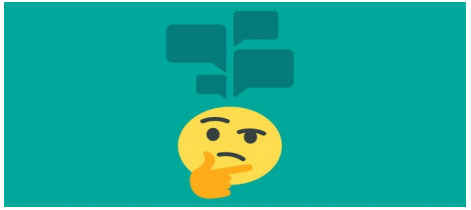
Dakae Tech

TEXT MINING

Par: Asmaa Mountassir, PhD


enterprise
infrastructure
technology
operations
scorecards
objectives
text mining
metrics
applications
solution
stakeholder


إضافة 60 دقيقة
إلى التوقيت الرسمي




Booool
Hate it!
I like it!
:) WOW
Best ever!
Great Speech!
Ehhh,
I dunno

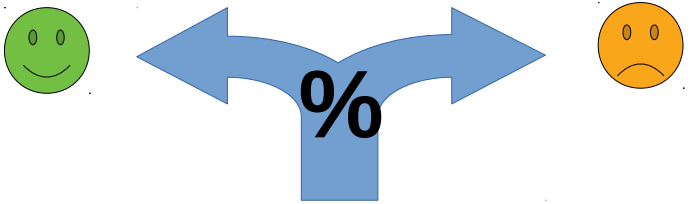
Encore une fois ?! Non !!
Vraiment j’en ai marre
c’est stressant...

Wellah ta 7chouma 3likom
wach makatfakkrouch lina ????

J’aime bien cet horaire
ça me convient très bien

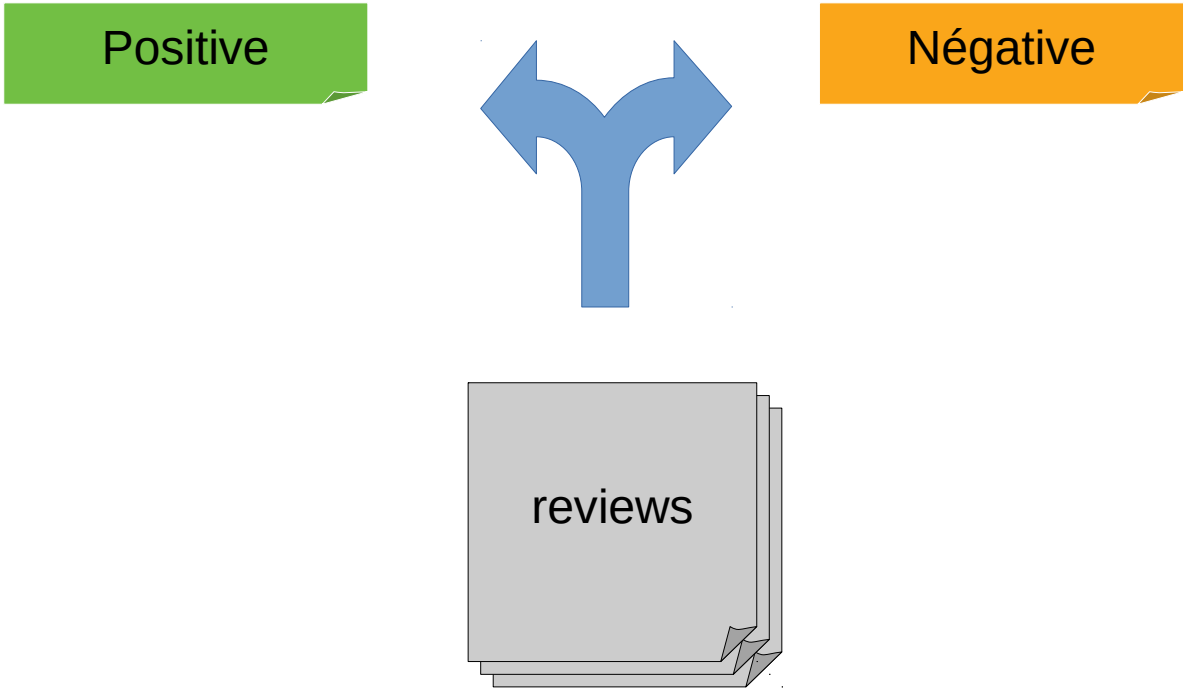
الله يهديكم وخلاص
را هدشي ماشي معقول
حسبنا الله ونعم الوكيل

Pfffff No Comment !!!

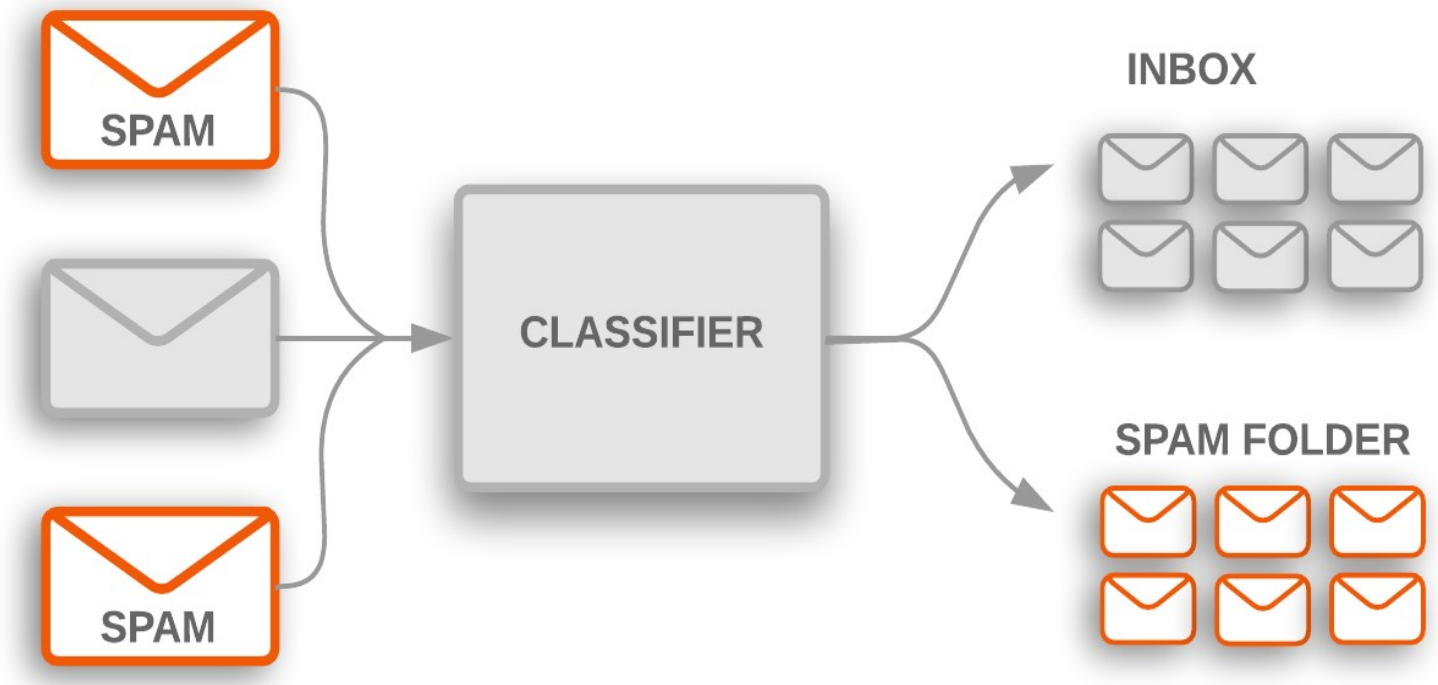


Problème de Classification

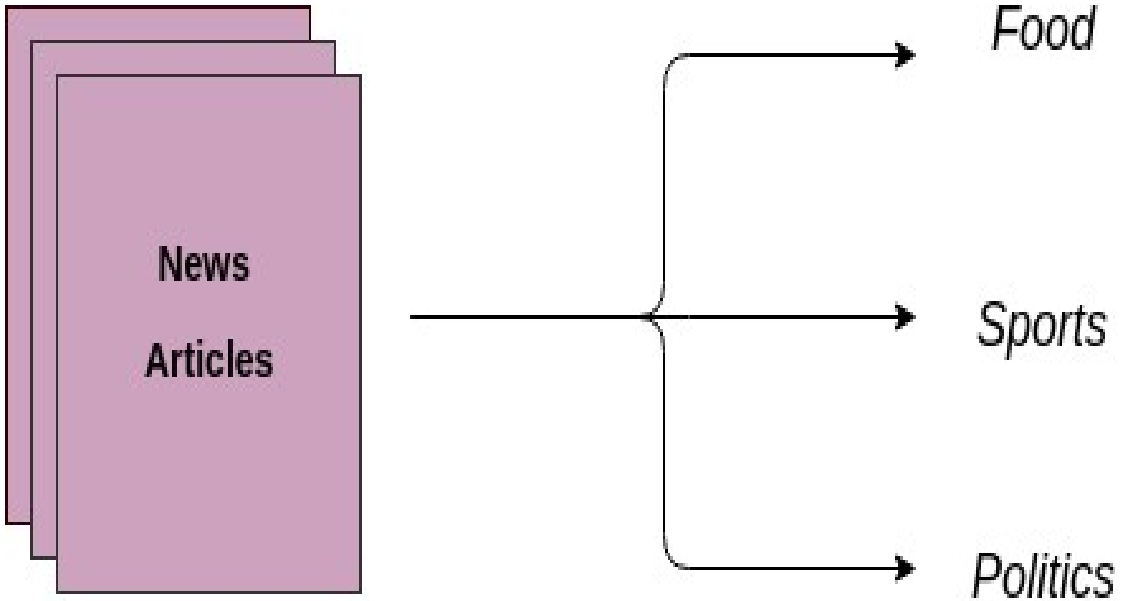
Dakae Tech



Opinion Mining
Sentiment Analysis



Spam Detection



Text Categorization

Dakae Tech

Machine Learning Supervisé

Est-il possible d'appliquer les Algos ML sur des textes ??

Algos ML : vecteurs numériques !!

Et si c’était des textes rédigés par des internautes ? Plein d’erreurs de ponctuation et de langue ?!

Encore une fois ?! Non !!
Vraiment j'en ai marre
c'est stressant...

Wellah ta 7chouma 3likom
wach makatfakkrouch lina ????

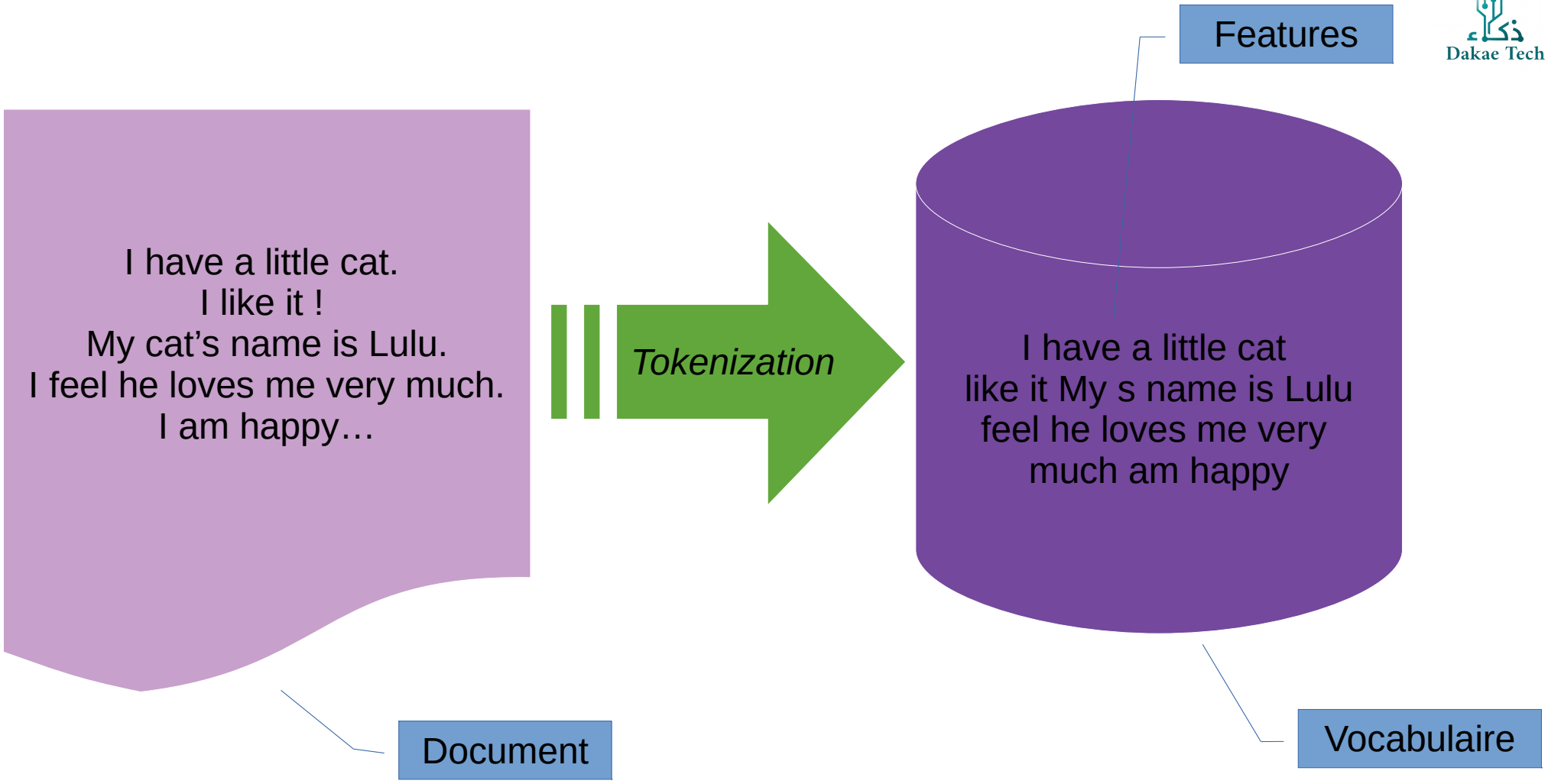
J aime b1 7 horaire
ca me convien tr b1

الله يهديكم وخلاص
را هدشي ماشي معقول
حسبنا الله ونعم الوكيل

Pfffff No Comment !!!


Multilingue
Erreurs syntaxiques, lexicales, grammaticales ,...
Mal ponctué
Challenges
Textes courts / Ironie
Divers Styles
Hors Sujets

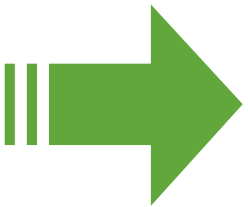
Sac de Mots


Features
I have a little cat.
I like it !
My cat’s name is Lulu.
I feel he loves me very much.
I am happy…
Tokenization
I have a little cat
like it My s name is Lulu
feel he loves me very
much am happy
Document
Vocabulaire

I have a little cat.
I like it !
My cat’s name is Lulu.
I feel he loves me very much.
I am happy…

My dog’s name is Jeff.
My mother does not like dogs.
But what can I do ?!

Cats do not love dogs !
But I have two kittens
and one dog.



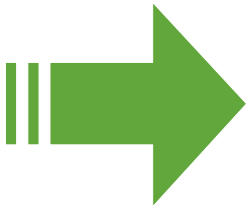
I have a little cat
like it My s name is Lulu
feel he loves me very
much am happy
dog Jeff mother does not
dogs But what can do
cats love two kittens
and one
……..

I have a little cat.
I like it !
My cat’s name is Lulu.
I feel he loves me very much.
I am happy…

My dog’s name is Jeff.
My mother does not like dogs.
But what can I do ?!

Cats do not love dogs !
But I have two kittens
and one dog.

……



I have a little cat
like it My s name is Lulu
feel he loves me very
much am happy
dog Jeff mother does not
dogs But what can do
cats love two kittens
and one
…….

I have a little cat.
I like it !
My cat’s name is Lulu.
I feel he loves me very much.
I am happy…

D1

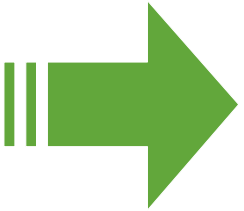
My dog’s name is Jeff.
My mother does not like dogs.
But what can I do ?!

D2

Cats do not love dogs !
But I have two kittens
and one dog.

D3

……



I have a little cat
like it My s name is Lulu
feel he loves me very
much am happy
dog Jeff mother does not
dogs But what can do
cats love two kittens
and one
…….

Dakae Tech

	I	have	a	little	cat	like	it	My	s	name	is	Lulu	feel	he	...	do	cats	two	kittens	and	one
D1																					
D2																					
D3																					

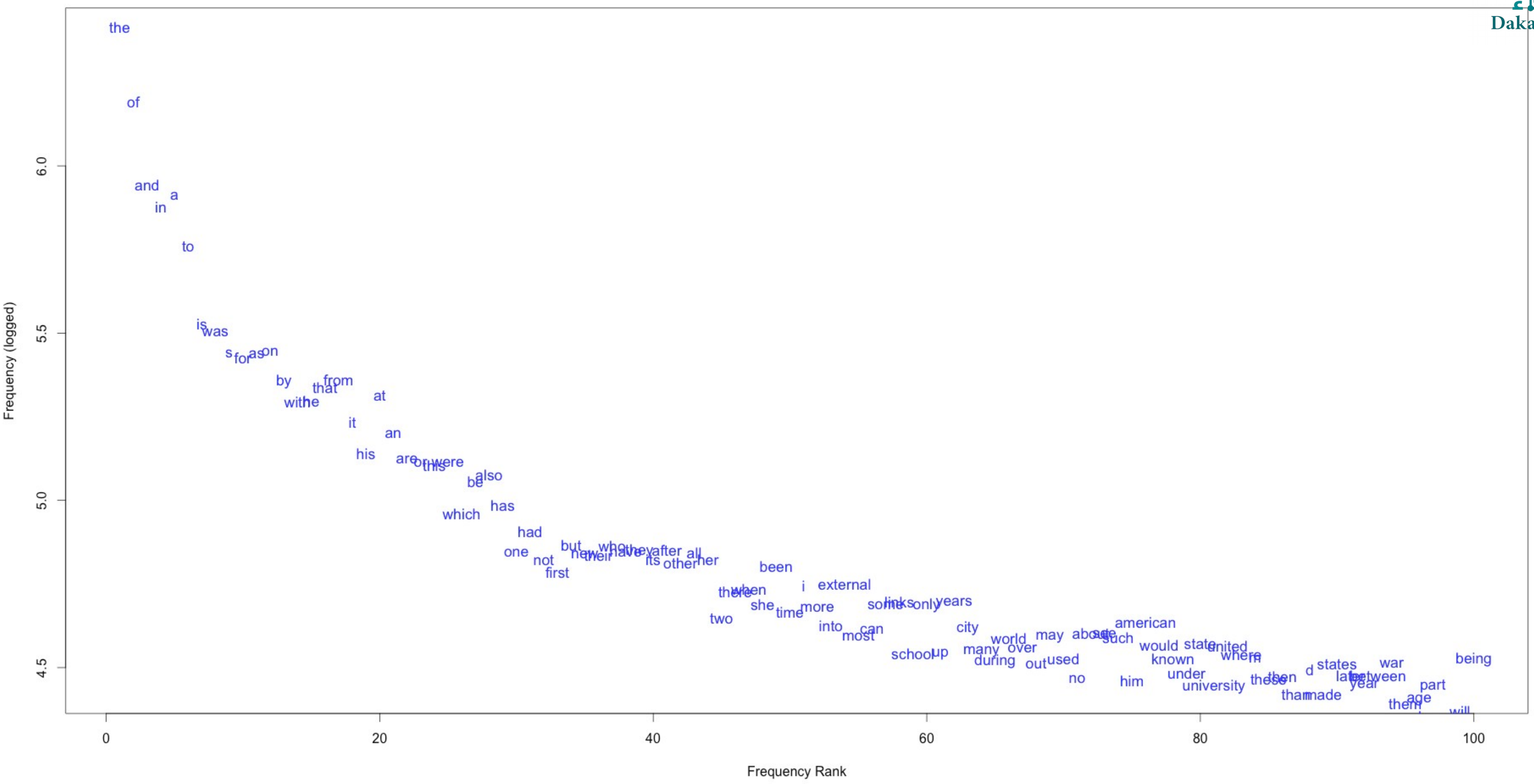
Document-Term Matrix
(3x36)

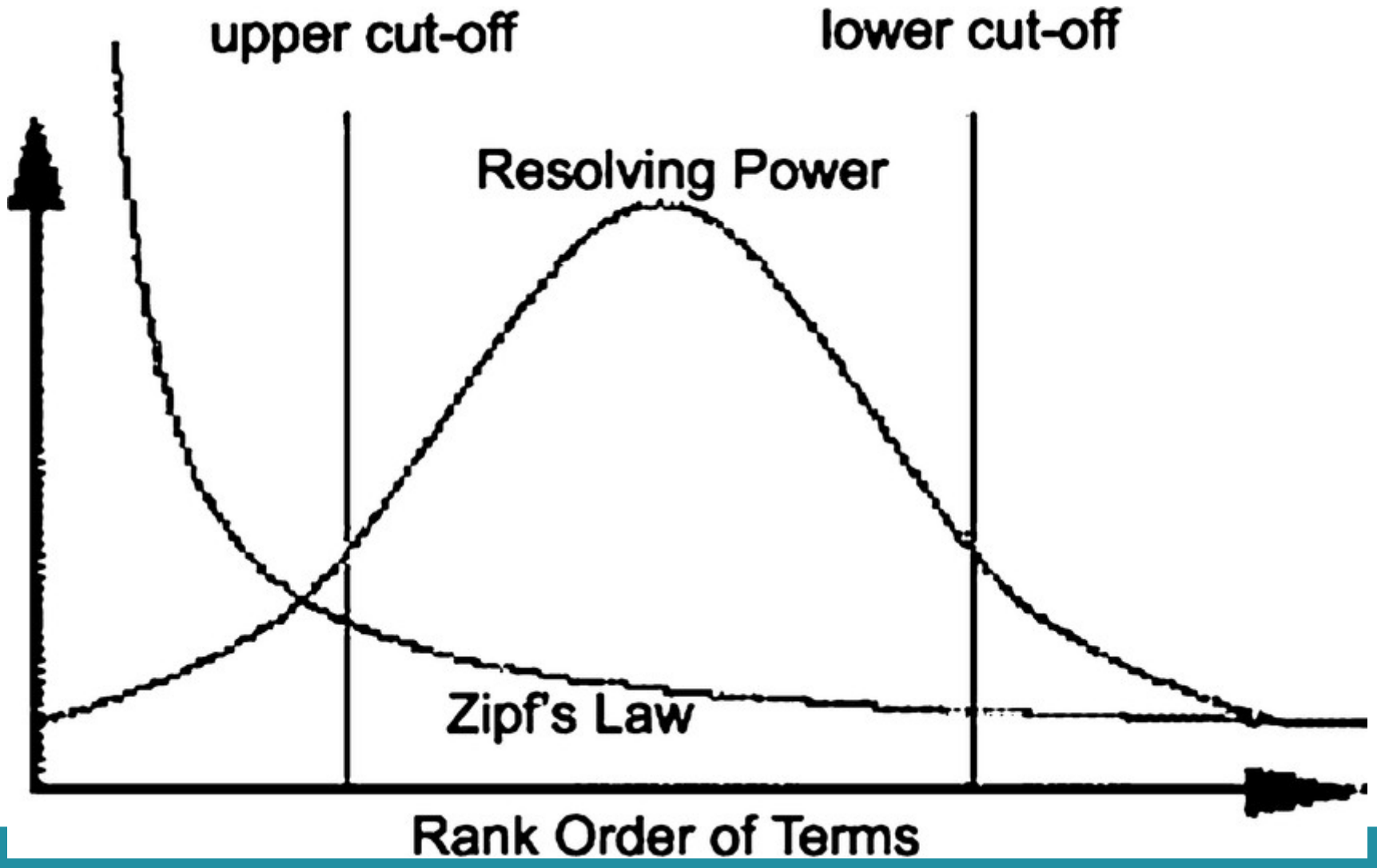
Est-ce que tous les termes seraient utiles ?
- Des mots qui se répètent souvent
- Des termes dérivant de la même racine
-Des mots qui apparaissent une fois
- Mêmes mots en minuscule/majuscule

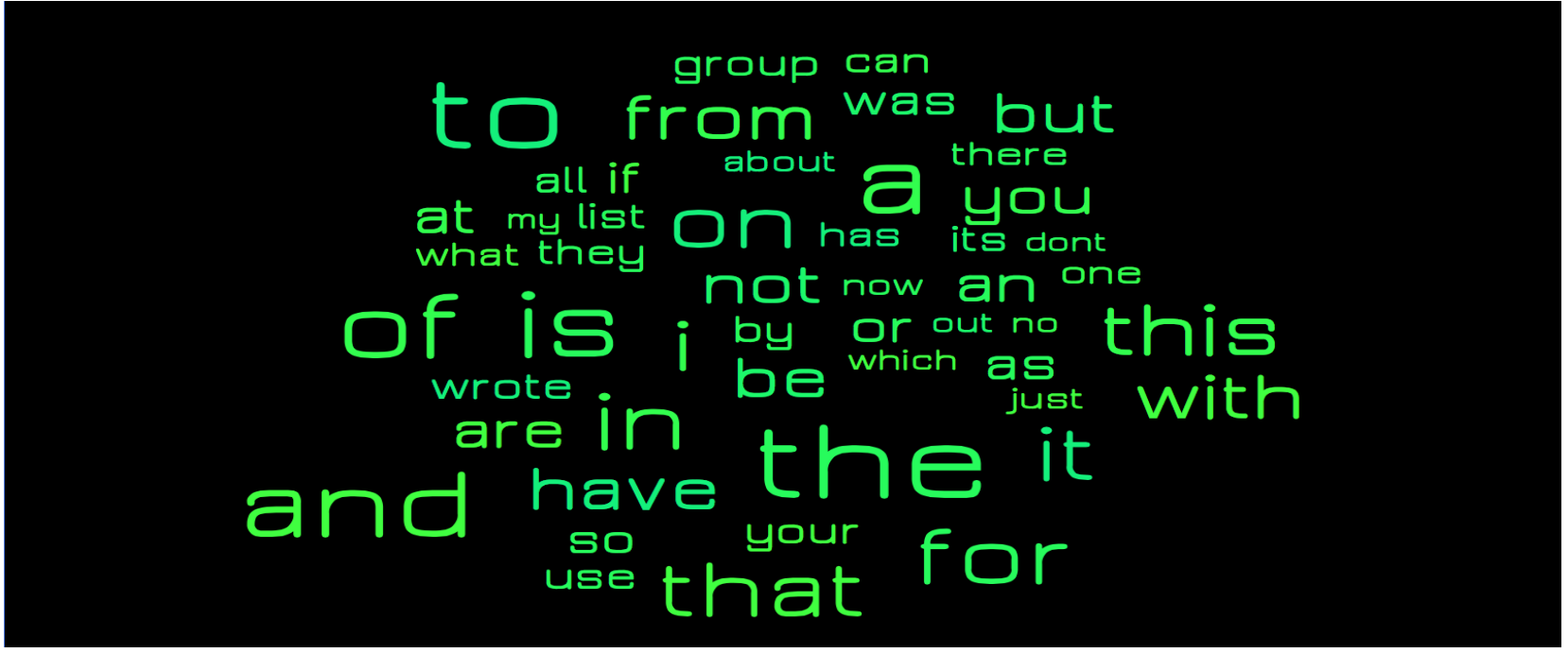
I have a little cat
like it My s name is Lulu
feel he loves me very
much am happy
dog Jeff mother does not
dogs But what can do
cats love two kittens
and one

Dakae Tech

100 Most Frequent Words in Wikipedia

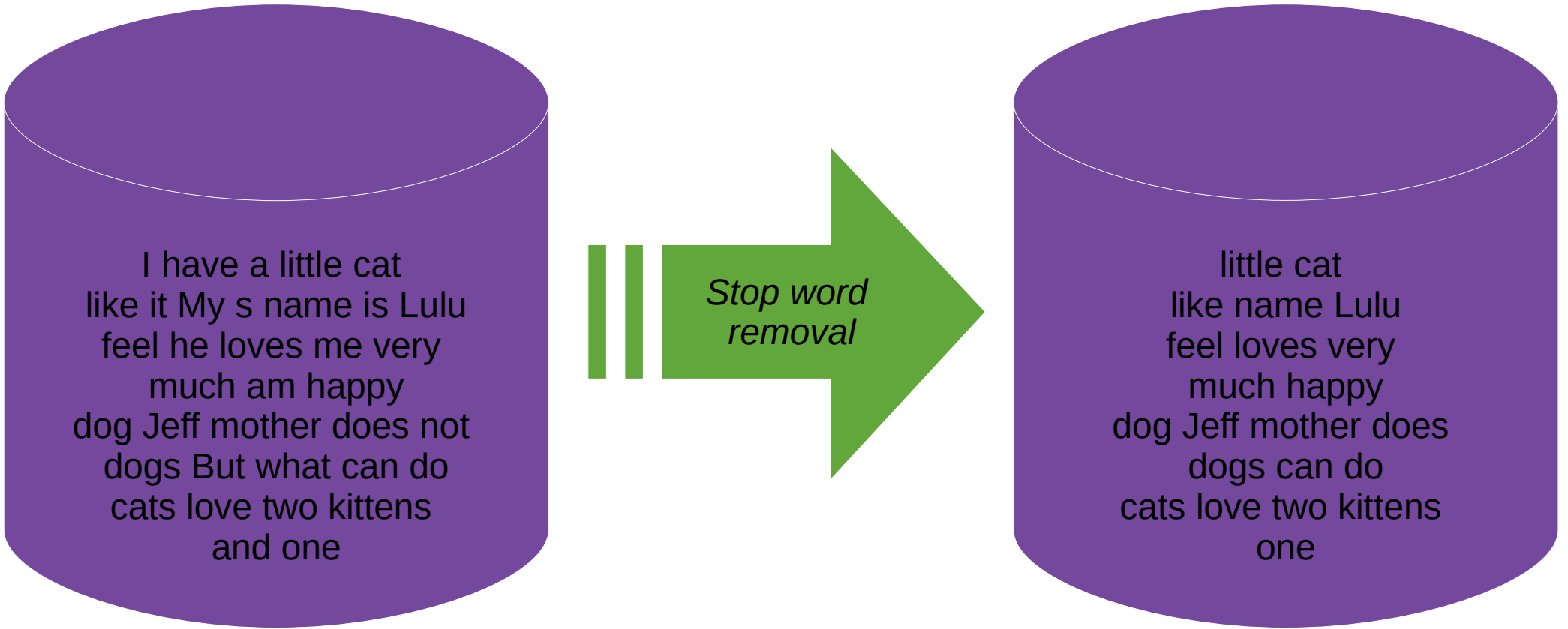


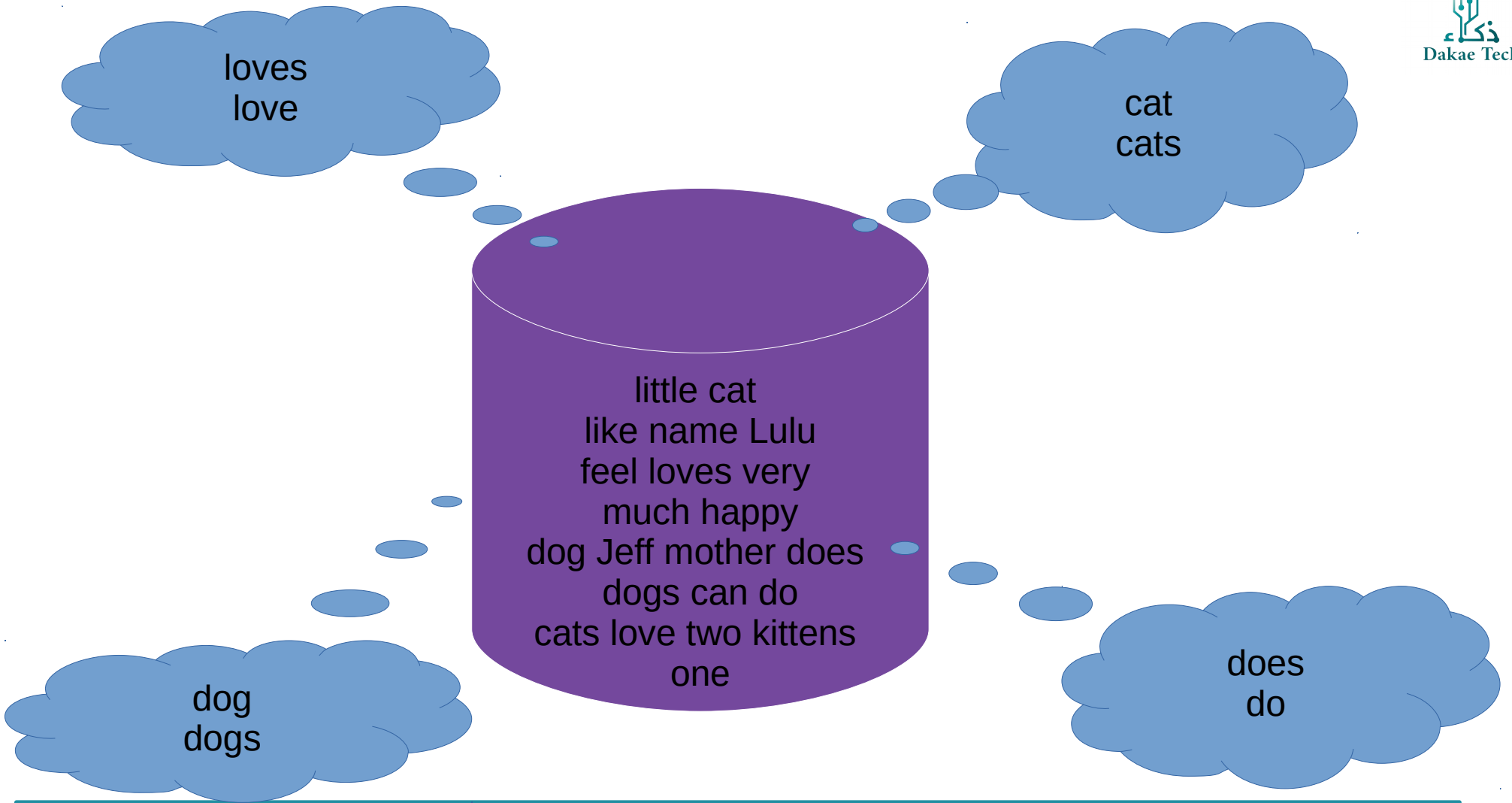

upper cut-off
lower cut-off
Resolving Power
Zipf's Law
Rank Order of Terms


group can
to from was but
about there
all if a you
at my list on has its dont
what they
not now an one
of is i by or out no this
which as
wrote be just with
are in the it
and have
so your
use that for


Stop Words
Common Words,
Determiners,
Conjunctions ...

- Cette liste est établie manuellement :
Chaque langue dispose de sa liste de mots vides !!
- Elle peut être construite automatiquement :
Les mots avec une fréquence qui dépasse un seuil sont considérés des mots vides


I have a little cat
like it My s name is Lulu
feel he loves me very
much am happy
dog Jeff mother does not
dogs But what can do
cats love two kittens
and one
Stop word removal
little cat
like name Lulu
feel loves very
much happy
dog Jeff mother does
dogs can do
cats love two kittens
one


loves
love
cat
cats
little cat
like name Lulu
feel loves very
much happy
dog Jeff mother does
dogs can do
cats love two kittens
one
dog
dogs
does
do

loves
love

cat
cats

dog
dogs

does
do

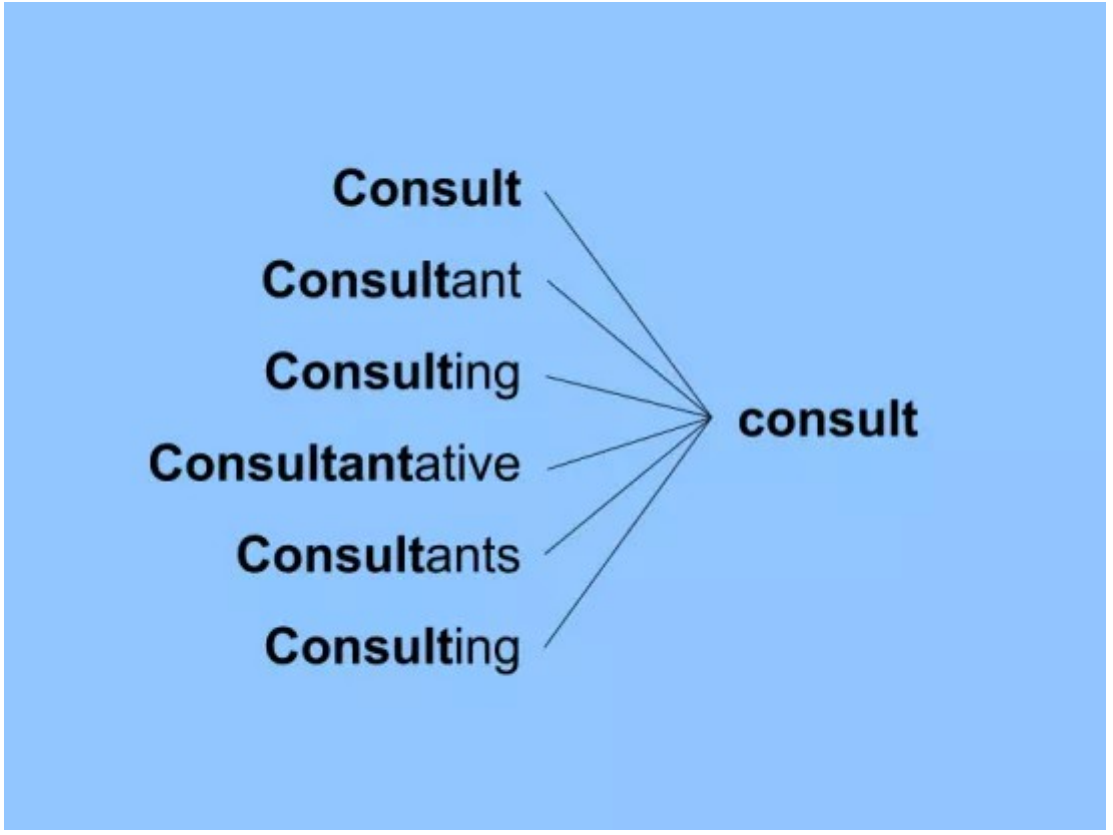
Stemming

STEMMING WORDS

Stemming reduces a word to its stem. The result is less readable by humans but makes the text more comparable across observations.

EXAMPLE: "Tradition" and "Traditional" have the same stem: "tradit"

ChrisAlbon

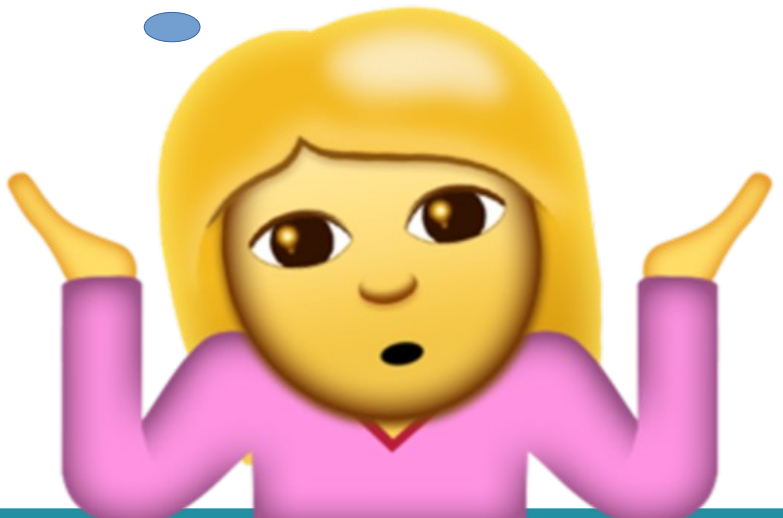

Consult
Consultant
Consulting
Consultantative
Consultants
Consulting
consult

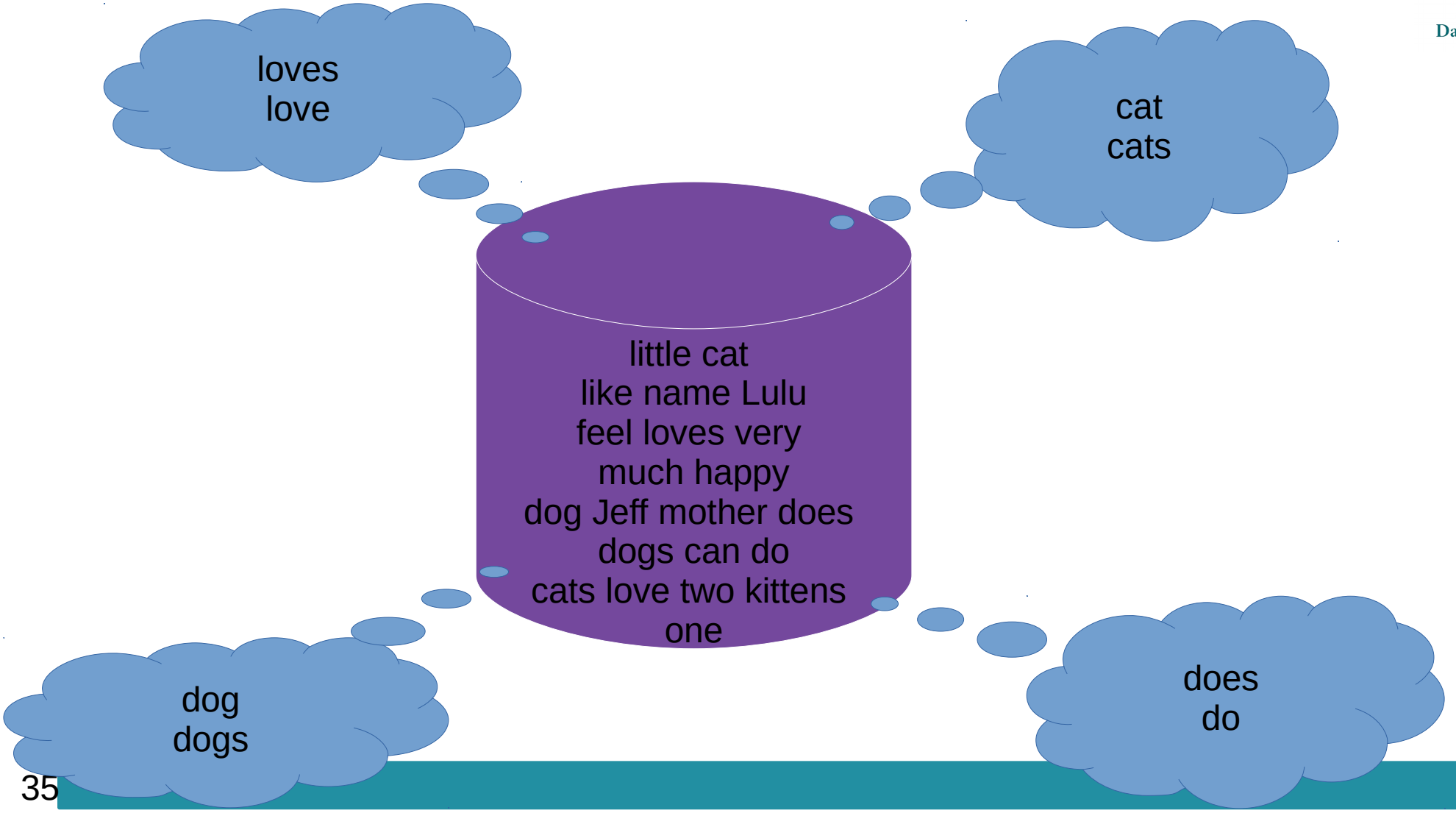
Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

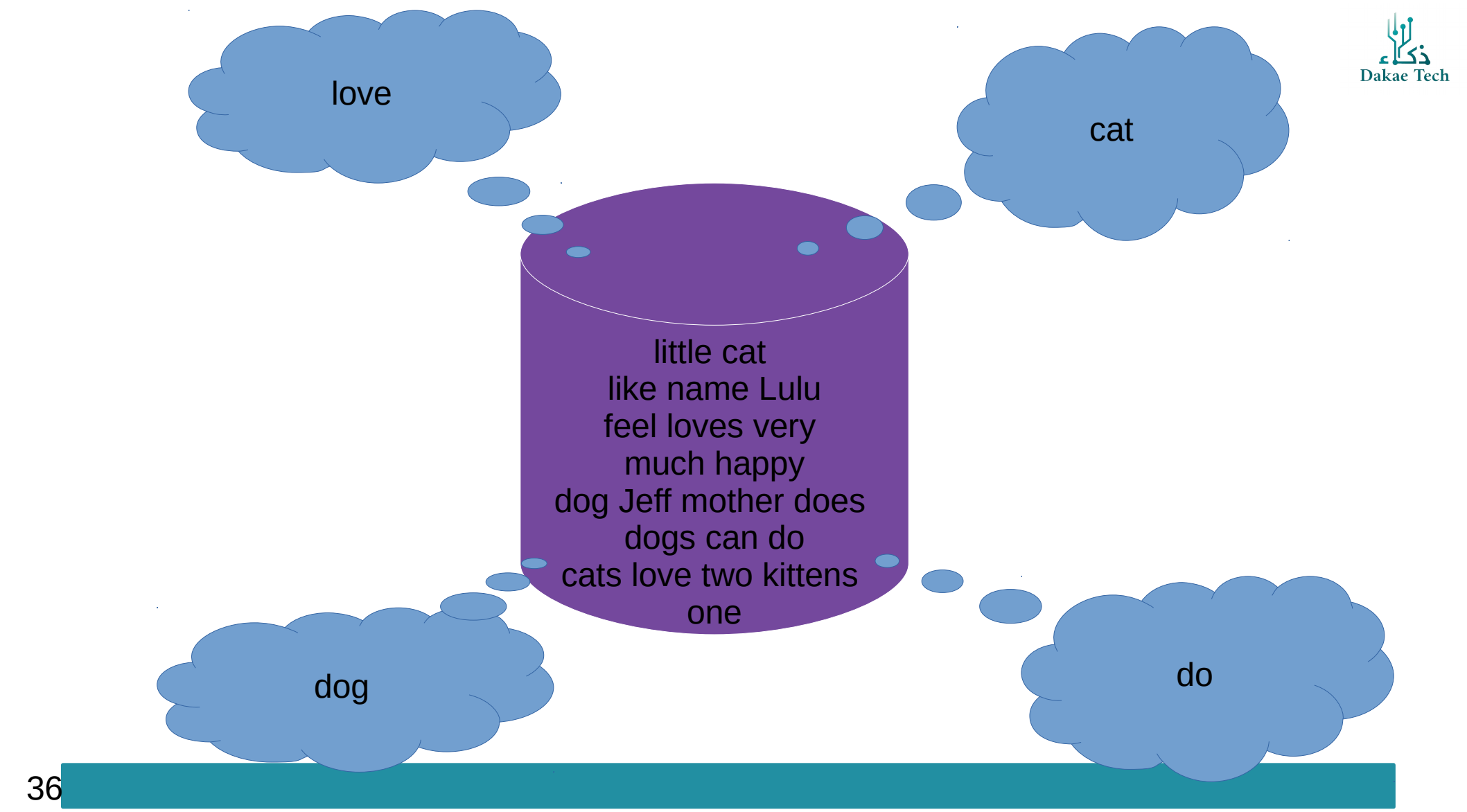
Dakae Tech

Il y a plusieurs algorithmes de stemming qui peuvent donner des résultats différents sur certains mots !

Word	*Porter1*	*Porter2*
cats	*cat*	*cat*
meekly	*meekli*	*meek*
pompous	*pompou*	*pompous*
thursday	*thursdai*	*thursday*
tied	*ti*	*tie*



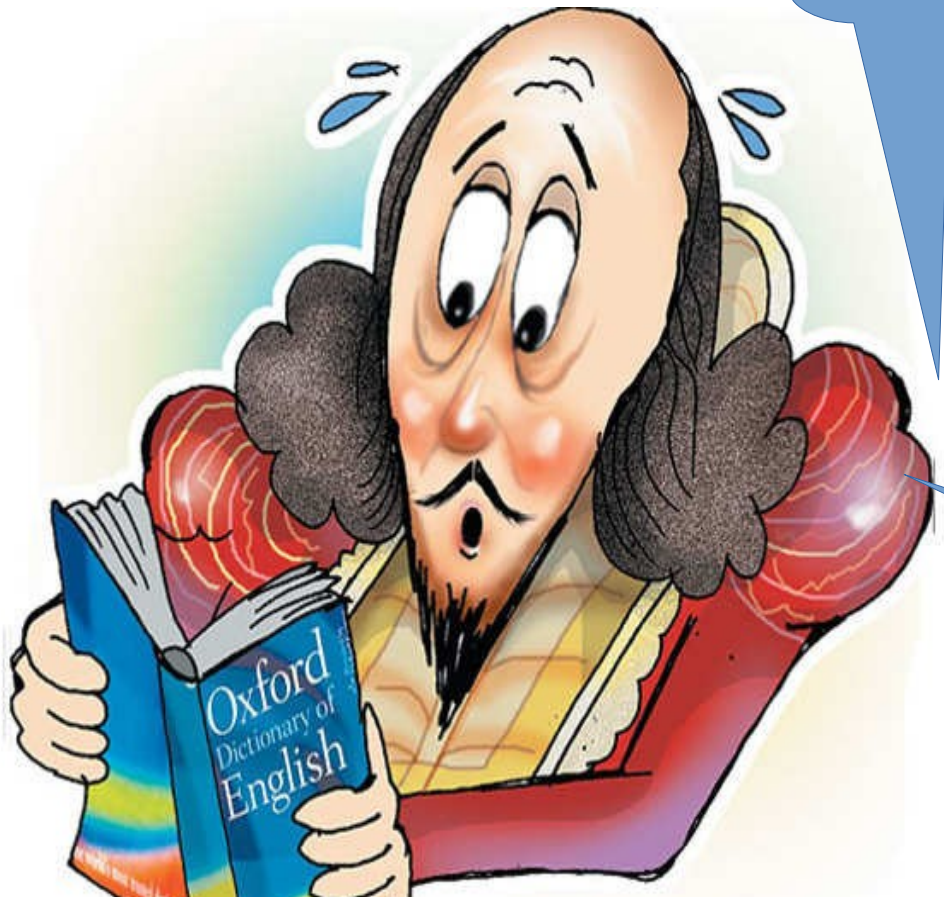

loves
love
cat
cats
little cat
like name Lulu
feel loves very
much happy
dog Jeff mother does
dogs can do
cats love two kittens
one
dog
dogs
does
do


love
cat
little cat
like name Lulu
feel loves very
much happy
dog Jeff mother does
dogs can do
cats love two kittens
one
dog
do

Réduction de 50 %

little cat
like name Lulu
feel very
much happy
dog Jeff mother
can do
love two kitten
one


On peut aller
même loin...



Like ~ Love

Very ~ Much

Kitten ~ Cat

Réduction de 60% !!

little cat
like name Lulu
feel
much happy
dog Jeff mother
can do
two one

Dakae Tech

Réduction de 60% !!

Noms propres
dont la fréquence=1

Réduction de 60% !!

little cat
like name ~~Lulu~~
feel
much happy
dog ~~Jeff~~ mother
can do
two one

	little	cat	like	feel	name	much	happy	dog	mother	can	do	two	one
D1													
D2													
D3													

Document-Term Matrix
(3x13)


Et si on est intéressé à analyser les sentiments ??

I have a little cat.
I like it !
My cat's name is Lulu.
I feel he loves me very much.
I am happy…

My dog's name is Jeff.
My mother does not like dogs.
But what can I do ?!

Cats do not love dogs !
But I have two kittens
and one dog.

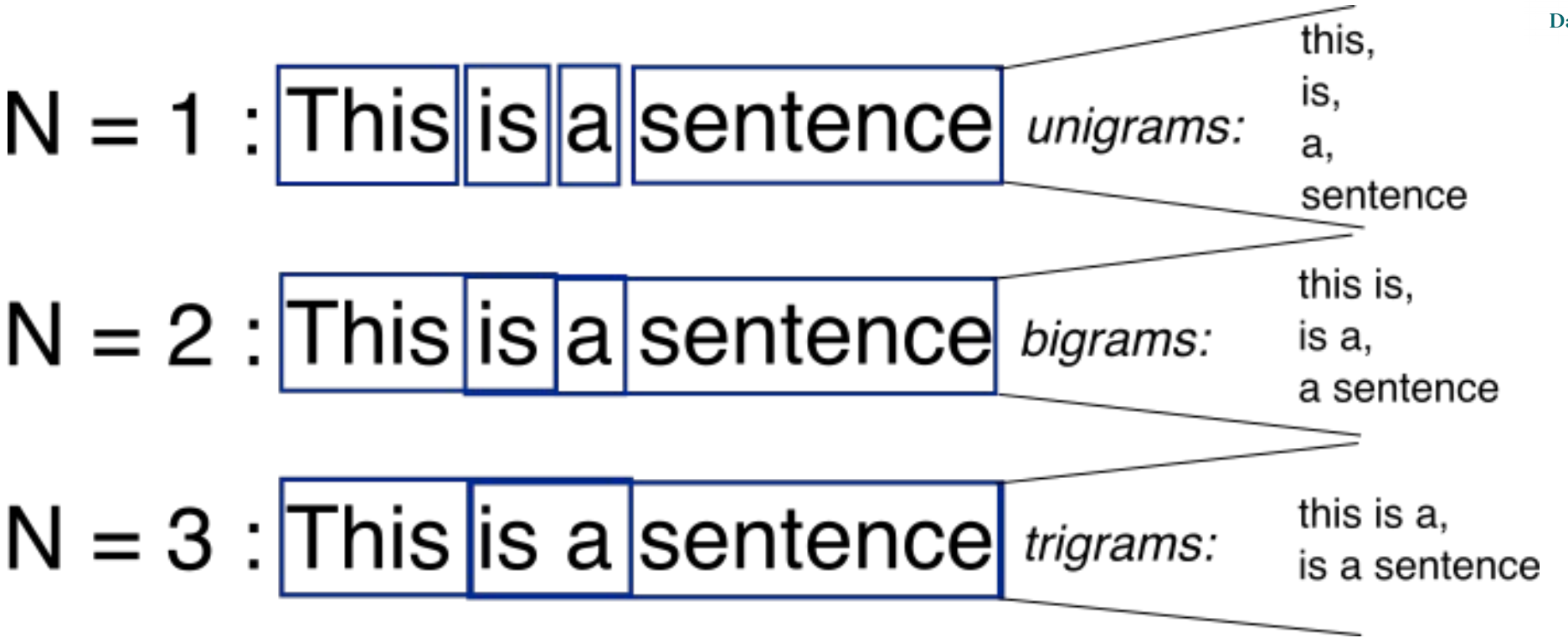
Dakae Tech

Mots n-grammes !

Dakae Tech

Mots n-grammes !

The official definition of an n-gram is that they are a subsequence of n items from a given sequence, where for search engines these sequences are words.



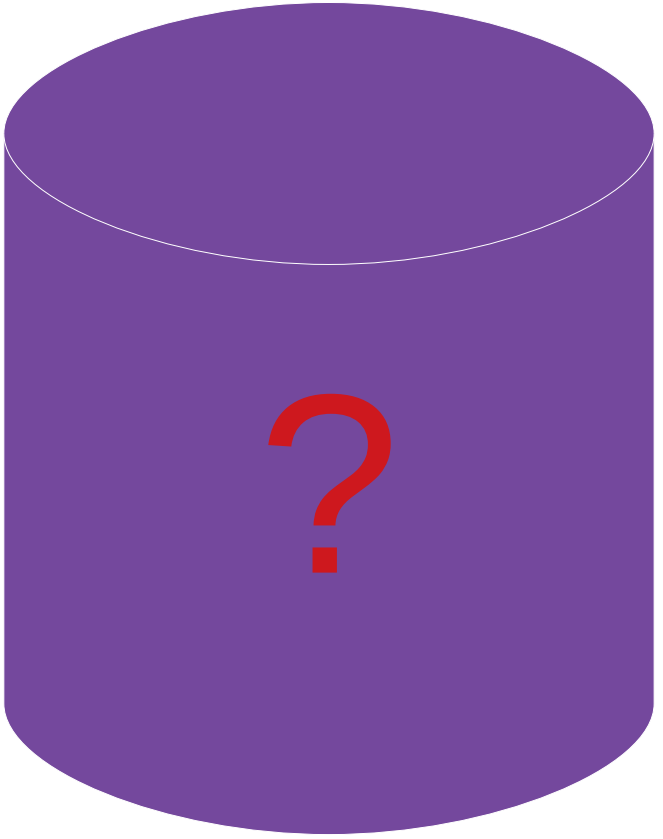
Mots n-grammes

Input Text	"Character N-gram"
Unigram (1-gram)	C, h, a, r, a, c, t, e, r, , N, -, g, r, a, m
Bigram (2-gram)	Ch, ha, ar, ac, ct, te, er, r , N, N-, -g, ...
Trigram (3-gram)	Cha, har, arc, rct, cte, ter, er , r N, N-, ...
Quadrogram (4-gram)	Char, harc, arct, rcte, cter, ter , er N, ...
...	...

N-grammes Charactères

I have a little cat.
I like it !
My cat's name is Lulu.
I feel he loves me very much.
I am happy…

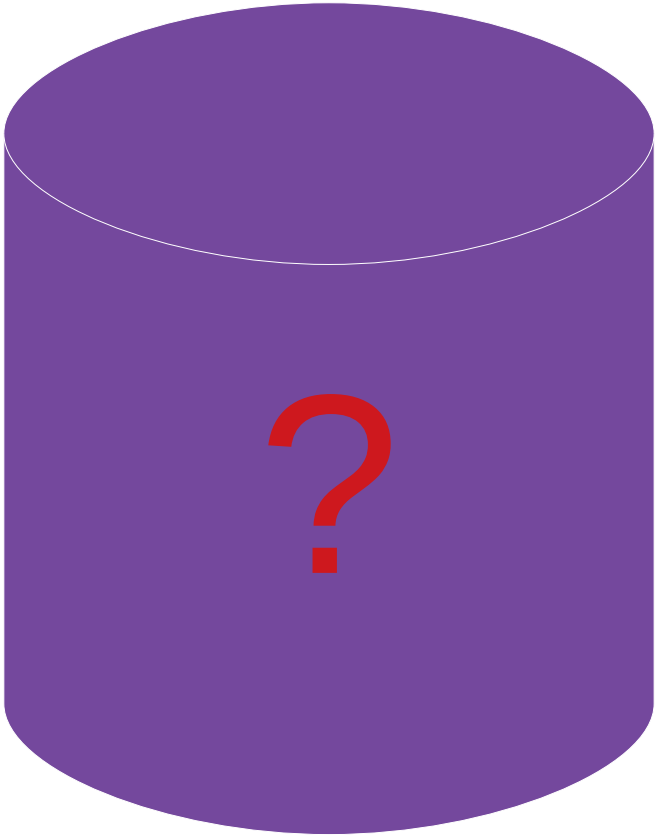
Mots n-grammes
n=2




Et si on supprime
les stop-words
avant de découper
en n-grammes ?

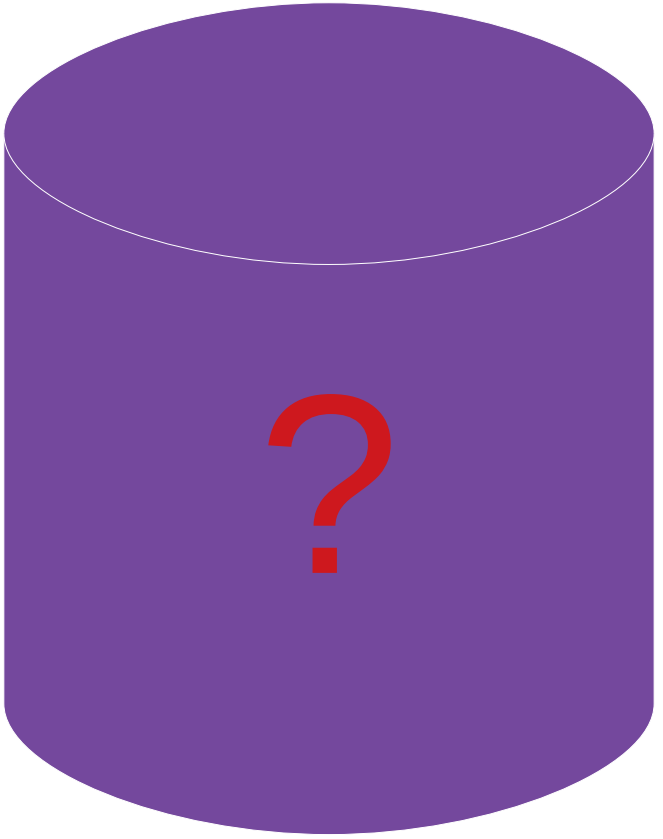
I have a little cat.
I like it !
My cat's name is Lulu.
I feel he loves me very much.
I am happy…

Stop-words
+
Mots n-grammes



I have a little cat.
I like it !
My cat's name is Lulu.
I feel he loves me very much.
I am happy…

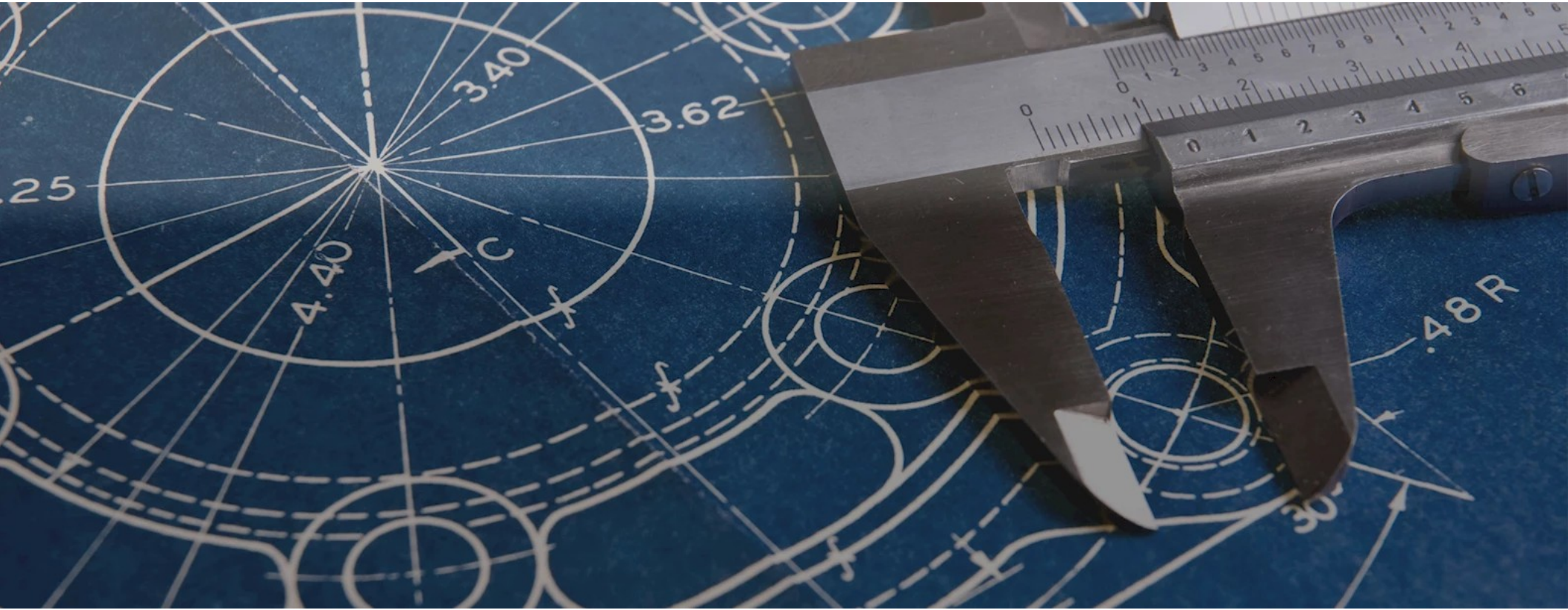
Stop-words
+
Mots n-grammes



Dakae Tech

Le choix du modèle et l’ordre des techniques de prétraitement peuvent donner des résultats tout à fait différents !!


DON'T MAKE ME
CHOOSE



Feature Selection methods

CHI-SQUARED

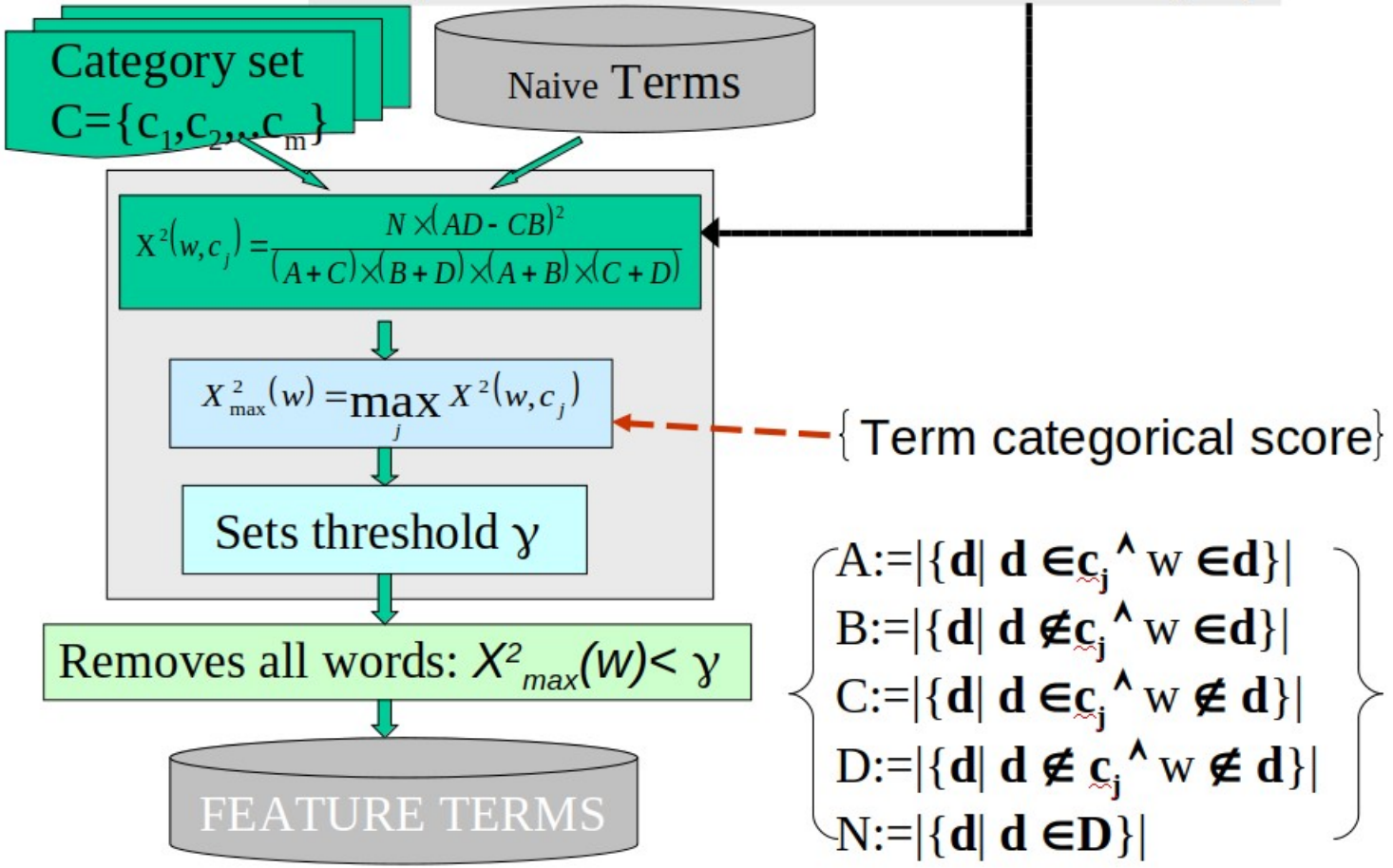
FOR FEATURE SELECTION

To use χ^2 for feature selection, we calculate χ^2 between each feature and the target, and select the desired number of features with the best χ^2 scores.

The intuition is that if a feature is independent to the target it is uninformative for classifying observations.

Dimension Reduction: X^2-statistic

- **Assumption**:a pre-defined category set for a training collection **D**
- **Goal**: Estimation independence between term and category



Features :

Happy : 23
Sad : 25
School : 3
...

	Pos	Neg
Feature present	A	B
Feature absent	C	D

Dakae Tech

Pour avoir un meilleur résultat de classification : essayer différentes combinaisons de settings, de prétraitement et de feature selection.

	little	cat	like	feel	name	much	happy	dog	mother	can	do	two	one
D1													
D2													
D3													

Dakae Tech

Variants of term frequency (TF) weight

weighting scheme	TF weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
log normalization	$1 + \log(f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

I have a little cat.
I like it !
My cat’s name is Lulu.
I feel he loves me very much.
I am happy…

	little	cat	like	feel	name	much	happy	dog	mother	can	do	two	one
D1	1	1	1	1	1	1	1	0	0	0	0	0	0
D2													
D3													

My dog’s name is Jeff.
My mother does not like dogs.
But what can I do ?!

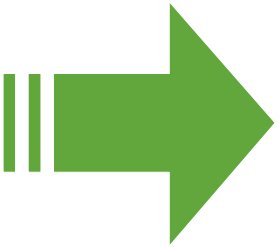
	little	cat	like	feel	name	much	happy	dog	mother	can	do	two	one
D1	1	1	1	1	1	1	1	0	0	0	0	0	0
D2	0	0	1	0	1	0	0	1	1	1	1	0	0
D3													

Cats do not love dogs !
But I have two kittens
and one dog.

	little	cat	like	feel	name	much	happy	dog	mother	can	do	two	one
D1	1	1	1	1	1	1	1	0	0	0	0	0	0
D2	0	0	1	0	1	0	0	1	1	1	1	0	0
D3	0	1	1	0	0	0	0	1	0	0	1	1	1

I have a little cat.
I like it !
My cat’s name is Lulu.
I feel he loves me very much.
I am happy…

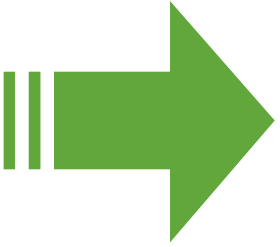
D1



1,1,1,1,1,1,1,0,0,0,0,0,0

My dog’s name is Jeff.
My mother does not like dogs.
But what can I do ?!

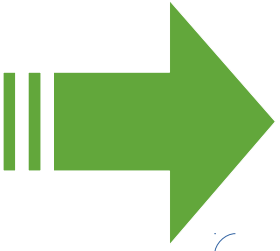
D2



0,0,1,0,1,0,0,1,1,1,1,0,0

Cats do not love dogs !
But I have two kittens
and one dog.

D3



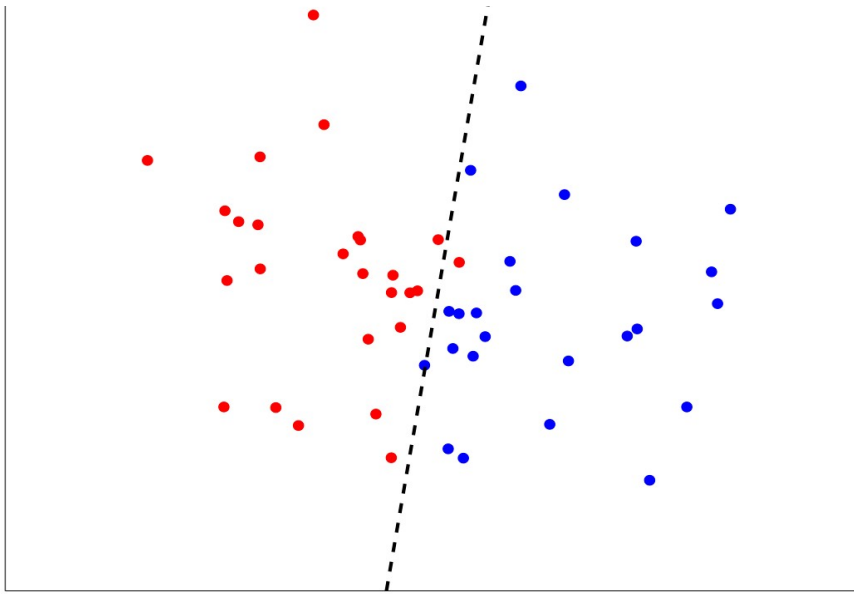
0,1,1,0,0,0,0,1,0,0,1,1,1

Représentation Binaire

1,1,1,1,1,1,1,0,0,0,0,0,0

0,0,1,0,1,0,0,1,1,1,1,0,0

0,1,1,0,0,0,0,1,0,0,1,1,1



Training Set

I am very happy !!
I found yesterday a little dog,
I liked it.
The poor has lost his mother

Comment classifier un nouveau document ??

I am very happy !!
I found yesterday a little dog,
I liked it.
The poor has lost his mother

D

	little	cat	like	feel	name	much	happy	dog	mother	can	do	two	one
D1	1	1	1	1	1	1	1	0	0	0	0	0	0
D2	0	0	1	0	1	0	0	1	1	1	1	0	0
D3	0	1	1	0	0	0	0	1	0	0	1	1	1
D	1	0	1	0	0	1	1	1	1	0	0	0	0

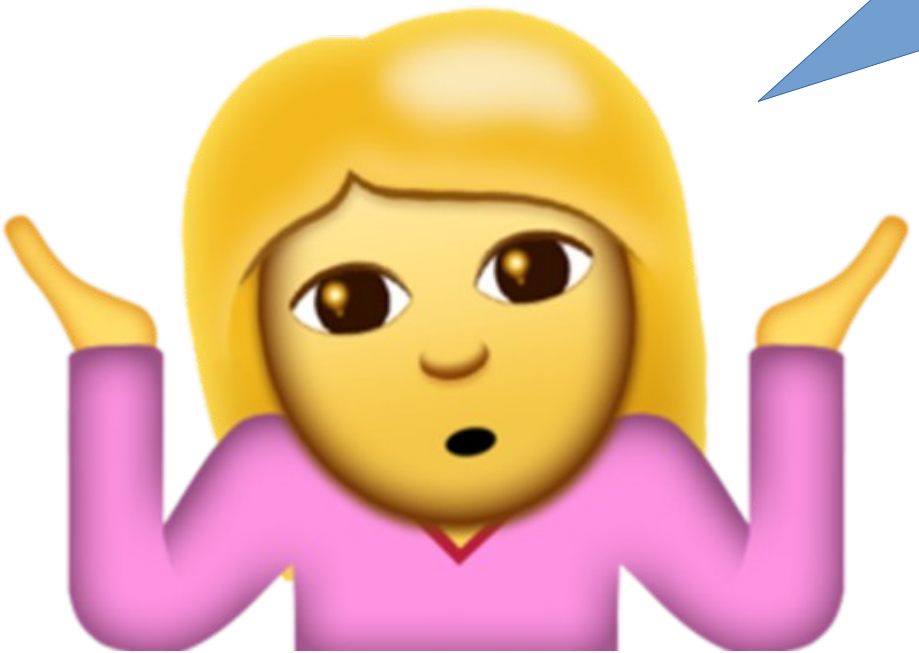
I am very happy !!
I found yesterday a little dog,
I liked it.
The poor has lost his mother

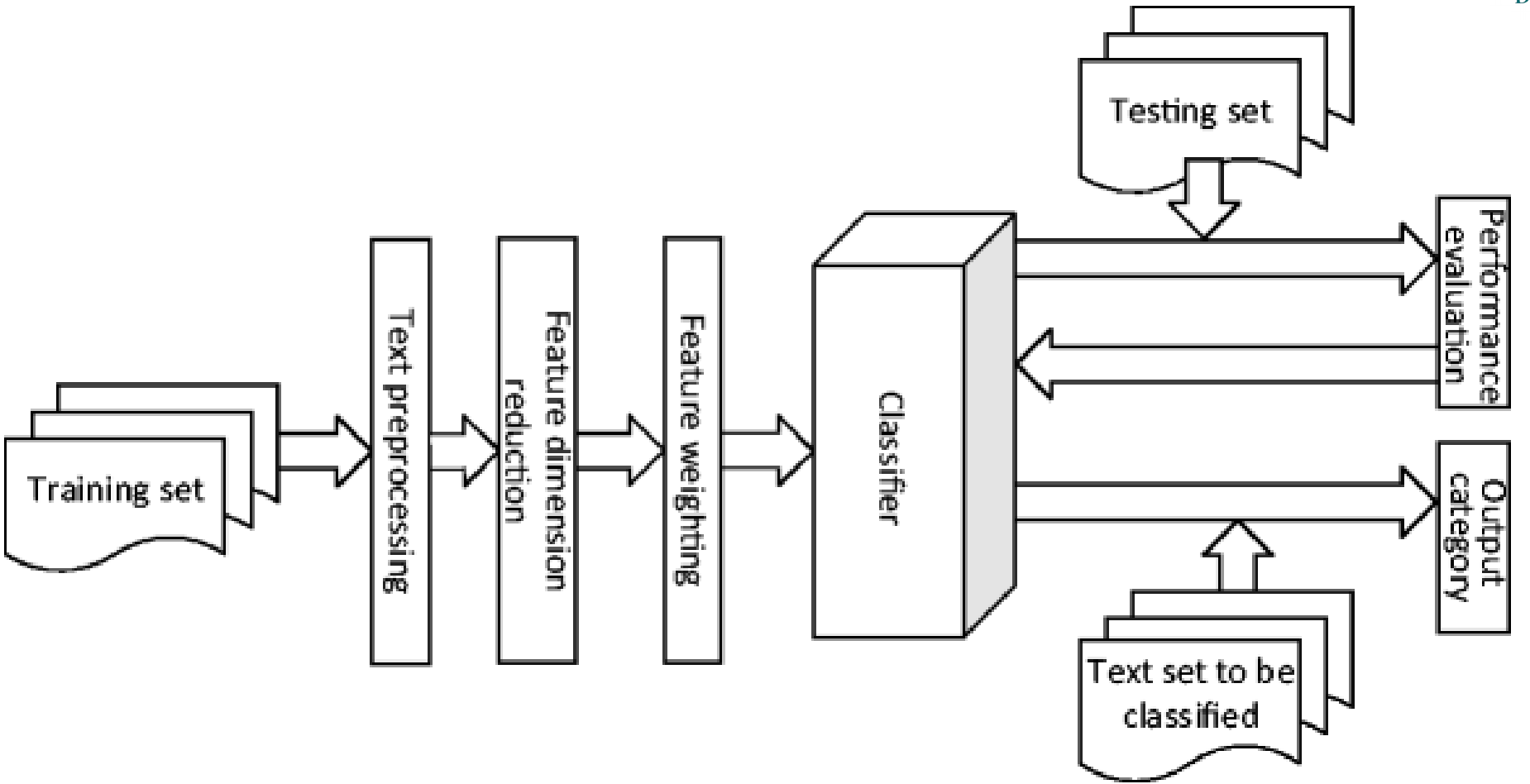
D

	little	cat	like	feel	name	much	happy	dog		two	one
D1	1	1	1	1	1	1	1			0	0
D2	0	0	1	0	1	0	0			0	0
D3	0	1	1	0	0	0	0			1	1
D	1	0	1	0	0	1	1		0	0	0

Matrix Sparsity !!!

Text Classification :
1. Matrix Sparsity
2. Curse of Dimensionality

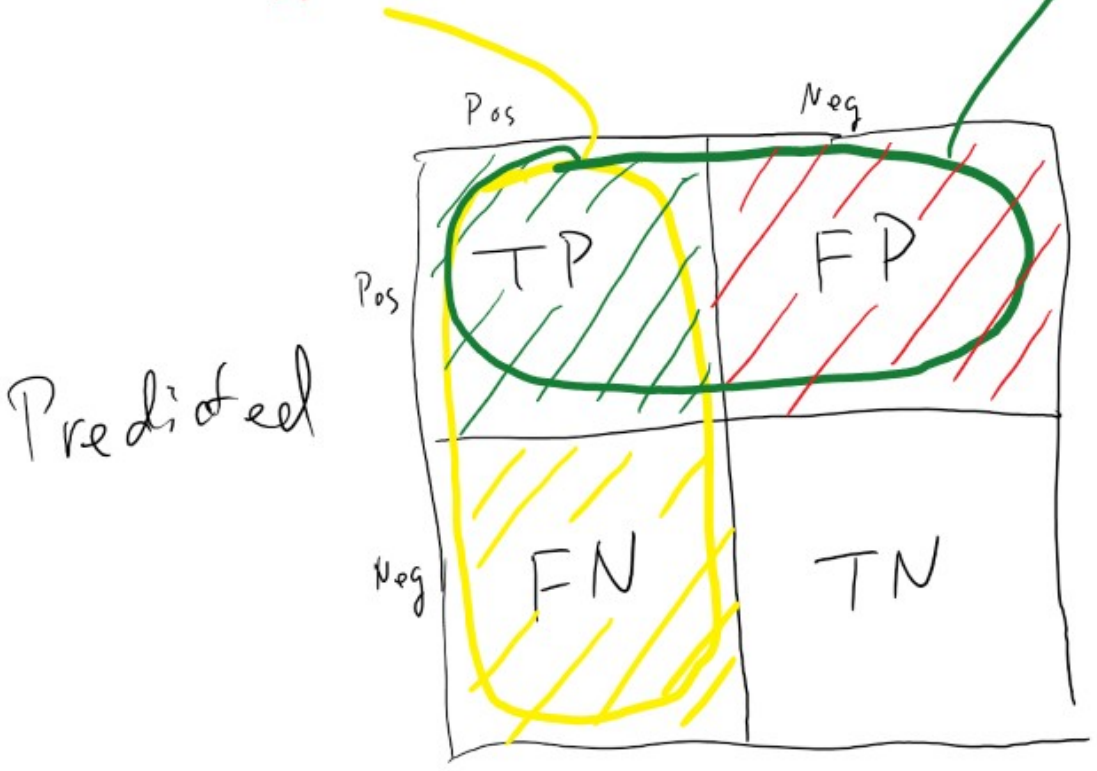



Training set
Text preprocessing
Feature dimension reduction
Feature weighting
Classifier
Testing set
Performance evaluation
Output category
Text set to be classified

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}}$$



Dakae Tech

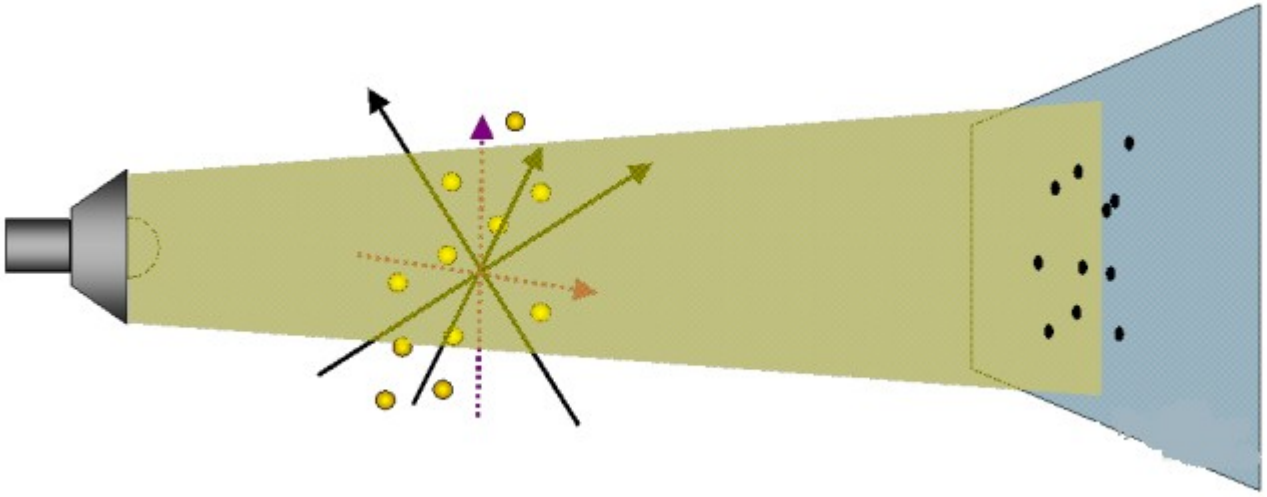
Analyse par Composants Principaux (ACP)

Dakae Tech

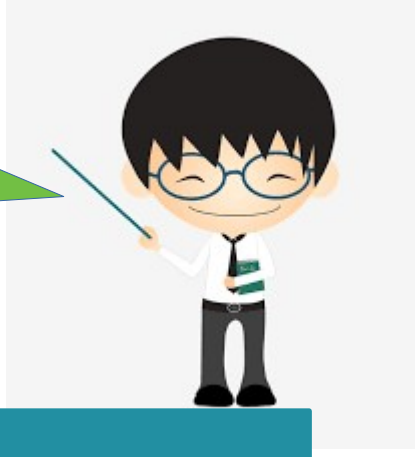
On passe d'un espace de dimension 3 à un espace de dimension 2

Quelle est la meilleure façon de photographier un dromadaire de la manière la plus similaire à la réalité ?

Le photographe cherche le meilleur angle pour prendre une photo qui reflète le maximum d'information...

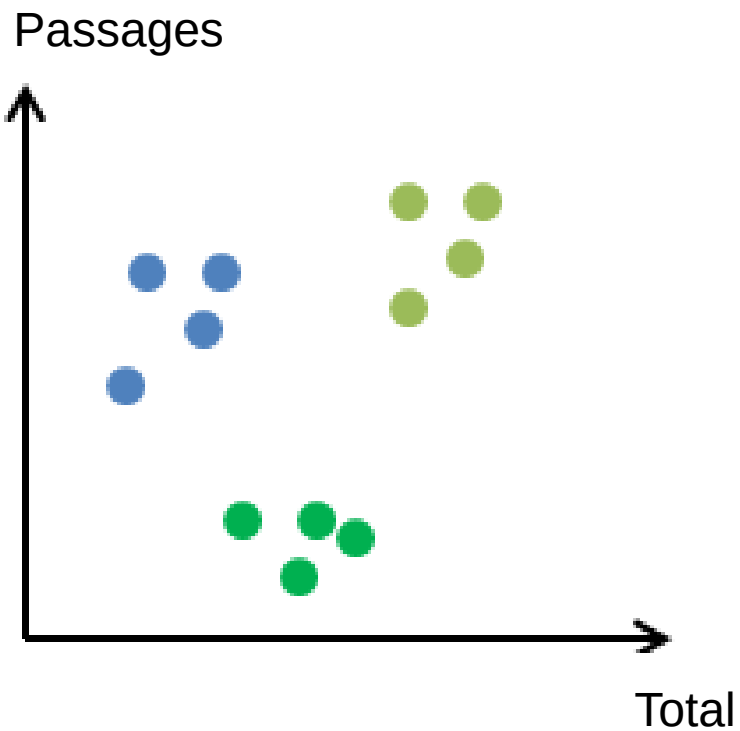


Il en est de même pour un nuage de points :
On cherche le meilleur espace de dimension inférieure au nombre de variables initiales pour y représenter nos données, sans perdre trop d'information.



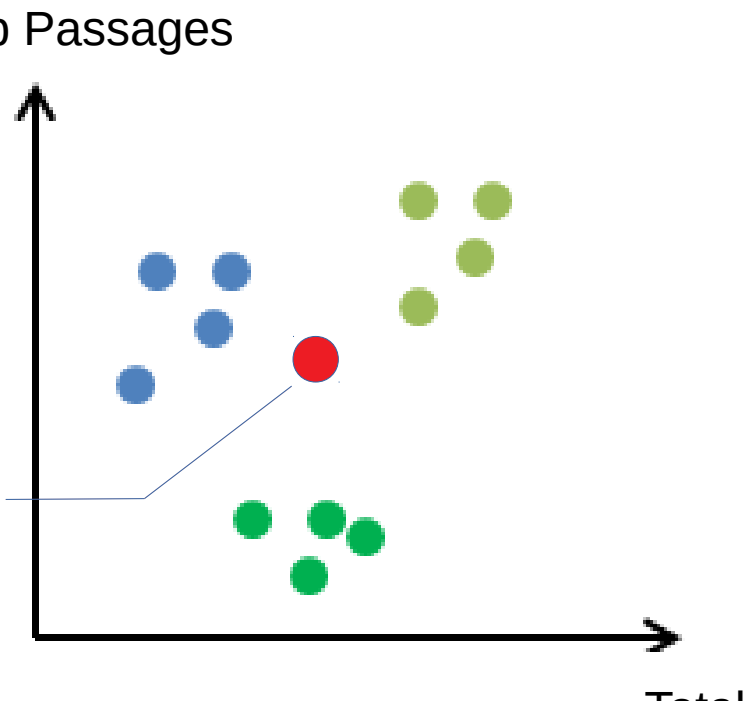
Dakae Tech

Prenons cet exemple : des clients représentés selon deux critères, le nombre de passages et le montant total acheté.
Supposons qu'on veut réduire la dimensionnalité, i.e représenter les clients dans un espace de 1 dimension.
Quelle serait la meilleure droite sachant qu'on veut garder le maximum d'information quant à la structure de notre data ?



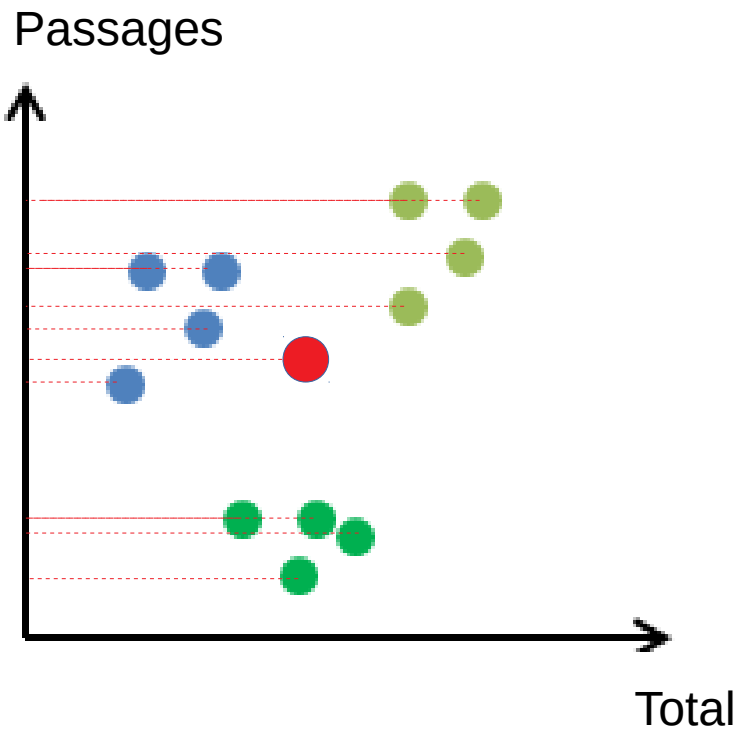
Peut-on se contenter de l'axe Nb Passages?
Qu'en est-il de l'axe Total ?
Pourquoi ?!

Centre de Gravité du nuage

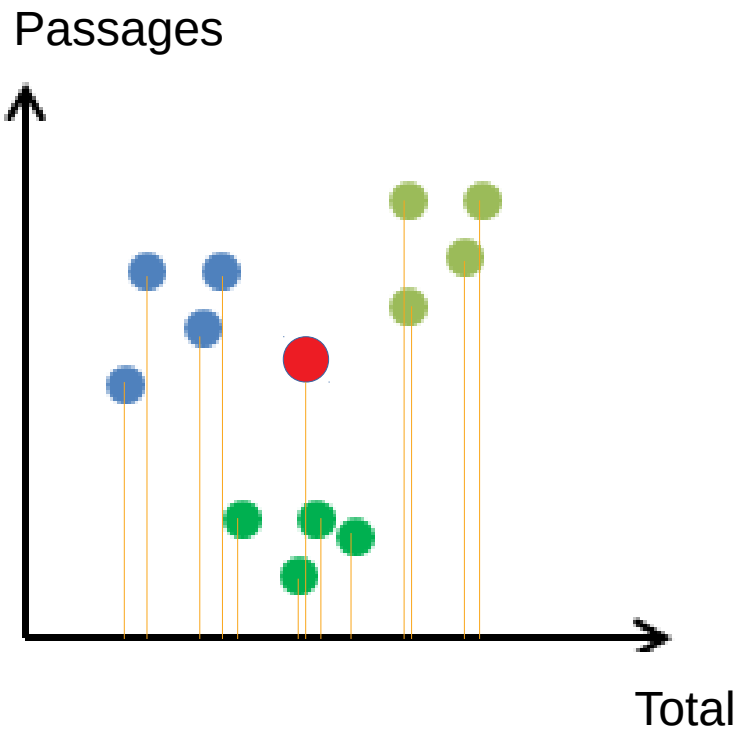


Dakae Tech

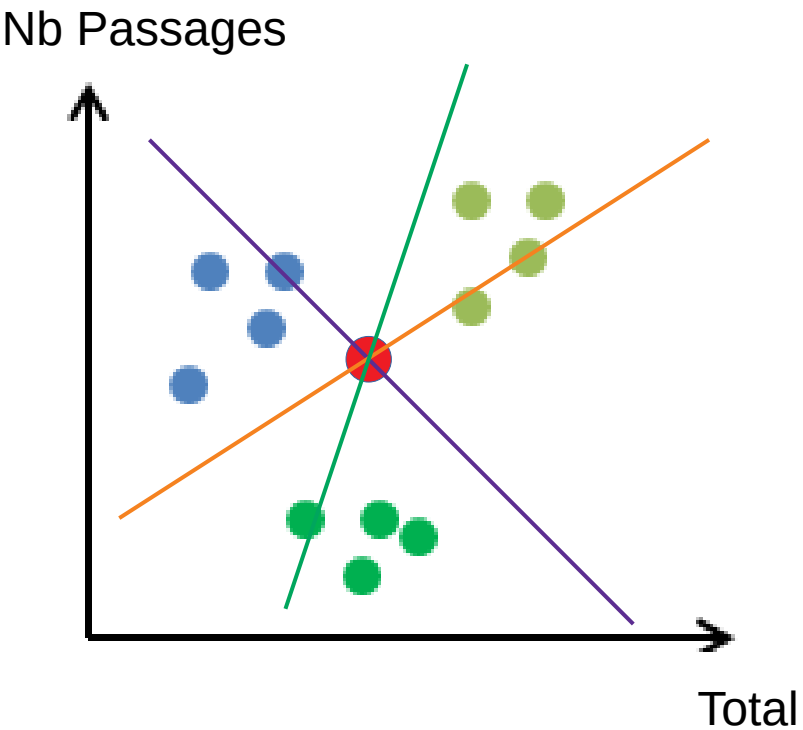
Peut-on se contenter de l'axe Nb Passages?
Qu'en est-il de l'axe Total ?
Pourquoi ?!



Peut-on se contenter de l’axe Nb Passages?
Qu’en est-il de l’axe Total ?
Pourquoi ?!
Est-ce que après la projection la structure reste la même ? Est-ce que les distances des points par rapport au centre d’inertie ne sont pas déformées ?

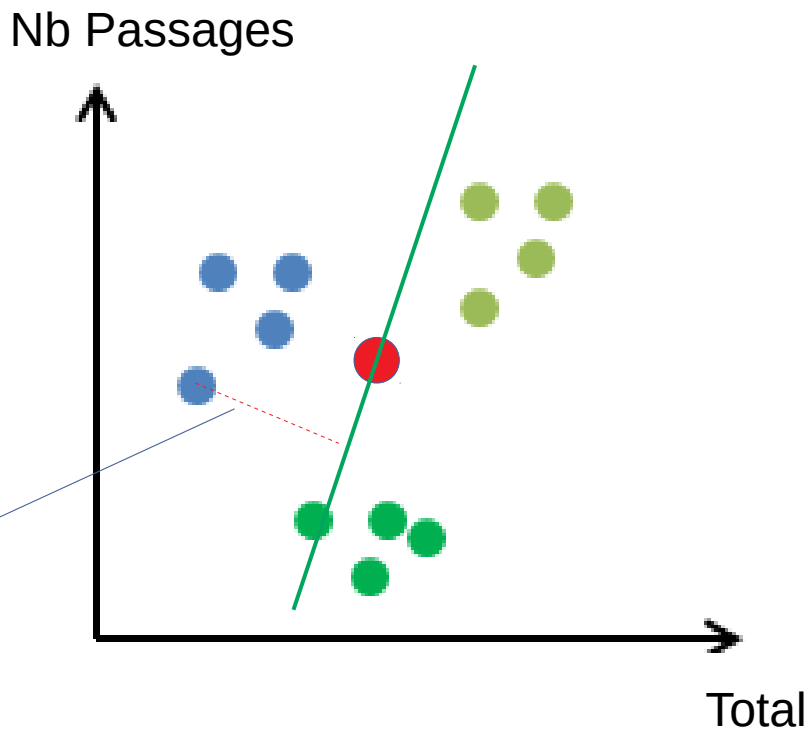


On va donc chercher la meilleure droite qui vérifie ces deux conditions :
- **Maximiser la variance** (distances des points projetés par rapport au centre de gravité)
- **Minimiser les résidus** (distances des points par rapport à la droite)



=> On dit qu'on pourra utiliser cette droite pour représenter notre nuage de point. On dit c'est un Facteur, **Composant Principal**.
Ce facteur est une combinaison linéaire des des deux variables.
=> On cherche à obtenir le résumé le plus pertinent des données initiales

Résidu



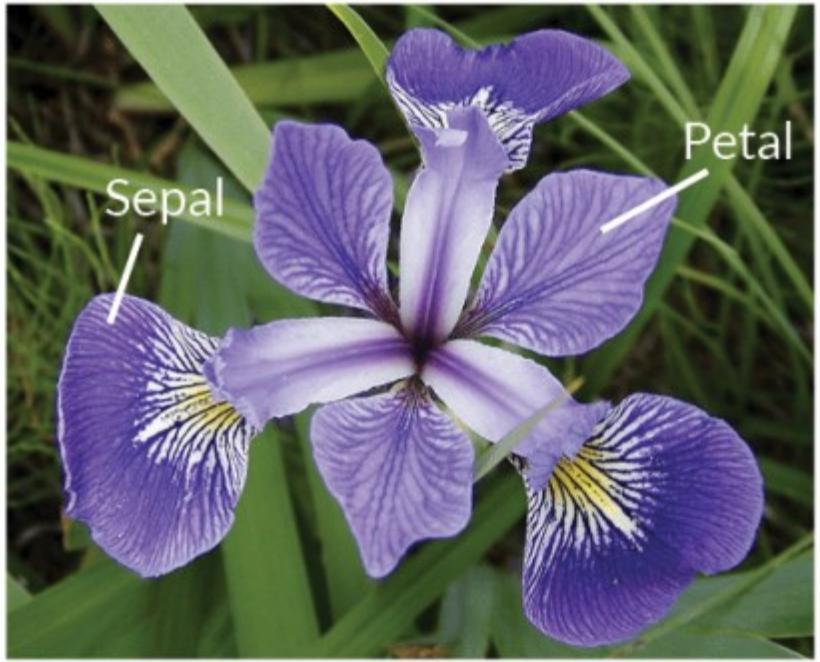
Dakae Tech

On veut réduire la dimensionnalité tout en gardant le maximum d'information.
Quel nombre choisir de facteurs à garder ?

Comment extraire ces Composants principaux ?

Au cas où on a 3 variables ou plus ?

Exemple d'application : IRIS Data

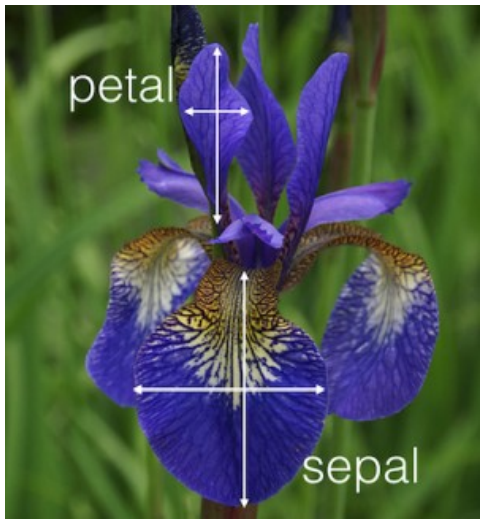


Iris Versicolor

Iris Setosa

Iris Virginica

Exemple d'application : IRIS Data

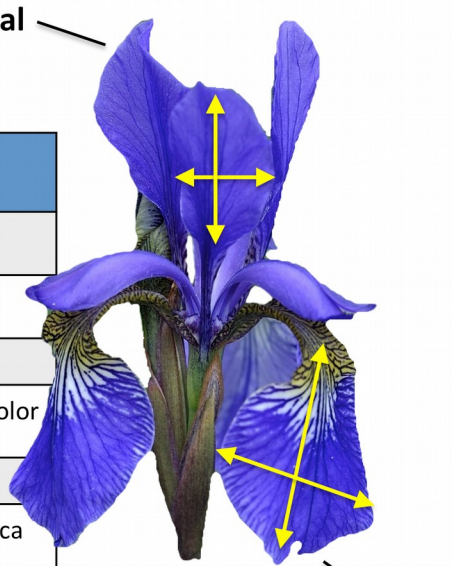


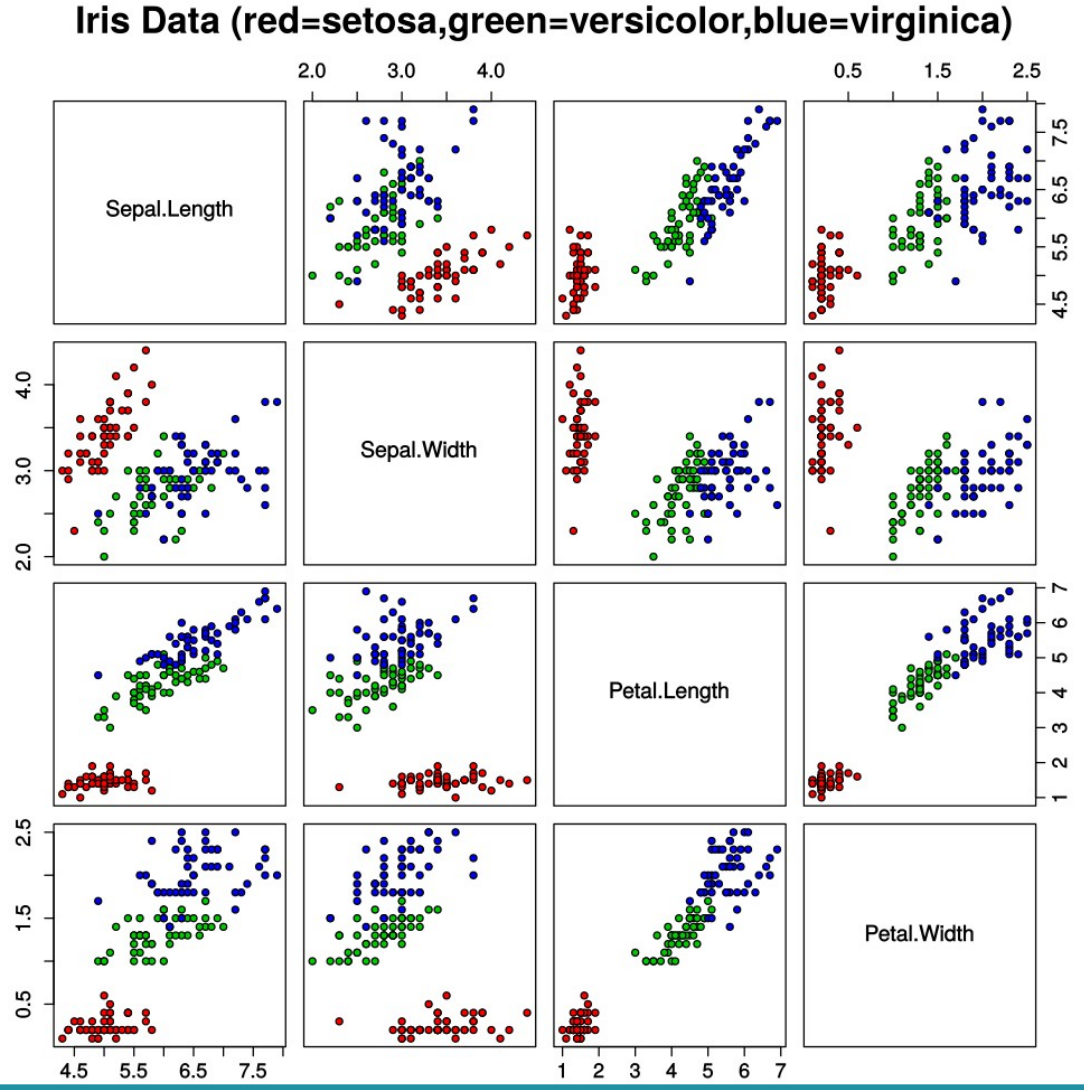
Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

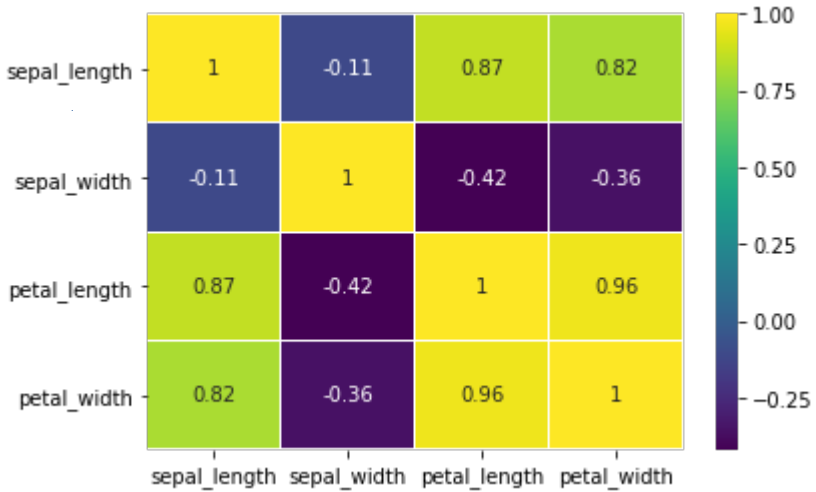
Class labels
(targets)

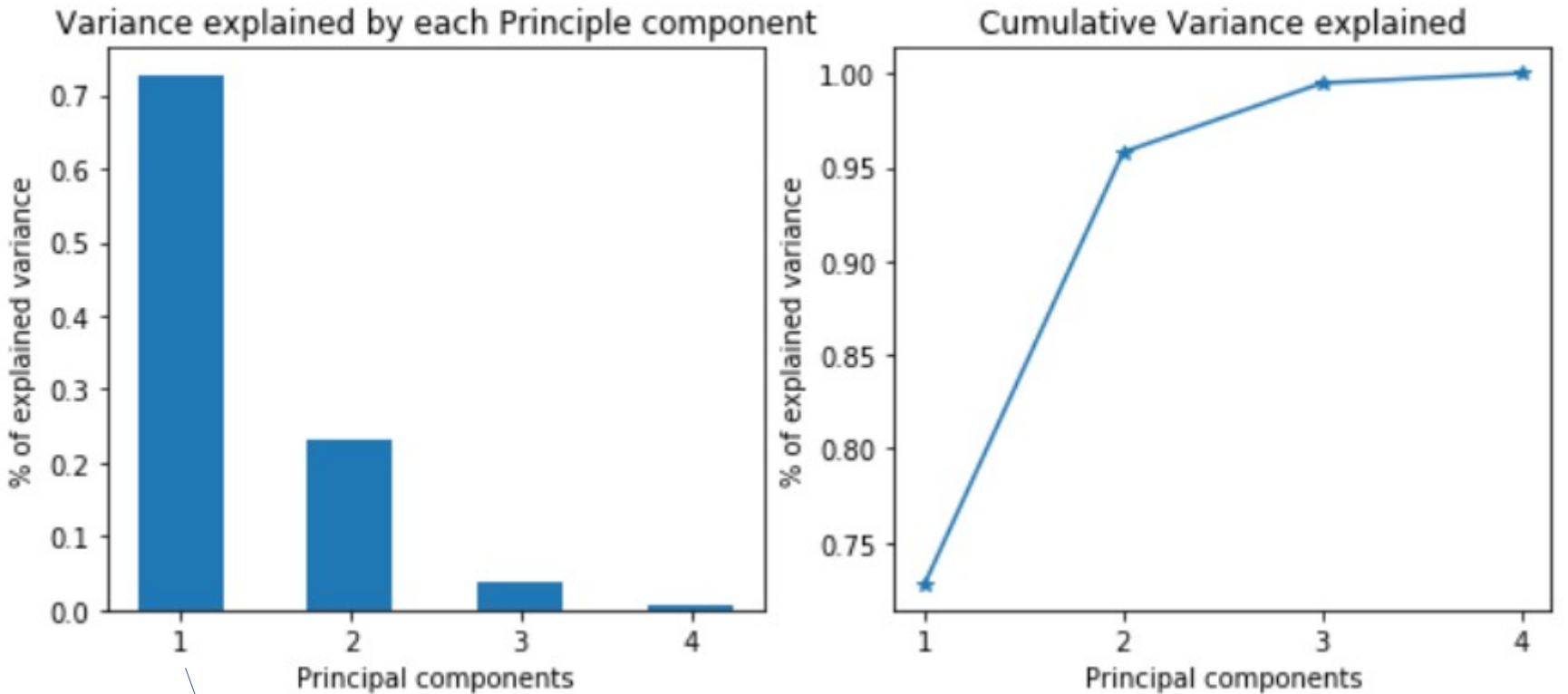



Iris Data (red=setosa,green=versicolor,blue=virginica)
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width

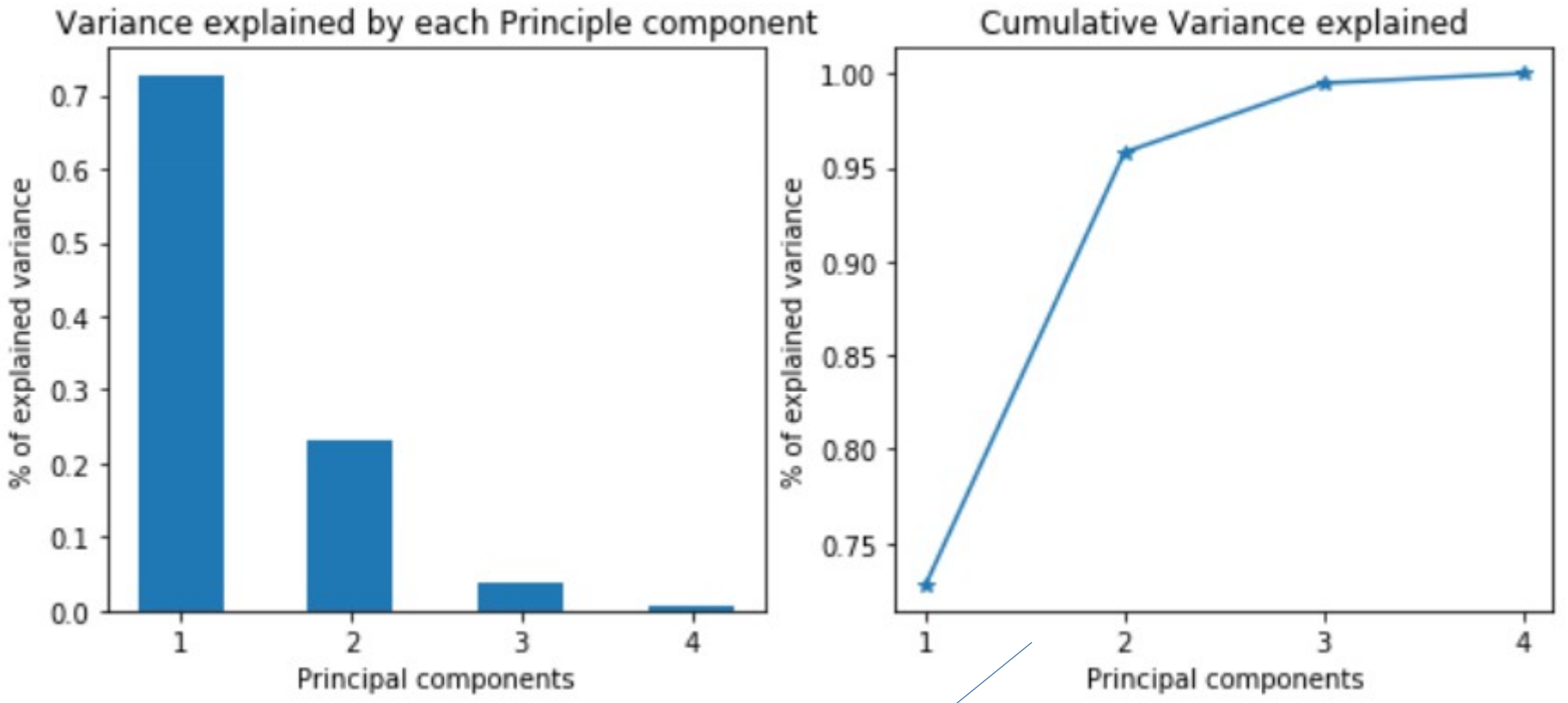
L’idée est de calculer la matrice de corrélation pour toutes les variables. Après on calcule les vecteurs propres et leurs valeurs propres Associées de cette matrice.

Le vecteur propre avec la plus grande valeur propre est dit Facteur 1. Sa valeur propre fait référence à la variance de la data projetée sur le facteur 1.

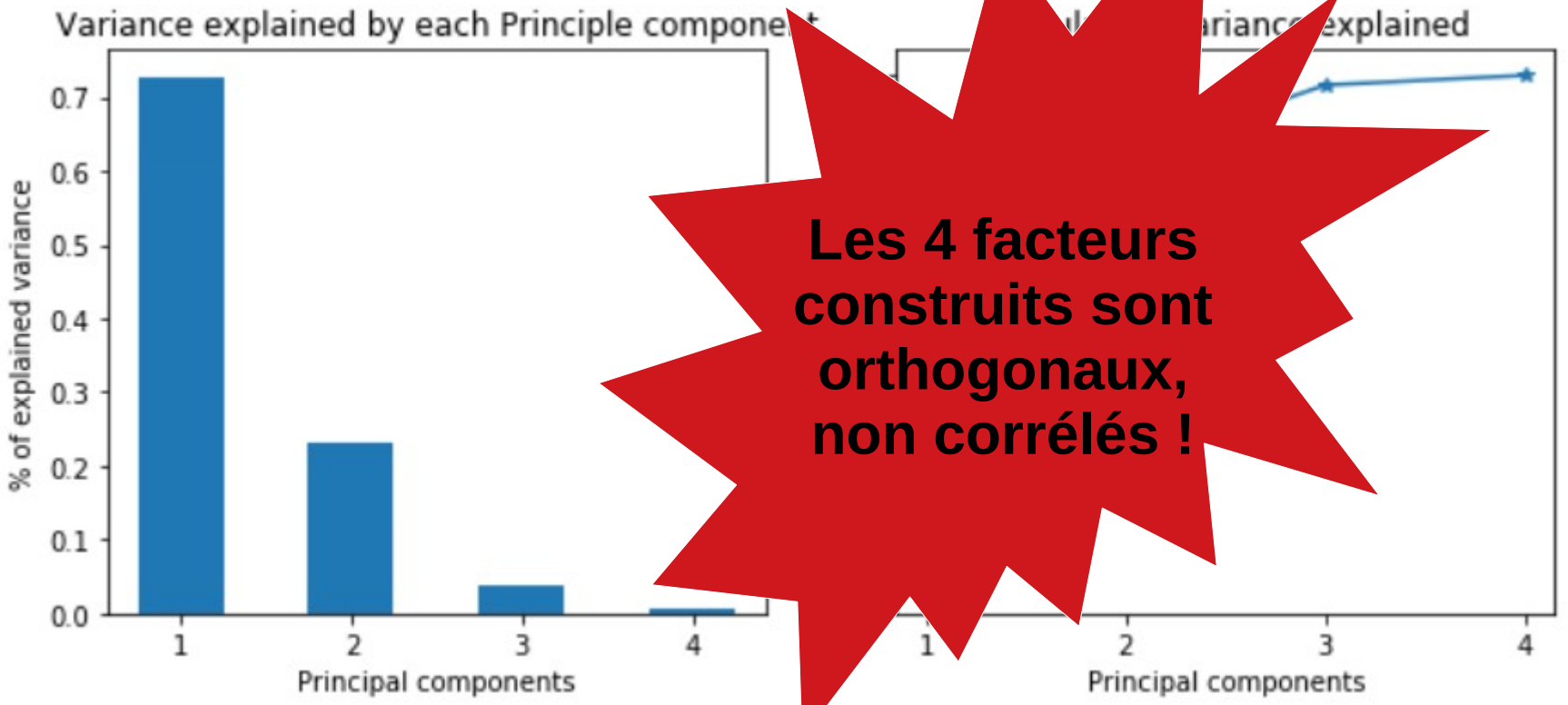



Variance explained by each Principle component
% of explained variance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
4
Principal components
Cumulative Variance explained
% of explained variance
0.75
0.80
0.85
0.90
0.95
1.00
1
2
3
4
Principal components

Facteur 1



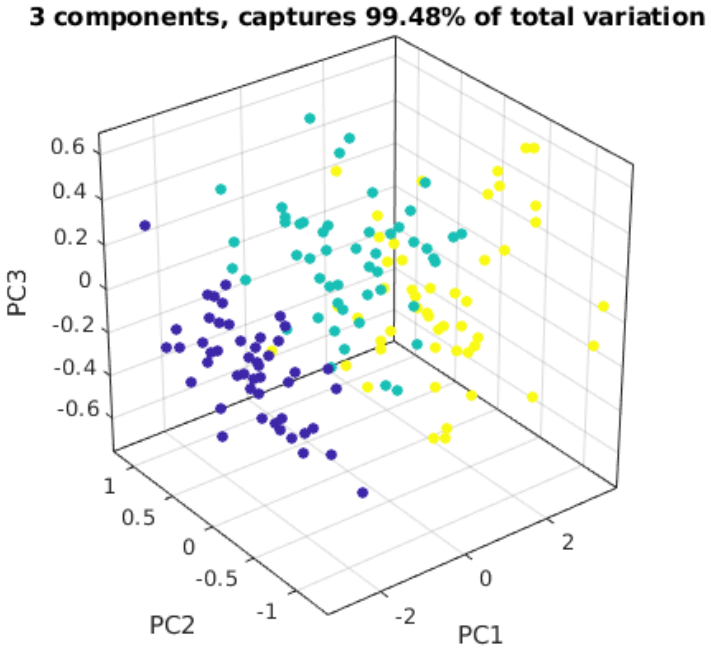
On se base sur ce graphe (Elbow) pour déterminer le nombre de facteurs à choisir, on s'arrête là où le gain en variance n'est pas aussi important.



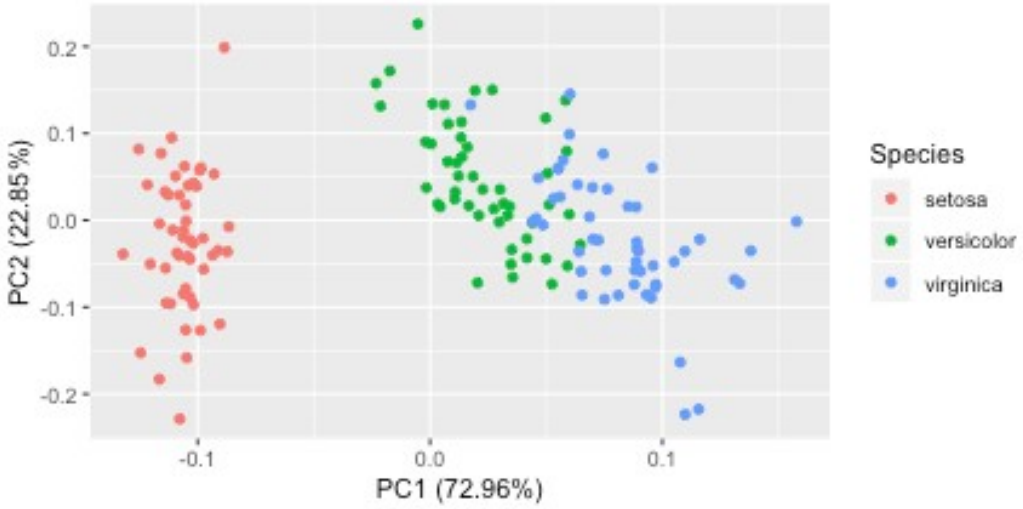
Les 4 facteurs construits sont orthogonaux, non corrélés !

	Sepal Length	Sepal Width	Petal Length	Petal Width
PC1	[0.52843794,	-0.23201227,	0.58394827,	0.5709011]
PC2	[0.3554837 ,	0.93369239,	0.00795684,	0.04226763]
PC3	[-0.72859925,	0.24803092,	0.15153118,	0.62021133]
PC4	[0.25204725,	-0.11344378,	-0.79748317,	0.53630521]

Chaque facteur est une combinaison linéaire des variables initiales. A partir de ces formules, on pourra donner des interprétations pour chacun des facteurs.



On peut visualiser notre data sur un espace de 3 ou 2 dimensions...



ÉTAPES DE l'ACP

Dakae Tech

1- Centrer-réduire les variables
2- Calculer matrice de corrélation
3- Calculer valeurs et vecteurs propres de la matrice de corrélation
4- Ne garder que les valeurs propres à grandes valeurs
5- Projeter les données dans ce nouvel espace propre