

# **Machine Learning, méthodes et solutions**

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)
- ...

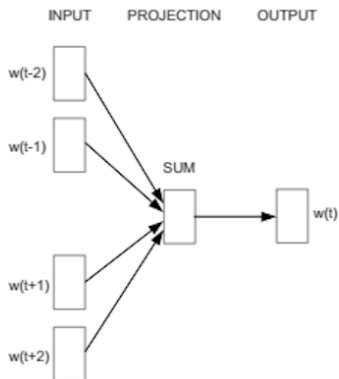
# Traitement du langage : word embeddings

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

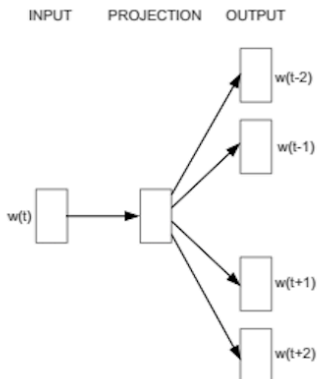
mot = vecteur “sémantique” (dimension  $< 1000$ )

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

# Traitement du langage : word embeddings

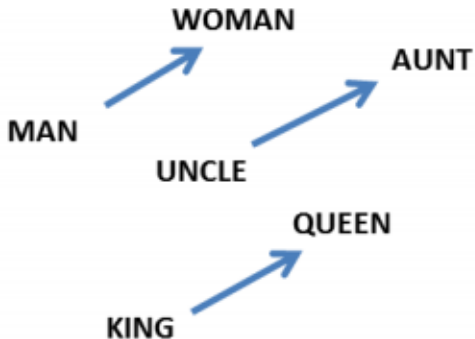


**CBOW**



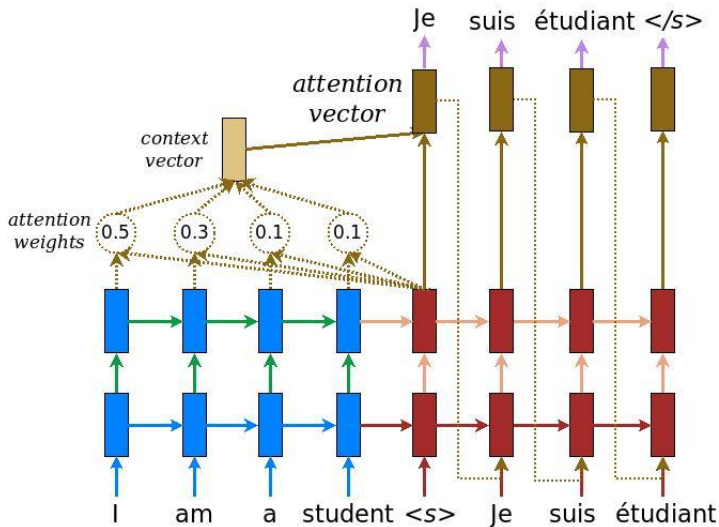
**Skip-gram**

# Traitement du langage : word embeddings

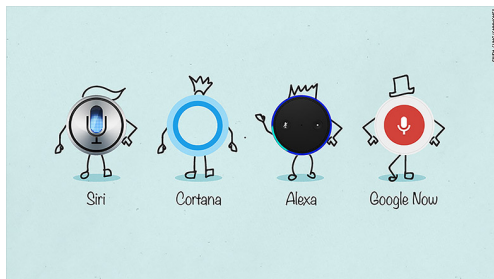


Démo dans l'espace word2vec  
Visualisation de l'espace word2vec

# Traitement du langage : Modèle à attention



## Transcription et synthèse de la parole

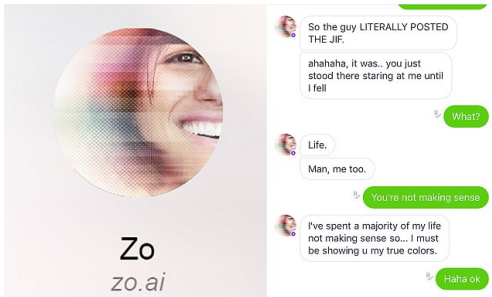




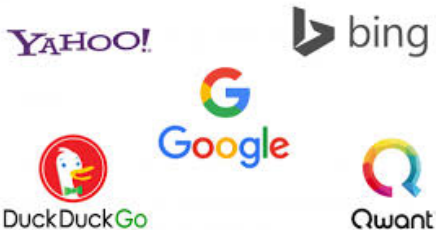
## Identification du locuteur



## Chatbot



## Moteur de recherche



## Extraction de données



## Analyse de sentiments



## Résumé

<b>Input: Article 1st sentence</b>	<b>Model-written headline</b>
metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year	mgm reports 16 million net loss on higher revenue
starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases	hainan to curb spread of diseases
australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday	australian wine exports hit record high in september

## Traduction

Traduire **français** (langue identifiée) ▾

Bonjour,

Nos filles souhaitent faire un échange ensemble dans le cadre du programme Brigitte Sauzay. Avant de passer au dossier administratif, nous souhaitons vous poser quelques questions pour s'assurer que nous envisagions cet échange entre nos filles de la même façon.

Le professeur d'allemand de Gaïa a gentiment accepté de traduire nos questions afin d'éviter des incompréhensions à cause de problèmes de traduction.

Concernant les dates, Emma nous a dit que vous souhaitiez que Gaïa soit repartie avant le 19 juin car vous partiez en vacances.

Traduire en **anglais** ▾

Hello,

Our daughters would like to do an exchange together as part of the Brigitte Sauzay program. Before moving on to the administrative file, we wanted to ask you a few questions to ensure that we consider this exchange between our daughters in the same way.

Gaïa's German teacher kindly agreed to translate our questions in order to avoid misunderstandings due to translation problems.

Regarding the dates, Emma told us that you wanted Gaïa to leave before June 19 because you were going on vacation.



## Language Technology

making good progress

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

### 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.

### Word sense disambiguation (WSD)

I need new batteries for my *mouse*.



### Parsing

I can see Alcatraz from the window!

### 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



Party  
May 27  
add

still really hard

### 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

### Dialog

Where is Citizen Kane playing in SF?



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



≈ 2015



# 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?

