Master 2 IPAC – Apprentissage numérique - TP3 - Chaines de Markov

vincent.thomas@loria.fr

13 octobre 2015

1 Le Chantomatic

Dans ce TP, vous allez construire un modèle (très simpliste) de langage à partir d'un corpus. Pour cela, vous allez construire une chaine de Markov de la manière suivante

- l'espace d'état de la chaine de Markov correspond à l'ensemble des mots du corpus
- la matrice de transition $p(s_t/s_{t-1})$ correspond à la probabilité d'avoir le mot s_t quand le mot précédent est le mot s_{t-1}

1.1 Implémentation

Le plus simple pour représenter la matrice de transition est de la stocker dans un objet de type `Map<String,Map<String,Integer>>`.

- la clef correspond au mot précédent s_{t-1}
- la map associée correspond à la distribution des occurences sur le mot suivant s_t
- la probabilite peut donc se construire en faisant le rapport des occurences du mot dont on cherche la probabilité sur les somme des occurences des mots.

Grâce aux tables, la construction des bigrammes se fera au fur et à mesure de la rencontre des mots.

1.2 Apprendre une chaine de Markov

1.2.1 Exemple simple

Supposons qu'on dispose du texte suivant pour apprendre les bigrammes

```
bonjour comment ca va bonjour ca va salut ca va bien salut comment ca va bonjour
```

Question 1
Décrire la matrice de transition obtenue après apprentissage de la chaine de Markov.

1.2.2 Texte fourni

On fournit sur arche différents fichiers textes

- un fichier `Beyonce.txt` qui contient les paroles de différentes chansons de Beyonce ;

- un fichier `Beyonce.txt.traite.txt` qui est le même fichier pré-traité, les mots sont disposés un par ligne et les caractères spéciaux ont été supprimés ;
- les fichiers `Manowar.txt` et `Manowar.txt.traite.txt` qui contiennent les paroles de différentes chansons de manowar.

Pour éviter des problèmes de début ou fin de fichier, au début de la construction du bigramme, on supposera que le premier mot est le mot "." et qu'a la fin, on fera une transition vers ce même mot ".". Le mot "." n'interviendra qu'une seule fois mais permet de faire boucler la chaine de markov.

Question 2
A l'aide des fichiers textes fournis sur arche, écrire une classe permettant de construire un modèle de langage sur les paroles de Beyonce ou de Manowar.

1.3 Estimer la probabilité d'émission

1.3.1 Exemple simple

Question 3
A partir de l'exemple simple-pretexte, quelle est la probabilité que le texte émis soit "bonjour ca va salut" sachant que le premier mot est "bonjour"

1.3.2 Application

Question 4
Dans la classe Bigramme construite, ajouter une méthode calculant la probabilité d'émettre une suite de mots.

Question 5
Calculer la probabilité que la phrase `"oh i asked you"` soit émise sachant que la phrase commence par `"oh"` ?

1.4 Générateur de chanson

Question 6
Ecrire un générateur de chanson en effectuant des échantillon sur les transitions.

2 Tracking dans labyrinthe

Le problème de tracking consiste à estimer la position d'un individu au cours de ses déplacements. La première brique d'un système de tracking à base d'inférence bayésienne est de disposer d'un modèle de l'individu.

Pour cela, cette partie vous propose de construire un modèle (très simpliste) d'un individu se déplaçant dans un environnement contraint. On suppose ainsi que le personnage considéré

se déplace d'une case vers une autre dans un environnement 2D discrétisé. Chaque case de cet environnement dispose d'un ensemble de directions données.

Lorsque l'individu doit choisir une direction, il choisit au hasard une direction parmi les directions proposées sur la case où il se trouve.

ATTENTION : l'individu ne fait jamais demi tour, il ne peut pas prendre une direction opposée à celle qu'il vient de suivre.

Question 7
Définir précisément l'espace d'état de la chaine de markov représentant l'évolution de l'individu.

On suppose que l'individu arrive dans la case centre du labyrinthe ci-dessous en venant de l'Ouest (représenté par l'endroit indiqué par le personnage dessiné - cf figure 1).

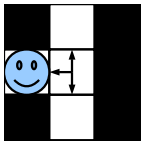

FIGURE 1 – Agent qui vient de l'Ouest

Question 8
donner la distribution $P(S_{t+1}/S_t = S_0)$ où S_0 correspond à son état courant (lorsqu'il est au centre du labyrinthe). Vous pouvez bien entendu définir les états atteignables et leur donner un nom.

On suppose désormais que l'individu vient du sud (cf figure 2)

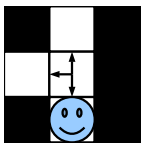

FIGURE 2 – Agent qui vient du Sud

Question 9
Donner la distribution $P(S_{t+1}/S_t = S_0)$ où S_0 correspond à son état courant.

Question 10
Les deux distributions sont elles les mêmes ? S_0 est-il le même ? Commentez éventuellement.