

Calcul de la probabilité d'une phrase

Calcul de la probabilité d'une séquence de mots

- Problème récurrent dans plusieurs applications :
 - reconnaissance de la parole
 - reconnaissance optique de caractères
 - traduction automatique
- Qu'attend-t-on du modèle ?
 - attribuer une bonne proba. aux séquences correctes :
 - *les traits très tirés.*
 - *les traits très tirées.*
 - *les très traient tiraient.*
 - *les très très t'iraient.*
 - attribuer une bonne proba. aux séquences attendues :
 - *les traits très tirés.*
 - *lettrés très tirés.*

Modèle unigramme

- On utilise la probabilité d'occurrence des mots de la séquence :

$$P(m_1 \dots m_n) = \prod_{i=1}^n P(m_i)$$

- $P(\text{les traits très tirés}) = P(\text{les}) \times P(\text{traits}) \times P(\text{très}) \times P(\text{tirés})$
- le modèle fait des **hypothèses** évidemment fausses !

séquence	log P
les très très tirés	-14.265
les trait très tiré	-15.3102
les traits très tiré	-15.4028
les trait très tirés	-15.8386
les traits très tirés	-15.9312

La bonne solution est en position 5

Modèle unigramme : estimation

- Comment donner une valeur à $P(m_i)$ avec $1 \leq i \leq |V|$?
- On prend une grande quantité de texte : $o_1 \dots o_n$ (o est une *occurrence* de mot)
- On calcule le nombre d'occurrences du mot m :

$$C(m) = \sum_{i=1}^n \delta_{o_i, m}$$

($\delta_{i,j}$ symbole de Kronecker, $\delta_{i,j} = 1$ si $i = j$, 0 sinon)

- Puis la fréquence relative de m :

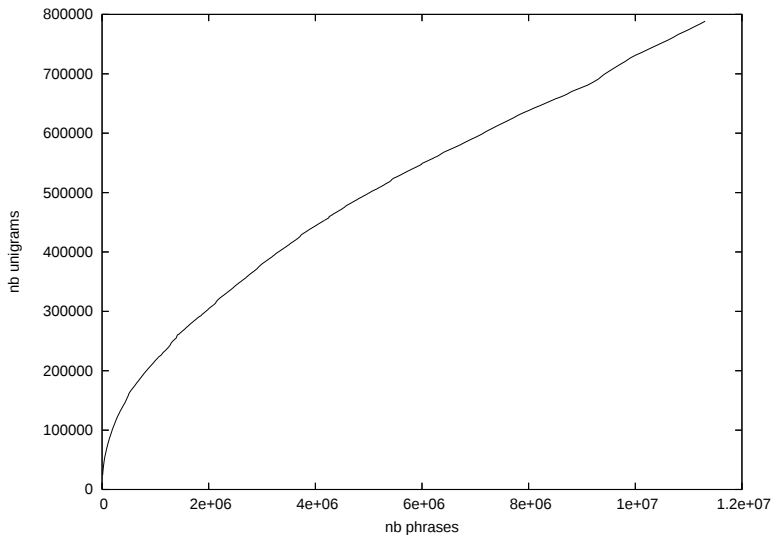
$$P(m) = \frac{C(m)}{\sum_{i=1}^{|V|} C(m_i)}$$

- Il y a $|V|$ probabilités à estimer.

Données d'apprentissage

- Journal Le Monde 1986-2002
- 16 479 270 phrases
- 370 005 285 occurrences

Nombre d'unigrammes différents



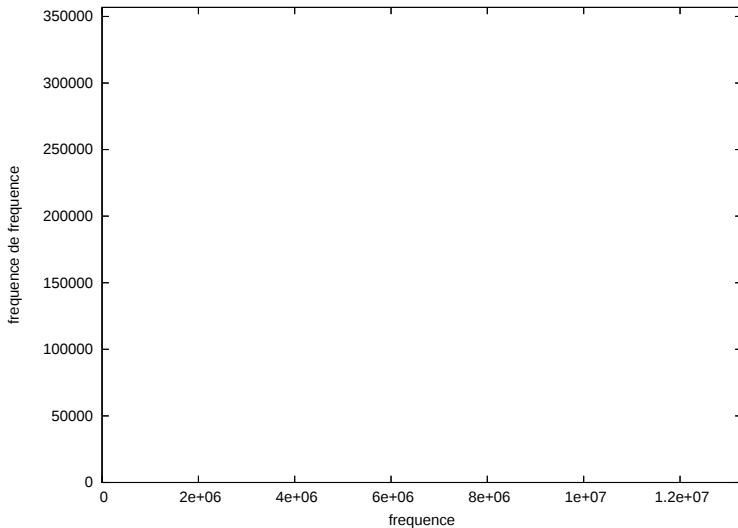
Rang, fréq et fréq. de fréq. des unigrammes

rang	fréq	f2f	unigramme
1	13 286 304	1	de
2	6 964 863	1	la
3	5 900 839	1	le
4	5 599 010	1	l'
5	5 017 018	1	et
6	4 762 293	1	les
7	4 208 264	1	des
8	3 856 293	1	d'
9	3 695 434	1	un
10	3 425 787	1	en

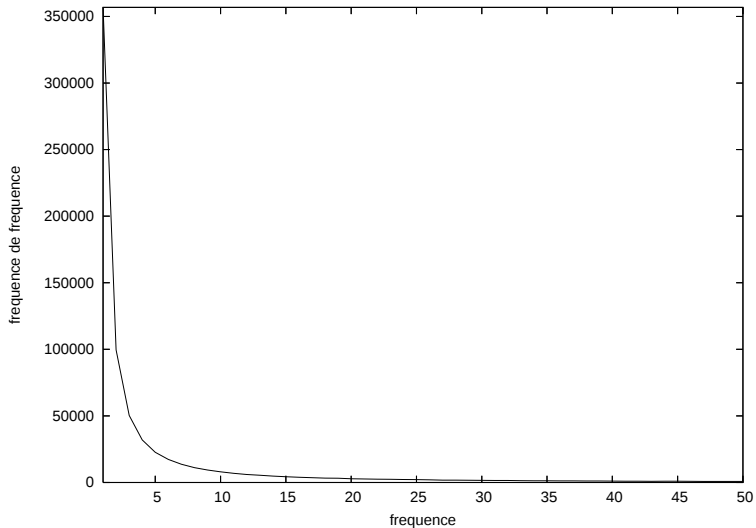
Rang, fréq et fréq. de fréq. des unigrammes

rang	fréq	f2f	unigramme
8532	10	7 992	abaissent, Abbott, antihypertenseur
8533	9	9 369	
8534	8	11 140	
8535	7	13 671	
8536	6	17 351	
8537	5	22 684	
8538	4	31 980	
8539	3	50 311	absolument, absorbable, Ambrozie
8540	2	99 581	
8541	1	356 780	

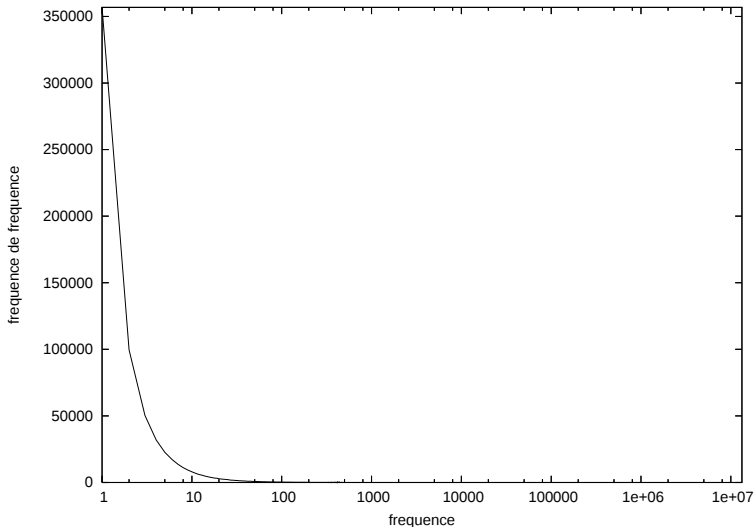
Fréquence des unigrammes, lin-lin



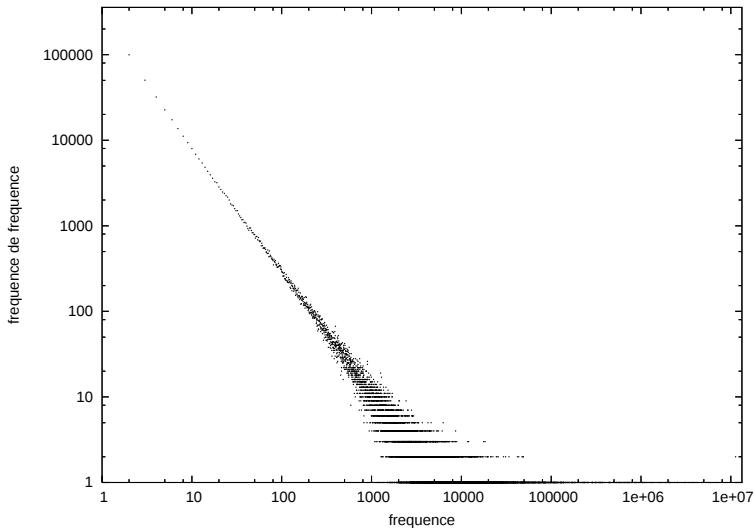
Fréquence des unigrammes, lin-lin



Fréquence des unigrammes, log-lin



Fréquence des unigrammes, log-log



La loi de Zipf

- Dans de nombreux phénomènes humains, il existe une relation linéaire entre la fréquence (f) et l'inverse du rang (r) :

$$f(r) = k \times \frac{1}{r}$$

- De manière grossière : peu de mots très fréquents beaucoup de mots peu fréquents.

Modèle bigramme

- On utilise la probabilité que le mot a suive le mot b :
 $P(m_{i+1} = a | m_i = b)$

$$P(m_1 \dots m_n) = P(m_1) \times \prod_{i=2}^n P(m_i | m_{i-1})$$

- $P(\text{les traits très tirés}) =$
 $P(\text{les}) \times P(\text{traits} | \text{les}) \times P(\text{très} | \text{traits}) \times P(\text{tirés} | \text{très})$

les très très tirés	-13.4767
les traits très tiré	-13.8494
les traits très tirés	-14.1414
les trait très tirés	-18.2462
les trait très tiré	-18.5381

La bonne solution est en position 3

Modèle bigramme : estimation

- On calcule le nombre d'occurrences de la séquence ab :

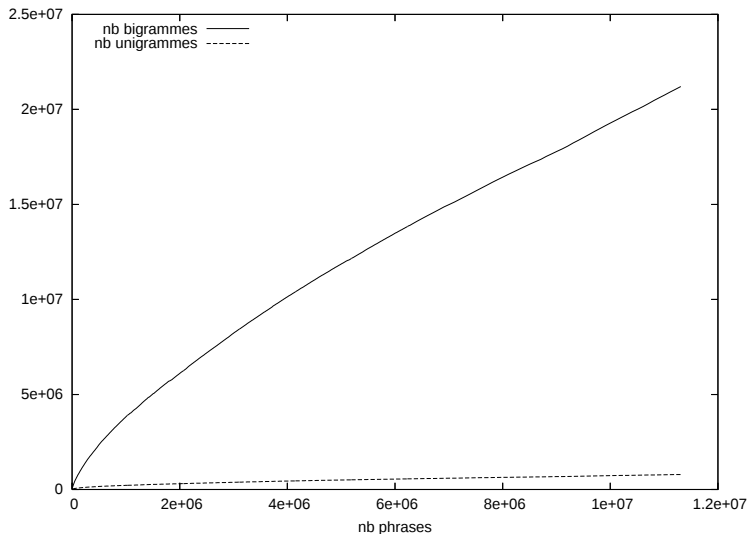
$$C(a, b) = \sum_{i=1}^{N-1} \delta_{o_i, a} \times \delta_{o_{i+1}, b}$$

- Puis la fréquence relative :

$$P(b|a) = \frac{C(a, b)}{C(a)}$$

- $|V|^2$ paramètres à estimer.

Nombre de bigrammes différents



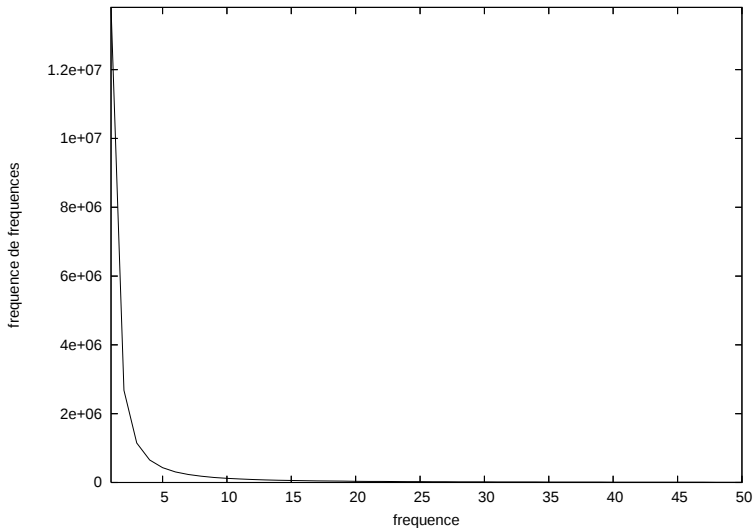
Fréquence des bigrammes

rang	fréq	f2f	bigramme
1	2 091 936	1	de la
2	1 563 496	1	de l'
3	1 156 551	1	neuf cent
4	1 139 331	1	mille neuf
5	921 010	1	cent quatre_vingt
6	744 220	1	d'un
7	578 140	1	d'une
8	571 205	1	c'est
9	556 919	1	et de
10	541 528	1	en mille

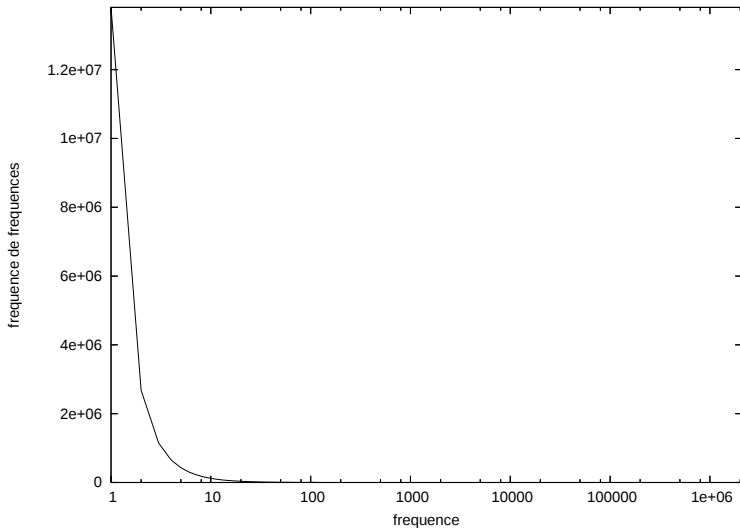
Fréquence des bigrammes

rang	fréq	f2f	bigramme
8817	10	121 362	
8818	9	146 471	
8819	8	181 721	
8820	7	231 738	
8821	6	307 461	
8822	5	429 606	
8823	4	655 244	
8824	3	1 148 479	
8825	2	2 680 674	
8826	1	13 808 569	

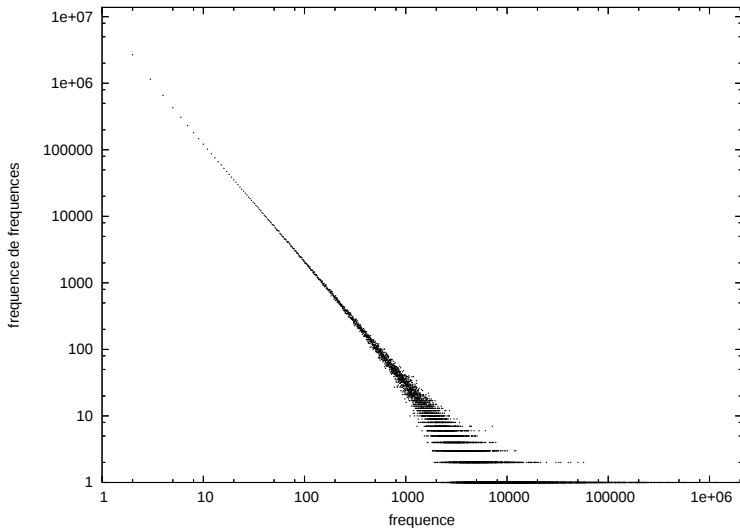
Fréquence des bigrammes, lin-lin



Fréquence des bigrammes, log-lin



Fréquence des bigrammes, log-log



Modèle trigramme

- On utilise la probabilité que c suive la séquence ab :

$$P(m_{i+2} = c | m_i = a, m_{i+1} = b)$$

$$P(m_1 \dots m_n) = P(m_1) \times P(m_2 | m_1) \times \prod_{i=3}^n P(m_i | m_{i-2} m_{i-1})$$

- $P(\text{les traits très tirés}) = P(\text{les}) \times P(\text{traits} | \text{les}) \times P(\text{très} | \text{les, traits}) \times P(\text{tirés} | \text{traits, très})$

les très très tirés	−24.0764
les traits très tirés	−28.134
les traits très tiré	−28.2369
les trait très tirés	−31.0759
les trait très tiré	−31.1788

La bonne solution est en position 2

Modèle trigramme : estimation

- On calcule le nombre d'occurrences de la séquence abc :

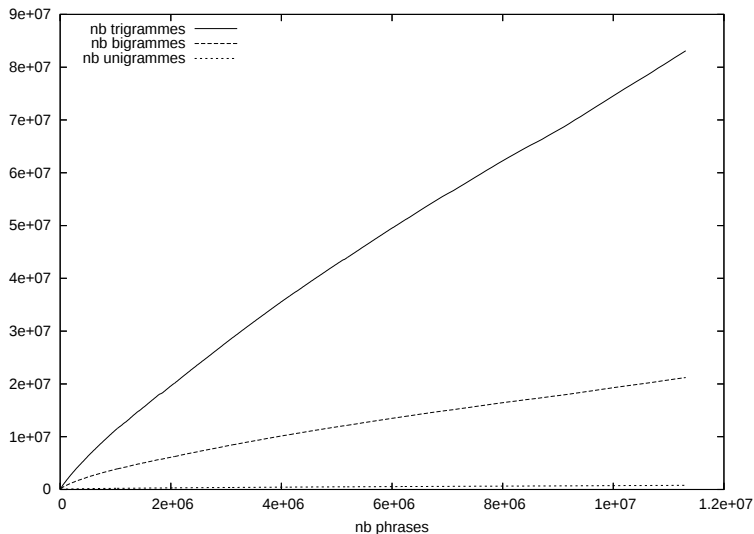
$$C(a, b, c) = \sum_{i=1}^{N-2} \delta_{o_i, a} \times \delta_{o_{i+1}, b} \times \delta_{o_{i+2}, c}$$

- Puis la fréquence relative :

$$P(c|a, b) = \frac{C(a, b, c)}{C(a, b)}$$

- $|V|^3$ paramètres à estimer.

Nombre de trigrammes différents



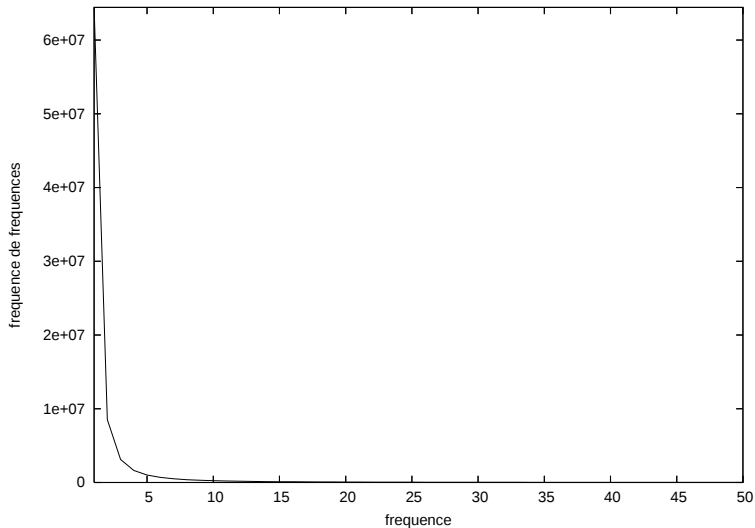
Rang, fréq et fréq. de fréq. des trigrammes

rang	fréq	f2f	trigramme
1	1	1 135 994	mille neuf cent
2	1	792 656	en mille neuf
3	1	379 902	cent quatre_vingt dix
4	1	191 324	n' est pas
5	1	184 626	neuf cent soixante
6	1	167 909	il y a
7	1	160 077	de mille neuf
8	1	145 121	n' a pas
9	1	95 683	et de la
10	1	95 201	cent soixante dix

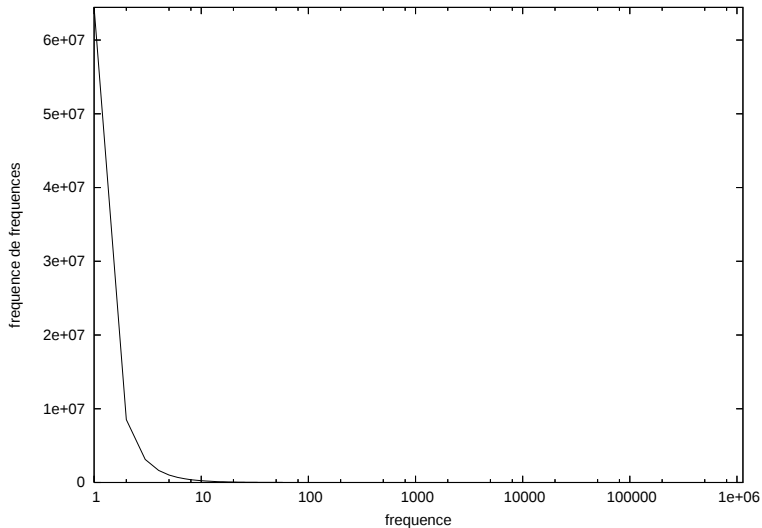
Rang, fréq et fréq. de fréq. des trigrammes

rang	fréq	f2f
5193	10	242 472
5194	9	299 876
5195	8	382 270
5196	7	502 171
5197	6	691 755
5198	5	1 013 295
5199	4	1 641 586
5200	3	3 124 196
5201	2	8 523 984
5202	1	64 425 232

Fréquence des trigrammes, lin-lin



Fréquence des trigrammes, log-lin



Fréquence des trigrammes, log-log

