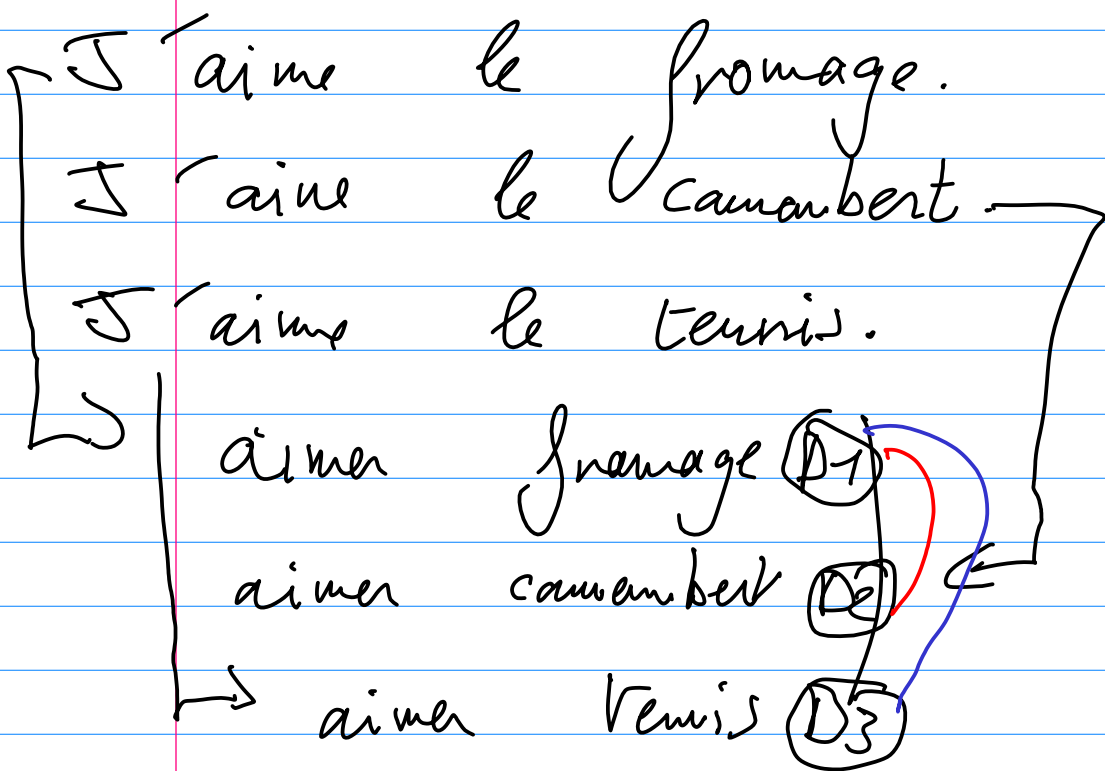


Plongements lexicaux

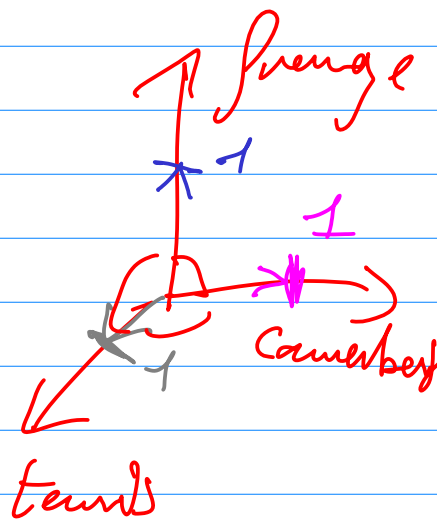


Lemmatisation → stop words / mots vides

$V = \{ \text{aimer}, \text{fromage}, \text{camembert}, \text{tennis} \}$

D1	{ 1	1	0	0 }
D2	{ 1	0	1	0 }
D3	{ 1	0	0	1 }

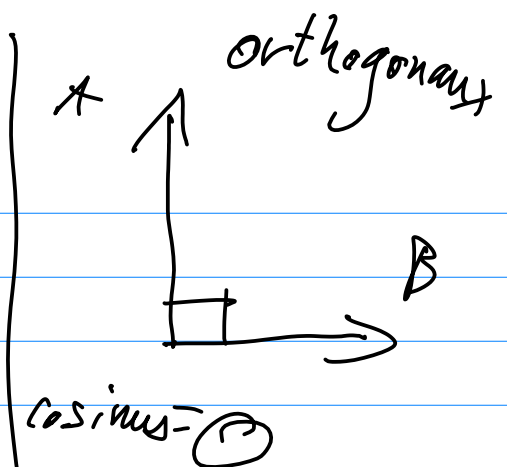
$n = 4$



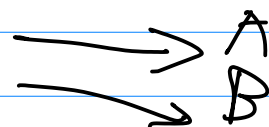
$$\cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|}$$

$$= \frac{\sum_{i=1}^n A_i \cdot B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}}$$

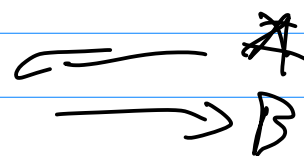
$$\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}$$



colinéaires



cosinus = 1



cosinus = -1

$$\cos(D_1, D_2) = \frac{1}{\sqrt{2} \sqrt{2}} = \frac{1}{2}$$

$$\cos(D_1, D_3) = \frac{1}{\sqrt{2} \sqrt{2}} = \frac{1}{2}$$

Sac de mots

Bag of Words BOW

→ Limite sémantique pas de

→ Vectors ^{sparse} |count| : bcp de 0
land en mémoire

bcp de paramètres pr les
algs d'apprentissage

⇒ Plongements lexicaux
word embedding

Plongements lexicaux contextuels

Transformers	/	BERT	/	CamemBERT
		anglais		Français
		(RoBERTa)		FlanBERT

Hypothèse distributionnelle

" words that occur in the same contexts tend to have similar meanings "

sens Harris 1954

in context	Je	mange	des pommes
	Je	cuisine	des pommes
	cuisine	mange	sens proche ?

"You shall know a word by the company it keeps" Firth 57

Je mange des pommes

co-occurrence
⇒ mange et pomme, proche en sens ?

• Gros corpus

→ Matrice de co-occurrence

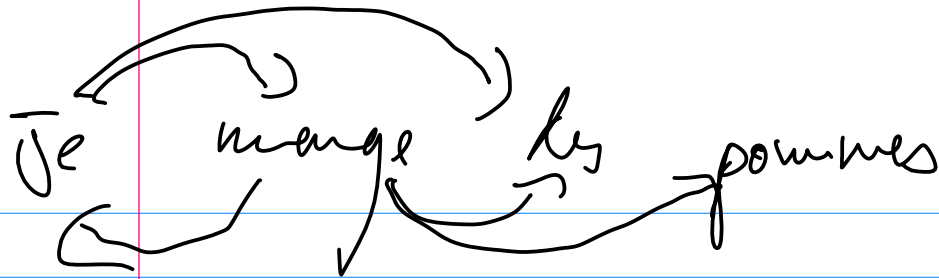
Paramètre : taille de la fenêtre

Je mange des pommes

$w=1$

	Je	mange	des	pommes
Je	0	1	0	0
mange	1	0	1	0
des	0	1	0	1
pommes	0	0	1	0

Symétrique



$W=2$

	Je	mange	les	pommes
Je	0	1	1	0
mange	1	0	1	1
les	1	1	0	1
pommes	0	1	1	0

$W =$ phrases

$W = (5)$ et ose/ car ponctuation

$|V| \times |V|$

PMI : Pointwise Mutual Information

$$PMI(x, y) = \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

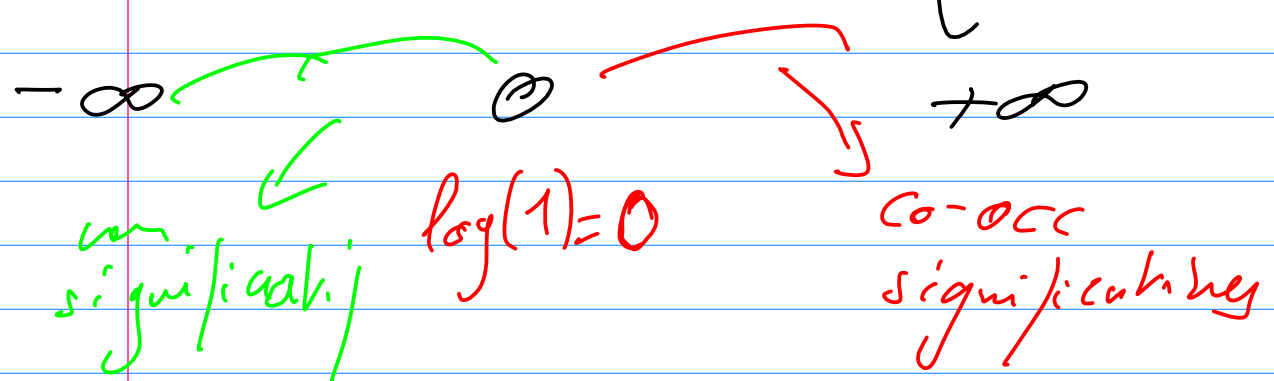
\downarrow mot1 \downarrow mot2 \downarrow fréquence mot1 \downarrow fréquence mot2
 fréquence de co-occurrence de mot1 avec mot2

	je mange des pommes			
je	0	1	1	0
mange	1	0	1	1
des	1	1	0	1
pommes	0	1	1	0

fréquences

$$PMI(x, y) = \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

$$PMI(mange, pommes) = \log \left(\frac{\frac{1}{5}}{\frac{1}{4} \cdot \frac{1}{4}} \right)$$



$$PPMI(x, y) = \max(0, PMI(x, y))$$

Positive PMI CSR

je	PPMI(je, je)	...	PPMI(mange, des)
mange	_____		_____
des	_____		_____
pommes	_____		_____

Analogie :

$\xrightarrow{\quad} \text{ROI} - \text{HOMME} + \text{FEMME} \xrightarrow{\quad}$
 $\xrightarrow{\quad} \text{le verbe le + proches} \xrightarrow{\quad}$
 REINE

Problème Matrice creuse

Matrice de co-oc \rightarrow PPMI (matrice)
analogie

\downarrow
 $\text{SVD (PPMI (matrice))}$
 \approx
 ACP

$\begin{matrix} \text{de} \\ \text{un} \\ \text{des} \\ \text{par} \\ \vdots \end{matrix} \begin{pmatrix} \times \end{pmatrix} \begin{matrix} |V| \times |V| \\ \hookrightarrow \text{voc} \end{matrix} \approx \begin{pmatrix} w \end{pmatrix} \begin{matrix} |V| \times 300 \end{matrix} \begin{pmatrix} s \end{pmatrix} \begin{matrix} 300 \times 300 \end{matrix} \begin{pmatrix} \theta \end{pmatrix} \begin{matrix} 300 \times |V| \end{matrix}$

$\begin{matrix} \text{de} \\ \text{un} \\ \text{des} \\ \text{par} \\ \vdots \end{matrix} \begin{pmatrix} w \end{pmatrix} \begin{matrix} 300 \end{matrix} \rightarrow \text{je} \begin{pmatrix} \end{pmatrix} \begin{matrix} 300 \end{matrix}$

Singularité

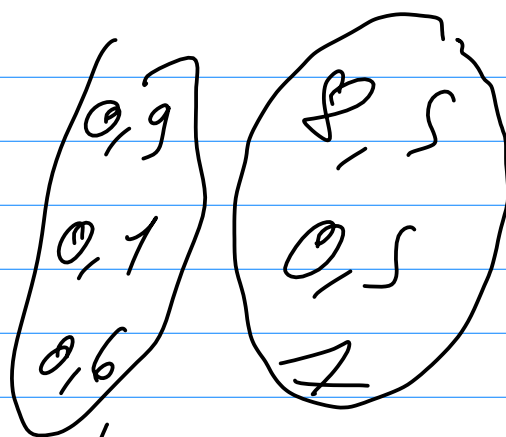
frange
frange
zoo

canard

zoo

girafe

casier



Glove

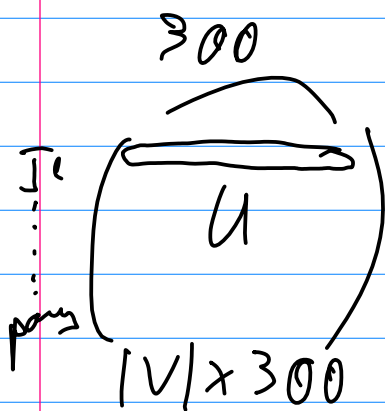
Pennington 2014
et al.

(X)

matrice de co-occurrences
extraite d'un gros corpus

U, V

initialisée aléatoirement
paramètres du modèle
2 matrices de plongement



$$\cos(u_i, u_j) \approx 1$$

si la mot i et le mot j ont un sens similaire

cos

$$\frac{\vec{u}_i \cdot \vec{u}_j}{\|\vec{u}_i\| \|\vec{u}_j\|}$$

$\hookrightarrow \vec{u}_i \cdot \vec{u}_j$ } \nearrow très élevée ou
très faible } mot similaires sens
non similaires

$$X \approx U \cdot V$$

X_{ij} très élevée $\approx U_i \cdot V_j$
 i et j sont similaires

$$\log(X) \approx U \cdot V$$

$$\Leftrightarrow \arg\min_{U, V} \log(X) - U \cdot V$$

$$\arg \min_{u, v} \sum_{i,j} \underbrace{f(x_{ij})}_{\text{pondération}} \underbrace{\left(u_i \cdot v_j - \log(x_{ij}) \right)^2}_{\text{erreur}}$$



u et v nos paramètres

Gradient

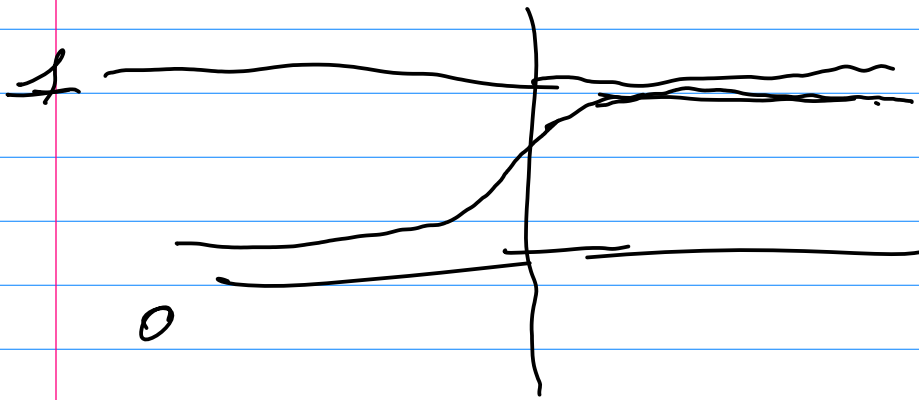
$$\frac{\partial}{\partial u} f(x_{ij}) \left(\log(x_{ij}) - u_i v_j \right)^2$$


$$f(x_{ij}) \frac{\partial}{\partial u} \left(\log(x_{ij}) - u_i v_j \right)^2 \quad (f^2)'$$

$$= f(x_{ij}) \times 2 v_j \times (\log(x_{ij}) - u_i v_j) \quad 2f'f$$

Word2Vec

Mikolov 2013
et al.




Sigmoidale

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

U, V matrices d'embeddings,
paramètres du modèle

u_i, v_j mot i et mot j

$$\sigma(u_i \cdot v_j) \rightarrow 1$$

\mathcal{E} un corpus

$(w_i, w_j) \in \mathcal{D}$, l' \mathcal{E} des co-occurrences
observées sur \mathcal{E}

$$\arg \max_{u, v} \prod_{(w_i, w_j) \in D} \sigma(u_i \cdot v_j)$$

$$\arg \max_{u, v} \sum_{(w_i, w_j) \in D} \log(\sigma(u_i \cdot v_j))$$

Échantillonnage négatif
 $(w_i^+, w_j^+) \in D^+$ paires de mots qui
 ne co-occurrent pas

$$\arg \max_{u, v} \left(\sum_{w_i, w_j} \log(\sigma(u_i \cdot v_j)) + \sum_{w_i^+, w_j^+} \log(1 - \sigma(u_i \cdot v_j)) \right)$$









