



```
{
  "categories": {
    "names": [ "Restaurant", "Hôtel", "Train", "Avion", "Parking", "Péage" ],
    "topics": [ {
      "Restaurant": [ "café", "jus", "resto", "chicken", "pizza", "sandwich", "frite", "glace", "table", "couvert", "eau",
                     "burger", "food", "boisson", "plat", "repas", "restau", "dessert" ]
    }, {
      "Hôtel": [ "room", "chambre", "nuit", "hôtel", "reception", "étage", "repas", "appartement", "résidence", "jour" ]
    }, {
      "Train": [ "train", "station", "arrivé", " métro", "classe", "depart", "gare", "banlieue", "cabine", "arrêt" ]
    }, {
      "Avion": [ "aéroport", "départ", "arrivé", "classe", "direct", "escale", "repas", "air", "vol", "passeport", "avion" ]
    }, {
      "Parking": [ "voiture", "matricule", "parking", "parc", "véhicule", "parc", "place", "stationnement" ]
    }, {
      "Péage": [ "péage", "véhicule", "autoroute" ]
    }
  }
}
```

Figure 22 : Fichier de classification de la facture

5.5 Extraction du montant

L'extraction du total est la tâche la plus délicate. La réponse de l'API vision contient une séparation du bloc de chaque mot trouvé dans la facture dont chaque bloc est caractérisé par le mot détecté et les 4 couples de coordonnées (x,y) proportionnelle au taille du facture.

En se basant sur la réponse de l'API vision, notre solution était une combinaison de deux algorithmes. Le premier consiste à jouer sur les mots clés qui peuvent être associés au total de la facture. Nous indiquons à l'algorithme la liste des mots clés à recherchés.

Ensuite, l'algorithme doit tester la présence de ces mots clés dans le texte de la facture. Pour chaque mot clés trouvés, l'algorithme va concaténer toutes les mots se trouvant sur la même ligne (en utilisant l'axe y) et devant le mot clés recherché dans une chaîne de caractère et appliquera la fonction regex pour détecter le réel trouvant dans cette chaîne de caractère qui peut être la valeur du total. Chaque réel trouvé est stocké dans un tableau. Enfin, le total de la facture doit être le maximum de ce tableau.

Mais, si ce n'est pas le cas et la facture ne contient pas aucune des mots clés recherchés la solution passera à utiliser le deuxième algorithme.

La réponse de l'API vision comprend une représentation du texte sous forme des blocs. Le deuxième algorithme consiste à parcourir l'image bloc par bloc. L'algorithme consiste à éliminer les blocs qui contiennent les mots comme tva, t.v.a, tél, km, numéro afin d'assurer d'éliminer les réels qui indique le taux de tva, le nombre des kilomètres, le numéro de la facture,