Machine Learning

Données à dimension variable : Traitement du langage

Traitement du langage : données

Classification :

- thème/genre (gutenberg.org : 57k livres)
- auteur (gutenberg.org : 57k livres)
- sentiment (Kaggle movie review : 222k commentaires rotten tomatoes)
- reconnaissance d'entités nommées (Kaggle Annotated Corpus for NER : 1.3M tags)
- ...

Traitement du langage : données

Compréhension :

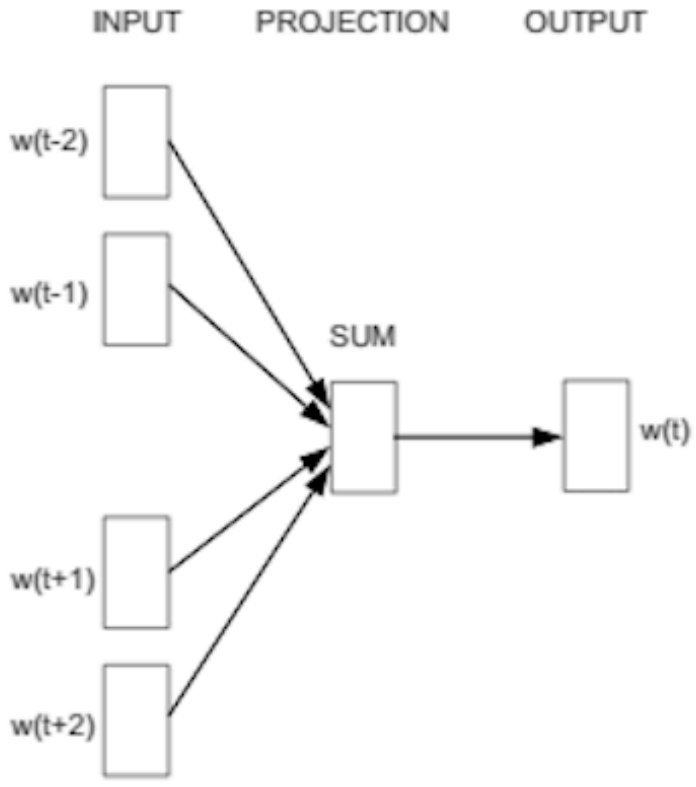
- Question/réponses (SQUAD : 150k questions)
- Traduction (europarl : 450k phrases alignées)
- ...

Traitement du langage : word embeddings

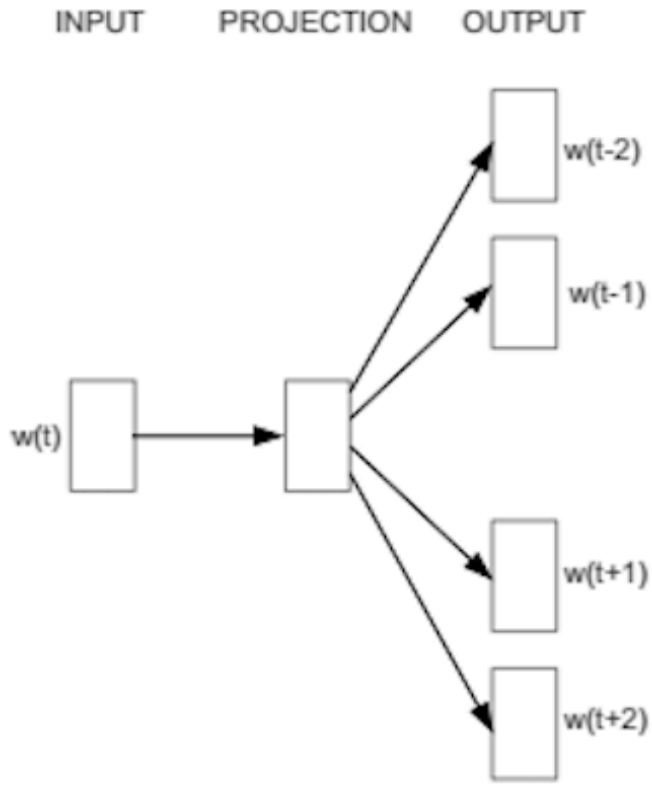
mot = indice dans un dictionnaire (dimension > 30000)
mot = vecteur “sémantique” (dimension < 1000)

- word2vec
- CBOW/Skip-Gram
- Thought vector (pour des phrases ou même des documents entiers)
- ...

Traitement du langage : word embeddings

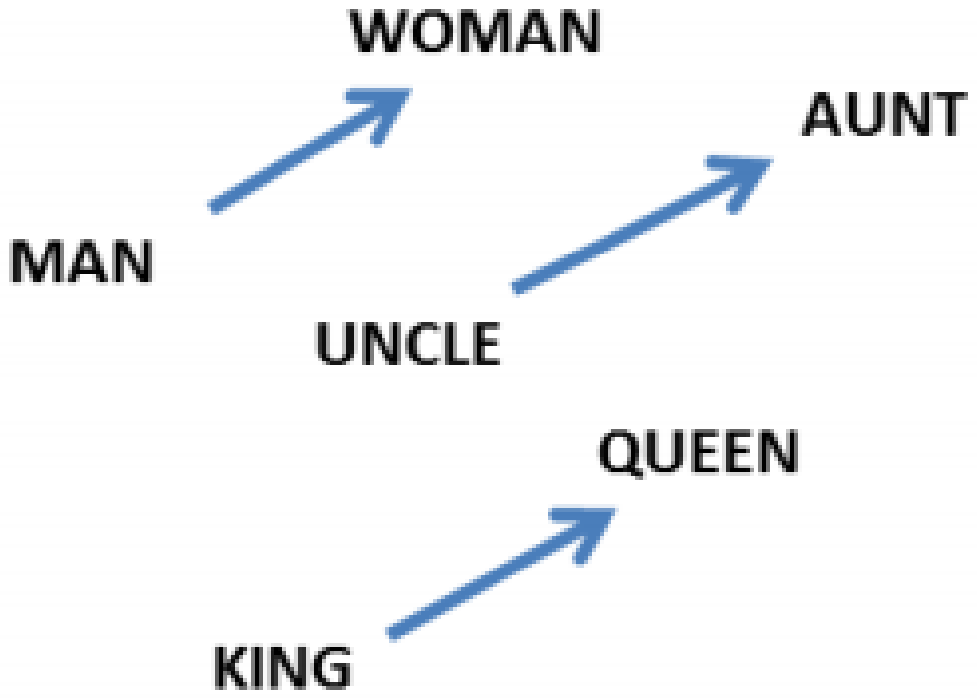


CBOW



Skip-gram

Traitement du langage : word embeddings



Visualisation de l'espace word2vec

Traitement du langage : Champs d'application