EPSI l'École d'ingénierie informatique

Big Data & Traitement par l'IA III

Jean-Claude Houbart

houbart.jc@gmail.com

EPSI l'École d'ingénierie informatique

Application au NLP

Jean-Claude Houbart

houbart.jc@gmail.com

Word2Vec

Tomas Mikolov Google, puis Facebook)

“Distributed Representations of Words and Phrases and their Compositionality”

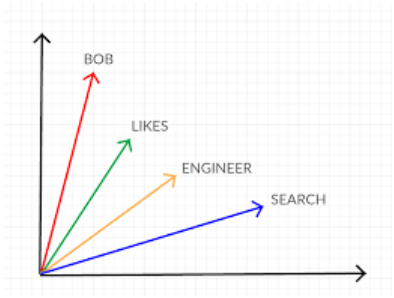
Mikolov et al. 2013

“Vous connaîtrez un mot par sa compagnie”
(J. R. Firth 1957: 11)

XXX est une école d’informatique qui propose des programmes de Bac à Bac+5.
XXX accompagne le développement des compétences informatiques.
XXX entretient des liens forts avec de nombreuses entreprises régionales et nationales.
XXX a obtenu en 2019 la qualification OPQF pour tous ses campus.

Transformation en vecteur

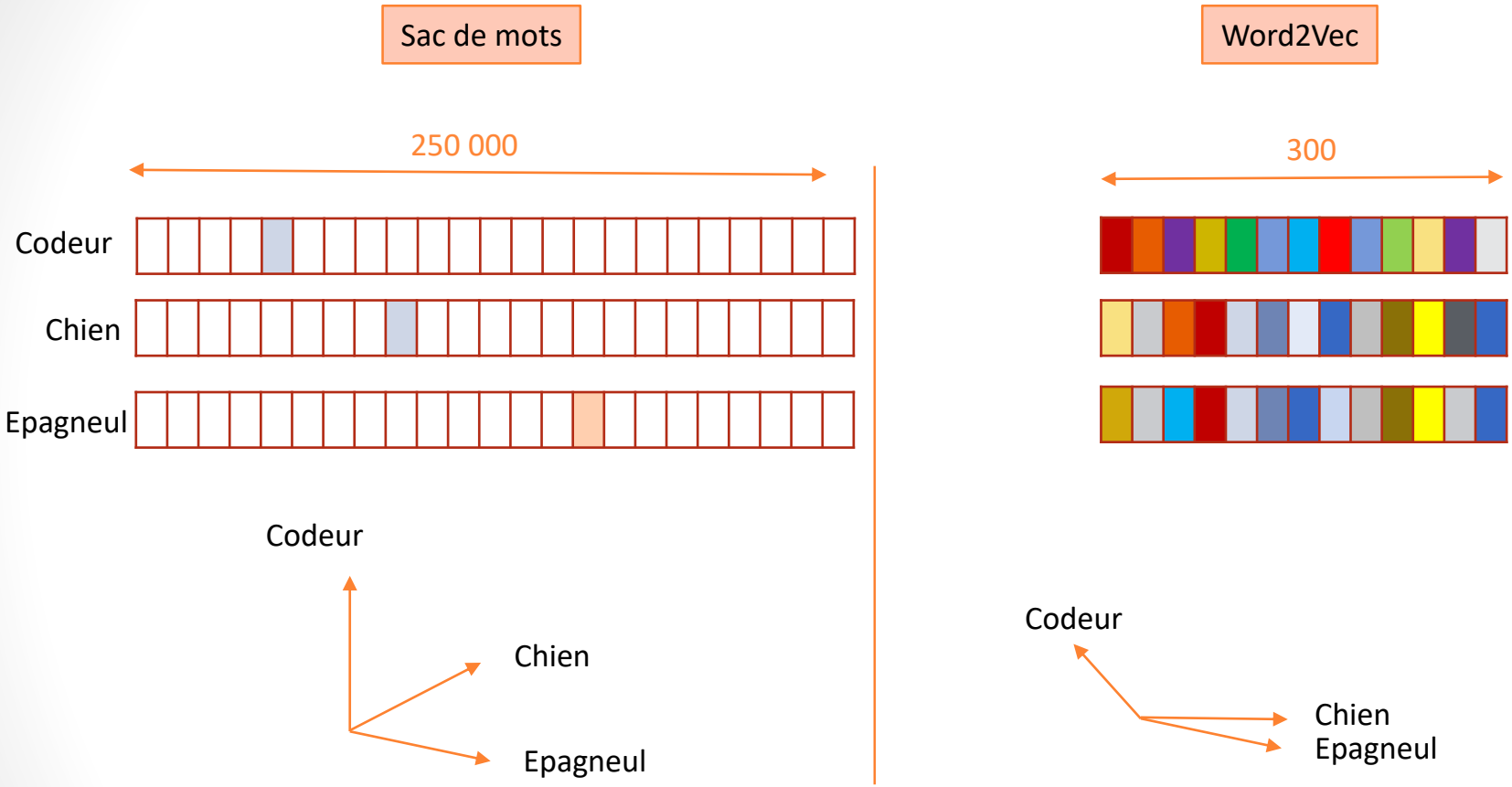
Un milliard de mots.
Vocabulaire de taille 692 000



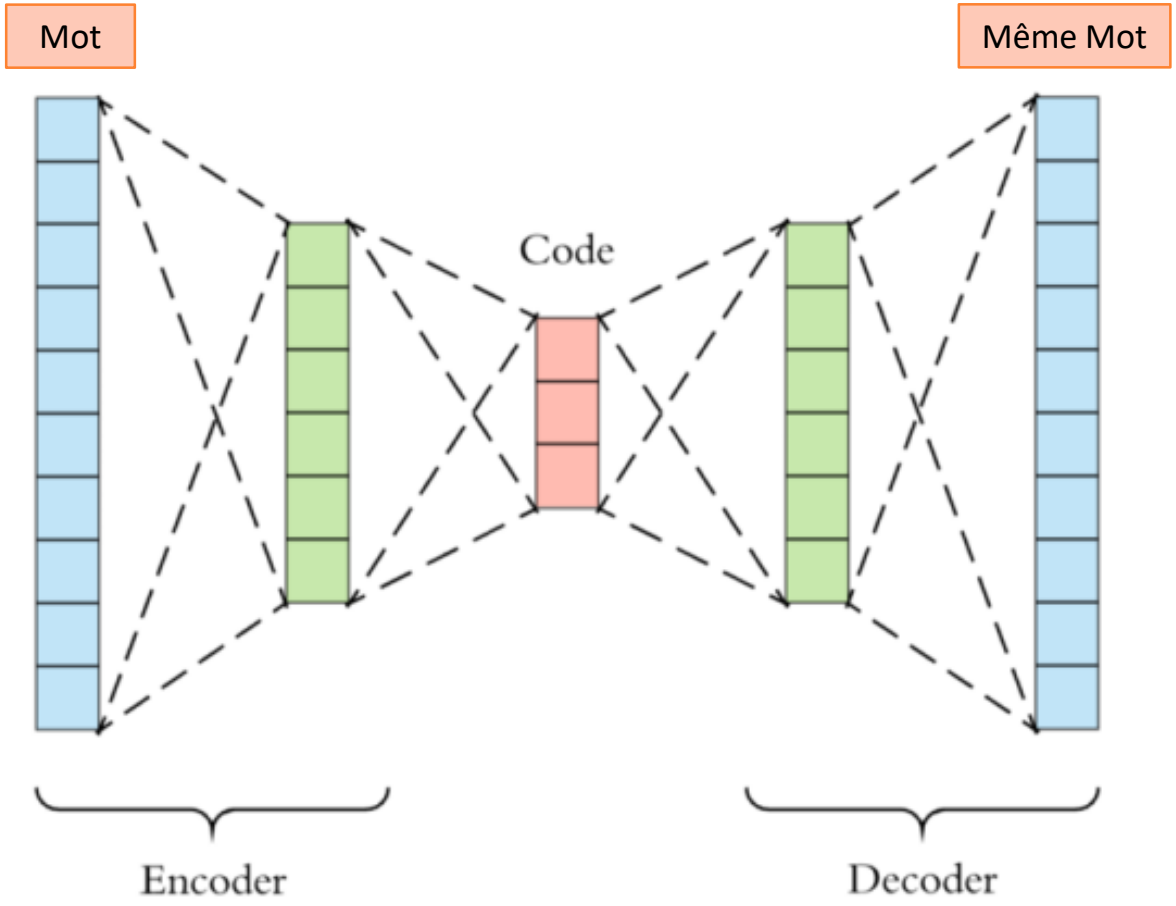
Espace Vectoriel.
(En général 300 Dimensions)

Objectif : 2 mots qui ont souvent les mêmes voisins doivent être géométriquement proches.

Réduction de la taille d'encodage

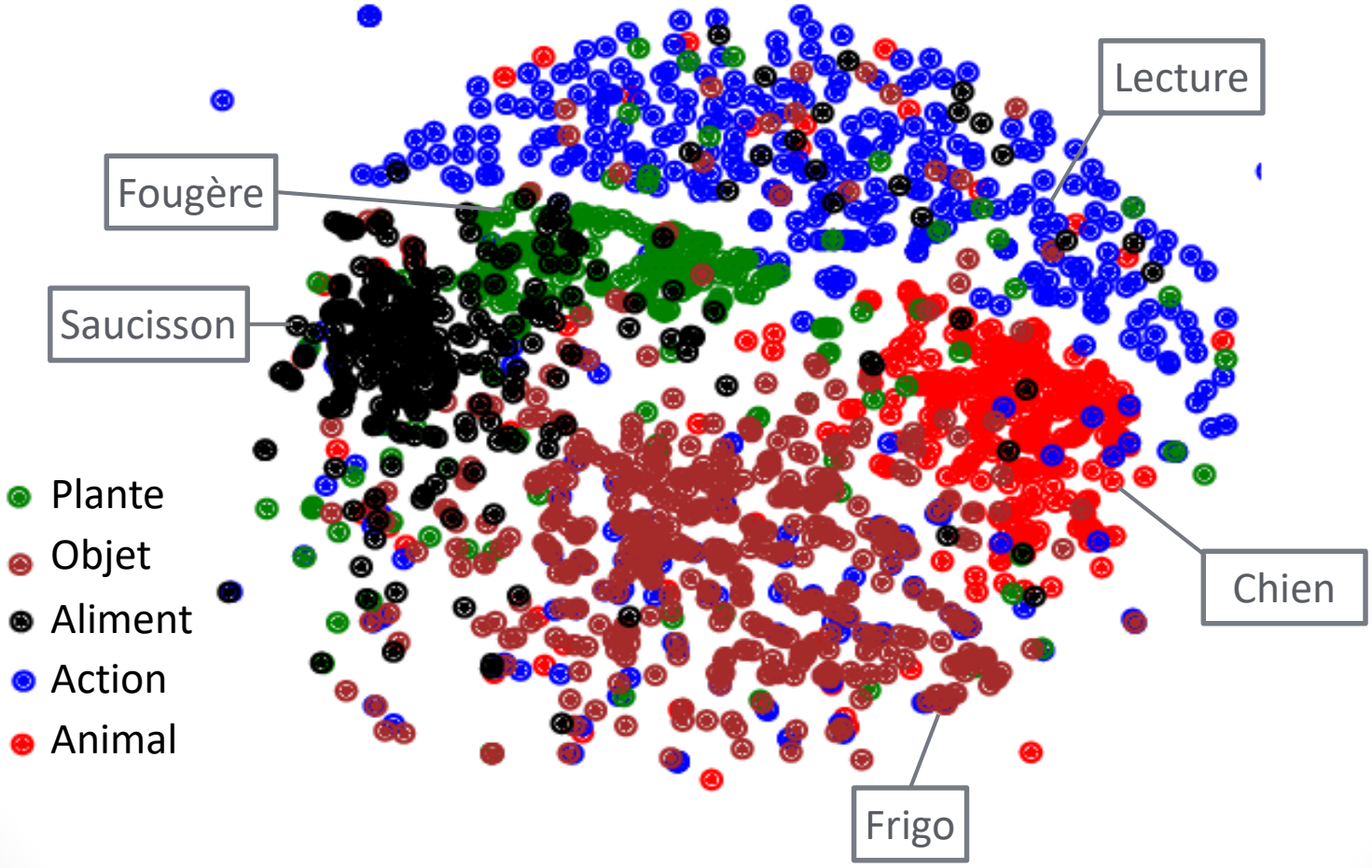


Auto-Encoder



Visualisation des mots

Roi – Homme + Femme = ?

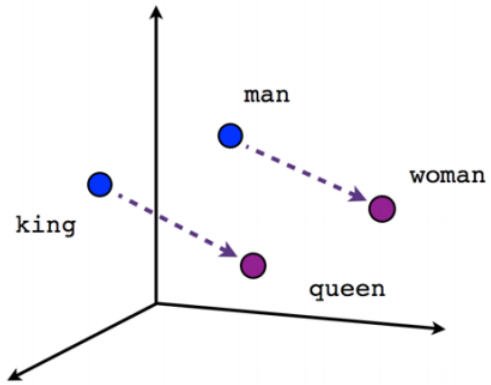


Opération sur les mots

Roi – Homme + Femme = ?

Paris – France + UK = ?

Mangeant – Manger + Nager = ?



EPSI l'École d'ingénierie informatique

Algorithme Incrémental (online)

Jean-Claude Houbart

houbart.jc@gmail.com

Définition d'un algorithme « en ligne »

Traite les données d'entrée morceaux par morceaux de manière séquentielle.

Les données sont traitées dans l'ordre d'arrivée par l'algorithme, sans qu'elles soient toutes disponibles dès le départ.

Exemples

Prévision de cours boursiers. L'apprentissage peut être réalisés au fil de l'évolution des cours.

Détection des spams. Chaque retour des utilisateurs est intégré au modèle.

Système de recommandation. Chaque note ou achat groupé alimente le modèle.

Exemples → Apprentissage → h(x)
Exemples → Apprentissage → h(x)
……
Exemples → Apprentissage → h(x)

Stock prices

Temps

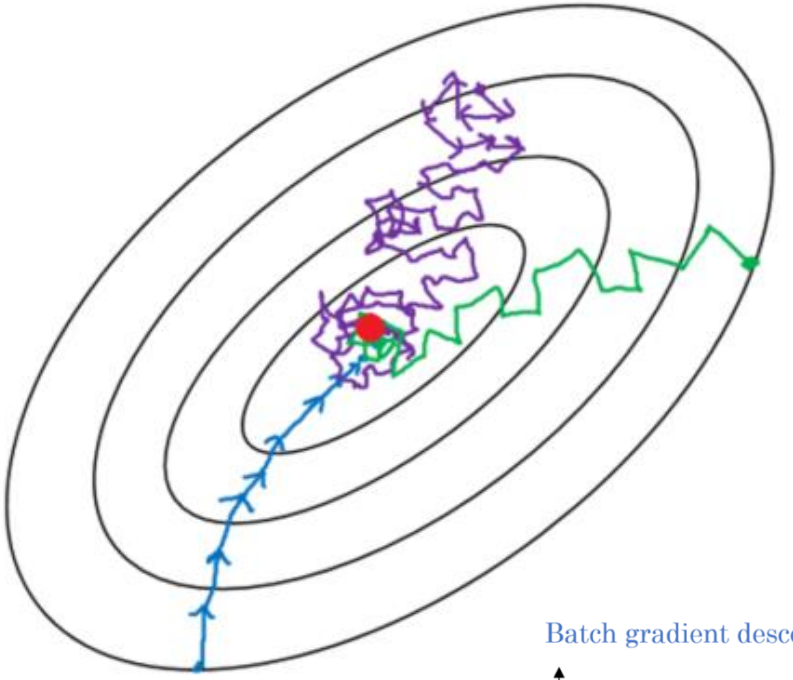
Data Processing

Prediction

Receiving Truth

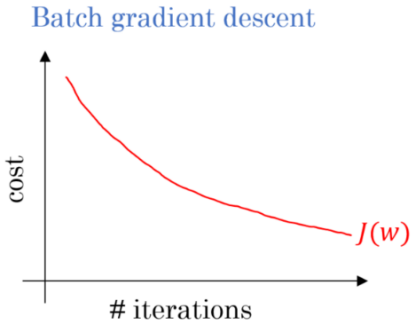
Update hypothesis h(x)

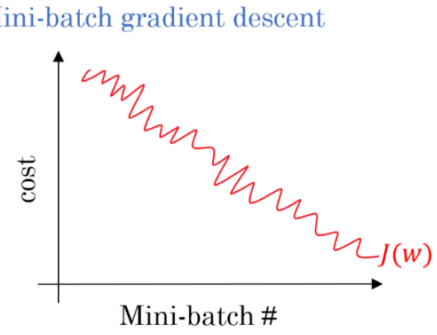
Un seul exemple à chaque itération
(Stochastic Gradient Descent)



Une partie des exemples à chaque itération
(Mini-Batch Gradient Descent)

Tous les exemples à chaque itération
(Batch Gradient Descent)





sklearn.cluster.MiniBatchKMeans

fit(self, X[, y, sample_weight])	Compute the centroids on X by chunking it into mini-batches.
fit_predict(self, X[, y, sample_weight])	Compute cluster centers and predict cluster index for each sample.
fit_transform(self, X[, y, sample_weight])	Compute clustering and transform X to cluster-distance space.
get_params(self[, deep])	Get parameters for this estimator.
partial_fit(self, X[, y, sample_weight])	Update k means estimate on a single mini-batch X.
predict(self, X[, sample_weight])	Predict the closest cluster each sample in X belongs to.
score(self, X[, y, sample_weight])	Opposite of the value of X on the K-means objective.
set_params(self, **params)	Set the parameters of this estimator.
transform(self, X)	Transform X to a cluster-distance space.

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html#sklearn.cluster.MiniBatchKMeans

sklearn.decomposition.IncrementalPCA

fit(self, X[, y])	Fit the model with X, using minibatches of size batch_size.
fit_transform(self, X[, y])	Fit to data, then transform it.
get_covariance(self)	Compute data covariance with the generative model.
get_params(self[, deep])	Get parameters for this estimator.
get_precision(self)	Compute data precision matrix with the generative model.
inverse_transform(self, X)	Transform data back to its original space.
partial_fit(self, X[, y, check_input])	Incremental fit with X.
set_params(self, **params)	Set the parameters of this estimator.
transform(self, X)	Apply dimensionality reduction to X.

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.IncrementalPCA.html#sklearn.decomposition.IncrementalPCA