2.2.3.2 Exemple détaillé (repris de Jurafsky et Martin (2019))

On calcule une matrice avec m lignes (les mots dont on cherche une représentation), et c colonnes (les contextes). La matrice peut-être carrée avec exactement les mêmes éléments en ligne et en colonne, mais ce n'est pas nécessaire, comme on le verra dans les exemples suivants.

Soit f_{ij} la fréquence d'apparition du mot $i \in [1, m]$ dans le contexte $j \in [1, c]$. Voici la matrice f que nous allons prendre comme exemple (fictif) :

f_{ij}	computer	data	pinch	result	sugar	Σ
apricot	0	0	1	0	1	2
pineapple	0	0	1	0	1	2
digital	2	1	0	1	0	4
information	1	6	0	4	0	11
Σ	3	7	2	5	2	19

La probabilité conjointe p_{ab} est définie comme le ratio entre la fréquence des éléments conjoints et le nombre total d'occurrences de tous les mots dans tous les contextes :

$$p_{ab} = \frac{f_{ab}}{\Sigma_{i=1}^{m}\Sigma_{j=1}^{c} f_{ij}}$$

La probabilité p_{a*} est la probabilité d'occurrence du mot a, on peut la calculer en déterminant le nombre total d'occurrences du mot divisé par le nombre total d'occurrences de tous les mots, mais on peut la calculer aussi en utilisant la fréquence f_{ij} (idem pour la probabilité d'avoir le contexte b) :

$$p_{a*} = \frac{\Sigma_{j=1}^{c} f_{aj}}{\Sigma_{i=1}^{m}\Sigma_{j=1}^{c} f_{ij}} \qquad p_{*b} = \frac{\Sigma_{i=1}^{m} f_{ib}}{\Sigma_{i=1}^{m}\Sigma_{j=1}^{c} f_{ij}}$$

Avec notre exemple, on obtient :

p_{ij}	computer	data	pinch	result	sugar	p_{a*}
apricot	0/19	0/19	1/19	0/19	1/19	2/19
pineapple	0/19	0/19	1/19	0/19	1/19	2/19
digital	2/19	1/19	0/19	1/19	0/19	4/19
information	1/19	6/19	0/19	4/19	0/19	11/19
p_{*b}	3/19	7/19	2/19	5/19	2/19	19/19

... ce qui donne en valeurs décimales :

p_{ij}	computer	data	pinch	result	sugar	p_{a*}
apricot	0,00	0,00	0,05	0,00	0,05	0,11
pineapple	0,00	0,00	0,05	0,00	0,05	0,11
digital	0,11	0,05	0,00	0,05	0,00	0,21
information	0,05	0,32	0,00	0,21	0,00	0,58
p_{*b}	0,16	0,37	0,11	0,26	0,11	

On peut finalement aboutir au tableau des *ppmi* (les zéros correspondent à des cas où le ratio des probabilités était inférieur à 1, les tirets aux cas où le nombre d'occurrences étant nul, la ppmi n'a pas de pertinence) :

ppmi	computer	data	pinch	result	sugar
apricot	-	-	2,25	-	2,25
pineapple	-	-	2,25	-	2,25
digital	1,66	0,00	-	0,00	-
information	0,00	0,57	-	0,47	-

Jurafsky et Martin (2019) ont repris l'exemple déroulé précédemment et ont calculé l'impact d'un lissage laplacien (de 2). Cela donne tout d'abord des nouveaux décomptes :

f	computer	data	pinch	result	sugar
apricot	2	2	3	2	3
pineapple	2	2	3	2	3
digital	4	3	2	3	2
information	3	8	2	6	2

... ce qui aboutit aux nouvelles probabilités :

p_{ij}	computer	data	pinch	result	sugar	p_{a*}
apricot	0,03	0,03	0,05	0,03	0,05	0,20
pineapple	0,03	0,03	0,05	0,03	0,05	0,20
digital	0,07	0,05	0,03	0,05	0,03	0,24
information	0,05	0,14	0,03	0,10	0,03	0,36
p_{*b}	0,19	0,25	0,17	0,22	0,17	

... et on peut alors comparer le tableau des ppmi calculé initialement avec celui qu'on obtient (toutes les étapes du calcul ne sont pas détaillées) avec un lissage laplacien +2, voir figure 2.9. On peut noter que les valeurs de ppmi sont « lissées », avec des taux d'information mutuelle qui semblent, au premier regard, bien plus comparables. Il faut garder en mémoire cependant que l'exemple que nous avons déroulé est un exemple « jouet ».

ppmi	computer	data	pinch	result	sugar
apricot	-	-	2,25	-	2,25
pineapple	-	-	2,25	-	2,25
digital	1,66	0,00	-	0,00	-
information	0,00	0,57	-	0,47	-
ppmi [add2]	computer	data	pinch	result	sugar
apricot	0,00	0,00	0,56	0,00	0,56
pineapple	0,00	0,00	0,56	0,00	0,56
digital	0,62	0,00	0,00	0,00	0,00
information	0,00	0,58	0,00	0,37	0,00

FIGURE 2.9 – Comparaison de mesures de *ppmi* sur l'exemple de (Jurafsky et Martin, 2019), avec ou sans lissage laplacien +2