**Rapport du PAN 2 – Module Classification**

Groupe 3.5

À l’attention de Monsieur Pierre-Alexandre Murena

Thierry Bécart et Thiziri Naït-Saada

**Rapport du PAN 2 – Objectifs fixés**

Nous avions établi pour le PAN 2 les objectifs suivants :

* Définir une liste définitive d’ambiances
* Se procurer un DataSet de textes affiliés à ces ambiances
* Vectoriser ces textes
* Obtenir un espace vectoriel de textes de dimension raisonnable
* Commencer à utiliser des classifieurs « classiques » sur notre DataSet
* En observer les premiers résultats

**Rapport du PAN 2 – État de l’avancement**

À ce jour, nous avons pu procéder aux étapes suivantes :

* Définition d’une liste définitive de **44 ambiances** par l’ensemble du groupe 3.5 (cf annexe)
* Obtention d’environ 70 à 100 textes pour chaque ambiance, permettant finalement d’avoir un DataSet constitué de plus de **3500 textes**
* **Vectorisation** de ces textes permettant d’établir le « vocabulaire » de notre classification ; ce dictionnaire étant constitué de plus de dizaine de milliers de mots

Pour réduire la dimension de l’espace vectoriel engendré, la vectorisation a fait appel à :

- une étape de pré-processing à l’aide de Stop-Words, de lemmatization, tokenization et d’application de la TF-IDF.

- une réduction de dimension par la PCA : la dimension retenue est n = 30

* Application de **classifieurs** « classiques » à notre jeu de données. Pour chacun d’eux, une batterie de tests a été réalisée en faisant varier chacun des paramètres d’entrée, ainsi que la proportion p du DataSet utilisée pour les tests. Pour chacun des classifieurs, nous avons retenu les paramètres d’entrée (et la proportion de DataTraining) qui permettait d’atteindre le maximum de taux de réussite.

Les classifieurs testés sont :

- KNN

- Linear SVC

- DecisionTree

- Naive Bayes

Toutes ces étapes sont détaillées dans la suite du rapport. Nous obtenons finalement un maximum de taux de réussite de 34.2% (à comparer à 2,2% que donnerait le hasard) grâce au classifieur LinearSVC, pour lequel les paramètres sont les suivants : c = 90 et p = 0.8

**Rapport du PAN 2 – Perspectives pour le PAN 3**

Nous espérons pouvoir améliorer ce taux de réussite grâce aux méthodes par Adaboost et si cela ne suffit pas, nous aimerions trouver des alternatives de classification plus adaptée à notre sujet. De plus, nous envisageons de mettre en place des indicateurs de réussites moins binaires. Nous prévoyions déjà d’obtenir les 3 meilleurs choix d’après le modèle et de pondérer la réussite par rapport à chacune de ces 3 choix. Aussi, nous allons regrouper les ambiances similaires afin de se rapprocher du nombre total d’ambiances que le projet pourra reconstituer. Ce nombre sera déterminé par le module consacré à leur émission.

**Rapport du PAN 2 – Articulation avec les autres modules**

Le seul module avec lequel le module Classification communiquera est le module Base de Données. Celui-ci dispose des textes sous format epub mais c’est lui-même qui se chargera d’importer les méthodes pour le convertir en format txt. Ainsi, nous prendrons en entrée un fichier sous format txt, auquel nous appliquerons la vectorisation puis la classification, et nous enverrons à ce même module un string indiquant l’ambiance retenue par le classifieur. En annexe, vous trouvez également un diagramme reflétant notre conception des articulations entre les différents modules.

*Diagramme représentant les interactions entre les différents modules*

