Introduction au Traitement Automatique des Langues

Examen du 23 juin 2017, de 15h30 à 18h30.
Les documents sont autorisés.

1 Expressions régulières

1. Définir l'expression régulière qui reconnaît les lignes commençant et finissant par un chiffre.
2. Définir l'expression régulière qui reconnaît les lignes qui ne contiennent ni a ni A.
3. Définir l'expression régulière qui reconnaît les mots qui commencent par une majuscule
4. Définir l'expression régulière qui reconnaît un nombre réel en notation scientifique.

2 Modèles de langue n-gramme

Soient les comptes suivant collectés sur un corpus de l'anglais:

w	$c(w)$	w_1, w_2	$c(w_1, w_2)$
person	123	person she	5
she	8907	she was	926
was	10209	was inferior	0
inferior	46	inferior to	9
to	24042	to both	12
both	418	both sisters	5

Considérons un modèle bigramme :

1. Donner les paramètres et la probabilité $P(s\text{he was inferior to both sisters})$ pour une estimation par maximum de vraisemblance.
2. Donner les paramètres et la probabilité $P(s\text{he was inferior to both sisters})$ pour une estimation par maximum de vraisemblance avec lissage "$+\lambda$" en posant $\lambda = 1$.

3 Modèles de Markov cachés

Soit la table suivante contenant les comptes de bigrammes d'étiquettes observées sur un corpus anglais, le Brown Corpus, qui indique par exemple que c(VB,BEZ) = 42 :

	AT	BEZ	IN	NN	VB	PERIOD	total
AT	0	0	0	48636	0	19	48655
BEZ	1973	0	426	187	0	38	2624
IN	43322	0	1325	17314	0	185	62146
NN	1067	3720	42470	11773	614	21392	81036
VB	6072	42	4758	1476	129	1522	13999
PERIOD	8016	75	4656	1329	954	0	15030
							223490

On suppose que le nombre total d'occurrences d'une étiquette peut s'obtenir en sommant sur une ligne (pour BEZ 1973 + 0 + 426 + 187 + 0 + 38)

Soit également la table suivante qui indique les comptes d'observations de certains mots en fonction des étiquettes :

	AT	BEZ	IN	NN	VB	PERIOD
bear	0	0	0	10	43	0
is	0	10065	0	0	0	0
move	0	0	0	36	133	0
on	0	0	5484	0	0	0
president	0	0	0	382	0	0
progress	0	0	0	108	4	0
the	69016	0	0	0	0	0
.	0	0	0	0	0	48809
total	69016	10065	5484	536	180	48809

Cette table indique entre autres que l'on a vu le mot *president* avec l'étiquette *NN* 382 fois.

1. Calculez selon l'estimation par maximum de vraisemblance les probabilités bigrammes P(AT|PERIOD), P(NN|AT), P(BEZ|NN), P(IN|BEZ), P(AT|IN)
2. Pour chaque étiquette t^k, calculez les probabilités suivantes (estimation par maximum de vraisemblance) :
   - P(bear|t^k), P(is|t^k), P(move|t^k), P(president|t^k), P(progress|t^k), P(the|t^k)
3. Calculez les probabilités des séquences d'étiquetages suivantes:
   - P(AT NN BEZ IN AT NN | the bear is on the move)
   - P(AT NN BEZ IN AT VB | the bear is on the move)

4 Analyse syntaxique

Soit la PCFG suivante $G = (N, T, S, P, q)$ avec:

- $N = \{S, NP, VP, DT, NBAR, NN, \}$,
- $T = \{$walks, the, champion, world, football$\}$,
- les règles de P avec leur poids q associés :

Règle	poids
$S \rightarrow NP\ \ VP$	1,0
$NP \rightarrow D\ \ NBAR$	1,0
$NBAR \rightarrow N$	0,7
$NBAR \rightarrow NBAR\ \ NBAR$	0,3
$VP \rightarrow walks$	1,0
$DT \rightarrow the$	1,0
$NN \rightarrow hampion$	0,1
$NN \rightarrow world$	0,2
$NN \rightarrow ootball$	0,7

1. Dérouler l'algorithme CKY et donner l'analyse la plus probable pour *la phrase the football world champion walks* ainsi que sa probabilité.
2. Proposez une modification de l'algorithme CKY (sans pointeur arrière) qui calcule non pas la probabilité de la meilleure analyse pour une phrase (i.e. $\max_{t \in \mathcal{D}_s} p(t)$) mais la probabilité de la phrase $p(s) = \sum_{t \in \mathcal{D}_s} p(t)$.