1									
3						1			1											
4																			1	
5																1				
6	1	1			1	1			1					1		1				1
7					1			1		1				1	1			1		
8																				
9											1									
10																				1
11																				1
12																				1
13															1			1		
14																				
15								1												
16	1																			
17			1		1															
18																1				
19																				
20																				
21			1																	
22	1	1	1	1		1	1		1			1					1		1	
23										1				1	1					
24												1							1	
25																				
26		1	1		1			1			1		1				1			
27																		1		
28																	1			
29					1					1			1		1	1				
30		1	1		1			1			1		1				1			
31									1											
32								1										1		
33											1									
34		1											1							
35								1												
36					1										1					
37	1	1				1			1			1								
38									1			1								
39					1				1											
40	1	1			1	1								1		1				
41																				
42																				
43											1									
44			1	1					1		1	1				1	1		1	1
45																				
46															1					
47												1								1
48	1		1	1		1	1		1		1	1					1		1	
49	1	1	1	1			1		1		1			1					1	
50					1			1												
51								1												
52	1					1														
53						1														
54		1			1								1		1					
55															1					
56																				
57										1	1									
58								1										1		
59	1							1						1						1
60								1												
61												1				1				
62		1			1								1		1			1		
63																				
$\lvert P_r \rvert$	9	11	8	4	13	8	3	11	10	4	11	9	6	6	9	7	6	6	6	8

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1			1	1															
															1				
1																			1
1		1																	
													1			1	1		
1			1	1		1				1		1							1
	1																		
	1	1																	
													1						
					1									1					
					1														
					1	1	1												
								1											
						1			1										
									1										
										1									
							1		1		1		1			1	1		
1					1														
							1						1						
		1	1			1				1									
1				1	1									1					1
	1	1	1							1				1					
		1	1			1				1									
1					1					1									1
	1	1			1														1
		1	1				1												
				1										1					
1			1	1	1									1					
		1							1						1	1		1	
																			1
													1			1	1		
											1								
						1													
						1	1		1						1	1	1		
						1													
1					1	1	1												
							1		1		1		1		1	1	1		
		1					1	1				1				1	1		
				1	1									1					
1																			
		1					1												
1			1	1										1					
1					1	1	1												
												1							
		1																	
1					1														
			1	1										1					
13	4	11	9	8	12	10	10	2	6	6	3	3	6	8	4	7	6	1	6

7.2 Classement moyen idéal des vins

En utilisant les valeurs de $|P_r|$ fournies dans la matrice de données en Annexe 7.1 et en sachant que le nombre total de repas vaut $nr = 40$, on obtient comme classement moyen idéal des vins :

$$\frac{\sum_{r=1}^{nr} \frac{|P_r|+1}{2}}{nr} = \frac{nr + \sum_{r=1}^{nr} |P_r|}{2 \cdot nr} = \frac{40 + \sum_{r=1}^{40} |P_r|}{2 \cdot 40} = 4.13.$$

Abstract : Ce mémoire analyse les performances d'un système de recommandation qui utilise des données textuelles exclusivement trouvées sur Internet pour suggérer des vins pour des plats spécifiques aux utilisateurs mais sans prendre en compte leurs préférences.

Ce système de recommandation, fondé sur le travail de Roald Schuring, est un système hybride basé d'une part sur le contenu en exploitant des données textuelles sur les vins et d'autre part sur le contexte en considérant les mets comme éléments contextuels. Un des avantages de ce système est l'utilisation de données non-structurées qui peuvent être collectées facilement et abondamment.

Notre contribution dans ce mémoire est triple : premièrement, nous conceptualisons et fournissons une explication détaillée de ce système de recommandation. Ensuite, nous implémentons une méthode d'évaluation du processus d'embedding du vin/des mets et montrons que celui-ci donne des résultats plus performants qu'un processus aléatoire. Finalement, nous implémentons deux méthodes d'évaluation des performances du système de recommandation et démontrons que ce système de recommandation de vins pour des mets spécifiques fournit de meilleures associations qu'un système aléatoire.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique
www.uclouvain.be/lsm