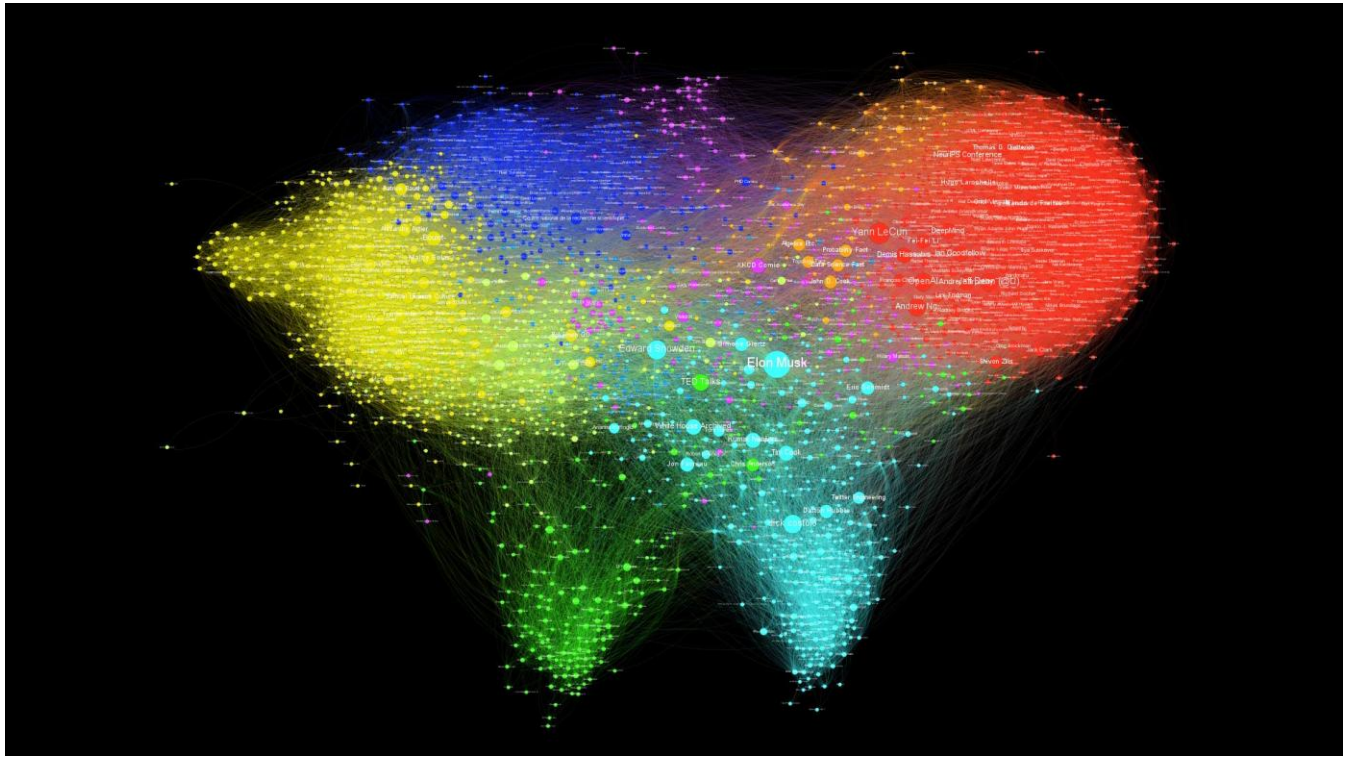
DÉTECTION DE COMMUNAUTÉS POUR LA REPRÉSENTATION ET L'ÉTUDE DES GRAPHES



FLORENT VIRELY

DÉTECTION DE COMMUNAUTÉ

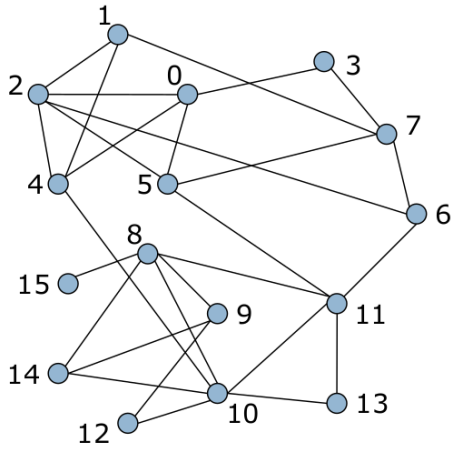
PRINCIPE DE L'ALGORITHME NODE2VEC

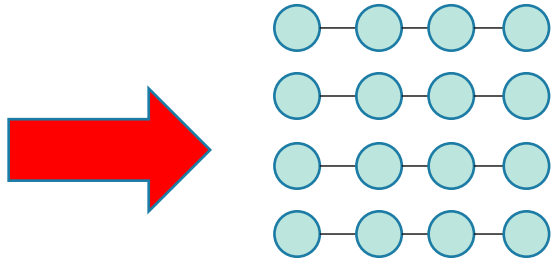
DÉTECTION DE COMMUNAUTÉ

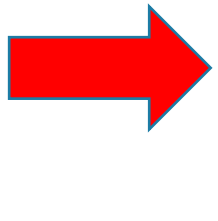
APPLICATION À WIKIPÉDIA

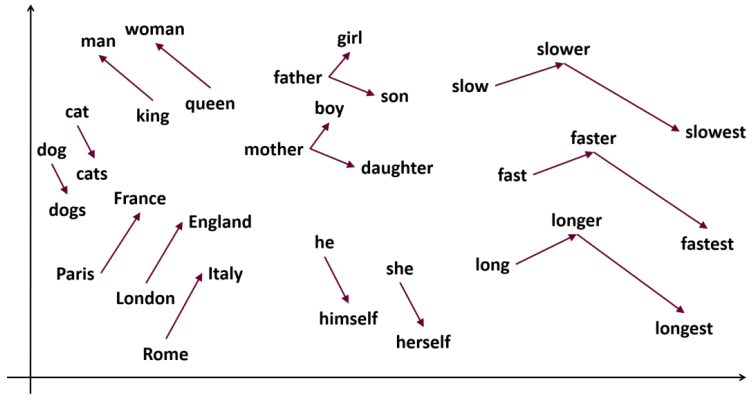
CONCLUSION

ALGORITHME NODE2VEC







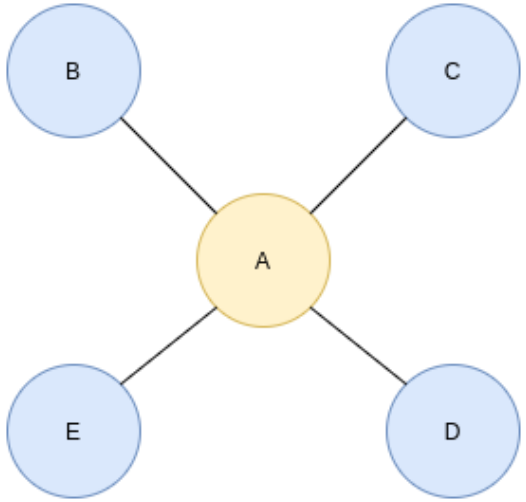


Marche aléatoire

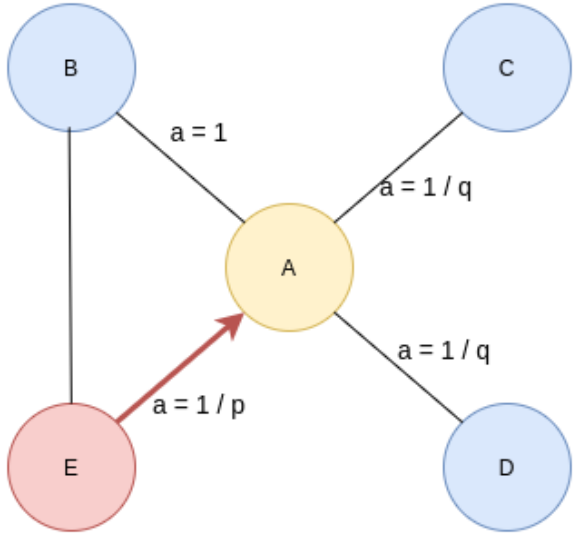
Word2Vec

MARCHE ALÉATOIRE

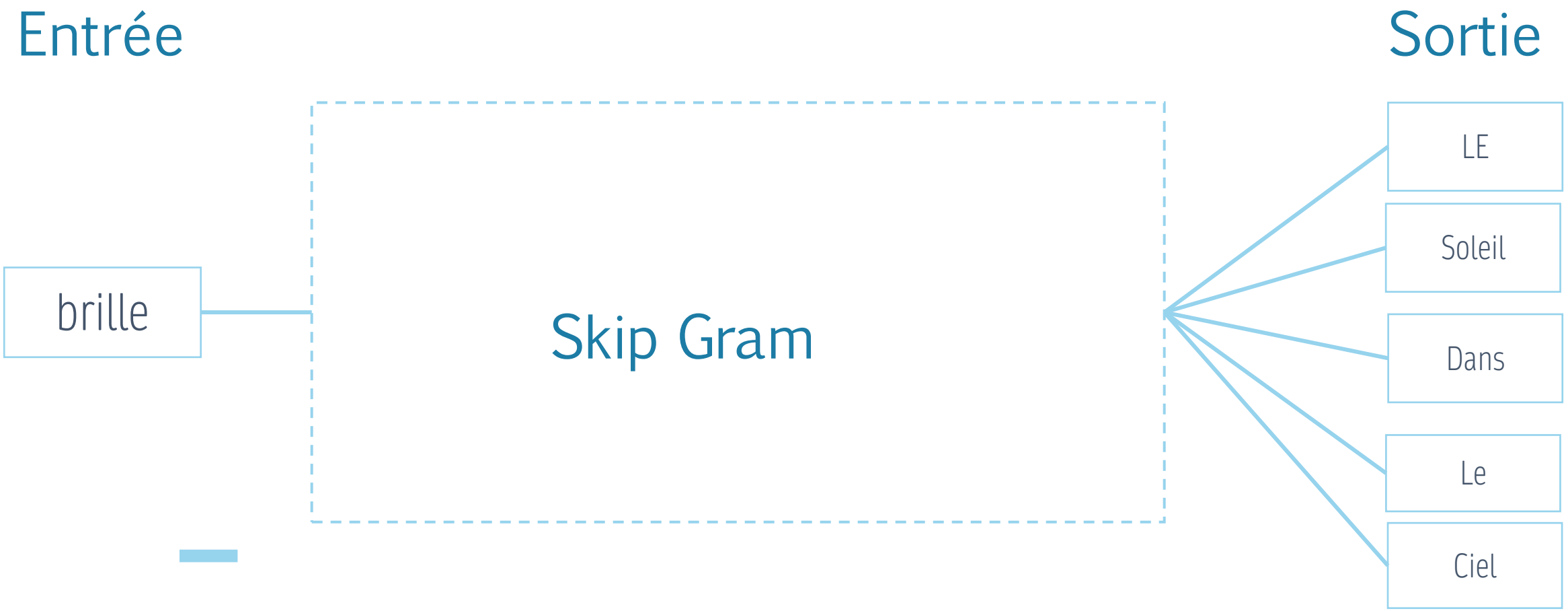
Marche aléatoire classique



Marche aléatoire biaisée



WORD2VEC



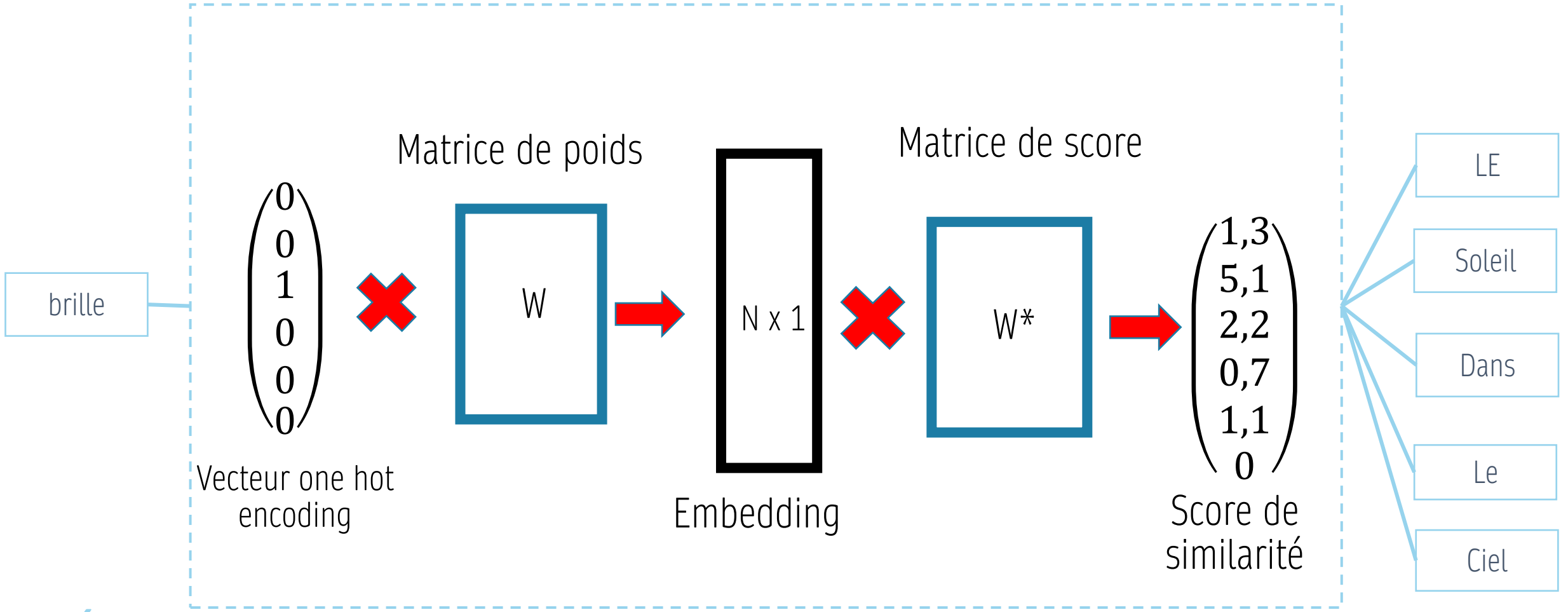
« Le soleil brille dans le ciel »

WORD2VEC

Skip Gram

Entrée

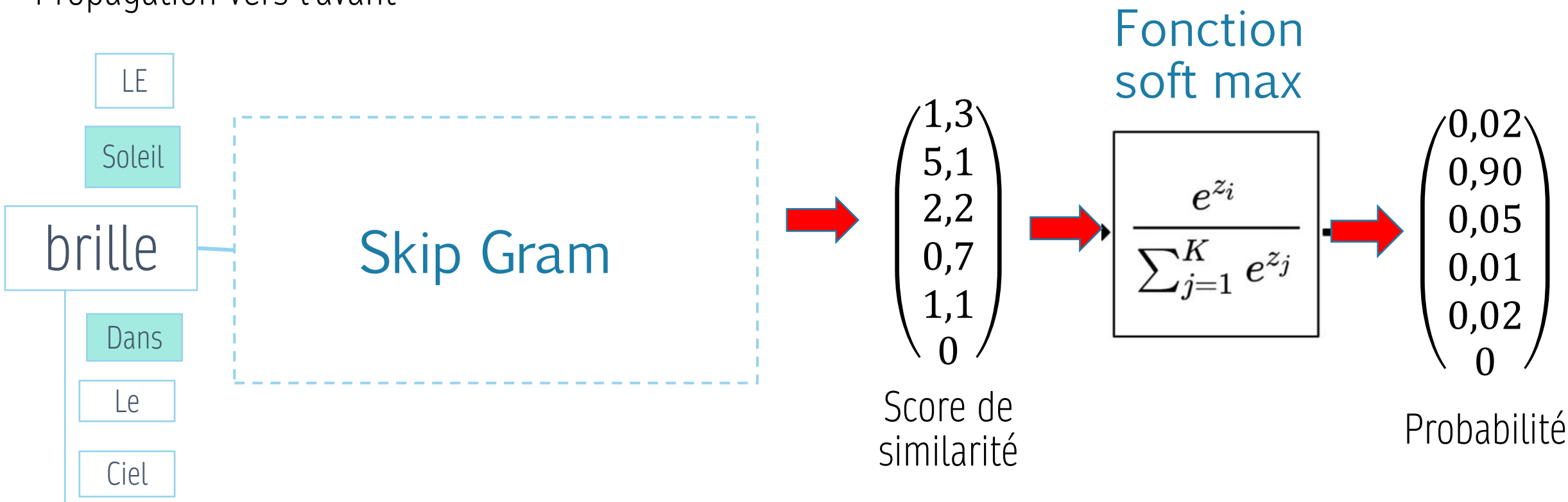
Sortie



« Le soleil brille dans le ciel »

ENTRAINEMENT SKIP GRAM

Propagation vers l'avant



Perte: $L = -\sum_{i=1}^{n} \log(P(m_c|m_p))$ = - log(0,9) – log(0,05)

Cas optimal: $P(m_k|m)$=1 $\Rightarrow \log(P(m_k|m)$ =0

ENTRAINEMENT SKIP GRAM

Propagation de l'erreur en sens inverse

h ✖ W ⟺ $\begin{pmatrix} a_1 & \dots & a_v \\ x_{1,1} & \cdots & x_{1,v} \\ \vdots & \ddots & \vdots \\ x_{n-1,1} & \cdots & x_{n,1,v} \end{pmatrix}^T \times \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = v_e$

v_e ✖ W* ⟺ $\begin{pmatrix} {v'}_{j_1} & \dots & {v'}_{j_v} \\ y_{1,1} & \cdots & y_{1,v} \\ \vdots & \ddots & \vdots \\ y_{n-1,1} & \cdots & y_{n-1,v} \end{pmatrix}^T \times v_e = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_v \end{pmatrix}$ avec $z_j = {v'_j}^T \times v_e$

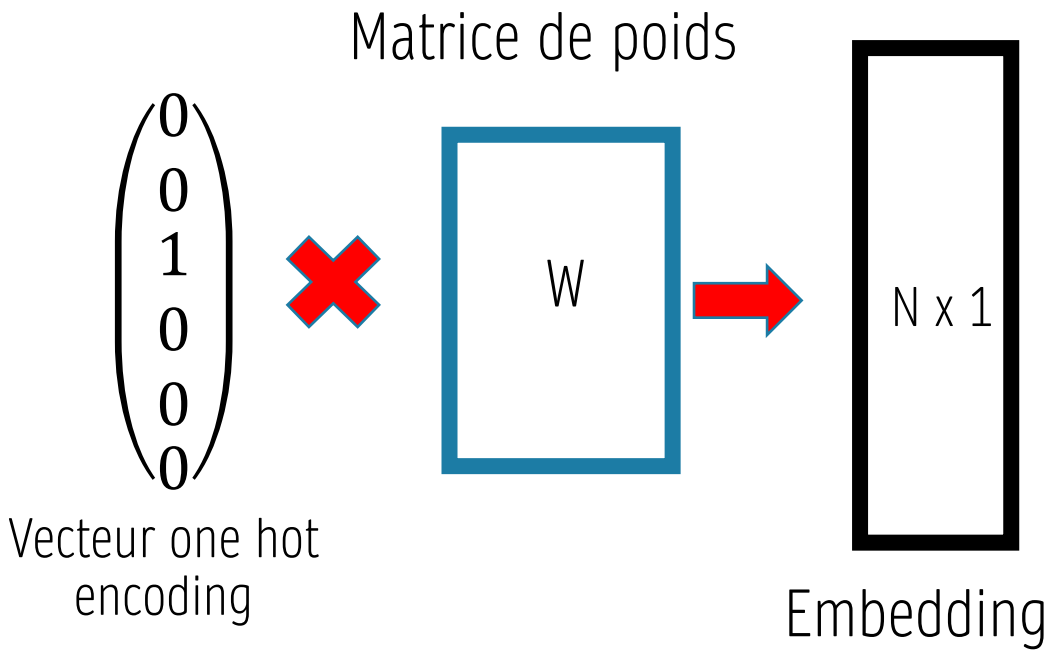
calcul du gradient: $\frac{\partial L}{\partial {v'}_j} = ({v'}_j - h_j) \times v_e$ $\qquad \frac{\partial L}{\partial v_e} = \sum_{j=1}^{v} ({v'}_j - h_j) \times {v'}_j$

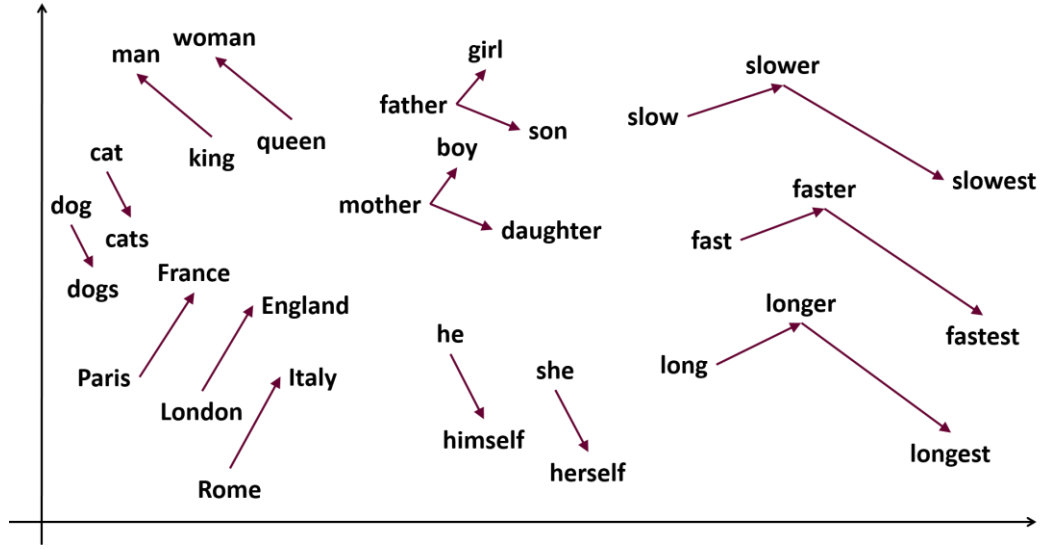
Descente de gradient: $v_e \leftarrow v_e - \eta \frac{\partial L}{\partial v_e}$ $\qquad {v'}_j \leftarrow {v'}_j - \eta \frac{\partial L}{\partial v'_j}$

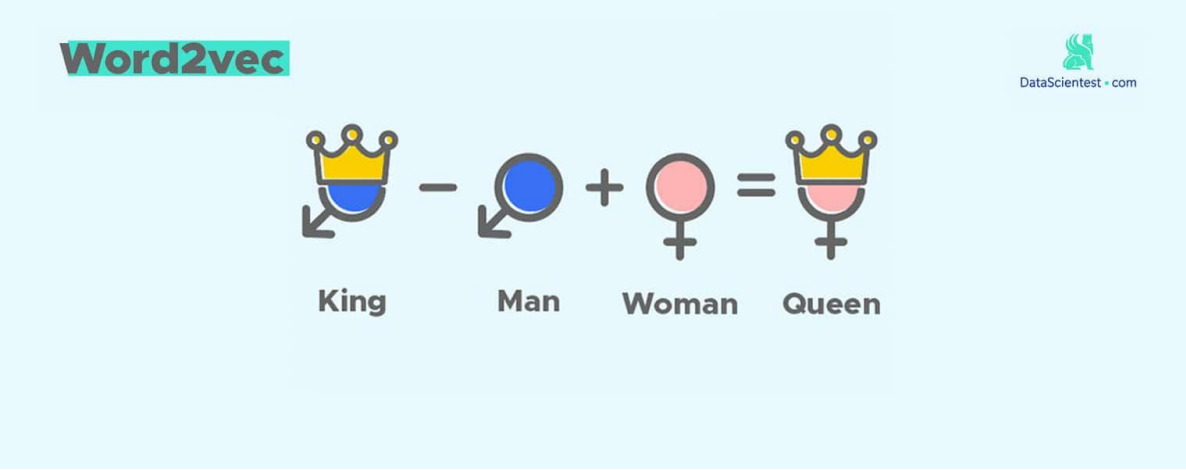
taux d'apprentissage: η

WORD2VEC

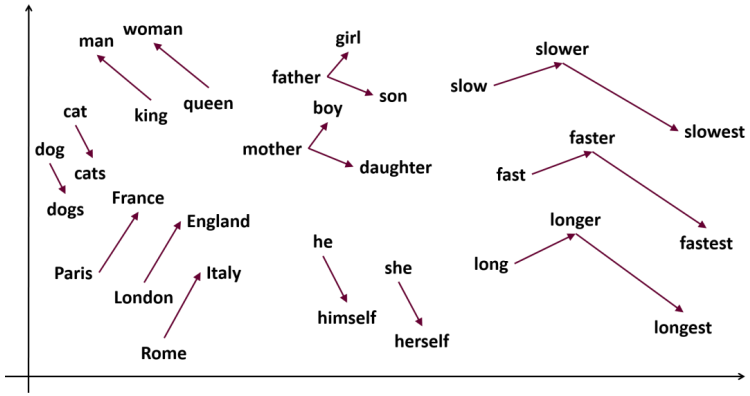
- Entrainement: Itération de Propagation vers l'avant et Back propagation
- Résultats:



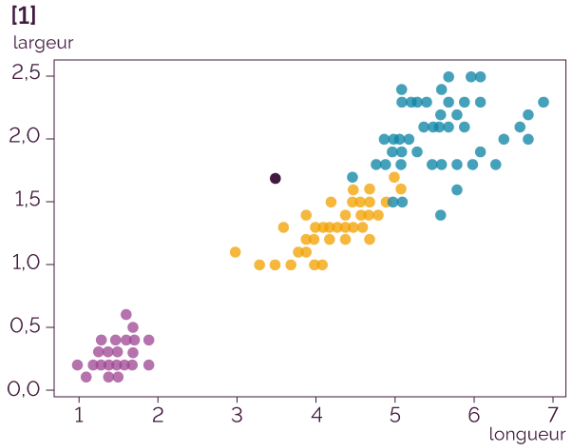




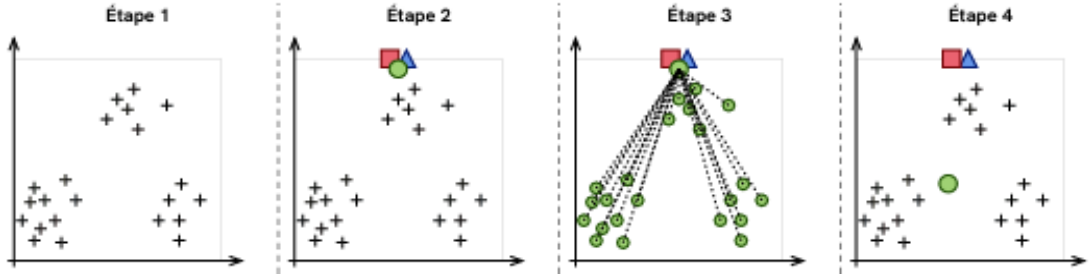
DÉTECTION DE COMMUNAUTÉ

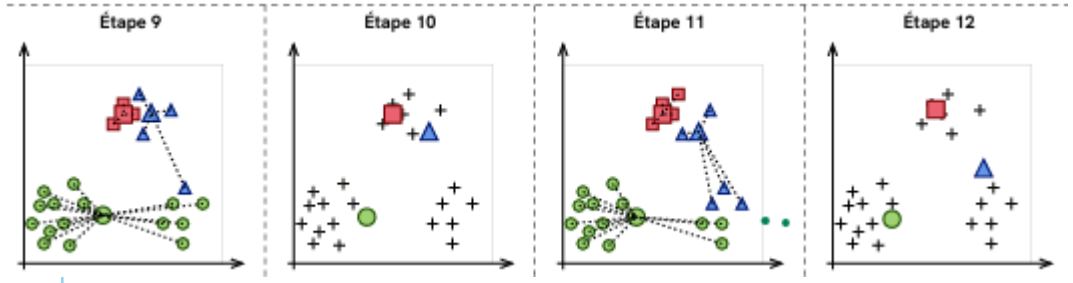


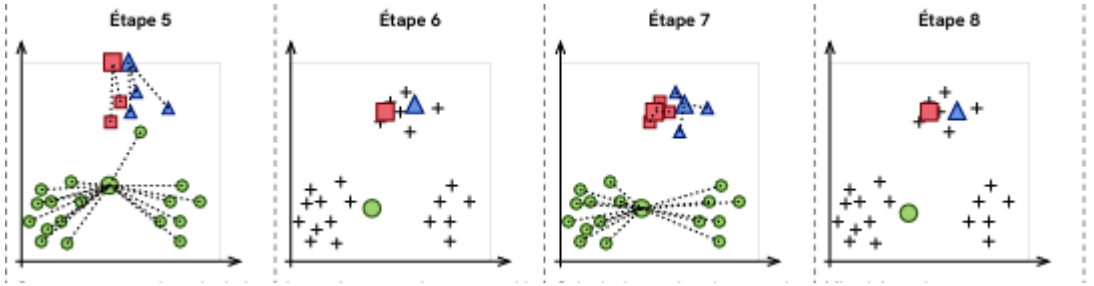
K-moyennes



ALGORITHME DES K MOYENNES



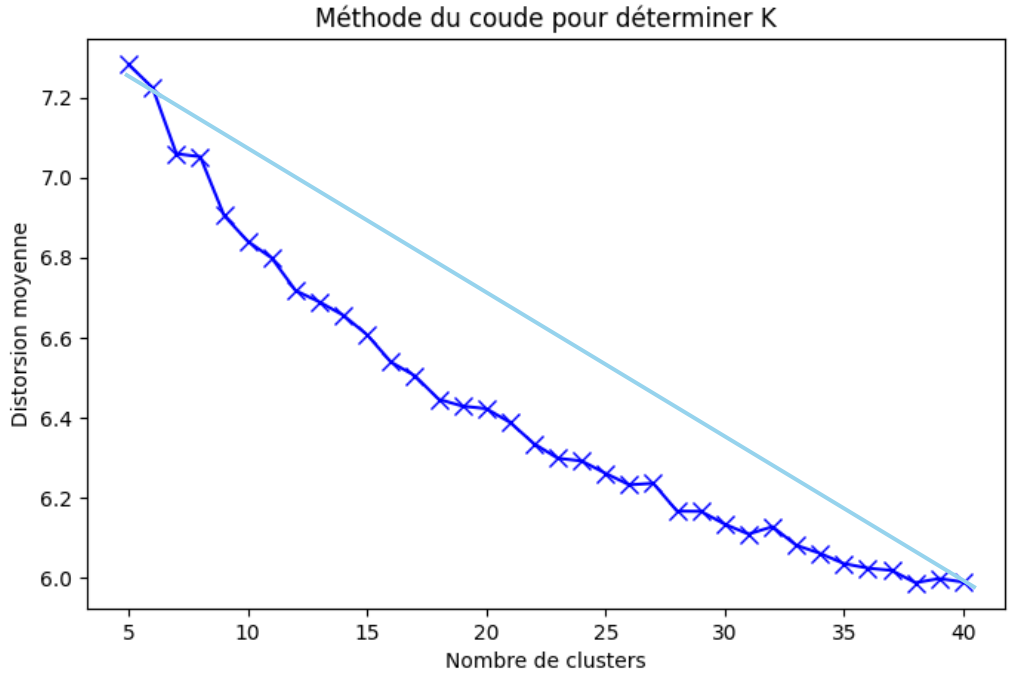




Détermination du K optimal: Méthode du coude

Détermination du coude:

Méthode de la distance maximale



APPLICATION: WIKIPÉDIA SIMPLE ENGLISH

PAGE WIKIPÉDIA

April

Page Talk Read Change Change s

From Simple English Wikipedia, the free encyclopedia

April (Apr.) is the fourth month of the year in the Julian and Gregorian calendars, and comes between March and May. It is one of four months to have 30 days.

April always begins on the same day of the week as July, and additionally, January in leap years. April always ends on the same day of the week as December.

The Month [change | change source]

April comes between March and May, making it the fourth month of the year. It also comes first in the year out of the four months that have 30 days, as June, September and November are later in the year.

April begins on the same day of the week as July every year and on the same day of the week as January in leap years. April ends on the same day of the week as December every year, as each other's last days are exactly 35 weeks (245 days) apart.

In common years, April starts on the same day of the week as October of the previous year, and in leap years, May of the previous year. In common years, April finishes on the same day of the week as July of the previous year, and in

FICHIER XML

```
33    <page>
34      <title>April</title>
35      <ns>0</ns>
36      <id>1</id>
37      <revision>
38        <id>9198094</id>
39        <parentid>9086769</parentid>
40        <timestamp>2023-11-18T09:15:54Z</timestamp>
41        <contributor>
42          <ip>149.50.163.70</ip>
43        </contributor>
44        <model>wikitext</model>
45        <format>text/x-wiki</format>
46        <text bytes="22010" xml:space="preserve">{{monththisyear|4}}
47  '''April''' (Apr.) is the fourth [[month]] of the [[year]] in the [[Julian calendar|Julian]]
48
49  April always begins on the same day of the week as [[July]], and additionally, [[January]] i
50
51  == The Month ==
52  [[File:Colorful spring garden.jpg|thumb|180px|right|[[Spring]] flowers in April in the [[Nor
53  April comes between [[March]] and [[May]], making it the fourth month of the year. It also c
54
55  April begins on the same day of the week as [[July]] every year and on the same day of the w
```

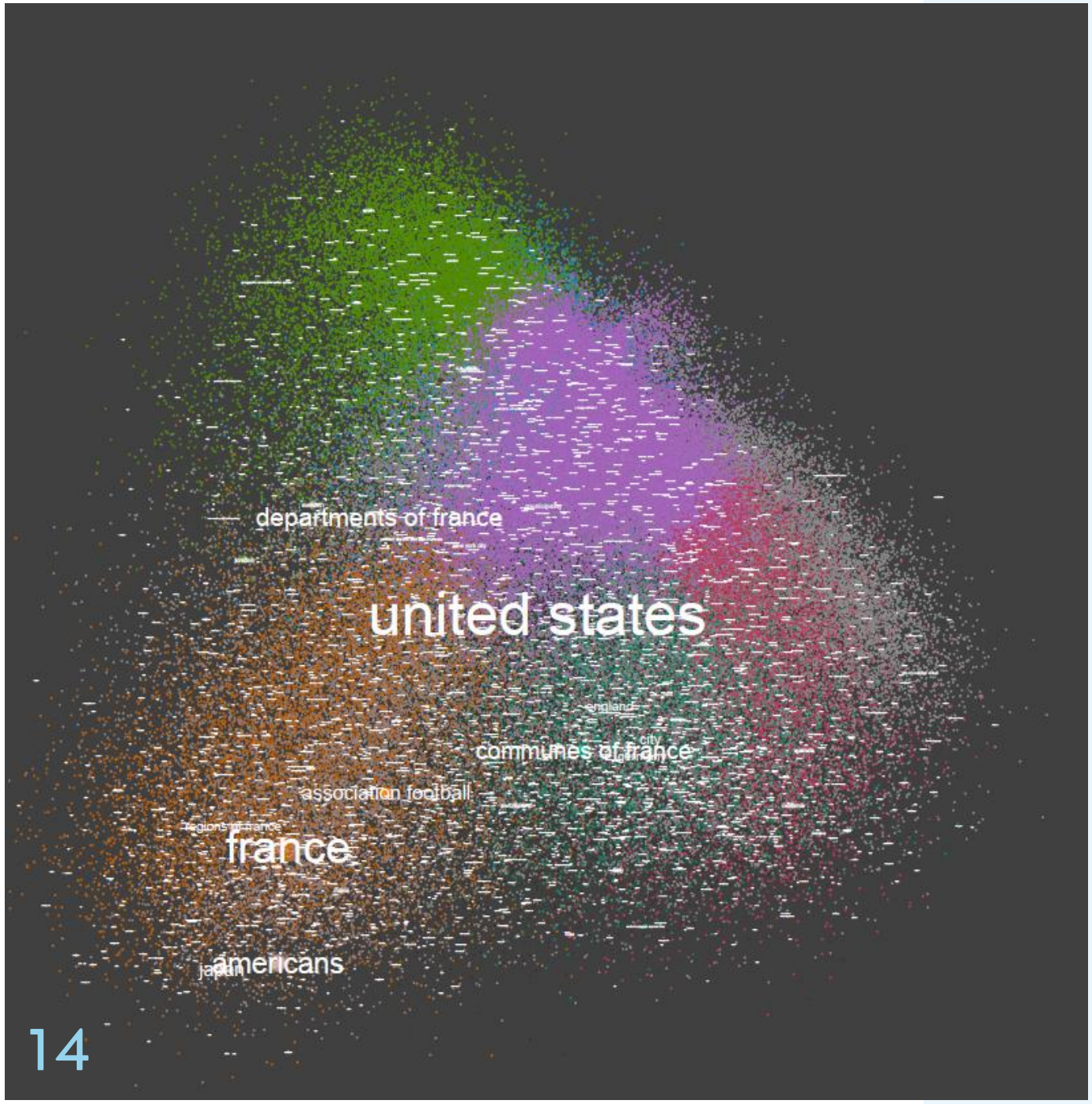
PROGRAMMATION EN PYTHON

IMPLÉMENTATION PERSONNELLE

```
> def embedding_mot(vocab_size,target_idx,W1): ···

> def sigmoid(x): ···

> def create_pairs(corpus, window_size): ···

> def negative_sampling(vocab_size, num_samples, context_idx):

> def forward(target_idx, context_idx, neg_samples, W1, W2): ···

> def backprop(target_idx, context_idx, neg_samples, h, u_pos,

> def f_word2vec_self_made(graphe,corpus,window_size,min_count
```

IMPORTATION DE LA FONCTION WORD2VEC

```
from gensim.models import Word2Vec
def f_Word2vec(graphe,corpus):

    # Défini le modèle
    print(corpus)
    model = Word2Vec(min_count=0, vector_size=100, epochs=10)

    # Construit le vocabulaire et entraine le model
    model.build_vocab(corpus)
    model.train(corpus, total_examples=model.corpus_count, epo
    return model
```



16 COMMUNAUTÉS

184 000 NŒUDS
POUR 1 786 000 LIENS

PAGES LES PLUS IMPORTANTES

-United States

-France

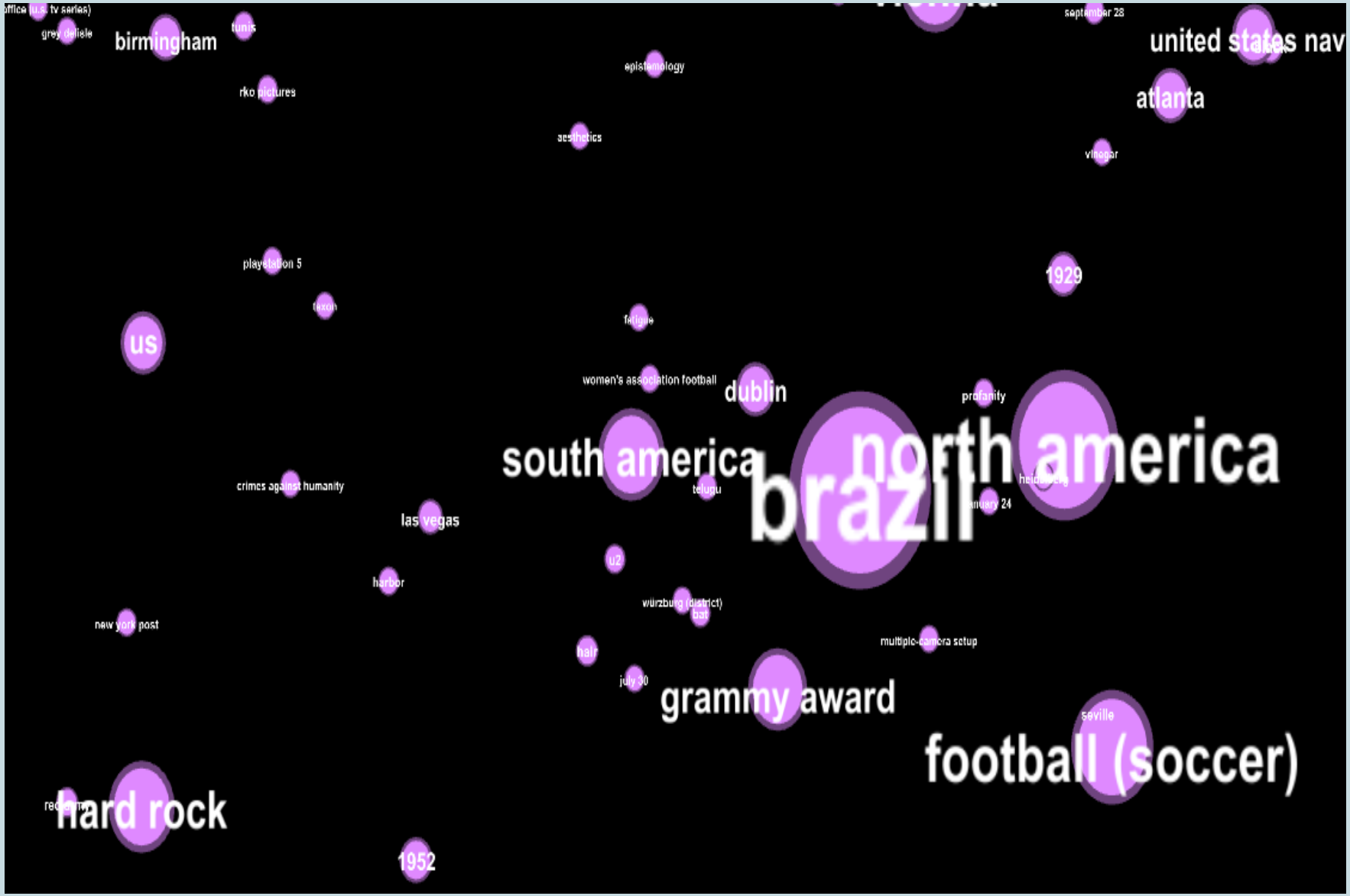
-Communes of France

PAGE LES PLUS PROCHES DE :

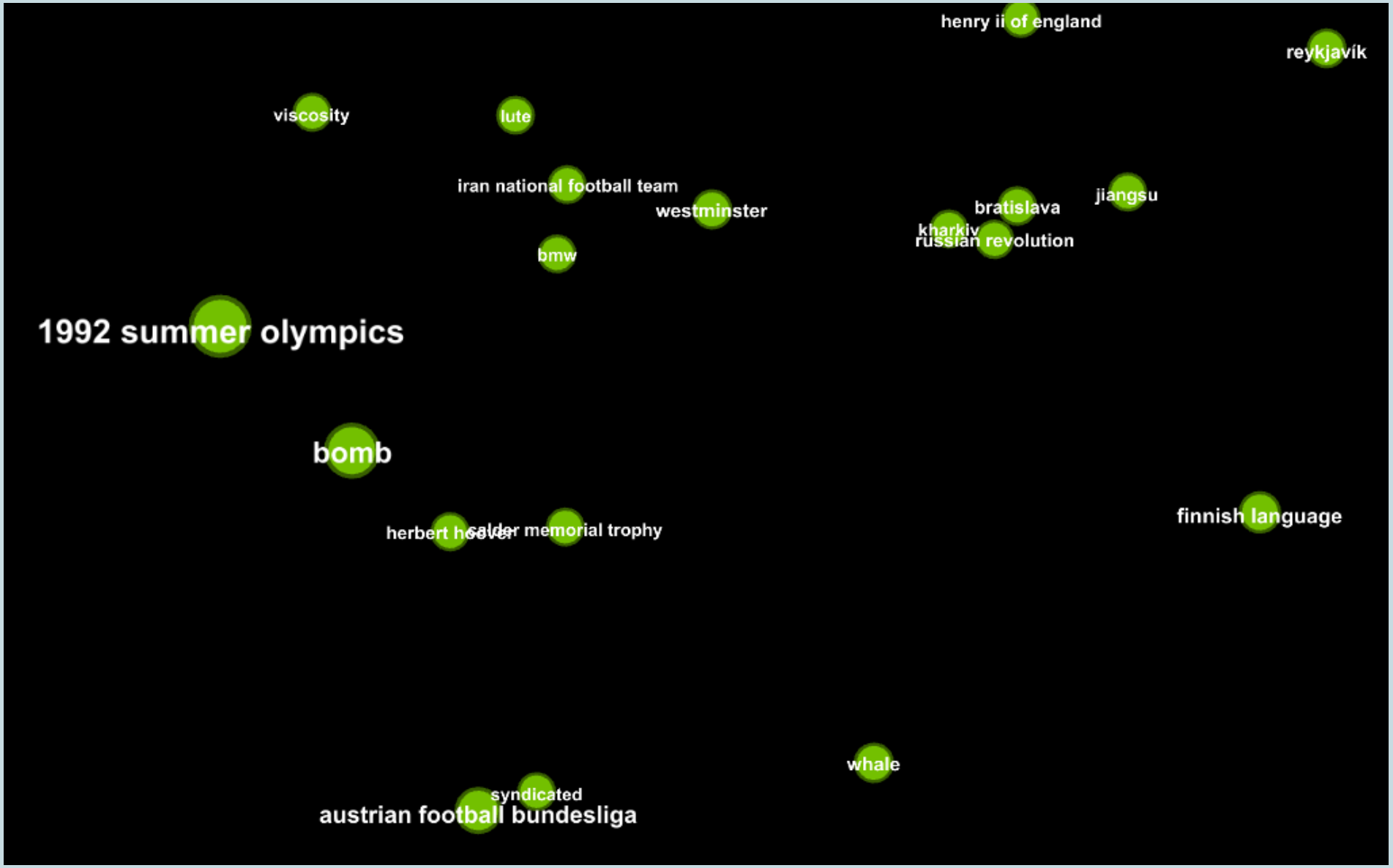
```
🔍 Mots les plus proches de 'science' :
  - facts (score: 0.7517)
  - theories (score: 0.7461)
  - observation (score: 0.6594)
```

```
🔍 Mots les plus proches de 'history' :
  - history of life (score: 0.8108)
  - israel-hamas war (2023-present) (score: 0.8027)
  - history of humans (score: 0.7957)
  - geological history (score: 0.7920)
  - chronology of the universe (score: 0.7914)
```

COMMUNAUTÉ « AMÉRIQUE »



COMMUNAUTÉ MÉLANGÉE



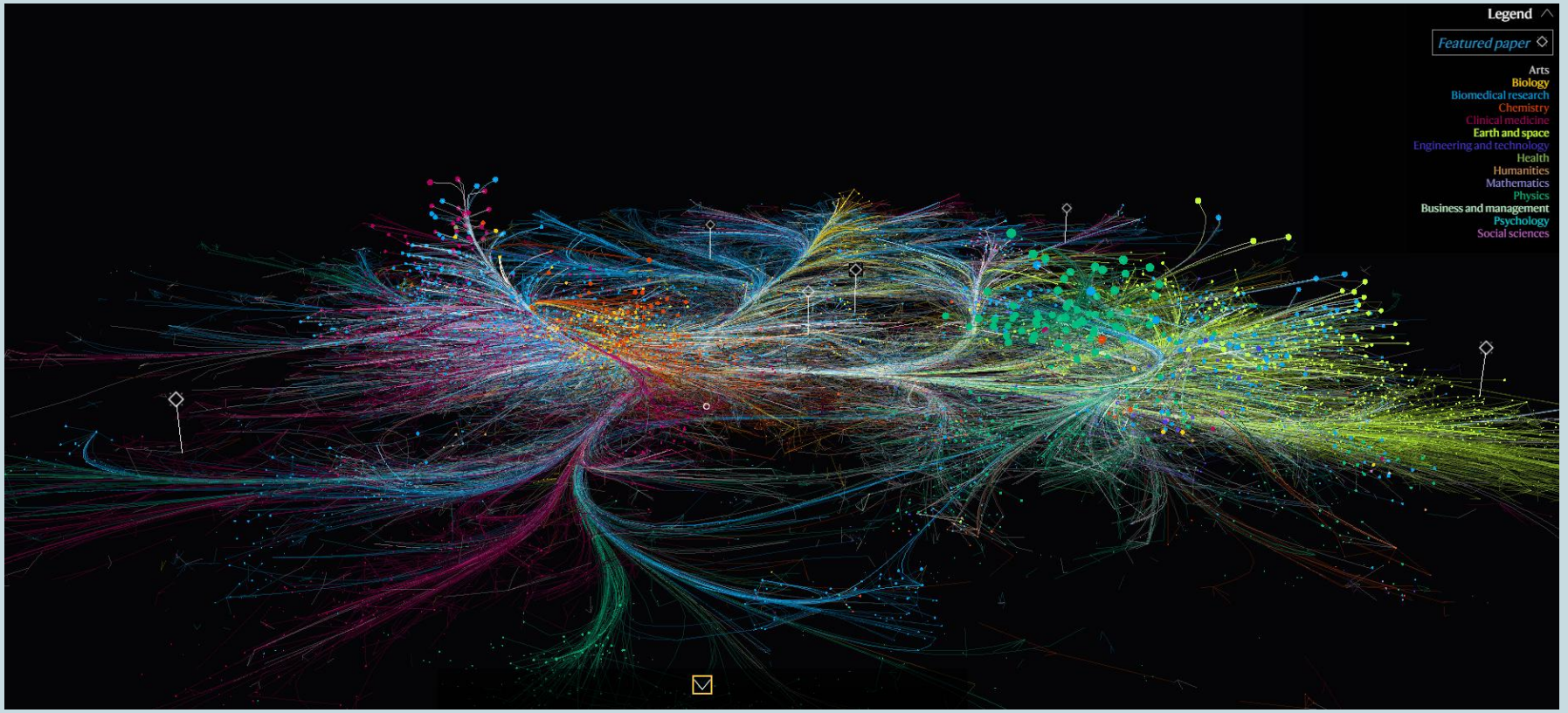
RECHERCHE DU MOT « COMPUTER »

titres	Community
thread (computer science)	1
pointer aliasing (computer programming)	4
word (computer architecture)	4
class (computer programming)	4
analogue computer	4
swedish transport agency's law-breaking ...	4
computer engineering	4
concurrency (computer science)	4
computer graphics	4
roger moore (computer scientist)	4
human–computer interaction	4
computer-aided design (cad)	4
list of words about computers	4
my computer	4
desktop computer	10
complex instruction set computer	10
computer model	10
computer mouse	10
whitespace (computer science)	10
computer science	10
class (computer science)	10
personal computer	12
kernel (computer science)	12
computer hardware	12
synchronization (computer science)	12

CONCLUSION

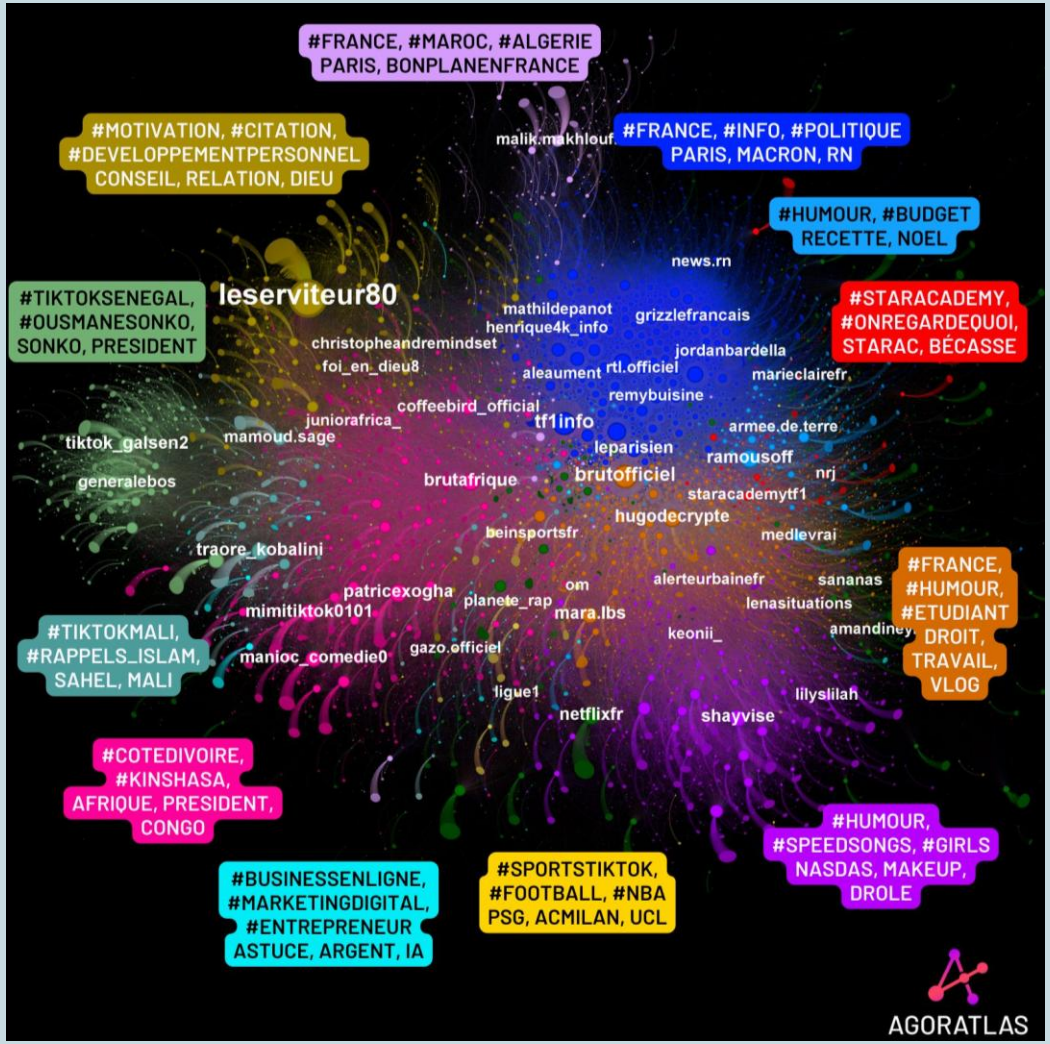
- Bonne identification des petits groupes
- Grandes communautés peu cohérentes
- Raisons : sous-entraînement ou données bruitées
- Algorithme prometteur pour de vraies applications

ARTICLES SCIENTIFIQUES

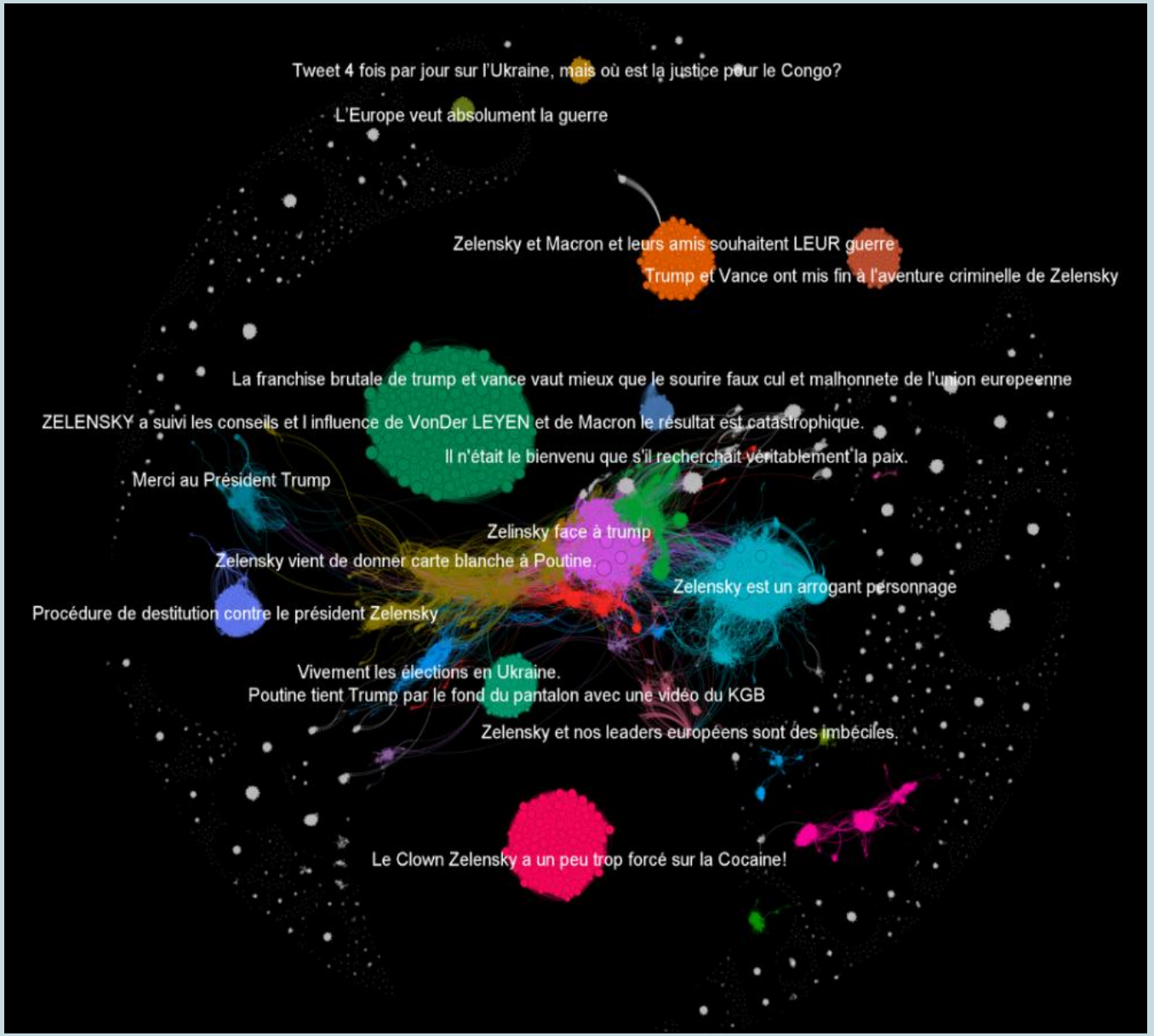


https://www.nature.com/immersive/d41586-019-03165-4/index.html

AGORATLAS



Comptes Tiktok français



Comptes Twitter après le débat
Trump-Zelensky

PROPAGATION DE L'ERREUR EN SENS INVERSE

Entrée: $(m_c : mot\ cible, m_o : mot\ de\ contexte)$
m_j un mot du vocabulaire
Objectif: **Calculer** le gradient de la perte par rapport à:

- vecteur de **sortie** v'_j
- vecteur **d'entrée** v_e

$$\begin{pmatrix} a_1 & \dots & a_v \\ x_{1,1} & \cdots & x_{1,v} \\ \vdots & \ddots & \vdots \\ x_{n-1,1} & \cdots & x_{n,1,v} \end{pmatrix}^T \times \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = v_e$$

$$\begin{pmatrix} {v'_j}_1 & \dots & {v'_j}_v \\ y_{1,1} & \cdots & y_{1,v} \\ \vdots & \ddots & \vdots \\ y_{n-1,1} & \cdots & y_{n-1,v} \end{pmatrix}^T \times \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_v \end{pmatrix} \text{ avec } z_j = {v'_j}^T \times v_e$$

$$L = -\log(P(m_o|m_c)) = -\log\left(\frac{e^{z_o}}{\sum_{j=1}^{v} e^{z_j}}\right) = -\log(r_o)$$

$$\frac{\partial L}{\partial z_j} = v'_j - h_j$$

calcul du gradient:

$$\frac{\partial L}{\partial v'_j} = \frac{\partial L}{\partial z_j} \times \frac{\partial z_j}{\partial v'_j} = (v'_j - h_j) \times v_e \qquad \frac{\partial L}{\partial v_e} = \sum_{j=1}^{v} \frac{\partial L}{\partial z_j} \times \frac{\partial z_j}{\partial v_e} = \sum_{j=1}^{v} (v'_j - h_j) \times v'_j$$

Descente de gradient:

taux d'apprentissage: η

$$v_e \leftarrow v_e - \eta \frac{\partial L}{\partial v_e} \qquad v'_j \leftarrow v'_j - \eta \frac{\partial L}{\partial v'_j}$$