Artex est un Autre Résumeur TEXtuel

Juan-Manuel Torres
2024

Université d'Avignon

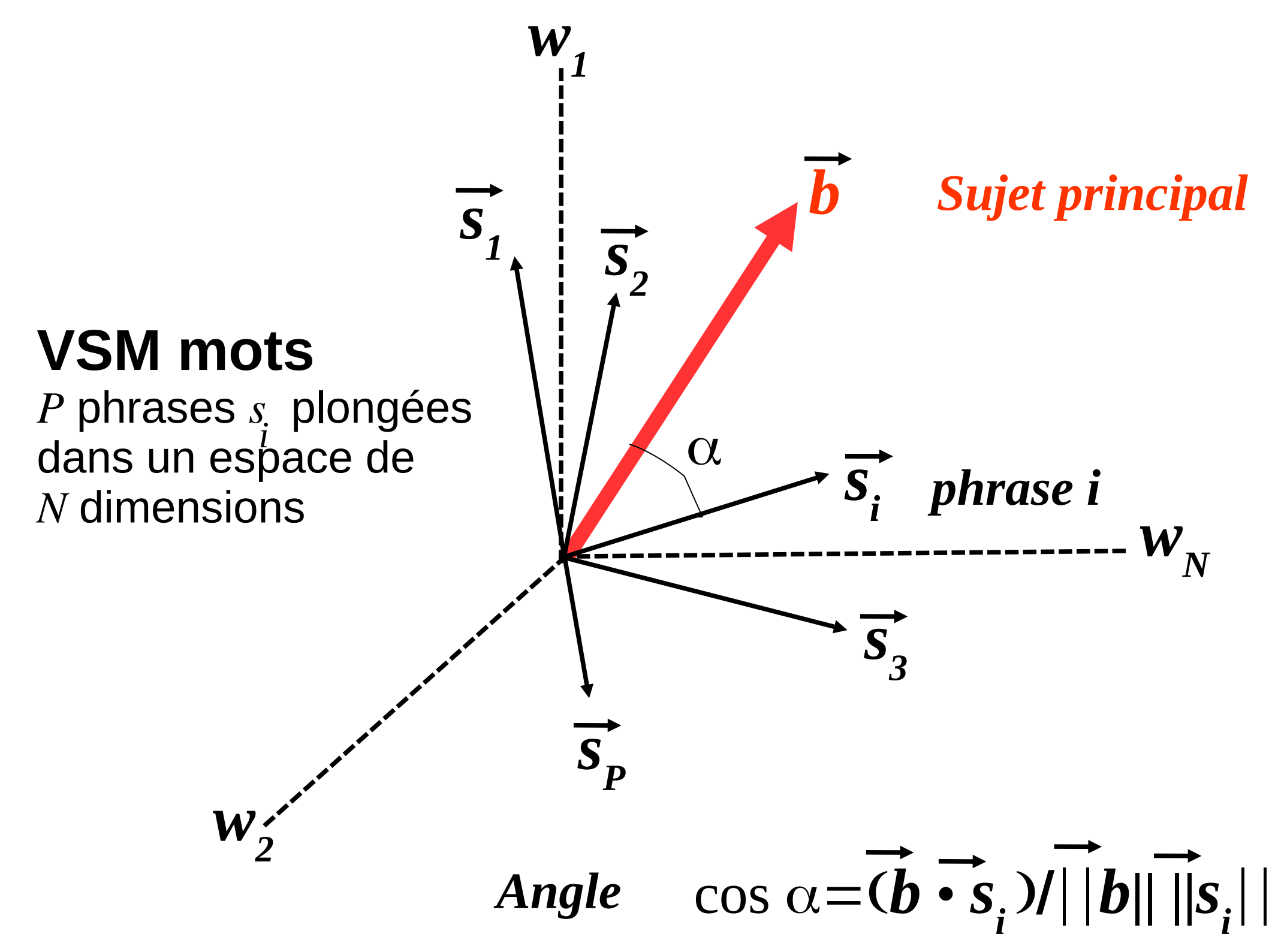
LABORATOIRE
INFORMATIQUE D'AVIGNON
Université d'Avignon et des Pays de Vaucluse

Objetifs

- **Poser la tâche de résumé comme un problème géométrique**
- **Indépendance de la langue**
- **Indépendance du domaine**

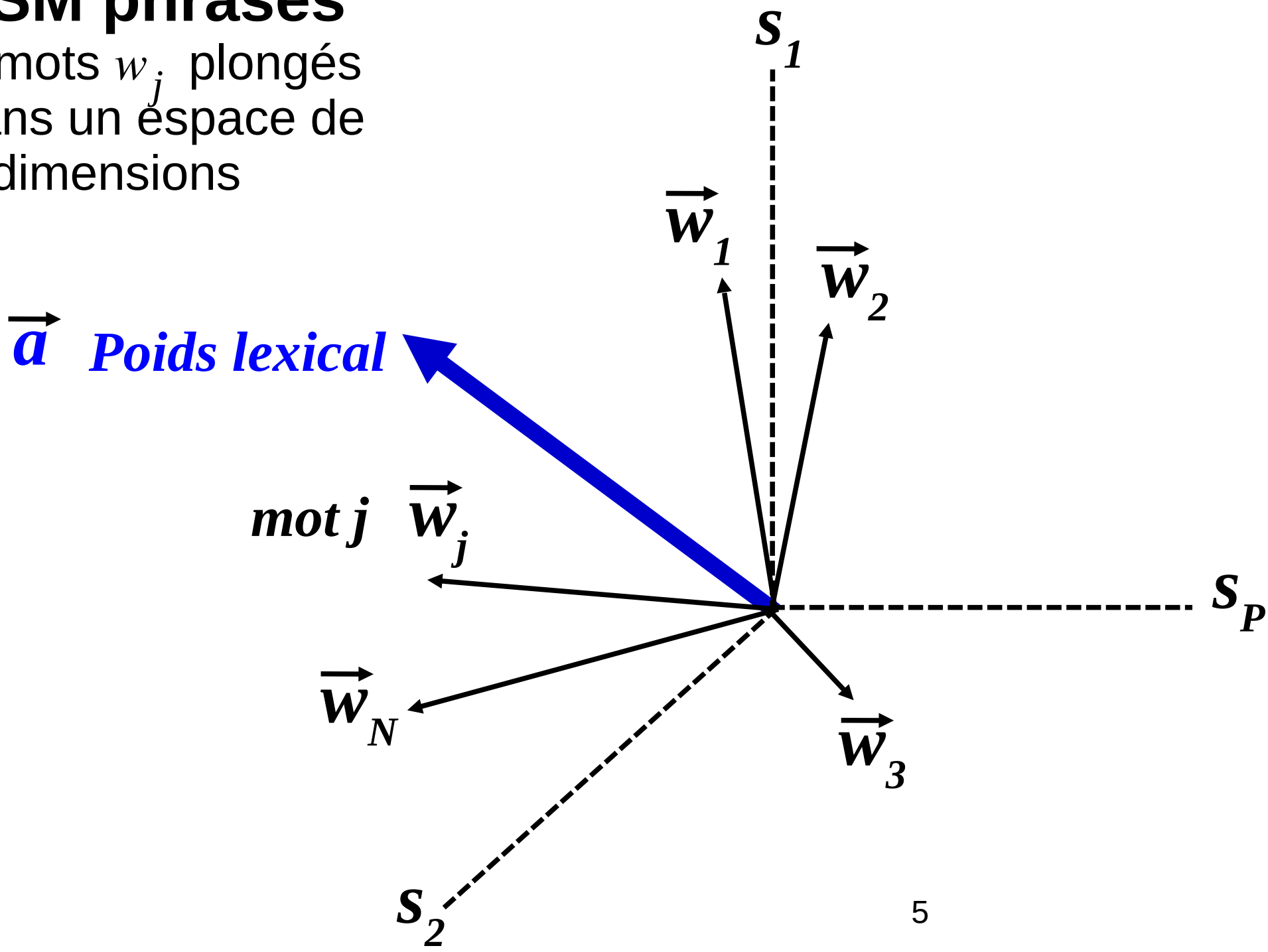
ARTEX

- **Pondération** et **tri** de phrases
- **Extraction** des phrases importantes
  - **Pre-traitement** classique
    - Segmentation
    - Filtrage
    - Lemmatisation/Stemming/Ultra-stemming
  - Calcul du **vecteur sujet principal a**
  - Calcul du **vecteur poids lexical b**
  - **Produit scalaire a.b**
  - **Post-traitement** de surface


w1
b
Sujet principal
s1
s2
VSM mots
P phrases si plongées
dans un espace de
N dimensions
α
si
phrase i
wN
s3
sP
w2
Angle
cos α=(b • si)/||b|| ||si||

VSM phrases

N mots w_j plongés dans un espace de P dimensions



Algorithme

$$(1) \quad a_i = \frac{1}{N} \sum_j s_{i,j}$$

Poids lexicale

$$(2) \quad b_j = \frac{1}{P} \sum_i s_{i,j}$$

Sujet principal

$$(3) \quad \text{score}(s_i) = \left(\vec{s} \times \vec{b}\right) \times \vec{a} = \frac{1}{NP} \left(\sum_j s_{i,j} \times b_j \right) \times a_i \, ;$$

P = 3 phrases
N= 5 mots

Artex

	avignon	pont	rhone	palais	rempart
Phrase 1	0	1	2	0	1
Phrase 2	1	0	0	1	1
Phrase 3	0	0	0	0	1

$a = 1/5 \times \begin{pmatrix} 4 \\ 3 \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{4/5} \\ \mathbf{3/5} \\ \mathbf{1/5} \end{pmatrix}$

$b = 1/3 \times \begin{pmatrix} 1 & 1 & 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} \mathbf{1/3} & \mathbf{1/3} & \mathbf{2/3} & \mathbf{1/3} & \mathbf{1} \end{pmatrix}$

Score(1) = 1/(N x P) $\begin{pmatrix} 0 & 1 & 2 & 0 & 1 \end{pmatrix}$ x $\begin{pmatrix} 1/3 & 1/3 & 2/3 & 1/3 & 1 \end{pmatrix}$ x **4/5**

= 1/15 x $\begin{pmatrix} 0 + 1/3 + 4/3 + 0 + 1 \end{pmatrix}$ x **4/5**

Score(1) = 0,06 x 2,67 x 0,8 = **0,143**

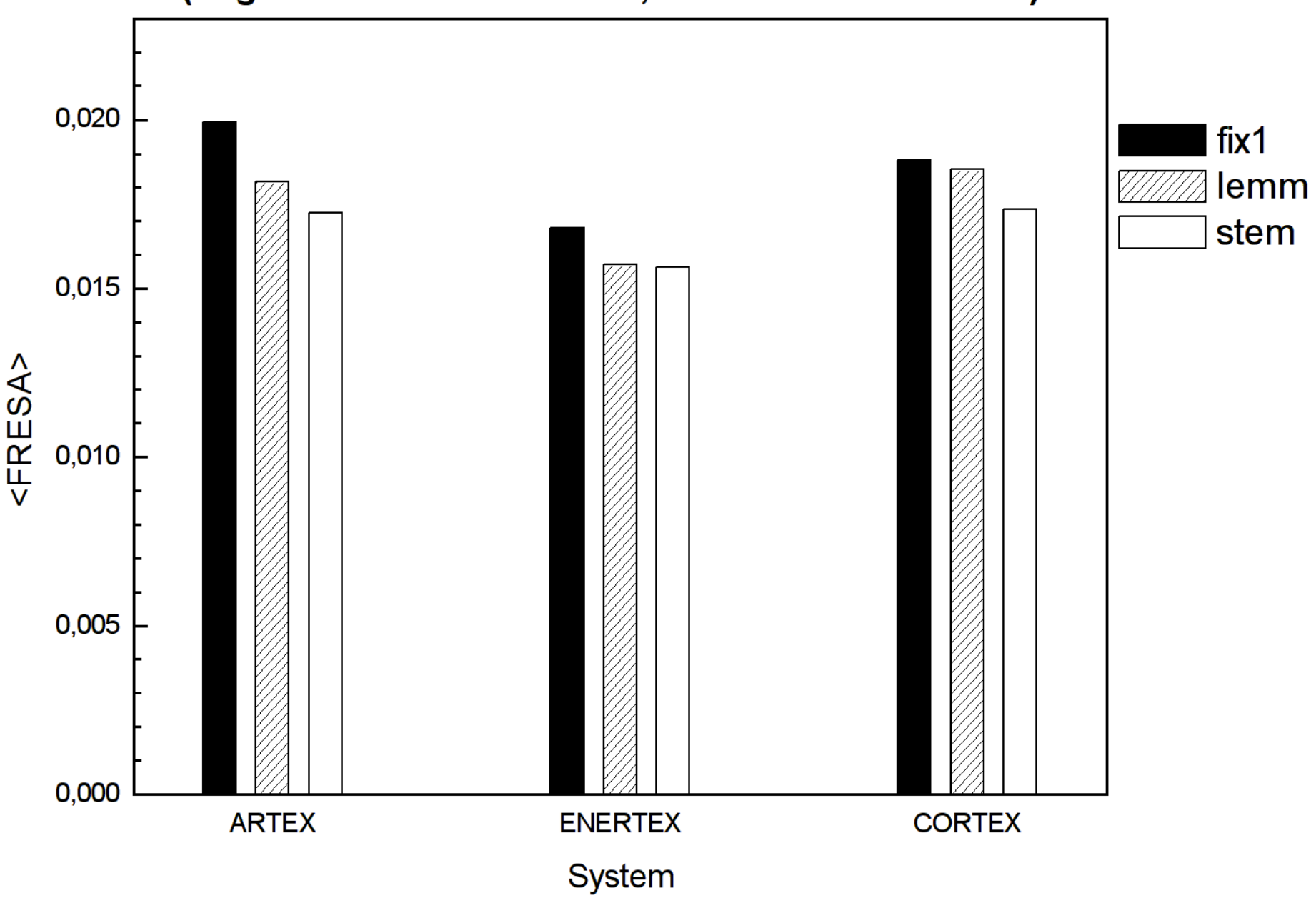
Score(2) = 1/15 x $\begin{pmatrix} 1 & 0 & 0 & 1 & 1 \end{pmatrix}$ x $\begin{pmatrix} \mathbf{1/3} & \mathbf{1/3} & \mathbf{2/3} & \mathbf{1/3} & \mathbf{1} \end{pmatrix}$ x **3/5**

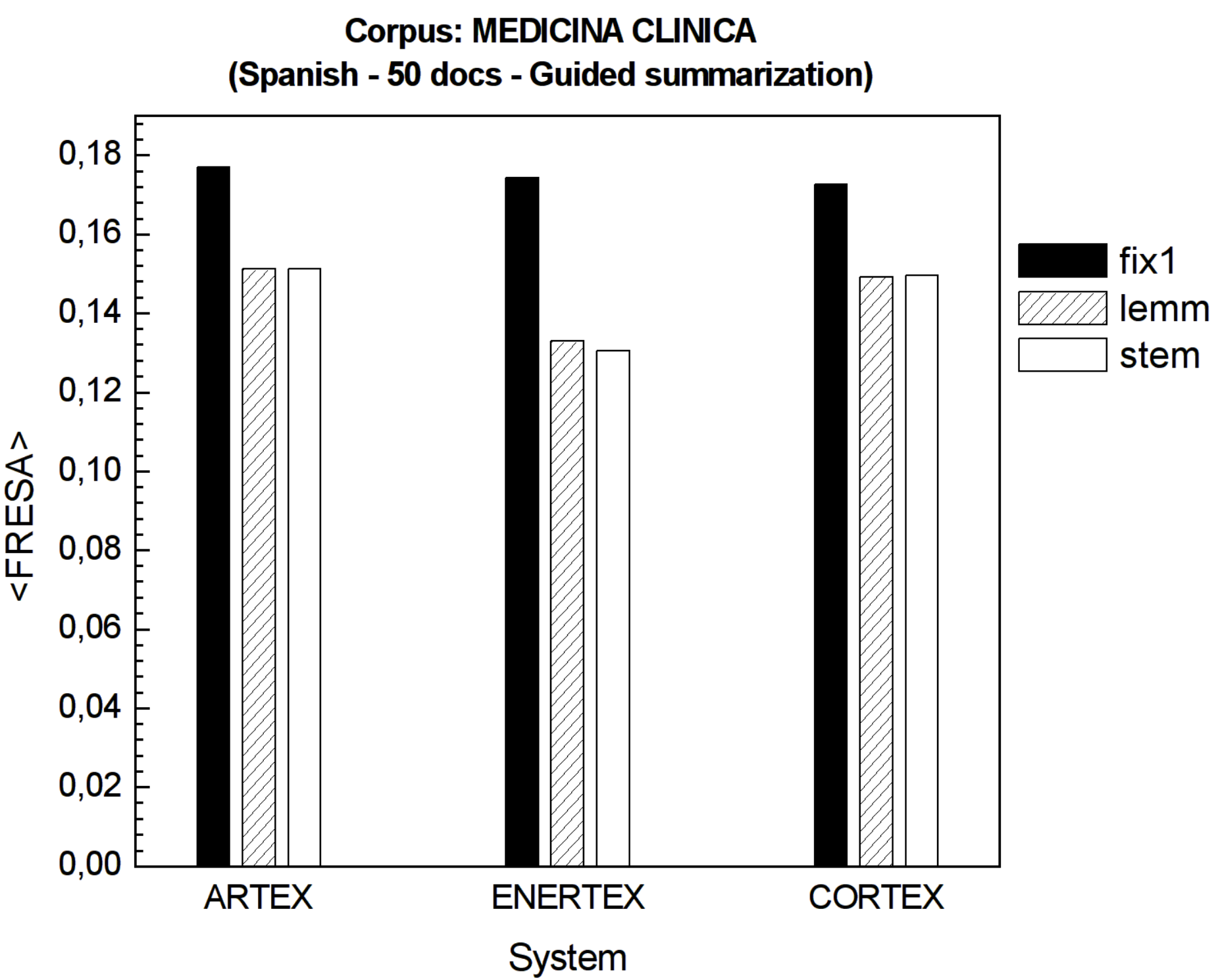
= 1/15 x $\begin{pmatrix} 1/3 + 0 + 0 + 2/3 + 1 \end{pmatrix}$ x **3/5**

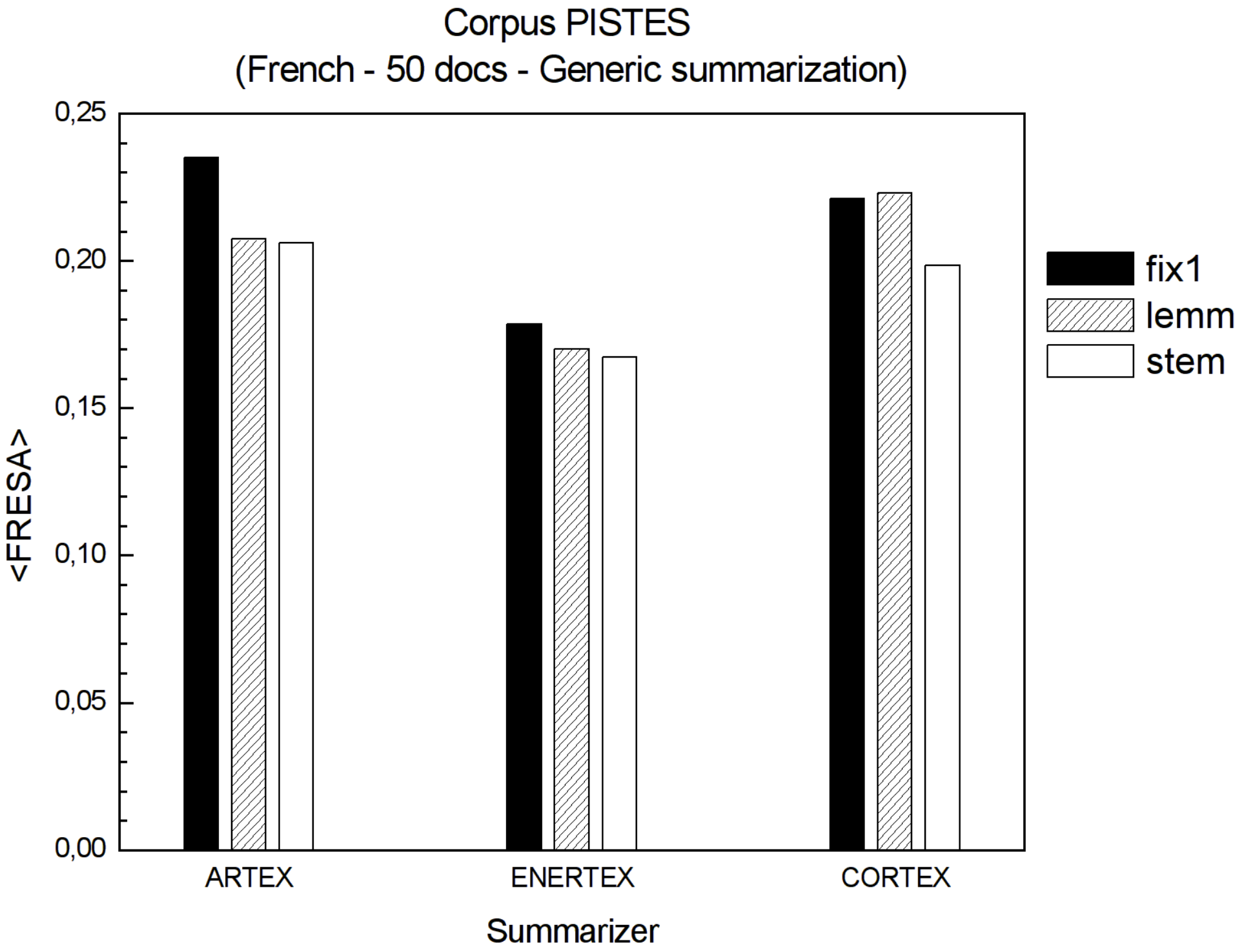
Score(2) = 0,06 x 2 x 0,6 = *0,08*

Score(3) = 1/15 x $\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \end{pmatrix}$ x $\begin{pmatrix} \mathbf{1/3} & \mathbf{1/3} & \mathbf{2/3} & \mathbf{1/3} & \mathbf{1} \end{pmatrix}$ x **1/5**

Score(3) = 0,06 x 1 x 0,2 = *0,01*


CORPUS: DUC'04
(English - 50 Clusters Task 2, Generic summarization)
<FRESA>
0,020
0,015
0,010
0,005
0,000
ARTEX
ENERTEX
CORTEX
System
fix1
lemm
stem


Corpus: MEDICINA CLINICA
(Spanish - 50 docs - Guided summarization)
<FRESA>
0,18
0,16
0,14
0,12
0,10
0,08
0,06
0,04
0,02
0,00
ARTEX
ENERTEX
CORTEX
System
fix1
lemm
stem


Corpus PISTES
(French - 50 docs - Generic summarization)
0,25
0,20
0,15
0,10
0,05
0,00
<FRESA>
ARTEX
ENERTEX
CORTEX
Summarizer
fix1
lemm
stem

Artex avec plogements de mots?

Des mots aux embeddings des mots...

Des mots et des textes...

Corpus

- *Ensemble de documents (textuels) ayant des caractéristiques intéréssantes (taille, domaine, format,...)*
- *Corpus vastes → Apprentissage automatique (modèles)*
- *Corpus spécialisées: évaluation, adaptation des modèles*

Corpus brut

- Variations de la source: PDF (horrible!), texte (encodage), etc
- Disponibilité
- Proprété

En IA il faut pré-traiter les documents → Corpus pré-traité

Pre-traitement de textes

Corpus

- *Avignon, la belle capitale du Vaucluse, jouit d'un climat vraiment bon*
- *De ce fait, les courageux étudiants de l'Université, sont ainsi récompensés dans leurs énormes efforts.*
- *Espérons alors qu'ils réussiront ce cours très intéressant !*

Filtrage/Lemmatisation (mots porteurs de sens)

- avignon capitale vaucluse jouir climat
- courage étudiant université récompense effort
- réussir cours

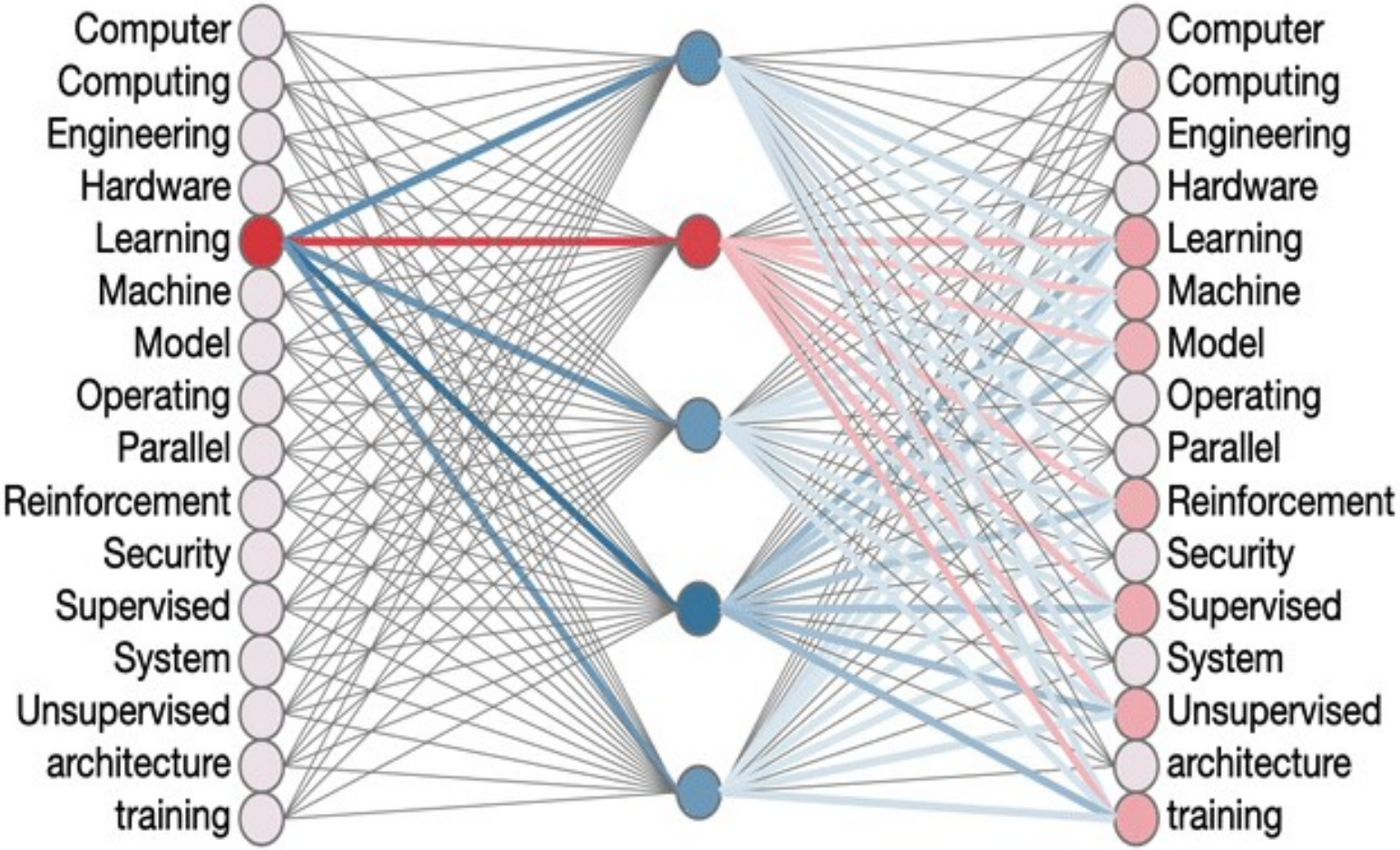
Representation de textes (vectoriel)

Corpus pré-traité:

avignon capitale vaucluse jouir climat
courage étudiant université récompense effort
ésperer réussir cours

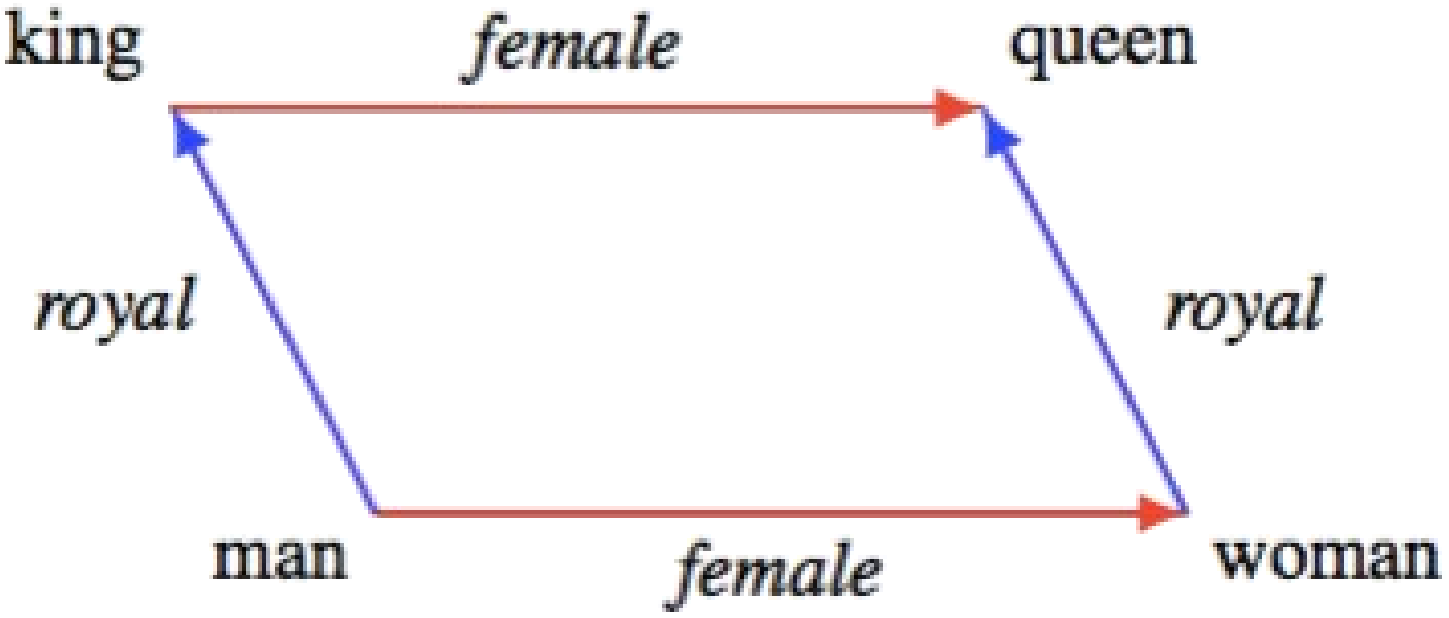
	avignon	capitale	vaucluse	jouir	climat	courage	étudiant	université	récompense	effort	reussir	cours
1	1	1	1	1	1	0	0	0	0	0	0	0
2	0	0	0	0	0	1	1	1	1	1	0	0
3	0	0	0	0	0	0	0	0	0	0	1	1

Représentation embeddings



https://medium.com/@evertongomede/embeddings-in-machine-learning-unleashing-the-power-of-representation-2402bab526fe

Word Embedding Analogies: Understanding King - Man + Woman = Queen



https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html

Artex + IA

- Construire Artex avec embeddings
- Focaliser sur des endroits “intéressants” des textes
- Zones intéressantes dans le cadre du défi = zones avec des EN (Noms des personnes)
- Explorer si vaut la peine de chercher sur tout le texte ou sur un résumé du texte...