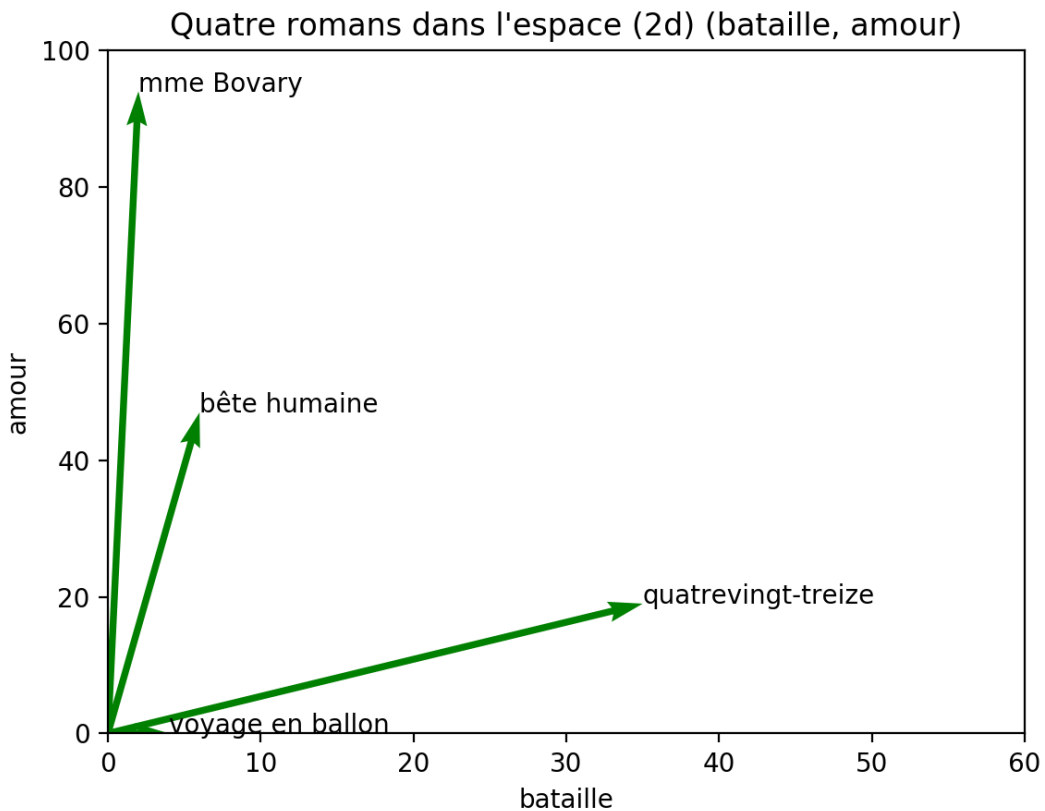
Matrice terme-document

	QuatreVT 119 Kw	Voyage Bal 82 kw	Bête Hum. 128 kw	Mme Bovary 117 kw
bataille	35	4	6	2
clair	105	26	96	52
facile	12	19	6	10
politique	11	0	9	5
voyage	17	196	94	44
idiot	2	1	2	6
amour	19	0	47	94

Quatrevingt-treize (Hugo)
Le voyage en ballon (Verne)
La bête humaine (Zola)
Mme Bovary (Flaubert)

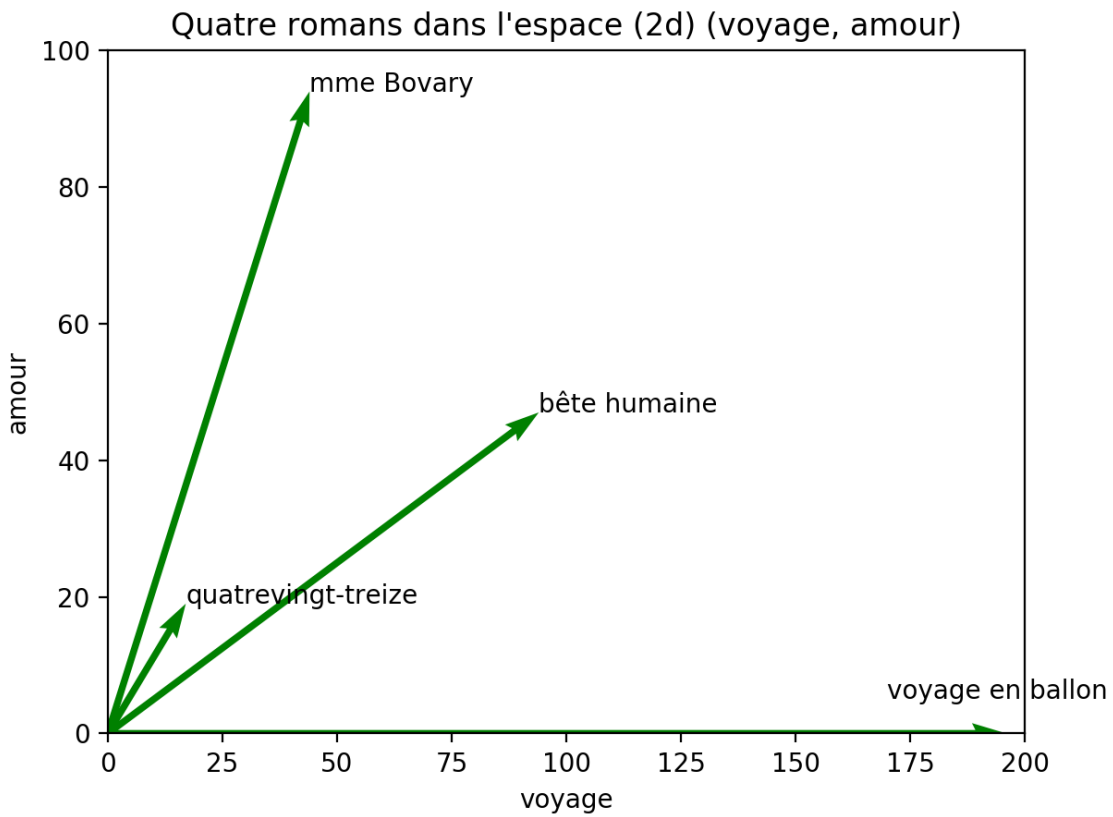
Documents comme vecteurs

	QuatreVT	Voyage Bal	Bête Hum.	Mme Bovary
bataille	35	4	6	2
amour	19	0	47	94



Documents comme vecteurs

	QuatreVT	Voyage Bal	Bête Hum.	Mme Bovary
voyage	17	196	94	44
amour	19	0	47	94



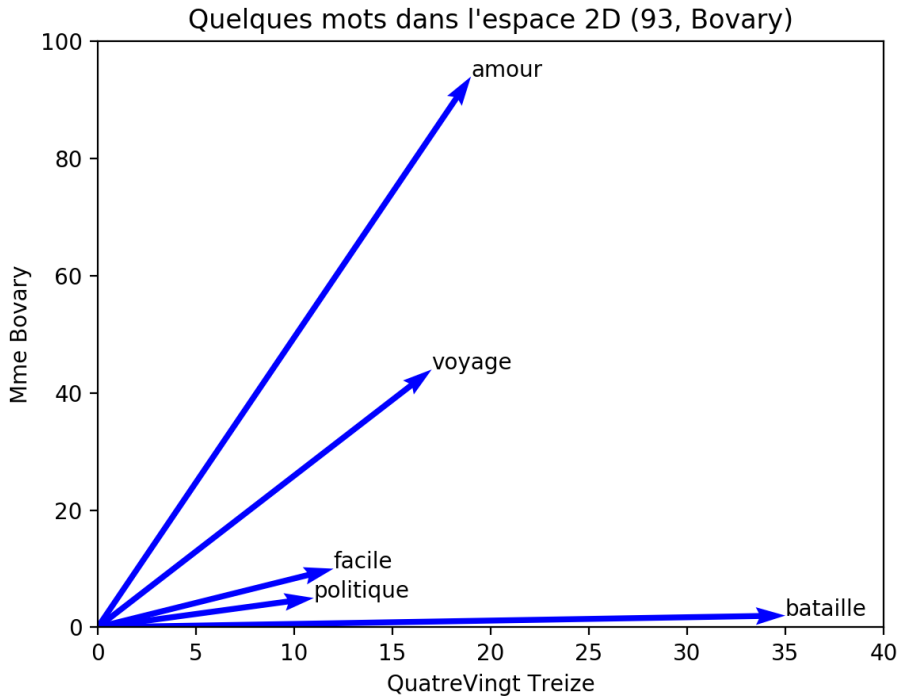
Vecteurs terme-document

On peut inverser la représentation : les dimensions sont maintenant les documents, les vecteurs permettent de décrire des mots.

	QuatreVT 119 Kw	Voyage Bal 82 kw	Bête Hum. 128 kw	Mme Bovary 117 kw
bataille	35	4	6	2
clair	105	26	96	52
facile	12	19	6	10
politique	11	0	9	5
voyage	17	196	94	44
idiot	2	1	2	6
amour	19	0	47	94

amour (comme *politique*)
est le genre de mot qui n'apparaît pas dans "Le voyage en ballon".

On peut visualiser les mots dans l'espace (Quatrevingt-treize, Mme Bovary) :



bataille	(35,2)
politique	(11,5)
amour	(19,94)
voyage	(17,44)

Comptages distributionnels

	arriver	tomber	habiller	mourir
bataille	246	100	2	180
voyage	470	83	4	116
homme	1 819	1 205	339	1 499
femme	890	660	384	1 088

Comptages distributionnels

	arriver	tomber	habiller	mourir	
bataille	246	100	2	180	55 331
voyage	470	83	4	116	208 520
homme	1 819	1 205	339	1 499	668 289
femme	890	660	384	1 088	346 093

Comptages distributionnels (normalisés)

	arriver	tomber	habiller	mourir
bataille	44	18	0	32
voyage	23	4	0	6
homme	27	18	5	22
femme	26	19	11	31

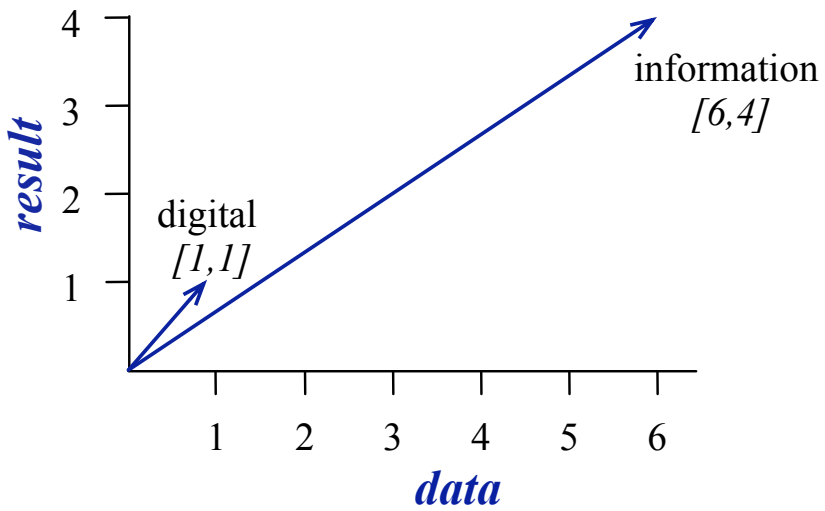
Matrice terme-terme

More common: word-word matrix (or "term-context matrix")

Two **words** are similar in meaning if their context vectors are similar

sugar, a sliced lemon, a tablespoonful of	**apricot**	jam, a pinch each of,
their enjoyment. Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer**.	In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the

	aardvark	computer	data	pinch	result	sugar	…
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	


4
3
2
1
result
digital
[1,1]
information
[6,4]
1
2
3
4
5
6
data

Reminders from linear algebra

$$\text{dot-product}(\vec{v},\vec{w}) = \vec{v}\cdot\vec{w} = \sum_{i=1}^{N} v_i w_i = v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

$$\text{vector length} \quad |\vec{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$

Cosine for computing similarity

Sec. 6.3

$$\text{cosine}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

v_i is the count for word v in context i

w_i is the count for word w in context i.

Cos(v,w) is the cosine similarity of v and w

$$\vec{a}\cdot\vec{b} = |\vec{a}||\vec{b}|\cos\theta$$

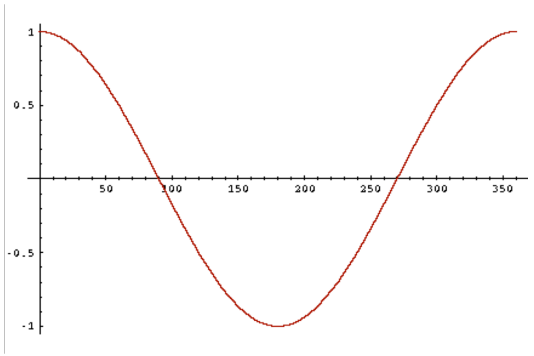
$$\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} = \cos\theta$$

Cosine as a similarity metric

-1: vectors point in opposite directions

+1: vectors point in same directions

0: vectors are orthogonal



Frequency is non-negative, so cosine range 0-1

$$\cos(\vec{v},\vec{w}) = \frac{\vec{v}\bullet\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\vec{v}}{|\vec{v}|}\bullet\frac{\vec{w}}{|\vec{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

	large	data	computer
apricot	1	0	0
digital	0	1	2
information	1	6	1

Which pair of words is more similar?

cosine(apricot,information) =

$$\frac{1+0+0}{\sqrt{1+0+0}\ \sqrt{1+36+1}} = \frac{1}{\sqrt{38}} = .16$$

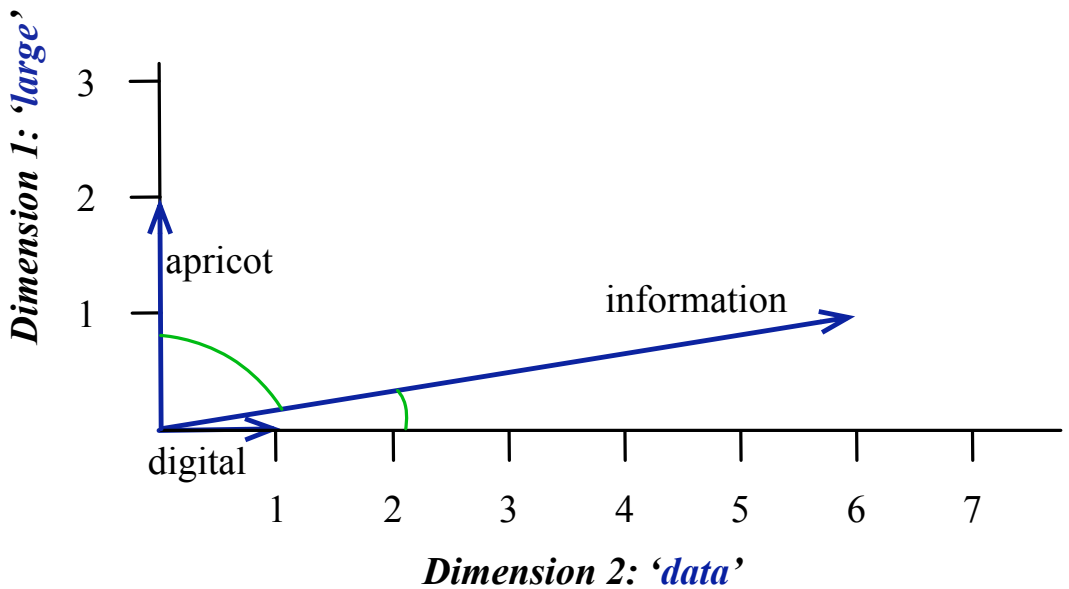
cosine(digital,information) =

$$\frac{0+6+2}{\sqrt{0+1+4}\ \sqrt{1+36+1}} = \frac{8}{\sqrt{38}\sqrt{5}} = .58$$

cosine(apricot,digital) =

$$\frac{0+0+0}{\sqrt{1+0+0}\ \sqrt{0+1+4}} = 0$$

Visualizing cosines (well, angles)



Exemple : clustering

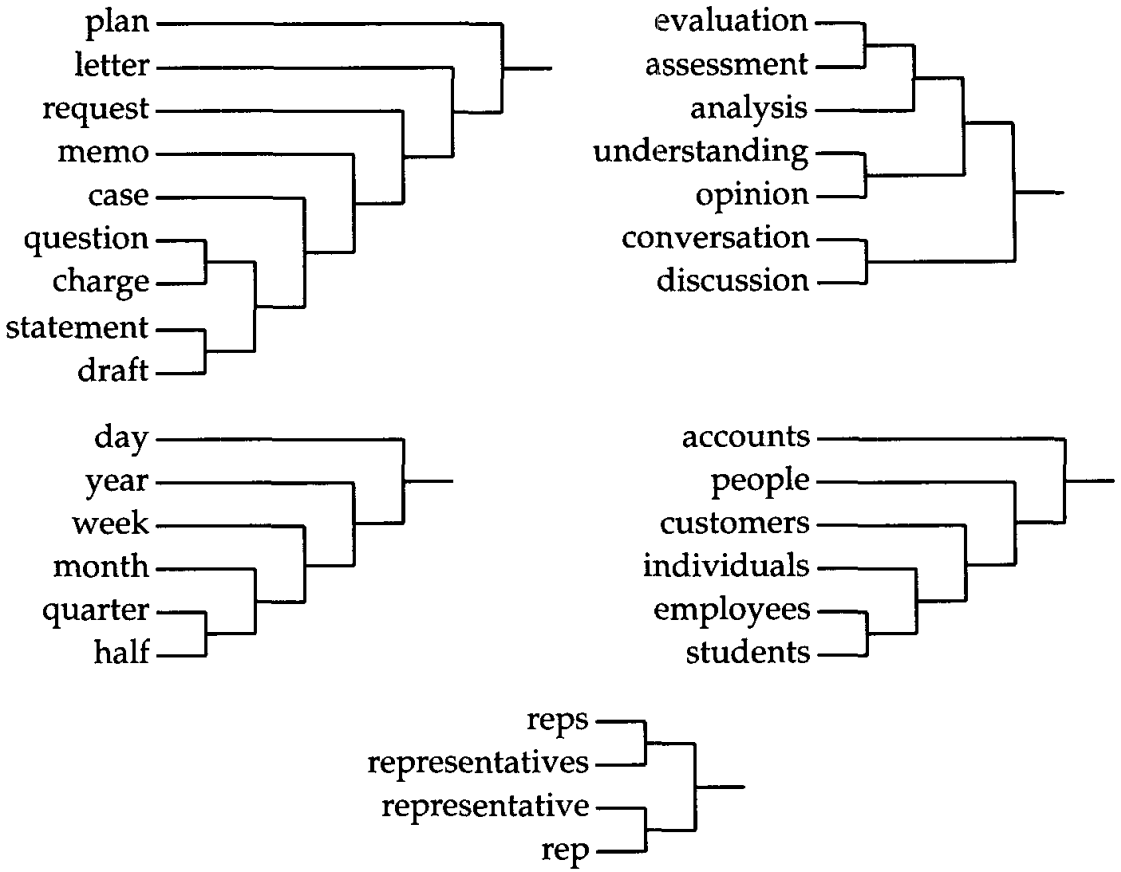


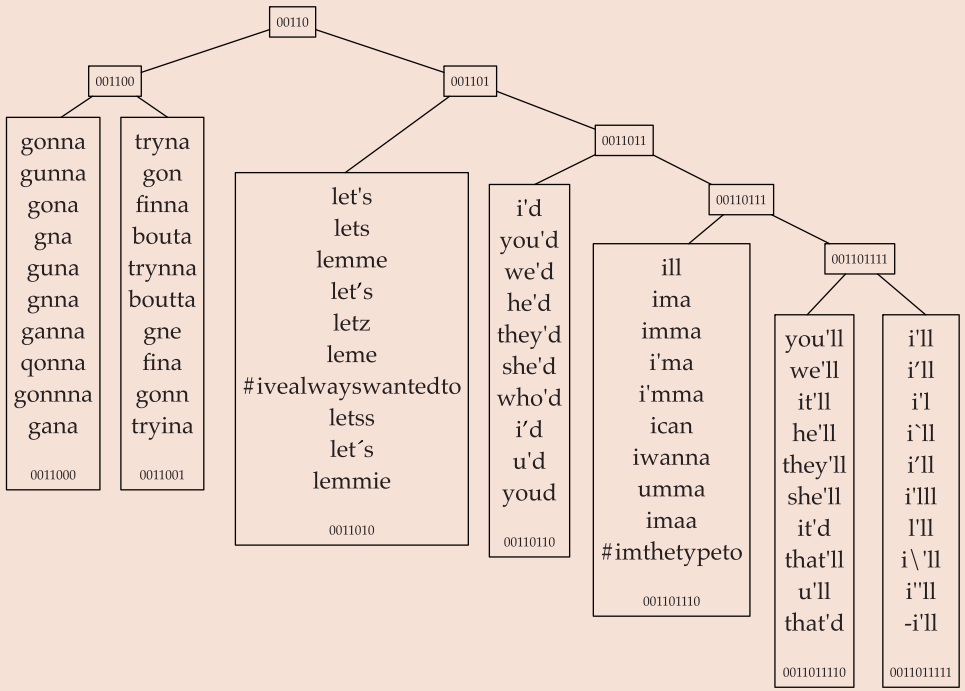
Figure 2
Sample subtrees from a 1,000-word mutual information tree.

Source : (Brown *et al.* , 1992)

Exemple : brown clusters sur worpus de tweets

Figure 1. Example Brown clusters.

These were derived from 56M tweets, see Owoputi et al..[26] for details. Shown are the 10 most frequent words in clusters in the section of the hierarchy with prefix bit string 00110. Intermediate nodes in the tree correspond to clusters that contain all words in their descendants. Note that differently spelled variants of words tend to cluster together, as do words that express similar meanings, including hashtags. The full set of clusters can be explored at http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html. Note that there are several Unicode characters that are visually similar to the apostrophe, resulting in different strings with similar usage.



Source : (Smith, 2020)