Programmation de modèles linguistiques (II)

L6SOPROG L3 SDL

Alice Millour

STIH EA 4509, Sorbonne Université Sorbonne Université

La séance d'aujourd'hui

- présentations par groupe de 3

Entraînement de word embeddings (plongements lexicaux)

- un peu de théorie
- pratique avec python et `gensim`

Présentations

Présentations : comme un cours

4 sujets / 4 groupes de 3 / 30 minutes de préparation / 3 minutes de présentation devant la classe

- 1. représentation vectorielle d'un mot : utilité, exemple
- 2. représentation vectorielle d'un texte : utilité, exemple
- 3. moteurs de recherche et systèmes de recommandations : méthodes, objectifs, similarités et différences
- 4. notions de programmation : listes vs dictionnaires / return vs print / txt vs json

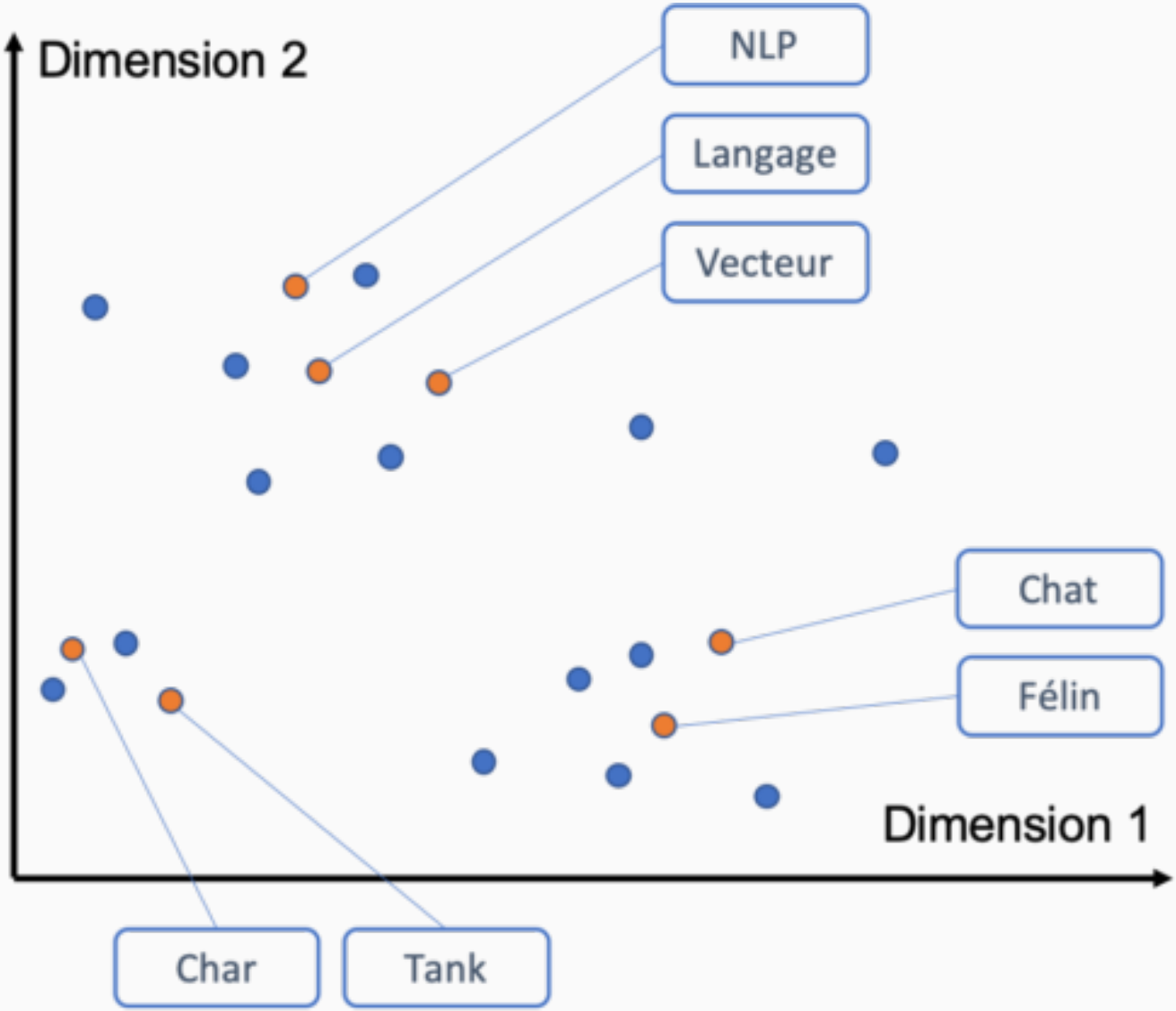
Représentation vectorielle de mots

Jusqu'à maintenant, les représentations vectorielles étaient déjà construites.

- le vecteur d'un mot peut être de taille variable (exemple de la polarité)
- le vecteur d'un mot dépend du texte à partir duquel il a été calculé

À partir des statistiques de co-occurrence, de position (relative et absolue) dans la phrase, on répartit les mots dans "l'espace du sens"

Exemple représentation vectorielle de mots



https://www.quantmetry.com/blog/comment-ia-apprend-a-lire/

Retour sur la démo de Word2Vec

plongement lexical = word embedding = vecteur

Demo Gensim

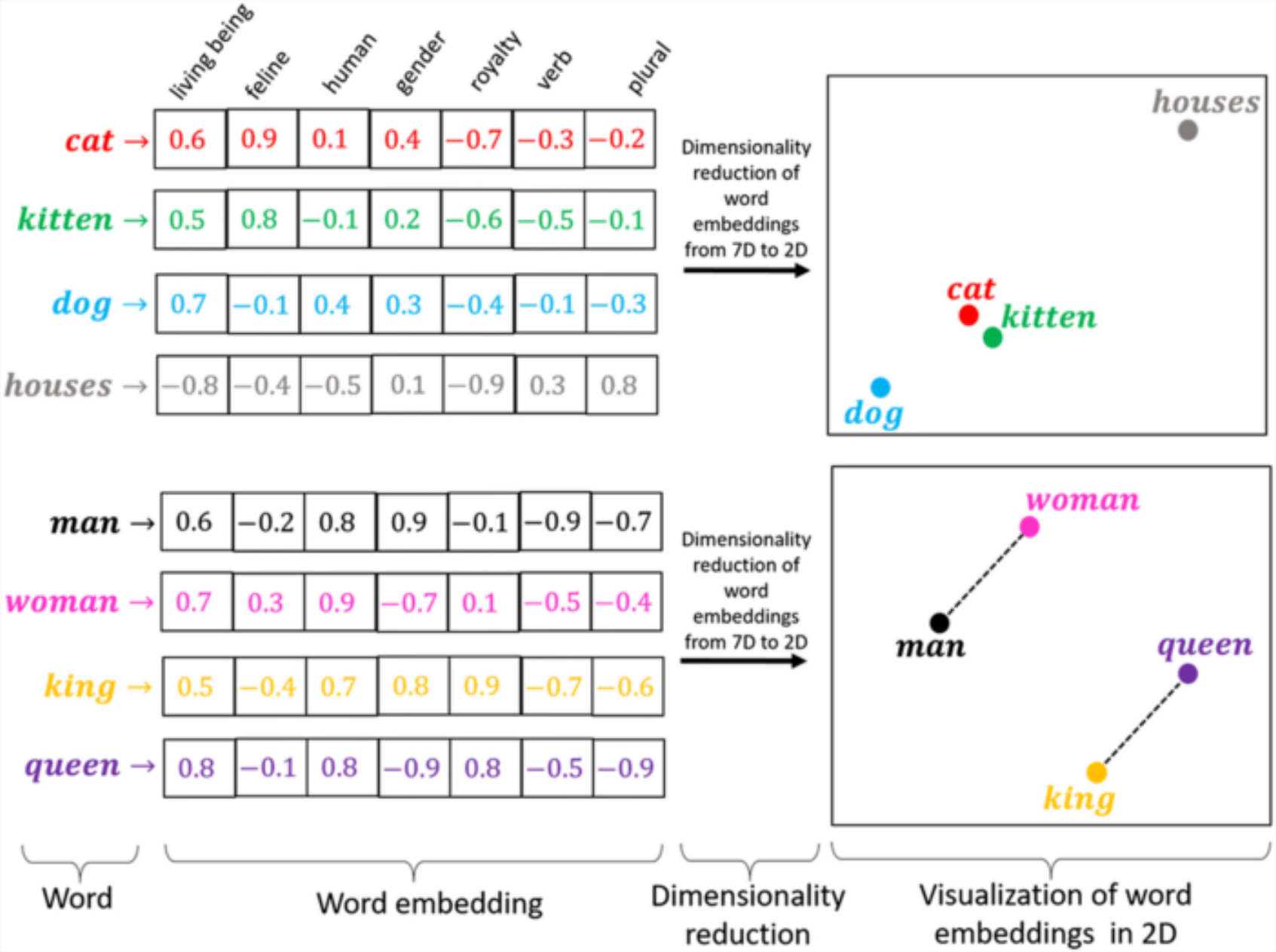
(source : https://web.stanford.edu/class/cs224n/materials/Gensim%20word%20vector%20visualization.html)

Quelles règles pour la mise à jour des vecteurs ?

3 types principaux de relations sémantiques

- relations de similarité
  - synonymie : rapprocher les représentations des mots de ces relations
- relations de dissimilarité
  - antonymie : éloigner les représentations des mots de ces relation
- relations de dissimilarité
  - relations d'ordre : modifier les représentations pour respecter les relations d'ordre

Construction des représentations vectorielles de mots



Construction des représentations vectorielles de mots

Comment sont calculés les vecteurs pour chaque mot ?

- initialisation aléatoire : les vecteurs sont aléatoirement répartis dans l'espace
- mise à jour itérative au fur et à mesure de la lecture du texte

`https://miro.medium.com/max/584/0*aidovtDCj-Hd3Z4g`

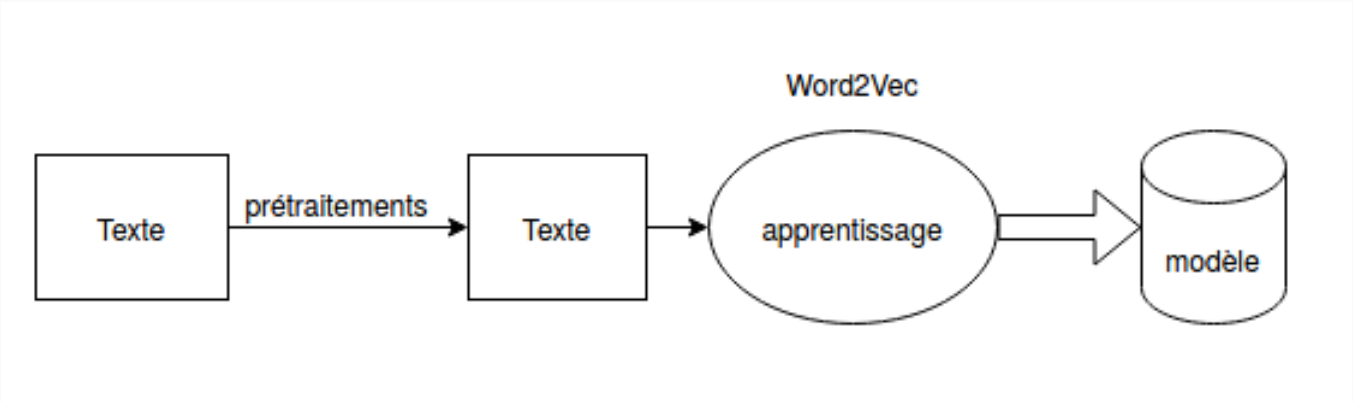
Commentez ce qu'on observe.

Les représentations varient au fil du temps !

https://www.youtube.com/watch?v=2uQ6bgemuLw

regarder à partir de la seconde 57.
Comment expliquez-vous ce qui est présenté ?

En pratique



que contient le modèle ?

TD4 : construction de modèles à partir de différents textes

- (maintenant) télécharger le dossier TD4 sur le moodle
- certaines librairies doivent être installées
- les exercices sont dans des fichiers séparés

(à rendre pour le 7 avril)