Déterminer la langue d'un texte*

LI323 — Statistique et informatique

2014-2015

Résumé

L'objectif de ce projet est de construire une fonction permettant de déterminer automatiquement à quelle langue appartient un mot, comme dans les deux examples suivants :

— `guess_language('pommme')` → `français`

— `guess_language('apple')` → `anglais`

Vous devrez rendre votre code et un compte-rendu de quelques pages après le TME 4 (et avant le TME 5 !). Le projet peut se faire en binôme. Le code doit être écrit en langage Java. Le rendu du projet doit suivre la démarche suivante :

— envoyez un e-mail à l'adresse nicolas.baskiotis@lip6.fr, le sujet de l'e-mail étant :
`[li323-projet1] nom-etu1 nom-etu2`

— l'e-mail doit contenir, en attachement, le compte rendu du projet (dans un format standard, .doc, .ods ou .pdf), et une archive contenant votre code Java.

1 Construction d'un modèle de langage

1.1 Principe

La méthode d'identification que nous proposons de mettre en œuvre est fondée sur l'observation que la fréquence d'apparition d'une lettre dépend de la langue dans laquelle on écrit. Par exemple un texte français comportera beaucoup plus de `e` qu'un texte écrit en anglais ou en italien.

Dans la suite de ce document, nous noterons $\mathbf{w}$ un mot composé de n lettres $(w_i)_{i=1}^n$. Nous supposerons que, pour une langue donnée, la probabilité d'observer un mot ne dépend que de la probabilité d'apparition de ses lettres :

$$p(\mathbf{w}|l) = \prod_{i=1}^{n} p(w_i|l)$$

La relation précédente est fondée sur l'hypothèse simplificatrice (mais fausse) que les lettres d'un mot sont mutuellement indépendantes.

*inspiré d'un sujet de Guillaume Wisniewski, Nicolas Usunier

Si l'on dispose d'un corpus (c.-à-d. un ensemble de documents) écrits dans la langue l, il est possible de déterminer automatiquement les probabilités $p(w_i|l)$ en la confondant avec la fréquence d'apparition de la lettre w_i dans le corpus :

$$p(w_i|l) = \frac{\text{nombre d'occurence de } w_i}{\text{nombre de caractères}} \tag{1}$$

les comptages se faisant sur l'ensemble des mots des documents rédigés dans la langue l.

1.2 Prédiction de la langue d'un mot

Nous allons utiliser les probabilités calculées au paragraphe précédent pour *prédire* la langue d'un mot. Pour cela, nous allons considérer la *fonction de décision* suivante :

$$l^* = \underset{l \in \mathcal{L}}{\operatorname{argmax}}\, p(l|\mathbf{w})$$

où $\mathcal{L}$ est l'ensemble des langues que notre système connaît, w le mot considéré et l^* est le résultat de la prédiction.

La probabilité $p(l|\mathbf{w})$ peut être déterminée grâce à la formule de Bayes :

$$p(l|\mathbf{w}) = \frac{p(\mathbf{w}|l) \cdot p(l)}{p(\mathbf{w})}$$

A quoi correspondent les différents termes de cette relation ? Peuvent-ils être tous calculer ? En particulier, est-il nécessaire de déterminer la probabilité $p(\mathbf{w})$? Quelle hypothèse faut-il faire pour $p(l)$?

2 Réalisation

On dispose d'un ensemble de documents écrits dans différentes langues [1]. Les documents ont été pré-traités de la manière suivante :

— toutes les majuscules ont été converties en minuscules ;

— toutes les lettres accentuées et ont été supprimées ;

— tous les signes de ponctuation ont été supprimés.

Les questions suivantes sont indicatives. Vous pouvez organiser vos classes de la manière qui vous convient. Faites attention cependant à la rapidité algorithmique de votre implémentation (en particulier, ne lire qu'une fois chaque corpus !).

Ecrivez une classe qui prend en charge la lecture d'un corpus pour une langue donnée et qui calcule son modèle de langage. En particulier, elle doit permettre de :

1. de lire un corpus d'une langue donnée.
2. Compter le nombre de fois qu'une lettre w est utilisée.
3. Calculez la probabilités $p(w|l)$ de la langue choisie pour la lettre w à l'aide de l'Équation 1.

1. Ces documents sont téléchargeables à l'adresse http://www-connex.lip6.fr/~usunier/li323/corpus.tar.gz.

4. Calculez la probabilité du mot $p(\mathbf{w}|l)$

Tester votre classe et déterminer les deux probabilités suivantes : $p(\text{statistics}|\text{anglais})$; $p(\text{probability})$.

Vous pouvez également tracer la loi de probabilité de chaque langue.

5. Ecrivez une classe permettant de déterminer la langue d'un mot.
6. Pour évaluer les performances du système, on considère la fonction d'erreur suivante :

$$\ell = 1 - \frac{\text{nombre de réponses fausses}}{\text{nombre de réponses}}$$

Programmez cette fonction et testez les performances de votre système sur la base d'exemples suivante :

mot	étiquette
fatta	italian
ora	italian
che	italian
dato	italian
volta	italian
by	english
other	english
mean	english
statistics	english
chocolate	english
president	english
thanks	english
patatoes	english
constitutionnellement	french
peter	english
pomme	french
daar	dutch

TABLE 1 – Liste des mots pour tester votre système de prédiction

Nous avons supposé, dans la partie précédente, que nous ne disposions pas d'information à priori permettant d'estimer $p(l)$. Pourtant dans de nombreux cas, cette hypothèse est fausse. Par exemple, si l'on souhaite prédire la langue d'un document choisi au hasard sur l'internet, on sait qu'il est plus probable que celui-ci soit rédigé en anglais plutôt qu'en français, la majorité des documents de l'Internet étant rédigé en anglais. Ce type de connaissance peut être introduit dans notre fonction de décision par l'intermédiaire de la probabilité à priori $p(l)$.

On suppose que les exemple de la Table 1 sont tirés selon $p(l)$. Il est alors possible de confondre la probabilité $p(l)$ avec la fréquence :

$$p(l) = \frac{\text{nombre de mots écrit dans la langue } l}{\text{nombre de mots dans la base}}$$

8. Évaluez les performances de votre système sur les exemples précédents en prenant en compte la probabilité à priori $p(l)$.

3 Amélioration du modèle

Selon le temps dont vous disposez, il est possible d'effectuer de nombreuses améliorations de la fonction de prédiction.

En particulier, vous pouvez :

— Augmenter la taille de données disponibles :

En téléchargeant des documents sur le Web en différentes langues (par exemple sur Wikipedia, Gutenberg), vous pouvez augmenter la taille de votre corpus de documents utilisé pour l'estimation, ainsi que la taille de l'ensemble de test (l'ensemble utilisé pour évaluer votre fonction).

7. Quel est l'intérêt d'agrandir la base de test ?
8. Comment évoluent les performances en fonction de la taille de l'ensemble d'estimation ?

— Utiliser un modèle plus compliqué :

L'hypothèse d'indépendance mutuelle des lettres d'un mot est extrêmement simplificatrice. Un modèle moins simpliste consiste à supposer que, connaissant la langue, la probabilité d'apparition d'une lettre dépend de la lettre précédente du mot (mais est indépendante des autres connaissant cette lettre).

1. Comment s'écrit la probabilité d'un mot sachant la langue avec ce nouveau modèle ?
2. Quelle règle utiliser pour estimer les probabilités nécessaire à calculer la probabilité d'un mot ?
3. Programmez la fonction de prédiction correspondant à ce nouveau modèle. Quelles sont ses performances en fonction de la taille de la base d'estimation ?

Avec de légères adaptations, vous pouvez également envisager d'autres applications (identification de l'auteur d'un texte par exemple, de quelle manière ?).

Quelques conseils sur le rapport

Votre rapport devra contenir au moins trois parties (en dehors de l'introduction et de la conclusion), la première expliquant la modélisation mathématique et la structure de votre programme (sans code ! ! !), la deuxième les expériences que vous avez conduites et une interprétation, la troisième les améliorations et les adaptations que vous avez envisagées. Rédigez-le avec soin !