

### 3.3 Mots suivants

On souhaite effectuer une analyse sur les mots d'un texte. Plus particulièrement, on se propose de savoir pour chaque mot  $m_i$  d'un texte donné, quels sont les mots  $m_{ij}$  qui le suivent. De plus, on souhaite associer à chaque  $m_{ij}$  un compteur  $c_{ij}$  du nombre d'occurrences du couple  $(m_i, m_j)$  dans le texte.

Une fois ces données calculées, on utilise un outil de visualisation de graphes : *graphviz* pour les représenter.

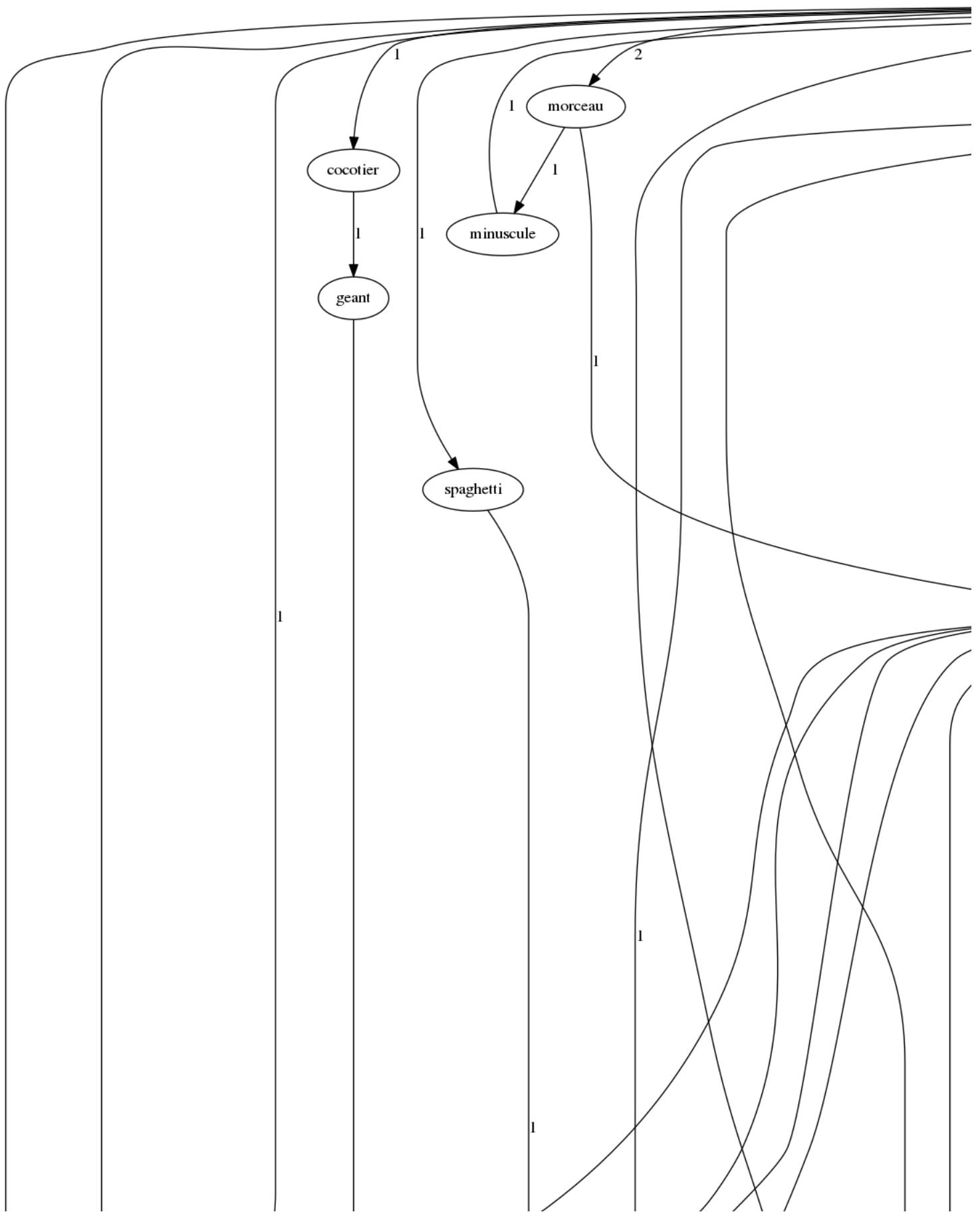
Par exemple, pour le texte suivant (Le serpent python, Charles Trenet) :

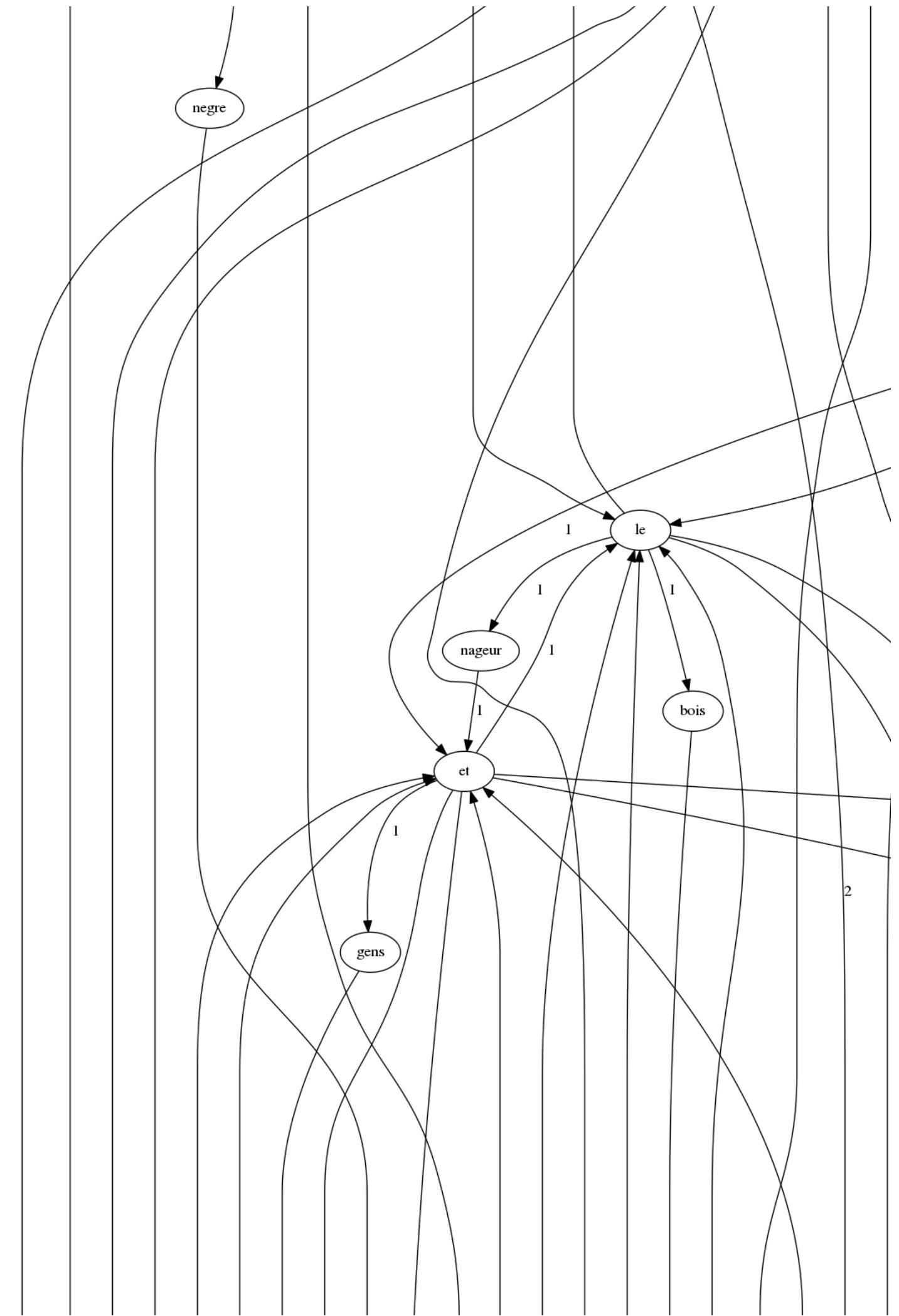
c'est un serpent python  
c'est un python serpent  
qui se promene dans la foret  
pour chercher a devorer  
un beau petit lapin  
ou bien un negre fin.  
car le serpent python a fain  
il a une faim sans fin !  
mais betes et gens sont partis hier  
loues par la Metro Goldwyn Mayer  
pour figurer dans un film de Tarzan  
qui doit rapporter beaucoup d'argent !  
et le serpent piteux  
est triste et se mord la queue  
car il comprend, o desespoir  
qu'il ne mangera pas ce soir.

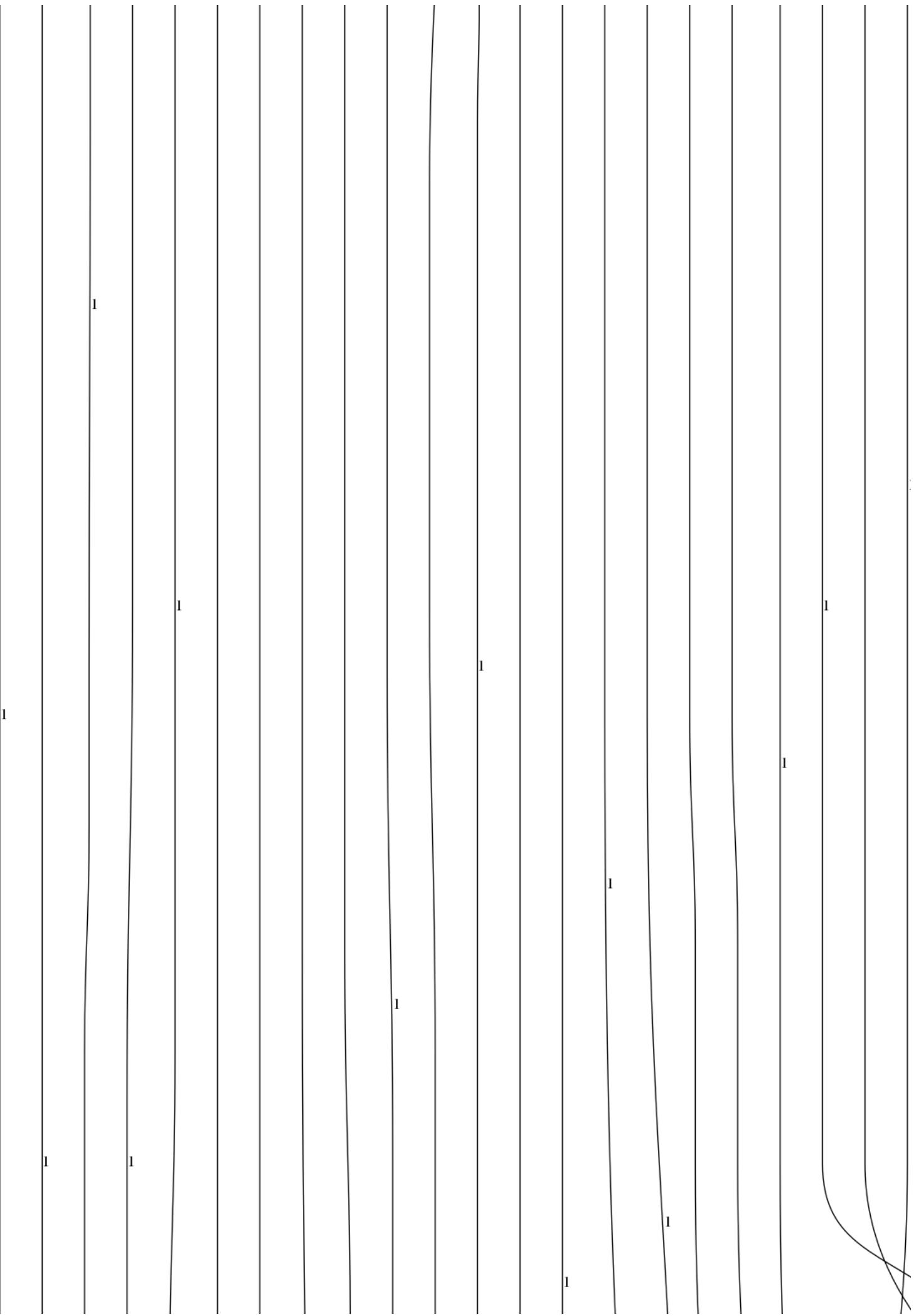
soudain le bois s'eveille  
arrivent des appareils  
de prises de vues de prise de son  
c'est la scene du grand frisson  
on lache des animaux  
des lions et des rhino-  
ceros qu'ont l'air feroce comm' tout  
mais sont doux comme des toutous  
notre serpent du haut d'une branche en l'air  
voit monsieur Johnny Weissmuller  
qui fait joujou avec un elephant  
quel joli tableau pour les enfants.  
mais tant de cinema  
ne remplit pas l'estomac  
du pauvre serpent qui n'aura pas  
qui n'aura pas de repas.

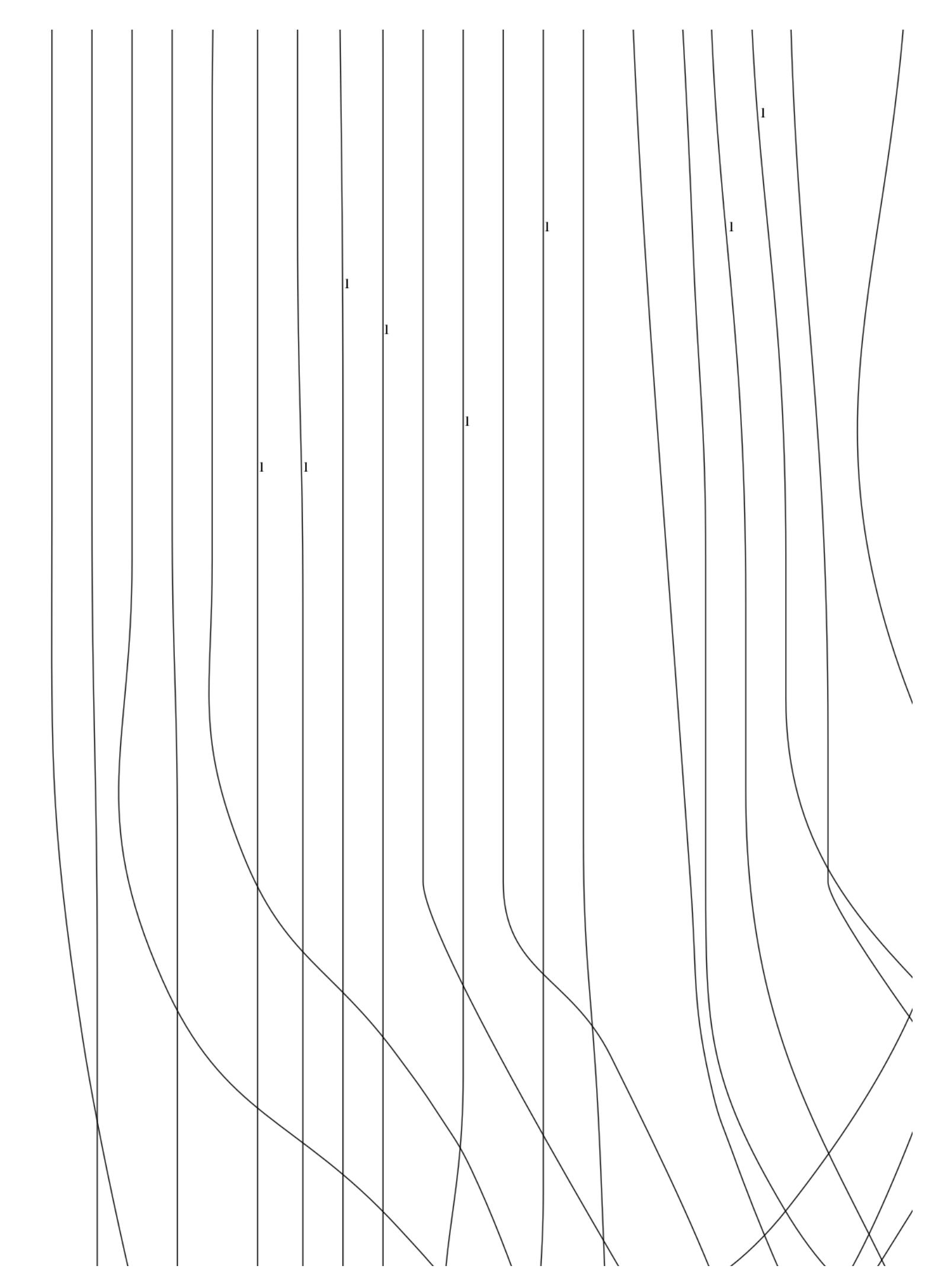
quand une idee subtile  
germe au coeur du reptile  
profitant d'une repetition  
voici qu'avec precaution  
dans l'ombre du crepuscule  
il avance il recule  
puis happe un morceau minuscuile  
un morceau de pellicule  
qui depassait d'une boite en fer  
c'etait la grande scene du Val d'Enfer  
tournee le matin dans une cloche a plongeur  
pour mieux voir evoluer le nageur.  
et comme un spaghetti  
le python en appetit  
avale deux cents metres a present  
des aventures de Tarzan !

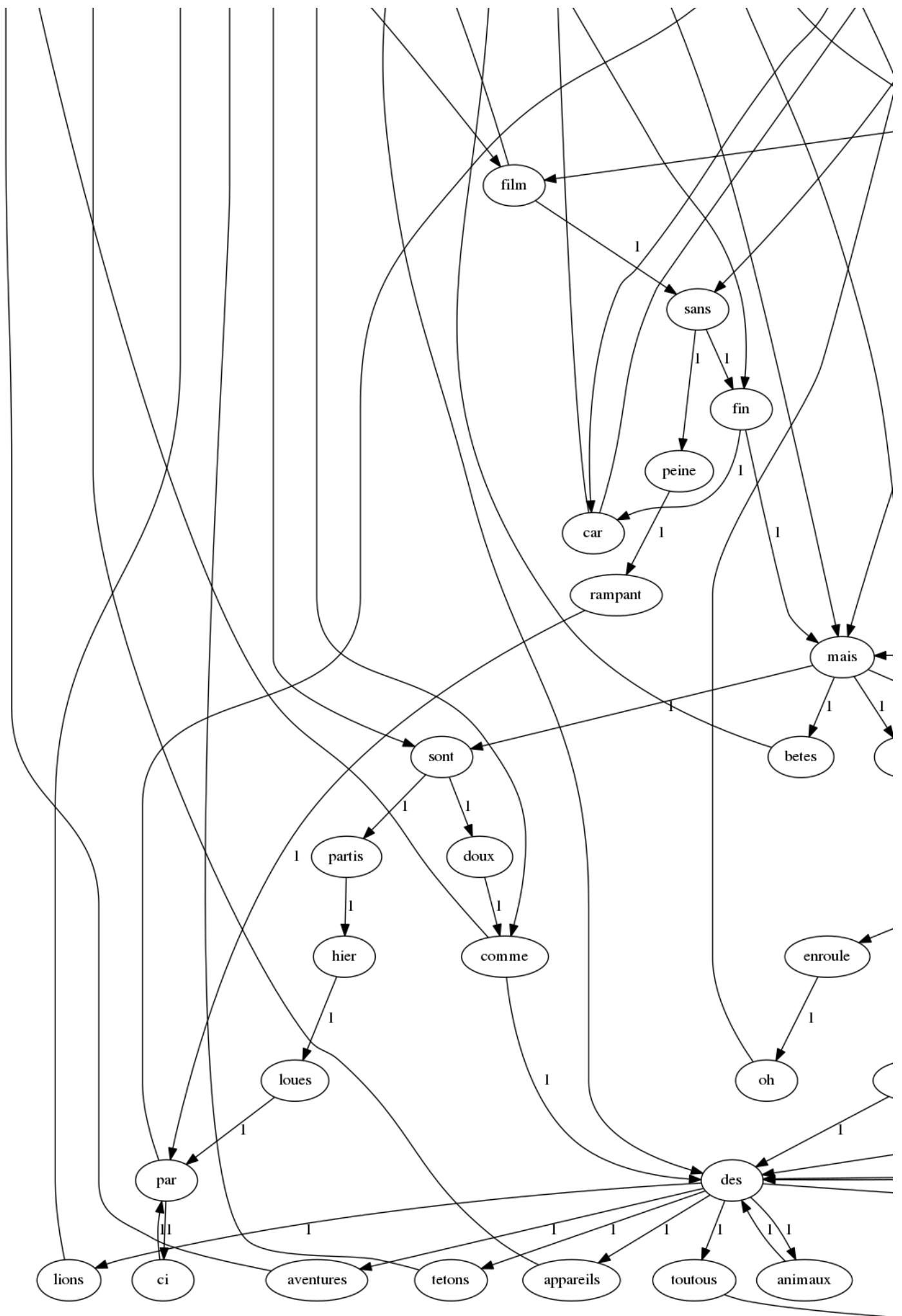
puis il s'en va joyeux  
pensant : " c'est merveilleux  
je vais dormir maintenant trois semaines  
digerer ce film sans peine. "  
rampant par-ci par-la  
il s'enroule, oh la la,  
autour d'un cocotier geant  
mais soudain s'ecrie : " j'ai en...  
j'ai envie de vomir c'est affreux, tu m'as  
empoisonne, cinema  
tarzan n'est pas pour les pauvres pythons  
j'en ai mal jusqu'au bout des tetons. "  
et la moralite  
du serpent depite  
c'est que parfois trop de cine parleur  
peut vous donner mal au coeur  
ou que les hommes digerent, dit-on,  
mieux que les serpents pythons.  
on obtient :











---

Comment stocker ces données ?

On utilise un *dictionnaire suivants*. Chaque clef est un mot  $m_i$ . À chaque clef, il faut associer un ensemble de mots, chacun muni d'un compteur. Pour ce faire, on utilise à nouveau un *dictionnaire*. Chaque valeur du dictionnaire suivants est donc un dictionnaire dont les clefs sont les  $m_{ij}$  et les valeurs sont les  $c_{ij}$ .

On vous demande de compléter le fichier [mot.py](#) qui réalise la lecture d'un fichier, son analyse, l'affichage du graphe et enfin la génération d'une petite phrase en se déplaçant aléatoirement dans le graphe.

En guise de documentation du format *.dot*, voici le fichier [python.dot](#) généré pour notre exemple.