MASTER OF SCIENCE IN ENGINEERING

# Cahier des charges - WEM - Wine Recommender

## Master MSE

Rédigé par : **Andrea Petrucci, Benjamin Pasquier, Laurent Hirschi**

**Groupe D**

Supervisé par : Dr. Laura Elena Raileanu

Assistants : Elena Najdenovska et Cédric Campos Carvalho

Version : 1.0

7 mai 2023

# Table des versions

| Version | Description |
|---|---|
| 1.0 | Version initiale |

# Table des matières

# Table des figures

# 1 Contexte et objectifs

Dans le cadre du cours de Web Mining, il est demandé de réaliser un projet de récupération et d'analyse de données. Le sujet choisi pour ce dernier est la création d'un système de recommandation de vin en fonction d'une recette.

## 1.1 Contexte

Lorsque nous choisissons de préparer une recette, nous nous demandons parfois quel vin accompagnerait le mieux ce plat. En règle générale, les descriptifs de vins viennent accompagnés d'une recommandation du style "boeuf", "poulet", "poisson", etc. Mais ces recommandations sont souvent trop générales et ne permettent pas de choisir un vin adapté à la recette. De plus, il est difficile de trouver des informations sur les vins, notamment sur les vins de qualité moyenne ou inférieure. C'est pourquoi nous avons décidé de créer un système de recommandation de vin en fonction d'une recette sous forme d'une application web.

Pour ce faire, deux sources de données seront utilisées :
- **Vivino** : pour la récupération des données sur les vins.
- **Marmiton** : pour la récupération des données sur les recettes.

Afin de traiter une quantité de données raisonnable, nous nous limitons aux plats principaux pour les recettes et aux vins rouges et blancs pour les vins.

## 1.2 Objectifs

La réalisation du projet est composé de trois parties principales :
- **Récupération des données** : récupération des données sur les vins et les recettes.
- **Création de l'API** : création d'une API REST pour servir les données.
- **Création de l'application web** : création d'une application web pour la recommandation de vin.

# 2 Données

Ce chapitre décrit les données que nous utiliserons dans ce projet et la façon dont elles seront extraites.

## 2.1 Sources des données

Deux sources de données sont considérés pour la réalisation du projet : le site web Vivino qui répertorie des vins et le site web Marmiton qui répertorie des recettes.

### 2.1.1 Vivino

Vivino est un site web et une application mobile qui permet de trouver des informations sur une grande quantité de vins. C'est une plateforme communautaire qui permet aux utilisateurs de donner leur avis sur les vins qu'ils ont dégustés, d'en trouver des similaires à ceux qu'ils ont appréciés ou en fonction de leur budget. Le site permet de filtrer les vins en fonction de leur couleur, de leur pays d'origine, de leur région, de leur cépage, etc. Pour notre projet, nous avons décidé de nous limiter aux vins rouges et blancs provenant de Suisse, d'Argentine, des États-Unis, de France, d'Italie et d'Espagne, afin de ne pas avoir une quantité trop importante de données à traiter.

### 2.1.2 Marmiton

Marmiton est un site web qui permet de trouver des recettes de cuisine. On peut y trouver des recettes de tous types de plats, de l'entrée au dessert. Notre projet se concentrant sur les plats principaux, nous avons décidé de nous limiter à cette catégorie.

## 2.2 Description des données

Les deux sources de données considérées contiennent une certaine quantité d'informations dont certaines ne sont pas pertinentes pour notre projet. Il est donc nécessaire de sélectionner les attributs qui nous seront utiles.

### 2.2.1 Vins

Les vins répertoriés sur Vivino possèdent une quantité importante d'attributs, nous avons choisi de garder les suivants :

- **Nom** : nom du vin.
- **Nom de la cave** : nom du producteur du vin.
- **Millésime** : année de production du vin.
- **Pays d'origine** : pays de production du vin.
- **Type** : type du vin, soit rouge, soit blanc.
- **Cépages** : liste des cépages utilisés pour la production du vin.
- **Prix** : prix d'une bouteille en CHF.
- **Liste d'accords** : liste des aliments qui s'accordent avec le vin.
- **Avis des utilisateurs** : liste des trois premiers avis jugés utiles par Vivino. Chaque avis est composé d'une note sur 5.0 et d'un commentaire textuel.

— **Lien** : lien vers le vin sur le site de Vivino.

### 2.2.2 Recettes

Les recettes répertoriées sur Marmiton possèdent une quantité importante d'attributs, nous avons choisi de garder les suivants :

— **Nom** : nom de la recette.
— **Type** : type de plat (entrée, plat principal, dessert, etc.).
— **Liste d'ingrédients** : liste des ingrédients nécessaires à la réalisation de la recette.
— **Lien** : lien vers la recette sur le site de Marmiton.

## 2.3 Extraction des données

Les données nécessaires à la réalisation de ce projet sont contenues sur les sites web Vivino et Marmiton. Il est donc nécessaire d'extraire ces données afin de pouvoir les utiliser.

### 2.3.1 Vivino

Pour extraire les données du site de Vivino, nous utilisons la librairie Python Selenium, qui permet de piloter un navigateur web. De cette manière, nous évitons d'être bloqué par le site, par exemple dans le cas où trop de requêtes sont envoyées. En effet, nous n'avons pas réussi à extraire des données avec la librairie Scrapy, Selenium semble donc être une bonne alternative.

La majorité des informations que nous voulons extraire se trouve sur la page du produit, il est donc nécessaire de récupérer les liens vers ces pages. La page répertoriant les vins ne contient que les 10 premier vins, il faut utiliser un "Infinite Scrolling" afin de charger les vins suivants. Nous chargerons donc entièrement la page en scrollant automatiquement jusqu'à sa fin, puis nous récupérerons les liens vers les pages des vins. Nous visiterons ensuite ces liens un par un pour récupérer les données manquantes.

### 2.3.2 Marmiton

Pour extraire les données du site Marmiton, nous utilisons la librairie Scrapy, vue en classe durant un laboratoire, qui permet de récupérer des données sur un site web. Nous avons également utilisé BeautifulSoup pour extraire les données de la page HTML, puisque Scrapy ne nous permettait pas de le faire comme nous voulions.

La majorité des informations que nous voulons extraire se trouve sur la page de la recette, il est donc nécessaire de récupérer le lien vers chaque recette. La page contenant les mets ne répertorie que les 30 premières recettes, nous avons dû extraire le lien vers toutes les autres pages afin de toutes les récupérer.

# 3 Technologies et méthodes

Ce chapitre décrit les technologies ainsi que les méthodes d'analyses qui seront utilisées dans le cadre de ce projet.

## 3.1 Technologies

Afin de mettre en place ce projet, différentes technologies vont être utilisées.

Pour la récupération des données, nous allons utiliser le langage de programmation **Python** avec les librairies **Scrapy**, **BeautifulSoup** et **Selenium**. Ces différentes données seront stockées dans une base de données **MongoDB**. Cela facilitera la gestion car elles seront stockées sous forme de documents JSON. Afin d'être utilisables ces données seront servies par une API REST en utilisant **FastAPI**. Pour finir une application web sera développée en le framework **ReactJS**.

En résumé, la stack technologique utilisée est la suivante :
- **Python** pour la récupération des données.
- **Scrapy**, **BeautifulSoup** et **Selenium** pour la récupération des données.
- **MongoDB** pour le stockage des données.
- **ExpressJS** pour la création de l'API REST.
- **ReactJS** pour la création de l'application web.

Le diagramme de la figure 3.1 illustre l'architecture globale du projet.

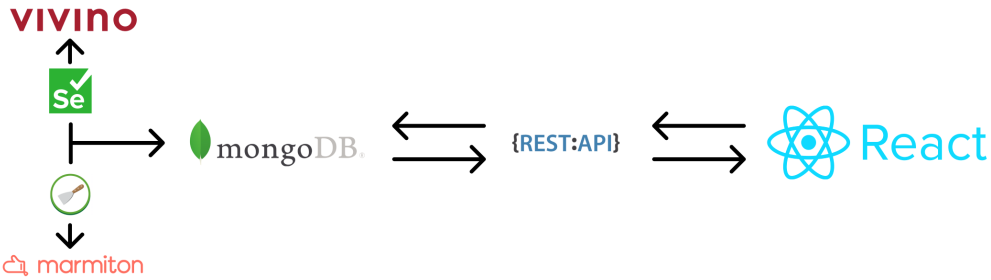


Figure 3.1 – Architecture globale du projet.

## 3.2 Méthodes d'analyse

Deux méthodes d'analyse sont envisagés dans ce projet. La première doit permettre recommander un vin pour une recette en comparant les listes d'accords des vins avec le nom et les ingrédients de la recette. La seconde doit permettre de déterminer si le vin est apprécié et les points qui le caractérisent en analysant les commentaires des utilisateurs sur les vins.

### 3.2.1 Recommandation de vin

Afin de recommander un vin suivant une recette, nous allons générer pour chaque recette une liste de vecteurs représentant son nom et ses ingrédients à l'aide d'un générateur d'embedding de mots comme

Word2Vec ou GloVe. Pour chaque vin, nous procéderons de la même manière en générant une liste de vecteurs représentant les plats avec lesquels il s'accorde. La figure 3.2 illustre la génération de ces vecteurs et leur stockage dans la base de données. Cette génération sera donc effectuée qu'une seule fois, sur toutes les données récupérées.

Ces vecteurs étant persistés dans la base, il sera possible de les récupérer une fois que l'utilisateur aura sélectionné une recette. Les similarités (par exemple la similarité cosinus) entre les vecteurs de la recette et ceux des vins seront ensuite calculées puis moyennées pour obtenir une note de recommandation. Une liste contenant les meilleurs vins s'accordant avec la recette sera alors retournée à l'utilisateur. La figure 3.3 illustre ce processus.

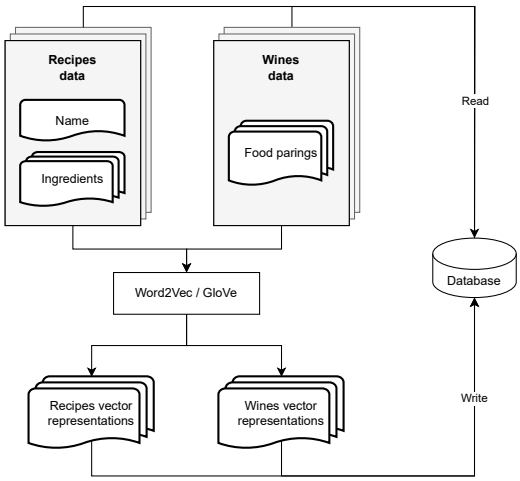


Figure 3.2 – Processus de génération et de stockage des vecteurs représentatifs.

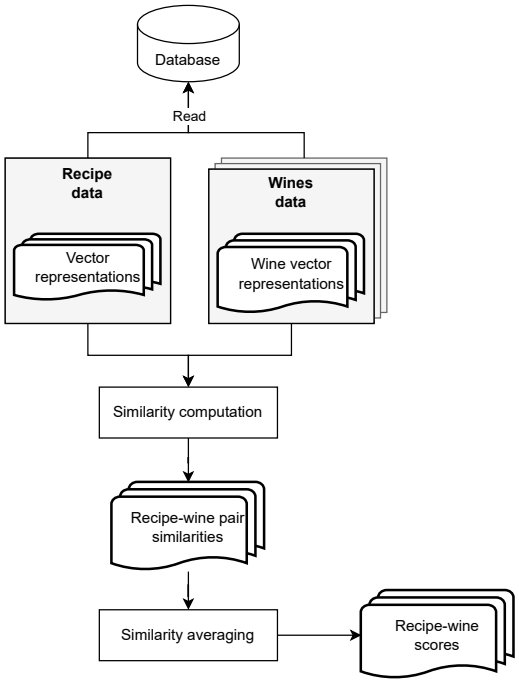


Figure 3.3 – Processus de recommandation d'un vin pour une recette.

### 3.2.2 Analyse des commentaires

Afin d'analyser les commentaires, nous allons utiliser la librairie Python "pysentimiento". Cette librairie se base sur des modèles d'apprentissage profond Transformeur pour réaliser des tâches de traitement du langage naturel. Elle permet notamment de classifier un texte (dans notre cas le cas un commentaire) selon s'il est de nature positive, neutre ou négative, et également de générer le sentiment principal qui en ressort comme la joie, la tristesse ou encore la colère.

# 4 Planning

La planification des tâches à effectuer pour la réalisation du projet est illustré par la figure 4.1. Chaque tâche est associée à un membre du groupe, représenté par une certaine couleur, à savoir :

- **Rouge** : Andrea Petrucci
- **Bleu** : Benjamin Pasquier
- **Orange** : Laurent Hirschi
- **Vert** : Tout le monde

Les jalons sont de couleur verte et les deliverables de couleur rouge.

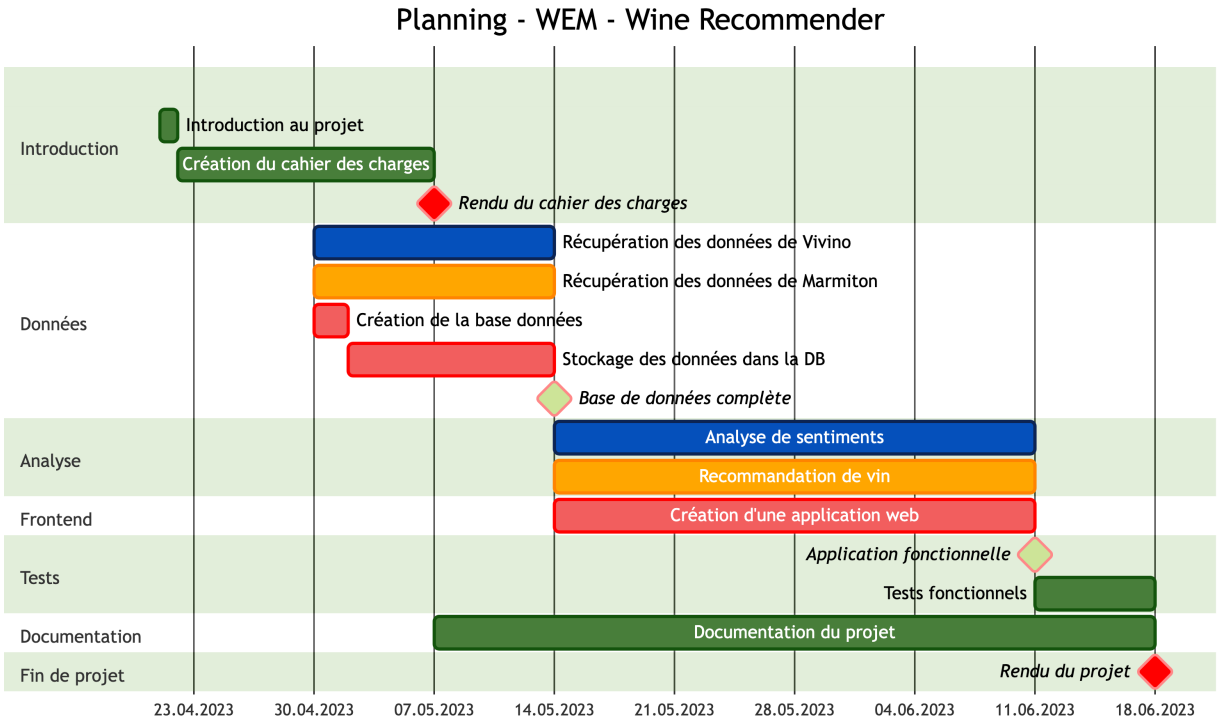


Figure 4.1 – Planification

# 5 Résultats attendus et risques

Ce dernier chapitre décrit les résultats attendus ainsi que les risques qui peuvent survenir lors de la réalisation de ce projet.

## 5.1 Résultats attendus

Le résultat final attendu est une page web qui permet à l'utilisateur de choisir une recette dans une liste ou en faisant une recherche. Une fois la recette choisie, l'utilisateur doit pouvoir voir les ingrédients nécessaires à la réalisation de la recette ainsi que les vins qui s'accordent avec cette dernière. Il doit également être possible de filtrer les recettes et les vins résultants en fonction de différents critères. Par exemple, on pourra filtrer les vins en fonction de leur prix, de leur pays d'origine, de leur type, etc. et les recettes en fonction des ingrédients nécessaires à leur préparation. La page web doit également retourner à l'utilisateur un sentiment décrivant les avis jugés utiles par Vivino. Ce sentiment doit être accompagné des mots les plus importants utilisés pour le jugement.

Bien entendu, il est attendu de l'application que les résultats retournés soient cohérents et que les vins proposés s'accordent bel et bien avec la recette choisie.

## 5.2 Risques

Quelques risques peuvent survenir lors de la réalisation de ce projet et ainsi influencer la qualité des résultats attendus.

### 5.2.1 Précision de la recommandation

Le premier risque est la précision de la correspondance des ingrédients. En effet, il est possible que la méthode de génération d'embedding et le calcul de similarité ne soit pas assez précis pour lier correctement les accords aux ingrédients de la recette. Dans ce cas, il sera nécessaire de trouver une autre méthode plus efficace.

Il est également possible que les accords des vins soient souvent les mêmes et que l'application recommande souvent les mêmes vins pour des recettes différentes. Dans ce cas, il est difficile d'adapter la méthode puisque la cause du problème est le manque de diversité des accords des vins. Il serait donc nécessaire de trouver une autre source de données pour les accords des vins.

### 5.2.2 Fiabilité de l'analyse des sentiments

L'analyse des sentiments comporte également un risque. En effet, il est possible que les résultats de l'analyse ne soit pas fiable et que les avis jugés utiles par Vivino ne soient pas représentatifs de l'avis général des utilisateurs. Dans ce cas, il faudrait considérer une plus grande quantité d'avis utilisateur.

### 5.2.3 Subjectivité du goût

La subjectvité du goût est un risque sur lequel il est difficile d'agir, bien qu'il soit important de le mentionner. En effet, il est possible que les vins qui s'accordent avec une recette ne soient pas au goût de

l'utilisateur. Il faudrait donc, si le temps le permet, ajouter une fonctionnalité permettant à l'utilisateur d'obtenir d'autres vins qui s'accordent avec la recette.