Word embedding

Traitement automatique du langage naturel

Université de Bordeaux / L3 Informatique / Épistémologie

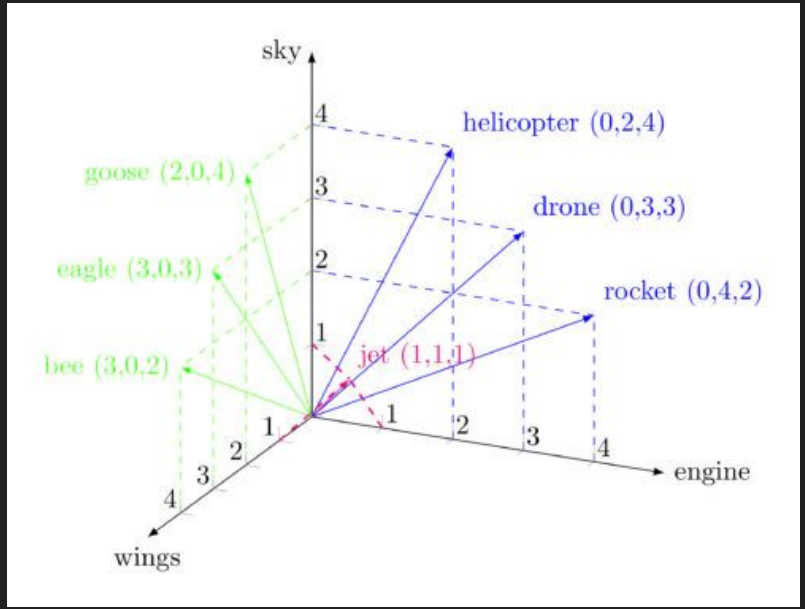
Václav Gregor, 2024

C'est quoi ?

Coder la sémantique des mots.

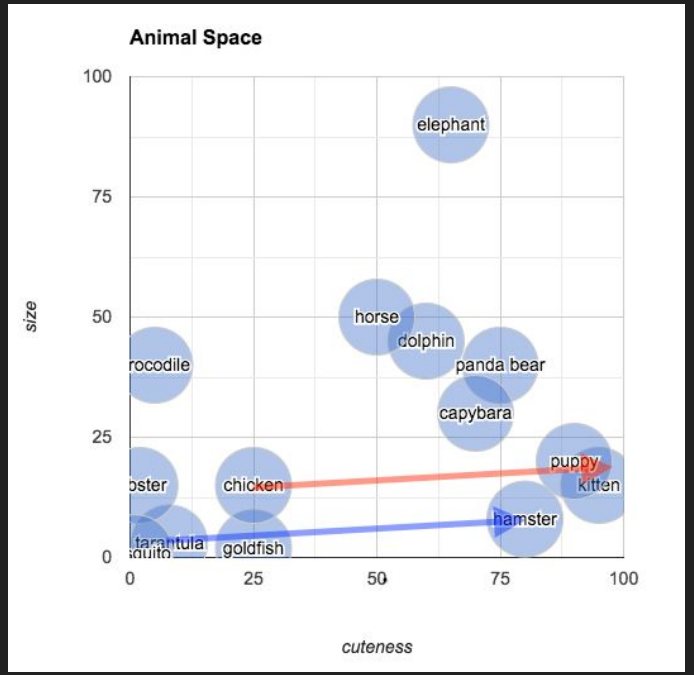
Moteurs de recherche, analyse de sentiment, similarité entre documents, analyse sémantique.

Apprentissage automatique.



Principe sur un exemple simple

	cuteness (0–100)	size (0–100)
kitten	95	15
hamster	80	8
tarantula	8	3
puppy	90	20
crocodile	5	40
dolphin	60	45
panda bear	75	40
lobster	2	15
capybara	70	30
elephant	65	90
mosquito	1	1
goldfish	25	2
horse	50	50
chicken	25	15



Sémantique distributionnelle

Pour un texte de taille arbitraire ?

Un mot est défini par ses voisins.

Il a fait bien **froid** hier.
Par contre, il va faire hyper **chaud** aujourd'hui.
Il fera très **beau** demain !
Est-ce qu'il va faire très **froid** mardi ?

Sémantique distributionnelle

It was the best of times, it was the worst of times.

Réduction de la dimensionnalité.

Count-based.

g	DÉBUT _ was	it _ the	was _ best	the _ of	best _ times	of _ it	times _ was	was _ worst	worst _ times	of _ FIN
it	1	0	0	0	0	0	1	0	0	0
was	0	2	0	0	0	0	0	0	0	0
the	0	0	1	0	0	0	0	1	0	0
best	0	0	0	1	0	0	0	0	0	0
of	0	0	0	0	1	0	0	0	1	0
times	0	0	0	1	0	0	0	0	0	1
worst	0	0	0	1	0	0	0	0	0	0

Word2Vec

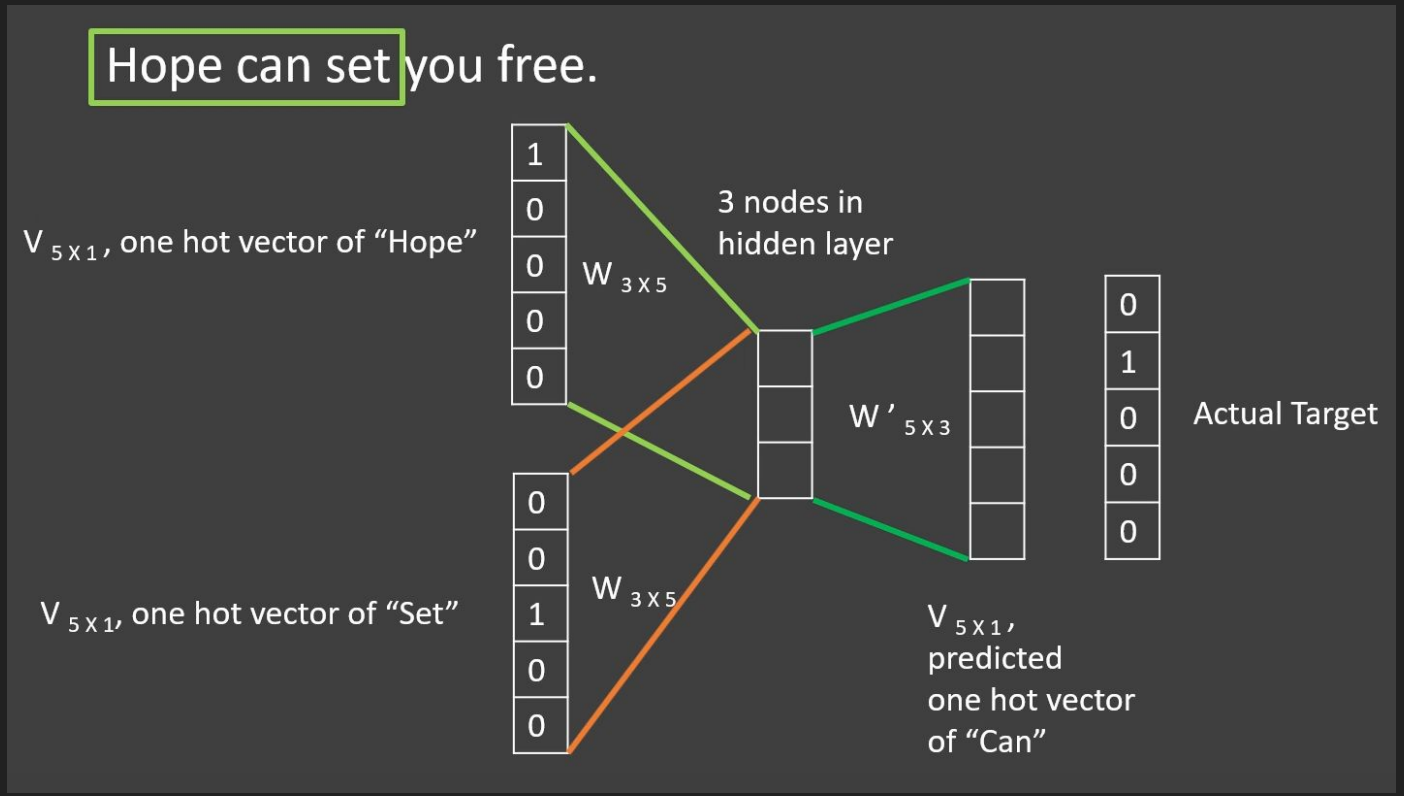
2013

Prediction-based.

2 modèles légers et puissants : CBOW et Skip-gram.

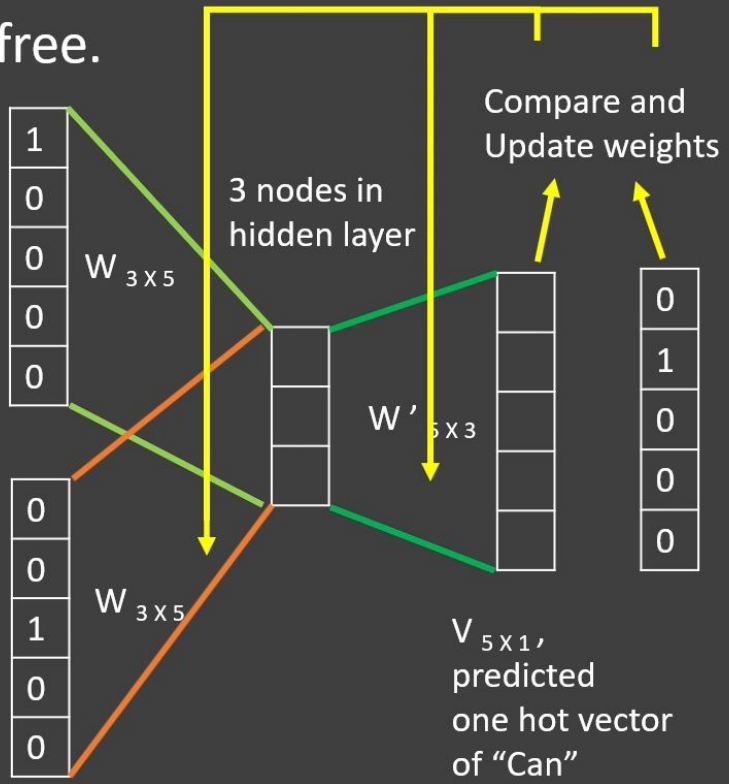
Utiles pour l'apprentissage profond.

CBOW


Hope can set you free.
V 5 X 1 , one hot vector of “Hope”
1
0
0
0
0
W 3 X 5
3 nodes in hidden layer
W ’ 5 X 3
0
1
0
0
0
Actual Target
V 5 X 1, one hot vector of “Set”
0
0
1
0
0
W 3 X 5
V 5 X 1 , predicted one hot vector of “Can”

CBOW

Hope can set you free.



V_{5X1}, one hot vector of "Hope"

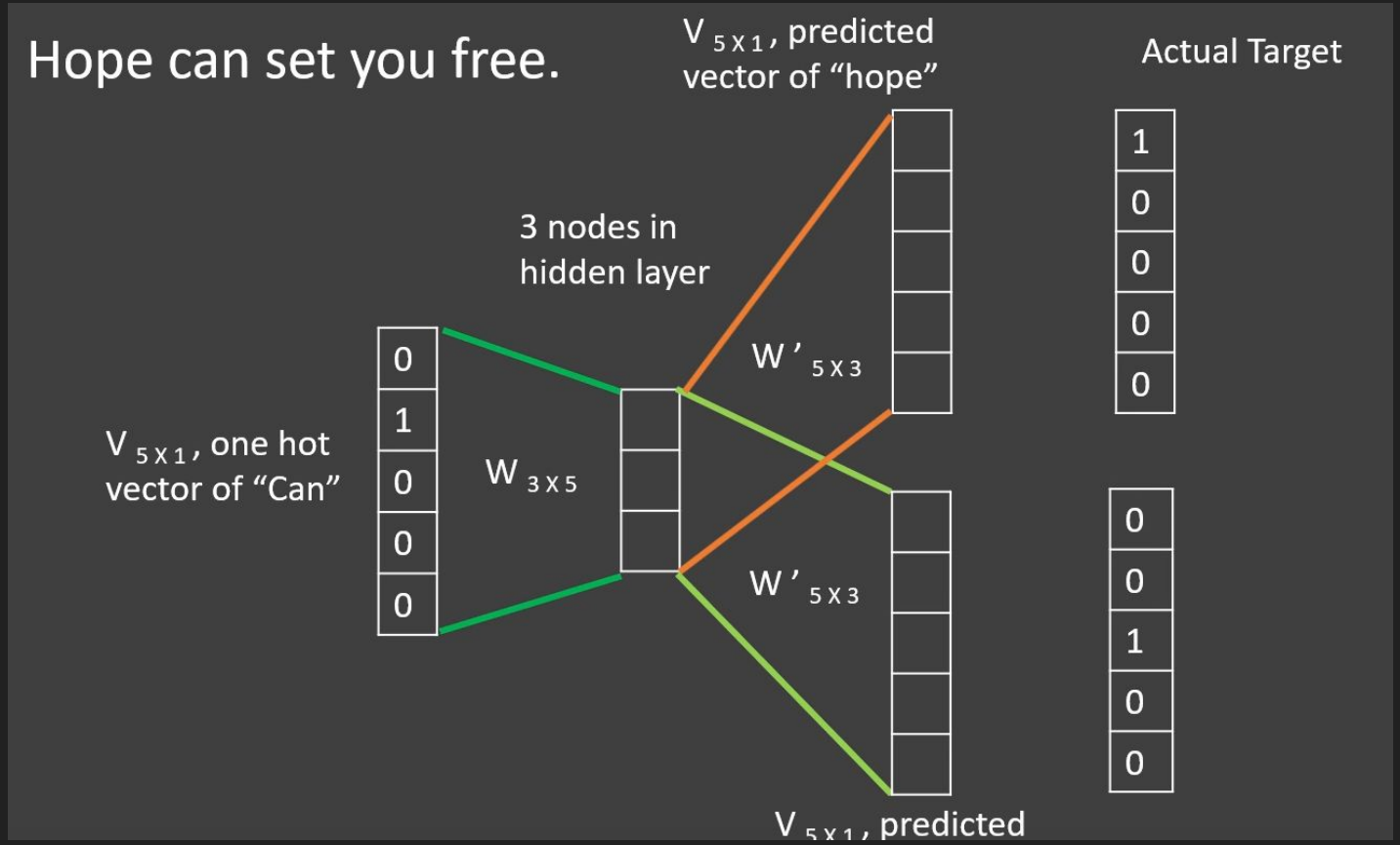
V_{5X1}, one hot vector of "Set"

V_{5X1}, predicted one hot vector of "Can"

w00	w01	w02	w03	w04
w10	w11	w12	w13	w14
w20	w21	w22	w23	w24

W_{3X5}

Skip-gram


Hope can set you free.
V 5 X 1 , predicted vector of “hope”
Actual Target
3 nodes in hidden layer
V 5 X 1 , one hot vector of “Can”
W 3 X 5
W ’ 5 X 3
W ’ 5 X 3
V 5 X 1 , predicted
0
1
0
0
0
1
0
0
0
0
0
1
0
0
0
0
0
0

Sources

Efficient Estimation of Word Representations in Vector Space, Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 2013

https://arxiv.org/abs/1301.3781

Understanding word vectors, Allison Parish

https://gist.github.com/aparrish/2f562e3737544cf29aaf1af30362f469

https://en.wikipedia.org/wiki/Word_embedding, consulté le 31 mars 2024
https://en.wikipedia.org/wiki/Distributional_semantics, consulté le 31 mars 2024
https://corpling.hypotheses.org/files/2018/04/3dplot-500x381.jpg l'image de word embedding
https://www.youtube.com/watch?v=UqRCEmrv1gQ, The Semicolon (pour les images de CBOW et Skip-gram)