Smart Data Project

Comparison of Large Language Models (LLMs) and a local NLP method : a case study on the citizen consultation “How can we protect and restore biodiversity together?” (Make.org)

ENSAI
École nationale
de la statistique
et de l'analyse
de l'information

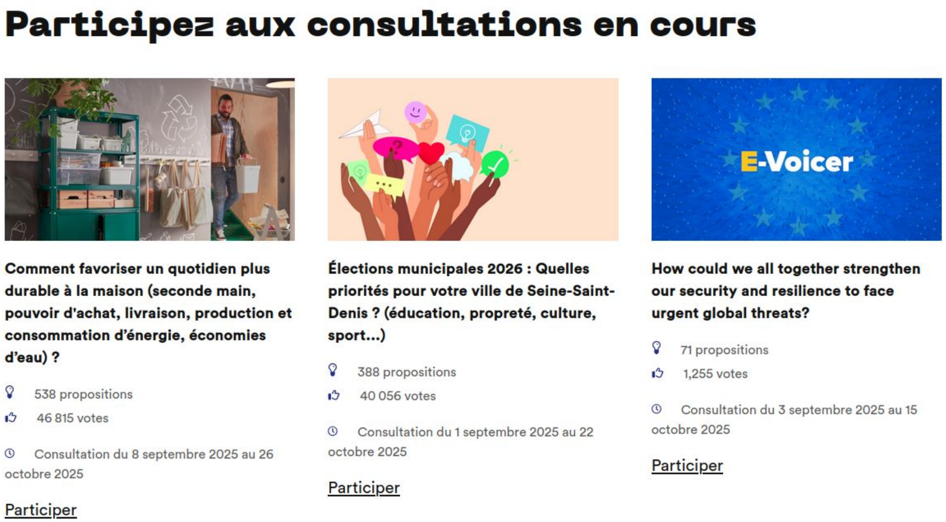
Anne-Cécile GAY
Data Scientist

Project objectives & Data

Data

- Data from a citizen consultation conducted by Make.org
  https://make.org

- Make.org : neutral and independent organization whose mission is to engage citizens and mobilize civil society to drive positive social change.



Participez aux consultations en cours

Comment favoriser un quotidien plus durable à la maison (seconde main, pouvoir d'achat, livraison, production et consommation d'énergie, économies d'eau) ?

538 propositions

46 815 votes

Consultation du 8 septembre 2025 au 26 octobre 2025

Participer

Élections municipales 2026 : Quelles priorités pour votre ville de Seine-Saint-Denis ? (éducation, propreté, culture, sport...)

388 propositions

40 056 votes

Consultation du 1 septembre 2025 au 22 octobre 2025

Participer

How could we all together strengthen our security and resilience to face urgent global threats?

71 propositions

1,255 votes

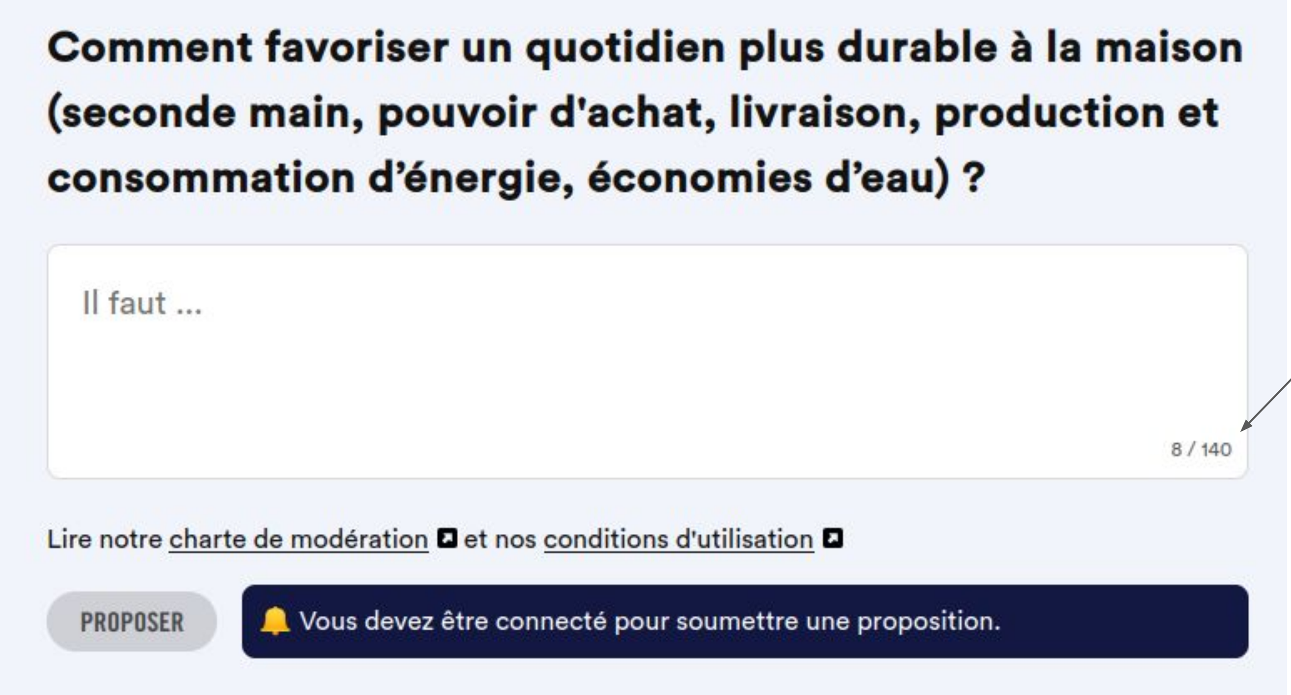
Consultation du 3 septembre 2025 au 15 octobre 2025

Participer

=> Any citizen can submit a proposal and vote on those submitted by others

Data

Example of how to submit a proposal :



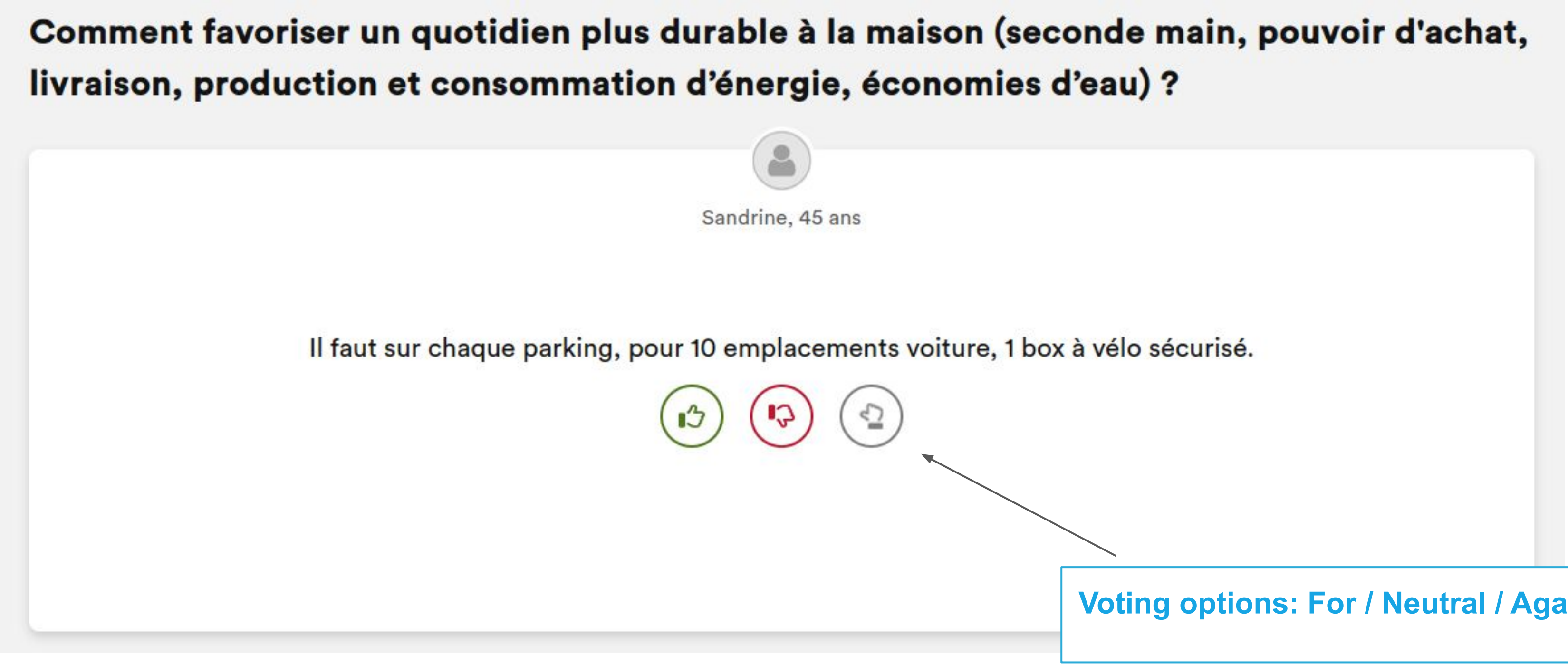
Comment favoriser un quotidien plus durable à la maison (seconde main, pouvoir d'achat, livraison, production et consommation d’énergie, économies d’eau) ?

Lire notre charte de modération ◘ et nos conditions d'utilisation ◘

🔔 Vous devez être connecté pour soumettre une proposition.

! important !
140 characters maximum

Data



Comment favoriser un quotidien plus durable à la maison (seconde main, pouvoir d'achat, livraison, production et consommation d’énergie, économies d’eau) ?

Sandrine, 45 ans

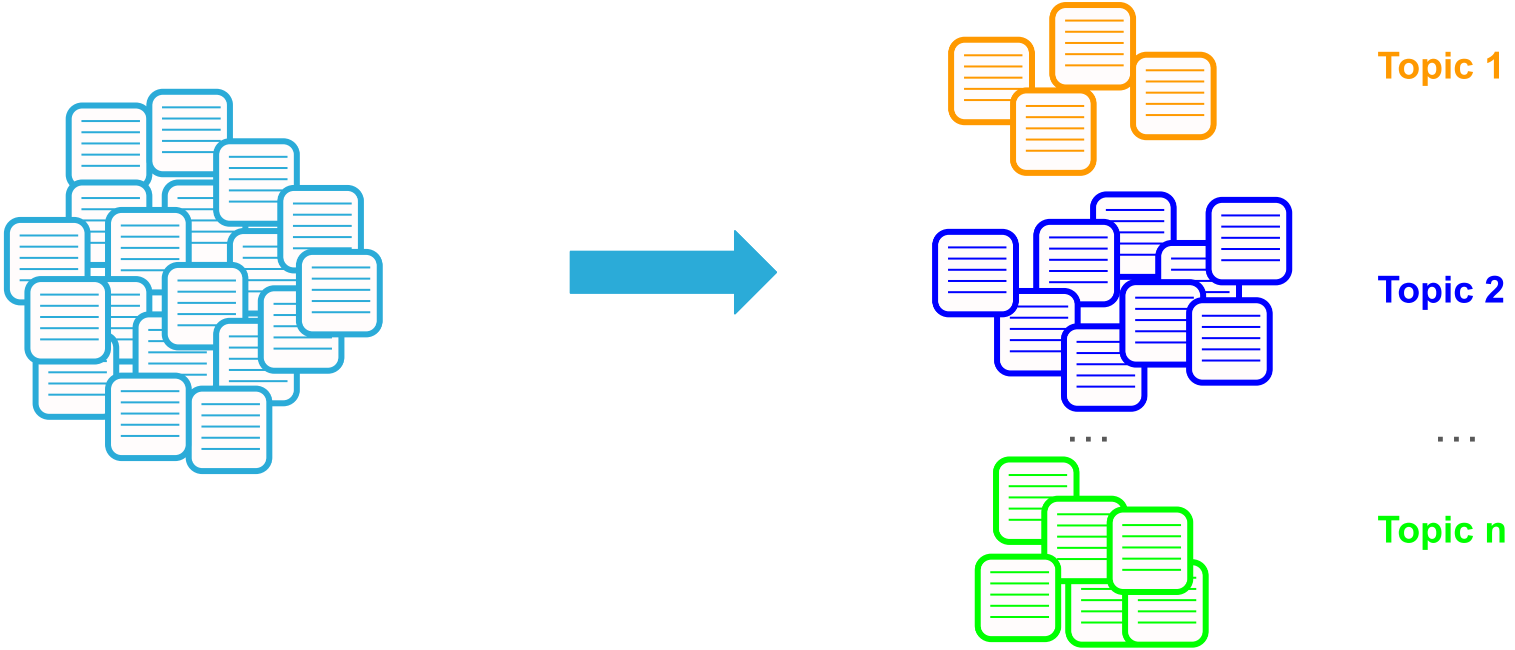
Il faut sur chaque parking, pour 10 emplacements voiture, 1 box à vélo sécurisé.

Voting options: For / Neutral / Against

Project objectives

=> to identify the best methodological approach for topic modeling these proposals

Topic Modeling: Automatic detection of themes in a text corpus



Project objectives

- Literature mainly focuses on **supervised tasks** with **pre-labeled datasets** (e.g., news or tweet rankings).

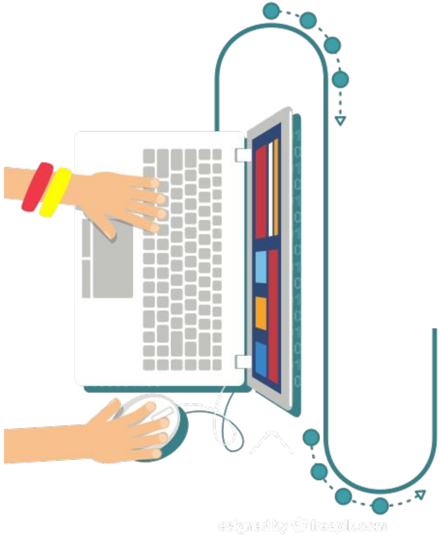
- **HOWEVER in Customer & citizen insights context**: very few labeled data → classical supervised approaches are limited.

➡ **This project: unsupervised problem.**

Descriptive statistics

Descriptive statistics

- Word count distribution per proposal
- Distribution of votes (for / against / neutral)
- Vocabulary
- etc.



NLP : data preparation

Data preparation

MAIN STEPS

1 **Data collection**

“I am happy with the ease, speed, and follow-up of my projects!”

2 **Tokenization:** splitting text into word units

[I,am,happy,with,the,ease,speed,and,follow−up,of,my,projects,!]

3 **Normalization:** removing capitalization, accents, punctuation, etc.

[i,am,happy,with,the,ease,speed,and,followup,of,my,projects]

4 **Stop-word removal:** removing common words

[happy,ease,speed,followup,projects]

5 **Lemmatization:** converting words to their base form

[happy,ease,speed,followup,project]

Data preparation

How to transform text into numerical representations ?

Unstructured data

Numerical data

Word vector representation

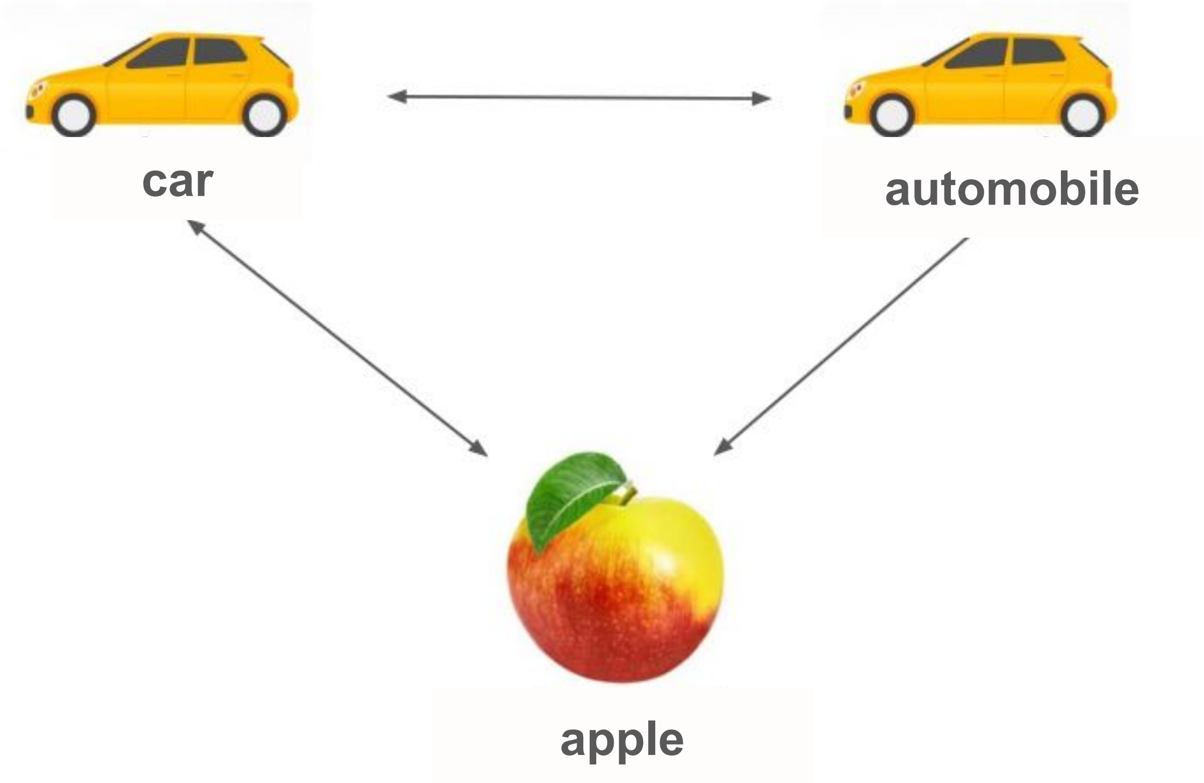
First idea : Bag Of Words

a **document**
=
a **vector**

a **corpus**
=
une **matrix**

$$\begin{array}{c|cccc} & \text{word\_1} & \text{word\_2} & \dots & \text{word\_n} \\ \hline \text{doc\_1} & 0 & 1 & \dots & 0 \\ \text{doc\_2} & 1 & 1 & \dots & 0 \\ \dots & & \dots & & \\ \text{doc\_m} & 1 & 0 & \dots & 1 \end{array}$$

BAG OF WORDS : limitations



car

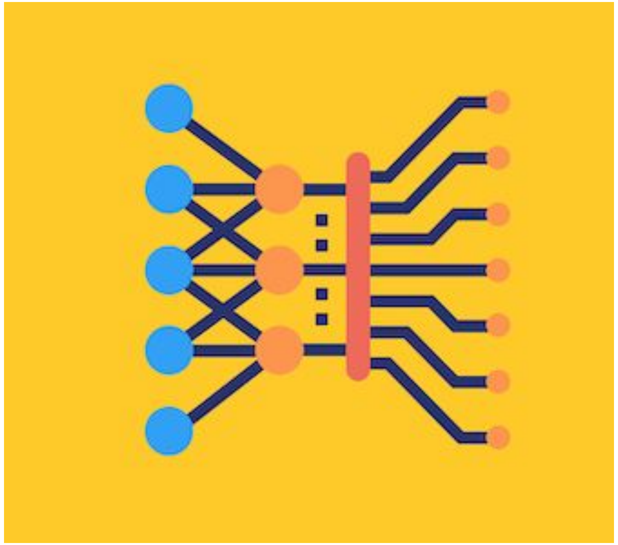
automobile

apple

Bag Of Words : limitations

Very simple method, but:

- **Sparse matrix** : inefficient for machine learning algorithms
- **Ignores meaning / context**: the word “automobile” is as distant from “car” as from any other word



Word embeddings

WORD2VEC method

OBJECTIVES

Projection of words into a **new space**

Similar words are close in terms of distance

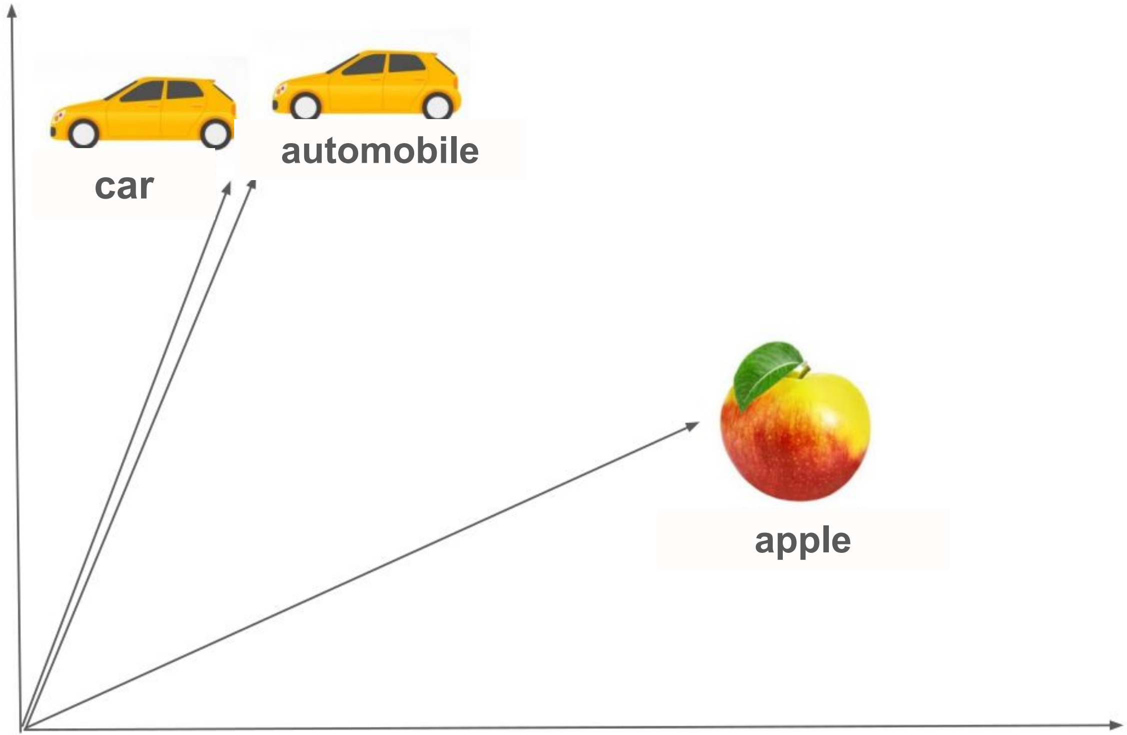
Lower-dimensional space (typically 20–100 dimensions)

Method based on **neural networks**

2 training méthodes : **CBOW & SKIP-GRAM**

CHARACTERISTICS

WORD2VEC method

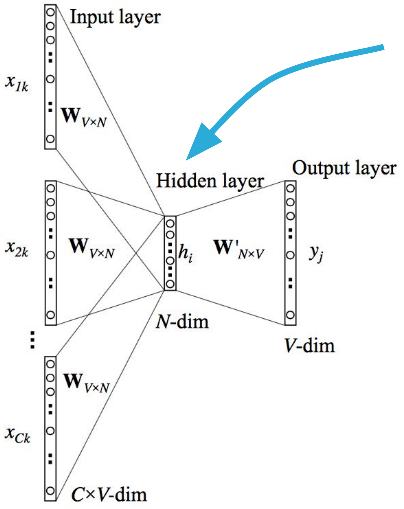


WORD2VEC method - CBOW

CBOW (Continuous Bag of Words)

- Predicts a word based on its context (i.e., surrounding words)
- Training: a neural network is trained to predict a word from its context (words before and after)

Input : window around the target word



dense vector

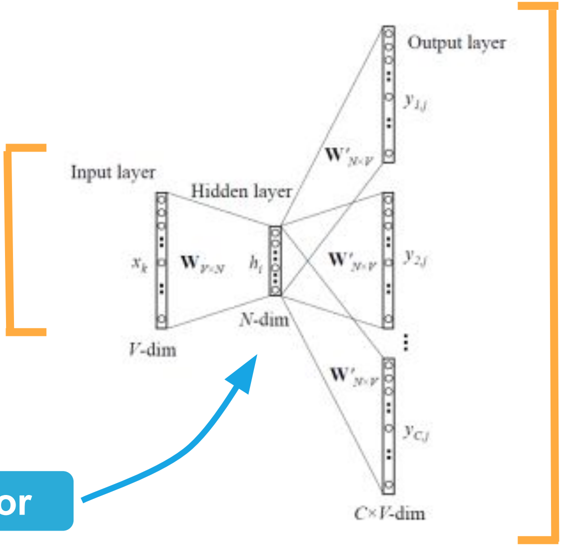
Output : word vector

WORD2VEC method - SKIP-GRAM

SKIP-GRAM :

- predict the context based on a target word
- training : a neural network is trained to predict surrounding words from a given word

Input : word vector



Output : window around the target word

dense vector

ADVANTAGES

CBOW - SKIP-GRAM : do not require large storage capacity

SKIP-GRAM : more accurate for rare words

CBOW : faster training than SkipGrams

Word2Vec : CBOW & Skip-Gram

CBOW - SKIP-GRAM : training time can be high

CBOW : more accurate for frequent words

SKIP-GRAM : slower thanCBOW

DISADVANTAGES

Word2Vec - training

On which data should a Word2Vec model be **TRAINED** ?

- **Several pre-trained embeddings** on very large corpora are available online: Google, Wikipedia

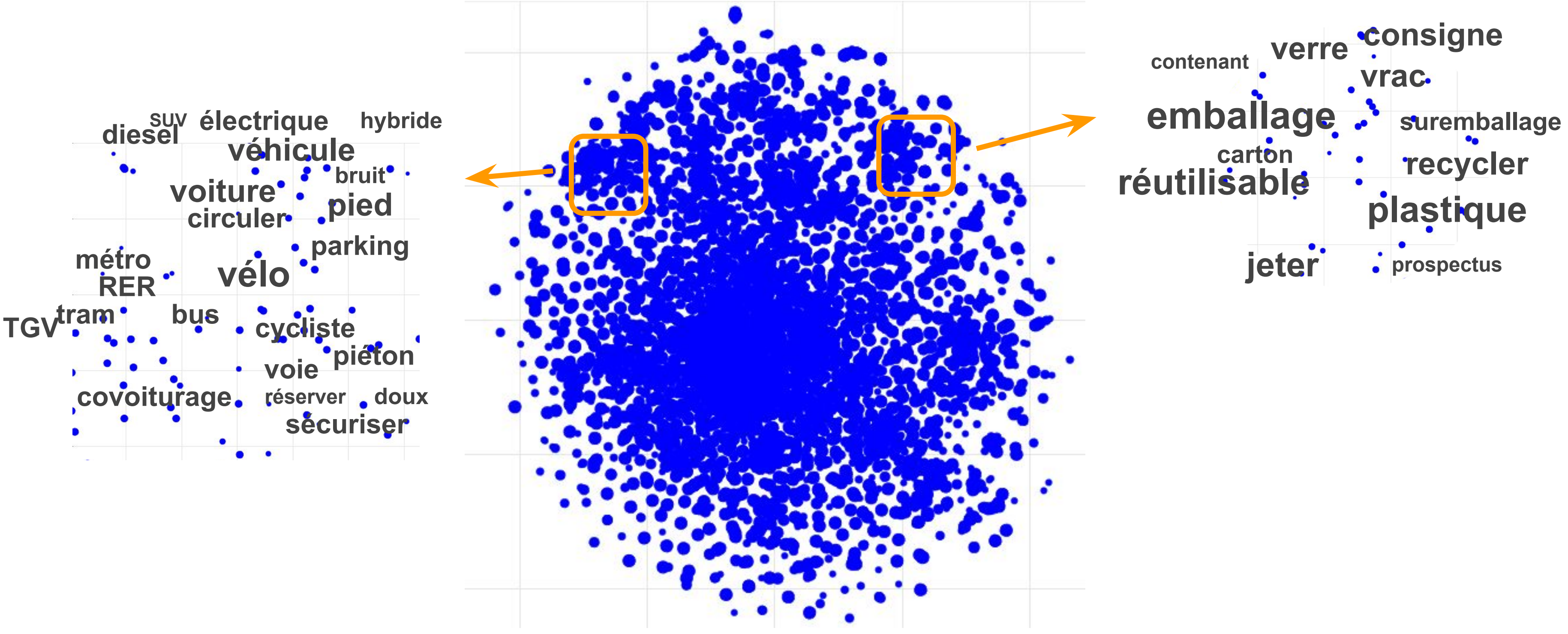
Very large and rich embeddings, BUT:

- **Not very accurate for a specific context**
- Developed on a **particular language register** (e.g., formal Wikipedia language ≠ customer questionnaire verbatims)

→ Real benefit of **TRAINING YOUR OWN EMBEDDING**
Specific to your own corpus => therefore more effective

Word embeddings

Example of visualization



Topic modeling

Topic Modeling

Word2Vec + clustering

transforming words into numerical vectors
ex : Word2Vec

+

applying a classification algorithm
example : k-means

Project steps

Project steps

1. Exploratory data analysis

2. Topic modeling approaches (3 pathways to explore)

A. Word2Vec + clustering

- Train W2V to obtain embeddings of proposal
- Project proposals into vector space
- Apply clustering methods to identify themes

B. LLM embedding + clustering

- Use a pre-trained LLM embeddings (e.g. via Hugging Face)
- Project proposals into vector space
- Apply clustering methods to identify themes
-

C. Direct LLM querying

- Query the LLM directly on the corpus to extract themes without using embeddings

3. Comparison of approaches

?

Any **questions ?**