

# Machine Learning

Données à dimension variable : Traitement du langage

---

## 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)
- ...

Compréhension :

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

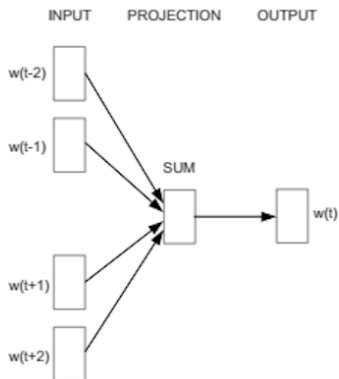
# Traitement du langage : word embeddings

mot = indice dans un dictionnaire (dimension  $> 30000$ )

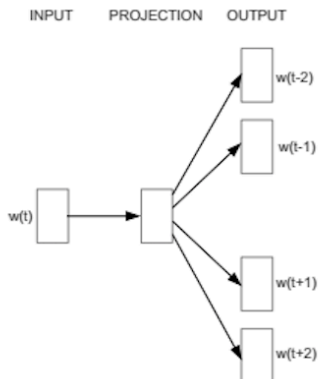
mot = vecteur “sémantique” (dimension  $< 1000$ )

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

# Traitement du langage : word embeddings

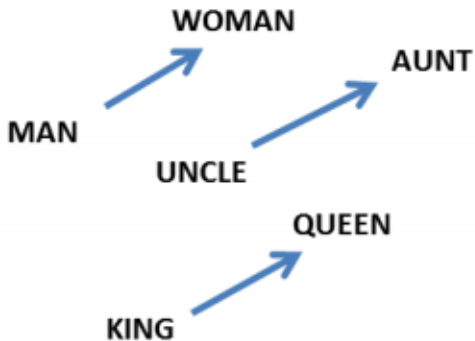


**CBOW**



**Skip-gram**

## Traitement du langage : word embeddings



Visualisation de l'espace word2vec

# Traitement du langage : Champs d'application

## mostly solved

### Spam detection

Let's go to Agra!



Buy VIAGRA ...



### Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

### Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

### Word sense disambiguation (WSD)

I need new batteries for my *mouse*.



### Machine translation (MT)

第13届上海国际电影节开幕...



The 13<sup>th</sup> Shanghai International Film Festival...

### Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



## making good progress

### Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



### Coreference resolution

Carter told Mubarak he shouldn't run again.

### Parsing

I can see Alcatraz from the window!

### Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

### Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

### Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

## still really hard

### Dialog

Where is Citizen Kane playing in SF?



Castro Theatre at 7:30. Do you want a ticket?

