Projet TAL - RI

L'objectif du projet est de réaliser un système de recherche d'information dans une collection de recettes de cuisine en anglais. Le projet se décompose en 3 étapes (voir ci-dessous pour le détail).

Le projet se fera par groupe de **2 étudiant.e.s** et le rendu final inclura :

- le code source (de préférence en Python, sous forme de notebooks jupyter)
- les données utilisées (ressources externes éventuelles, corpus, modèles, etc.)
- un rapport de 6 pages qui expliquera votre méthode, les expériences réalisées et les scores obtenus, la répartition du travail au sein du groupe, une bibliographie (références / sites consultés pour réaliser le travail)

Date limite pour rendre le projet : **24 mai 2020**

Etape 1

Vous avez à votre disposition sur Moodle un fichier CSV (`train_cuisine.csv`) de recettes de cuisine[1] en anglais étiquetées avec leur origine géographique (colonne 1, appelée 'cuisine' ci-dessous). Les recettes ont été pré-traitées de manière à ne conserver que les ingrédients (colonne 2, appelée 'ingredients' ci-dessous). Lorsqu'un ingrédient inclut plusieurs mots, les mots sont joints à l'aide du caractère '_' (*underscore*). La figure ci-dessous montre un extrait du fichier :

cuisine	ingredients
greek	romaine_lettuce black_olives grape_tomatoes ga...
southern_us	plain_flour ground_pepper salt tomatoes ground...
filipino	eggs pepper salt mayonaise cooking_oil green_c...
indian	water vegetable_oil wheat salt
indian	black_pepper shallots cornflour cayenne_pepper...

La première recette a pour origine géographique 'greek' (grecque) et contient la liste d'ingrédients 'romaine_lettuce black_olives grape_tomatoes garlic pepper purple_onion seasoning garbanzo_beans feta_cheese_crumbles'

L'objectif est d'entraîner un outil de classification automatique des recettes par origine géographique, à l'aide de la liste d'ingrédients. L'apprentissage se fera à l'aide de scikit-learn. Vous pourrez comparer

1 Source : https://www.kaggle.com/c/whats-cooking

différents algorithmes de classification et l'utilisation de différents types de traits déduits de la liste des ingrédients (ingrédients, nombre d'ingrédients, utilisation de la lemmatisation ou de la désuffixation, application de différents seuils de fréquence, etc. : à vous de trouver et de tester des traits éventuellement pertinents).

Etape 2

Le meilleur modèle obtenu à l'étape 1 devra être appliqué à un nouveau corpus de recettes, afin de prédire automatiquement leur origine géographique.

Vous trouverez sur moodle trois corpus de recettes en format JSON[2]. Vous en choisirez un et lui appliquerez le meilleur modèle. Gardez l'origine géographique prédite pour chaque recette, car vous en aurez besoin pour l'étape 3. Vous pourriez par exemple ajouter dans chaque recette un nouveau champ contenant la valeur prédite. La figure ci-dessous montre le format des recettes.

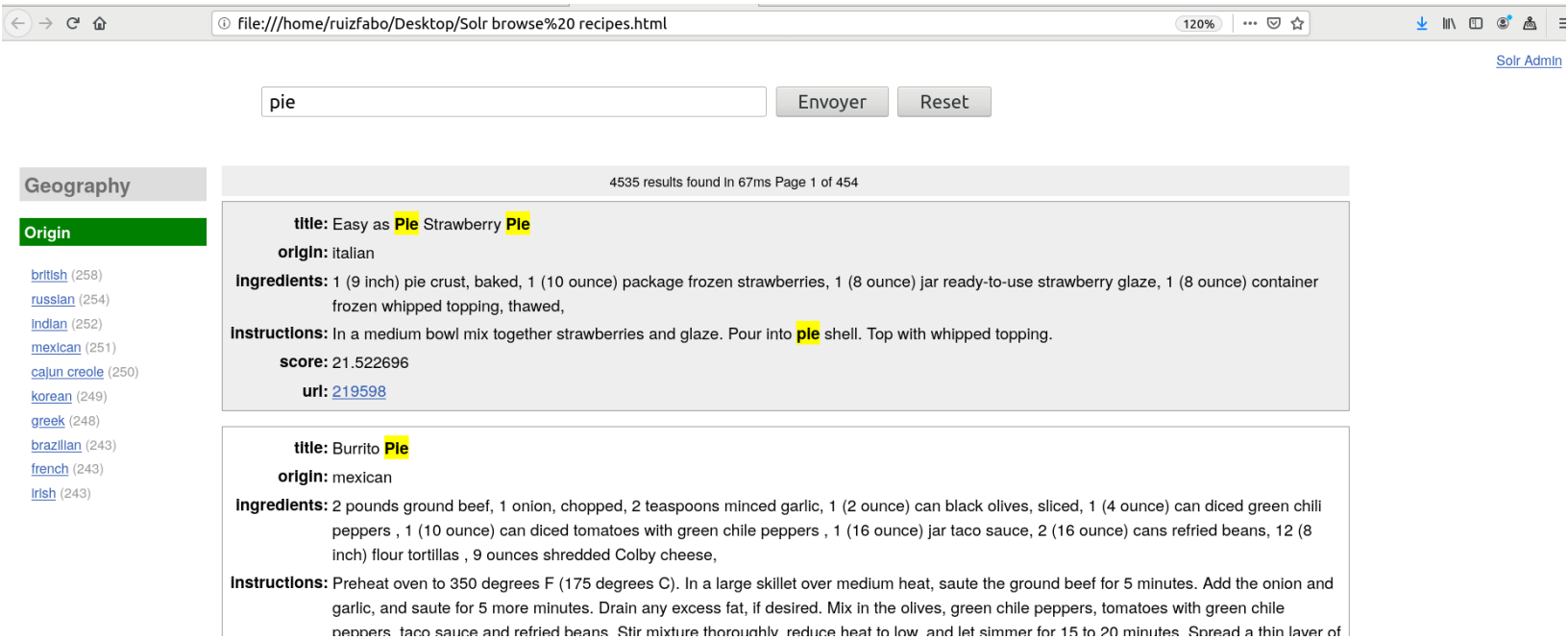

```
  "I6n88sCMwovOQVAXLLJXh9137stmW66": {
    "title": "Baked Tortilla Chips",
    "ingredients": [
      "1 (12 ounce) package corn tortillas ADVERTISEMENT",
      "1 tablespoon vegetable oil ADVERTISEMENT",
      "3 tablespoons lime juice ADVERTISEMENT",
      "1 teaspoon ground cumin ADVERTISEMENT",
      "1 teaspoon chili powder ADVERTISEMENT",
      "1 teaspoon salt ADVERTISEMENT",
      "ADVERTISEMENT"
    ],
    "instructions": "Preheat oven to 350 degrees F (175 degrees C).\nCut each tortilla i
nto 8 chip sized wedges and arrange the wedges in a single layer on a cookie sheet.\nIn
a mister, combine the oil and lime juice. Mix well and spray each tortilla wedge until s
lightly moist.\nCombine the cumin, chili powder and salt in a small bowl and sprinkle on
 the chips.\nBake for about 7 minutes. Rotate the pan and bake for another 8 minutes or
until the chips are crisp, but not too brown. Serve with salsas, garnishes or guacamole.
\n",
    "picture_link": "n/e.8yf.h3ffr7tNZRUfMtI4lk0f3ae"
  },
```


Etape 3

Dans cet étape, il s'agit d'indexer un minimum de 10 000 recettes, parmi celles dont vous avez prédit l'origine géographique dans l'étape 2, dans une instance du serveur de recherche Solr[3]. Vous allez ensuite créer une interface simple de recherche en texte intégral, qui fournira également la possibilité de filtrer les résultats selon l'origine géographique des recettes. La fonction de recherche par facettes de Solr sera exploitée pour ce filtrage. Voir ci-dessous pour un exemple.

2 Source : https://storage.googleapis.com/recipe-box/recipes_raw.zip

3 L'outil Solr aisi que son interface graphique de recherche par défaut seront présentés dans la deuxième partie du cours.



Exemple de l’interface qui peut être créée. Dans cet exemple, la configuration Solr et les templates du TP ont été adaptés pour afficher une facette avec les origines des recettes.