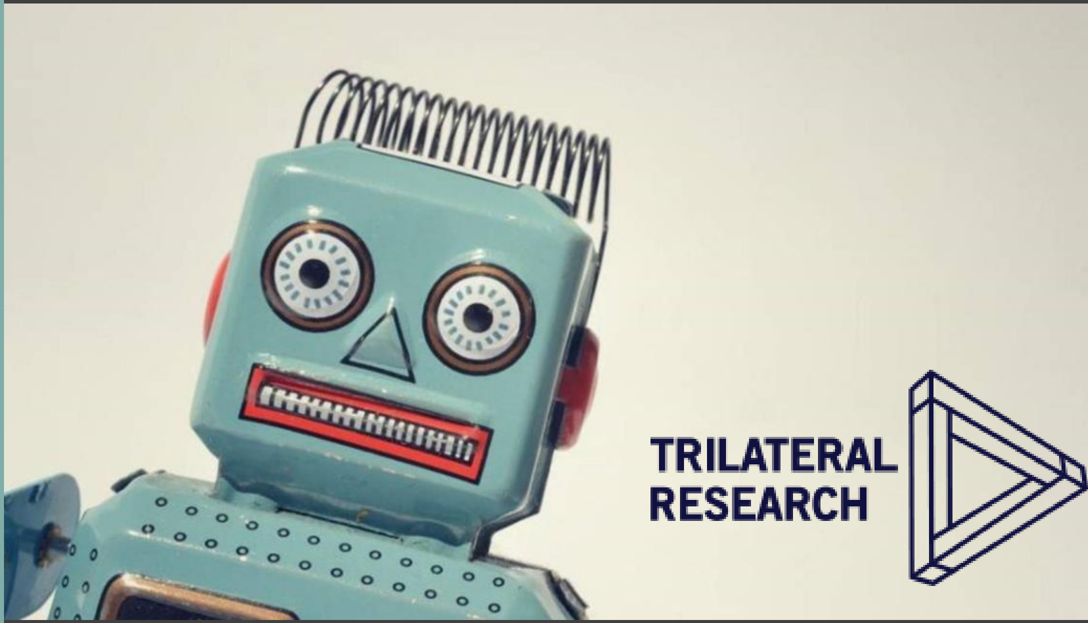
Online citizen participation & Natural language processing

Auditing the urban planning process in Decidim Barcelona

Ana Valdivia
Data Scientist

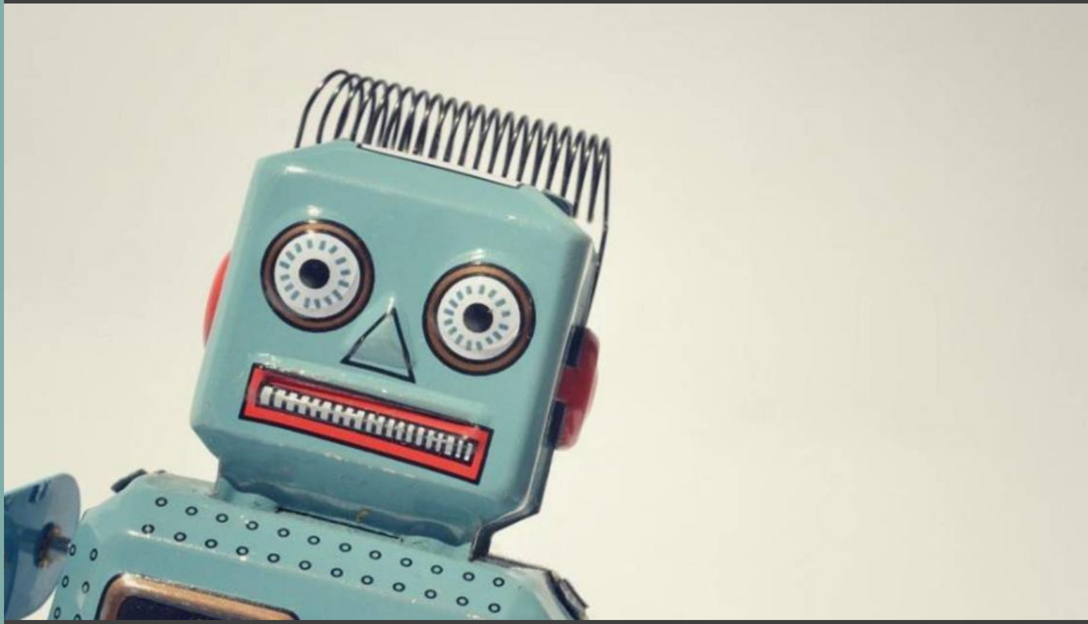
| *PyData Salamanca, 28 de Noviembre de 2019*



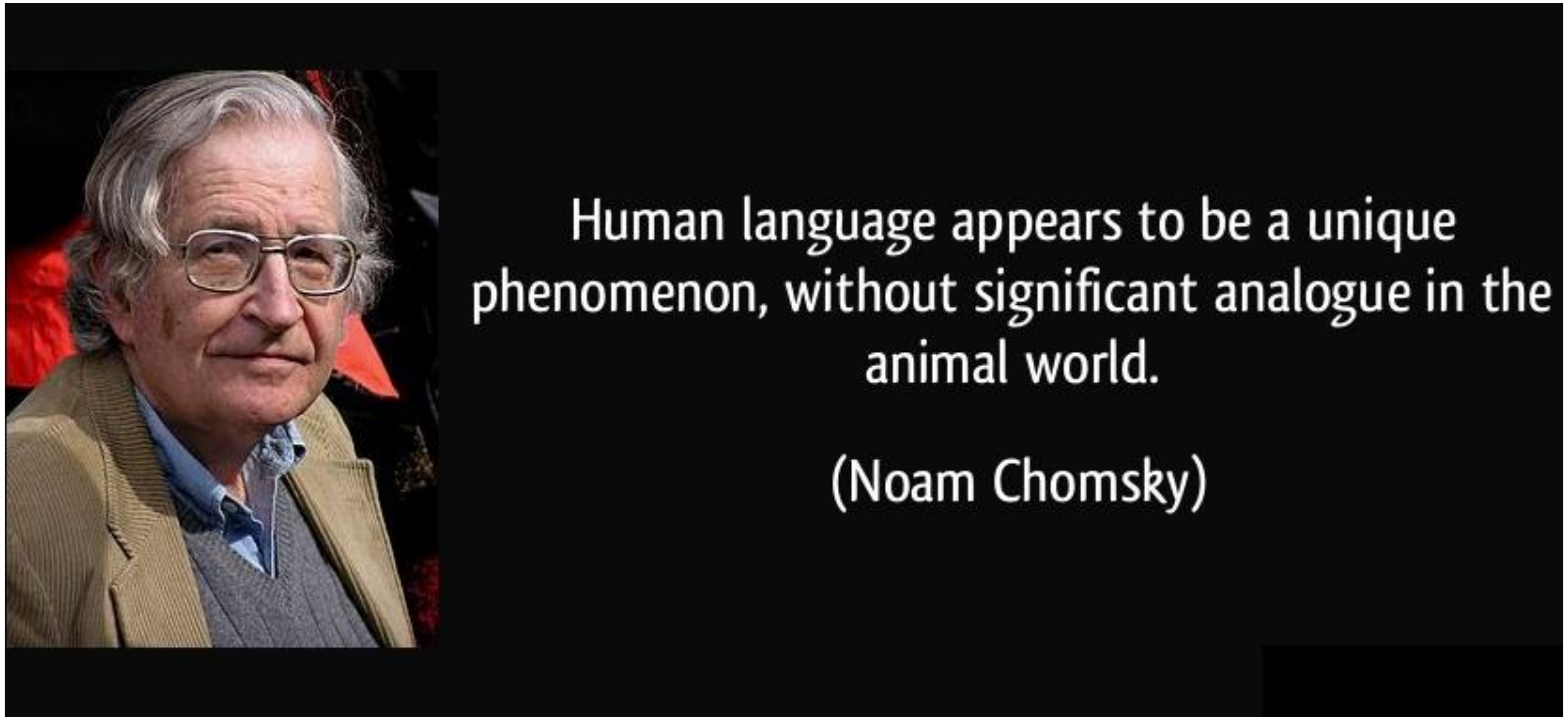
ROAD MAP

1. Intro to NLP
2. Word embeddings
3. Decidim Barcelona

Natural Language Processing



Human Language

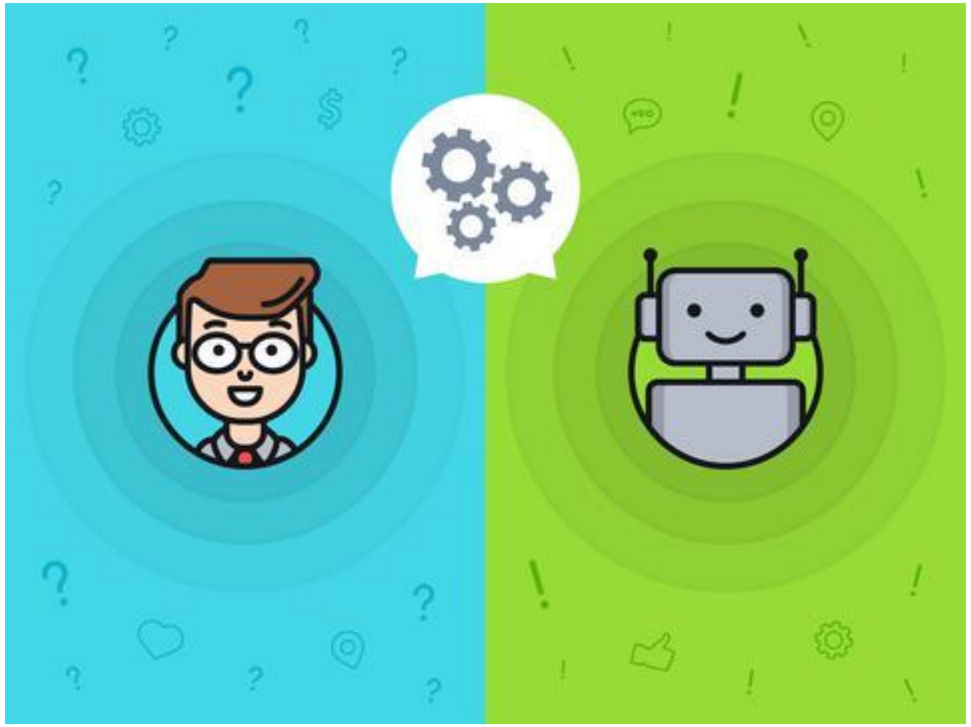


Human language appears to be a unique phenomenon, without significant analogue in the animal world.

(Noam Chomsky)

Natural Language Processing (NLP)

Natural Language Processing (*NLP*) is a sub-field of ***Artificial Intelligence*** that is focused on enabling computers to “understand”, process and generate human languages, to get computers closer to a human-level understanding of language.



N
L
P

Natural Language Processing (NLP)

Natural Language Processing (*NLP*) is a sub-field of *Artificial Intelligence* that is focused on enabling computers to “understand”, process and generate human languages, to get computers closer to a human-level understanding of language.

Tasks:

- github.com/Kyubyong/nlp_tasks
- nlp.lsi.upc.edu/freeling

LOW

- Part of Speech Tagging
- Named Entity Recognition
- Spam Detection

MEDIUM

- Syntactic Parsing
- Word Sense Disambiguation
- Sentiment Analysis
- Topic Modeling

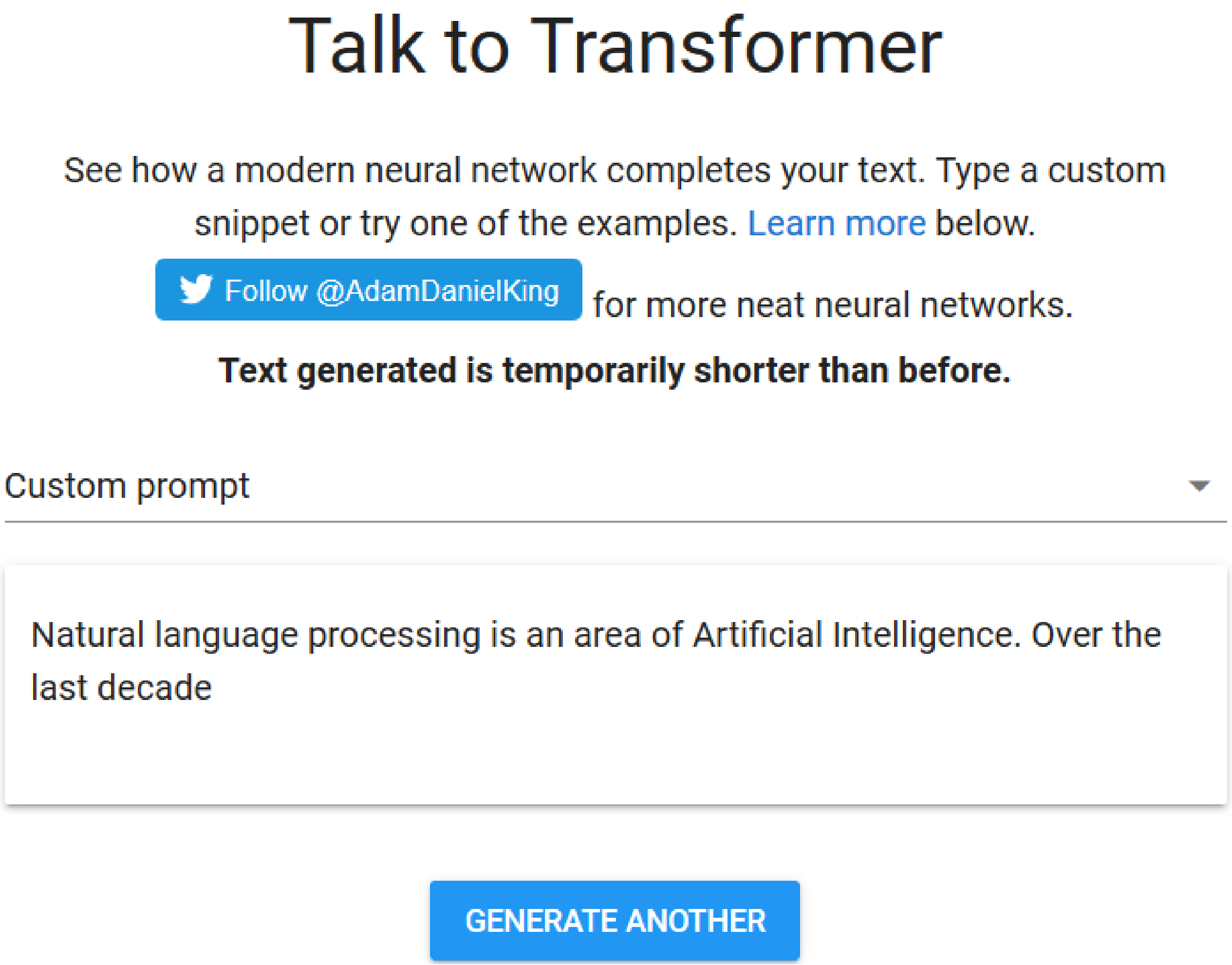
HIGH

- Machine Translation
- **Text Generation**
- Automatic Summarization
- Question Answering

NLP

HIGH: Text Generation

https://talktotransformer.com/



Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. Learn more below.

for more neat neural networks.

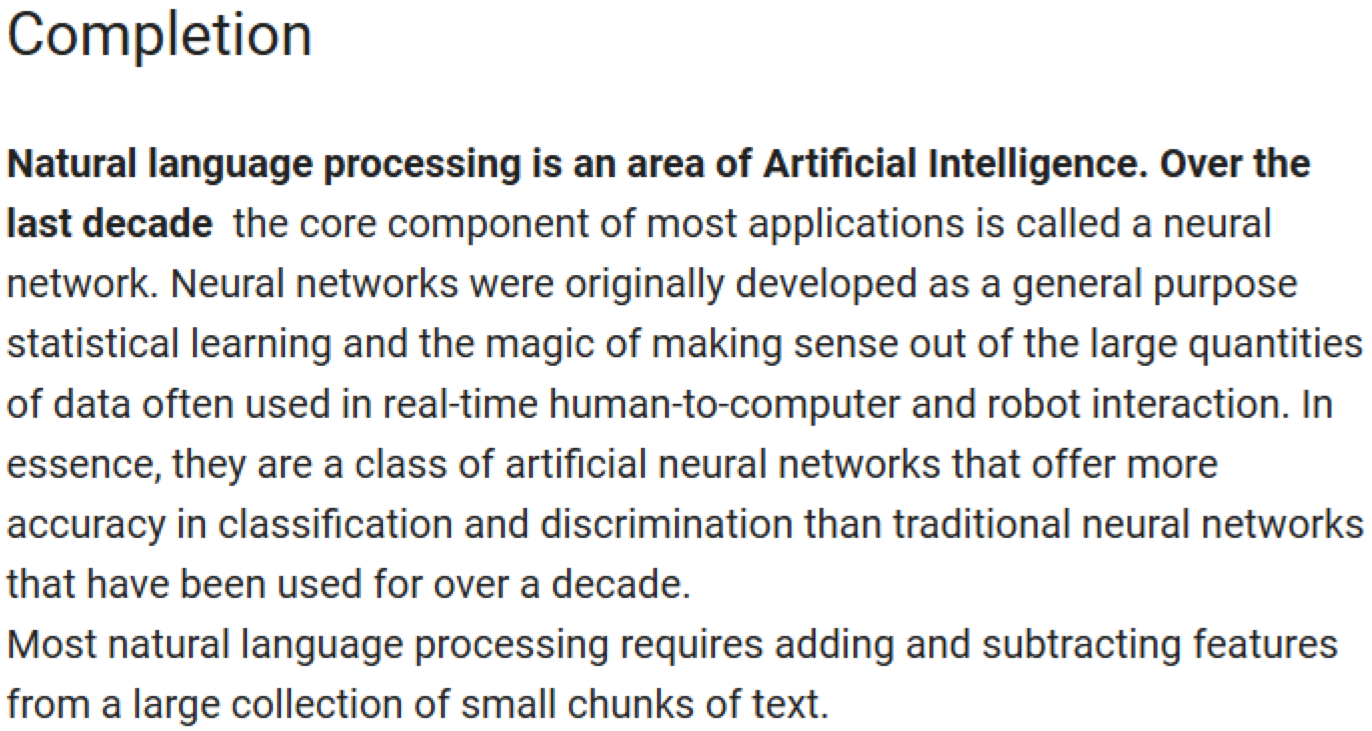
Text generated is temporarily shorter than before.

Custom prompt

Natural language processing is an area of Artificial Intelligence. Over the last decade

HIGH: Text Generation

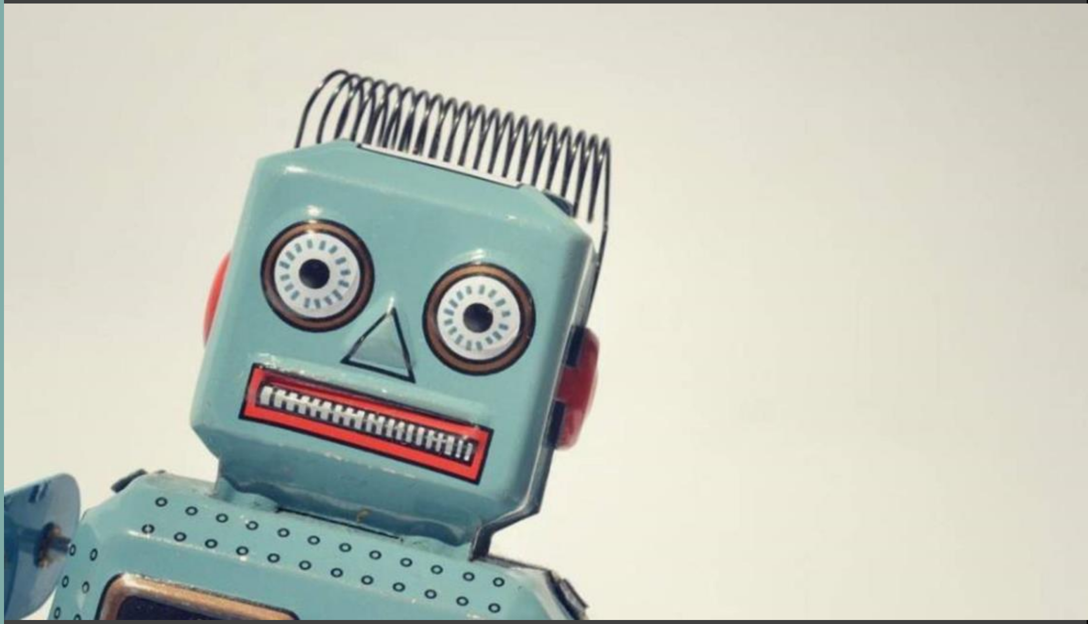
https://talktotransformer.com/



Completion

Natural language processing is an area of Artificial Intelligence. Over the last decade the core component of most applications is called a neural network. Neural networks were originally developed as a general purpose statistical learning and the magic of making sense out of the large quantities of data often used in real-time human-to-computer and robot interaction. In essence, they are a class of artificial neural networks that offer more accuracy in classification and discrimination than traditional neural networks that have been used for over a decade.
Most natural language processing requires adding and subtracting features from a large collection of small chunks of text.

Word Embeddings



How to transform a word into a number?

WORD EMBEDDINGS

How to transform a *word* into a *number*?

one-hot encoding

```
dog = [ 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 ]

cat = [ 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ]
```

How to transform a *word* into a *number*?

Lack of information:

How a model will know that these two words are related/similar?

```
airplane = [ 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 ]
flight   = [ 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ]
```

WORD EMBEDDINGS

How to transform a *word* into a *number*?

word embeddings

Word embedding is the collective name for a set of language modeling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers.

CONTEXT.

- “What time is your **airplane** scheduled?”
- “The pilot marked the cruise speed on our **airplane**'s flight.”
- “The engine of an **airplane** uses the propulsion force to take off.”
- “Many passengers are afraid to fly, even though the **airplane** is the safest mode of transportation.”

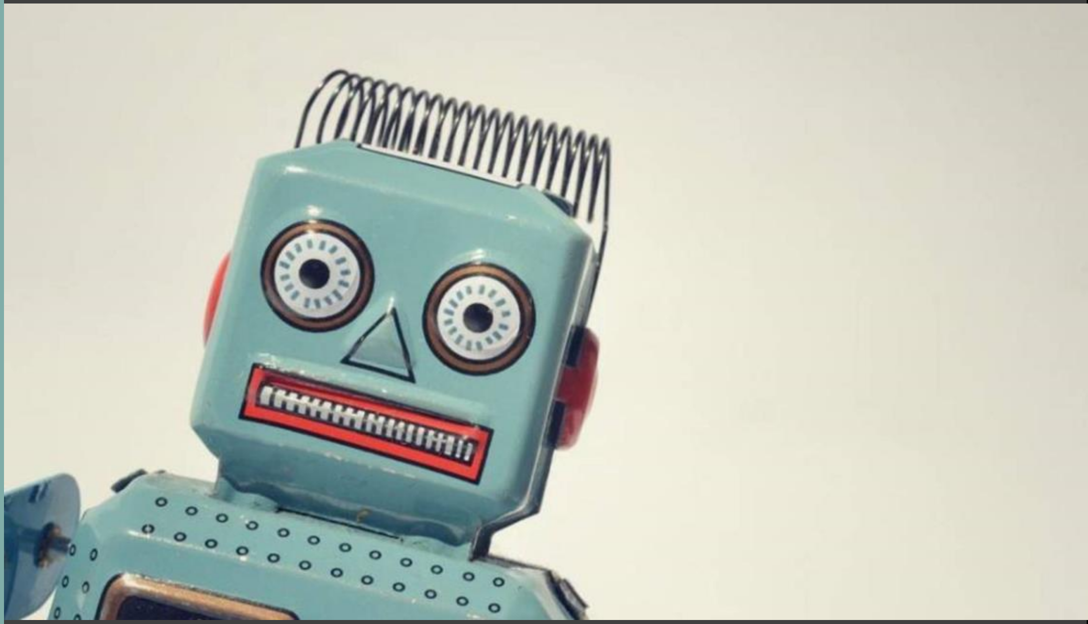
- **Airplane** is related with **scheduled, cruise speed, pilot, fligh, take off, passengers**, etc.

How to transform a *word* into a *number*?

word embeddings

- Matrix representations: LDA, GloVe
- Neural networks: **word2vec**, ELMo

word2vec



How to transform a *word* into a *number*?

word2vec (Mikolov et. al. 2013)

- Two **neural networks** with two **layers**:
  - *input*: one-hot vectors.
  - *hidden* layer: lineal.
  - *output*: softmax function.

- **CBOW** (Conditional Bag of Words): given the **context** $\{w_{t\text{-}C} \dots , w_{t\text{-}1} , w_{t+1} , \dots , w_{t+C}\}$, predict the **central word** w_t.

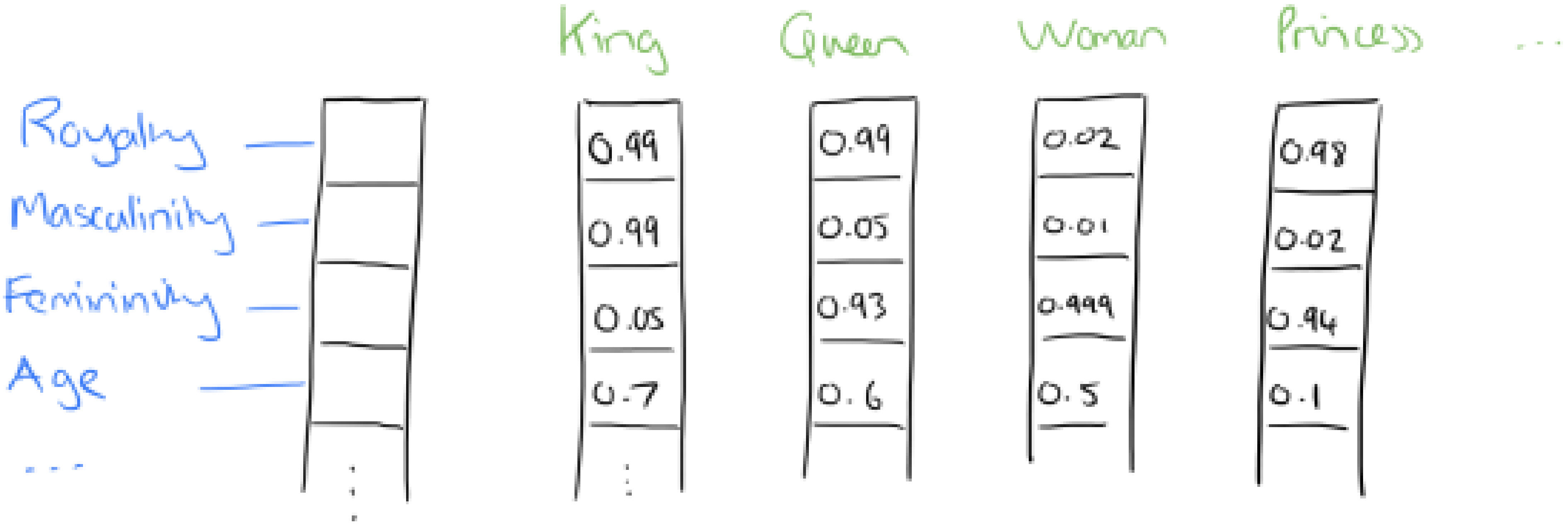
- **Skip-gram**: given the **central word** $w_{t,}$ predict the **context** $\{w_{t\text{-}C} \dots , w_{t\text{-}1} , w_{t+1} , \dots , w_{t+C}\}$.

- **Weights** of the hidden layer are the embeddings representations.

- Its performance depends on the size of the corpus (the bigger, the better).

How to transform a *word* into a *number*?

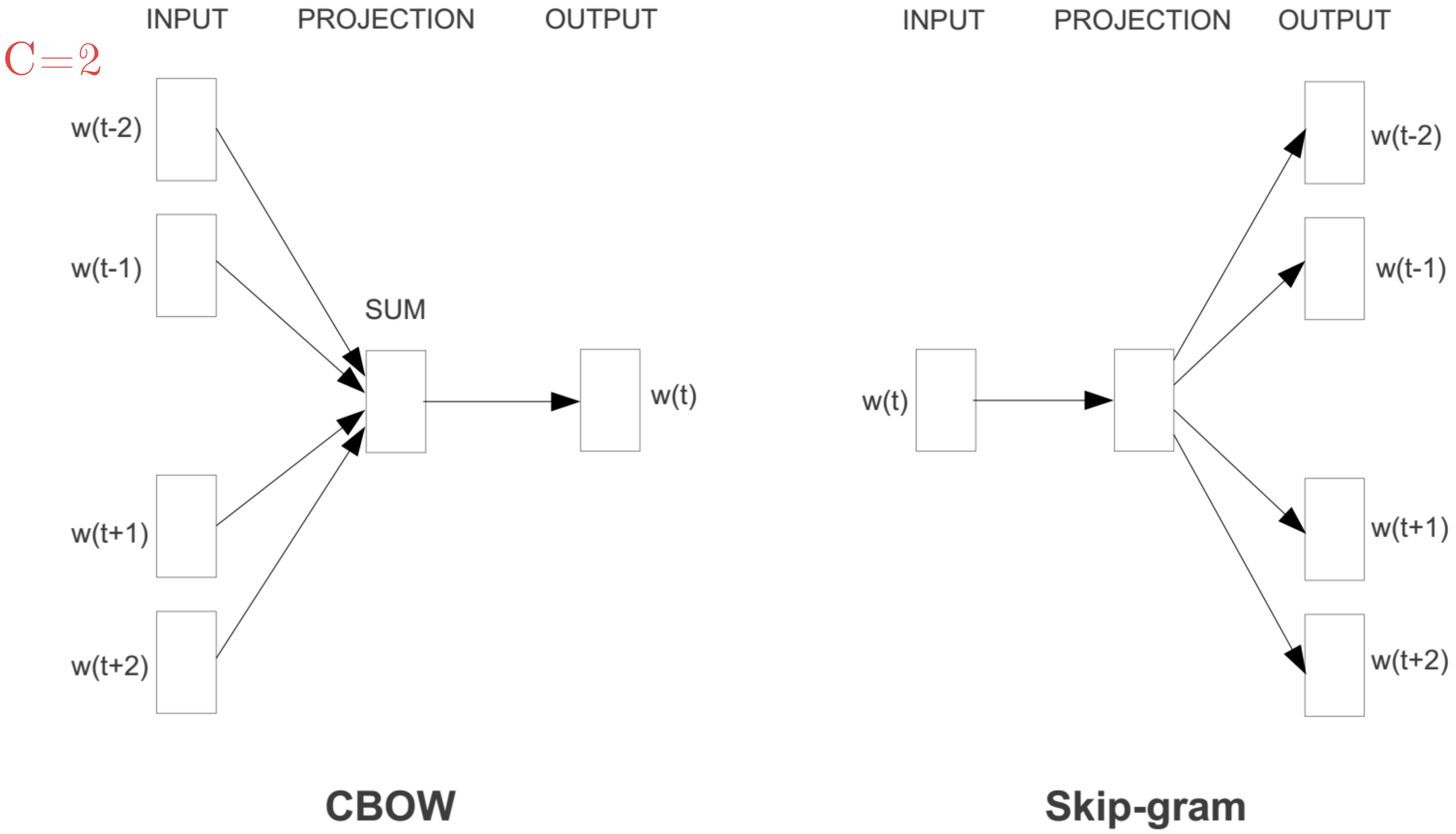
word2vec (Mikolov et. al. 2013)



	King	Queen	Woman	Princess	...
Royalty	0.99	0.99	0.02	0.98	
Mascalinity	0.99	0.05	0.01	0.02	
Femininity	0.05	0.93	0.999	0.94	
Age	0.7	0.6	0.5	0.1	
...	⋮				

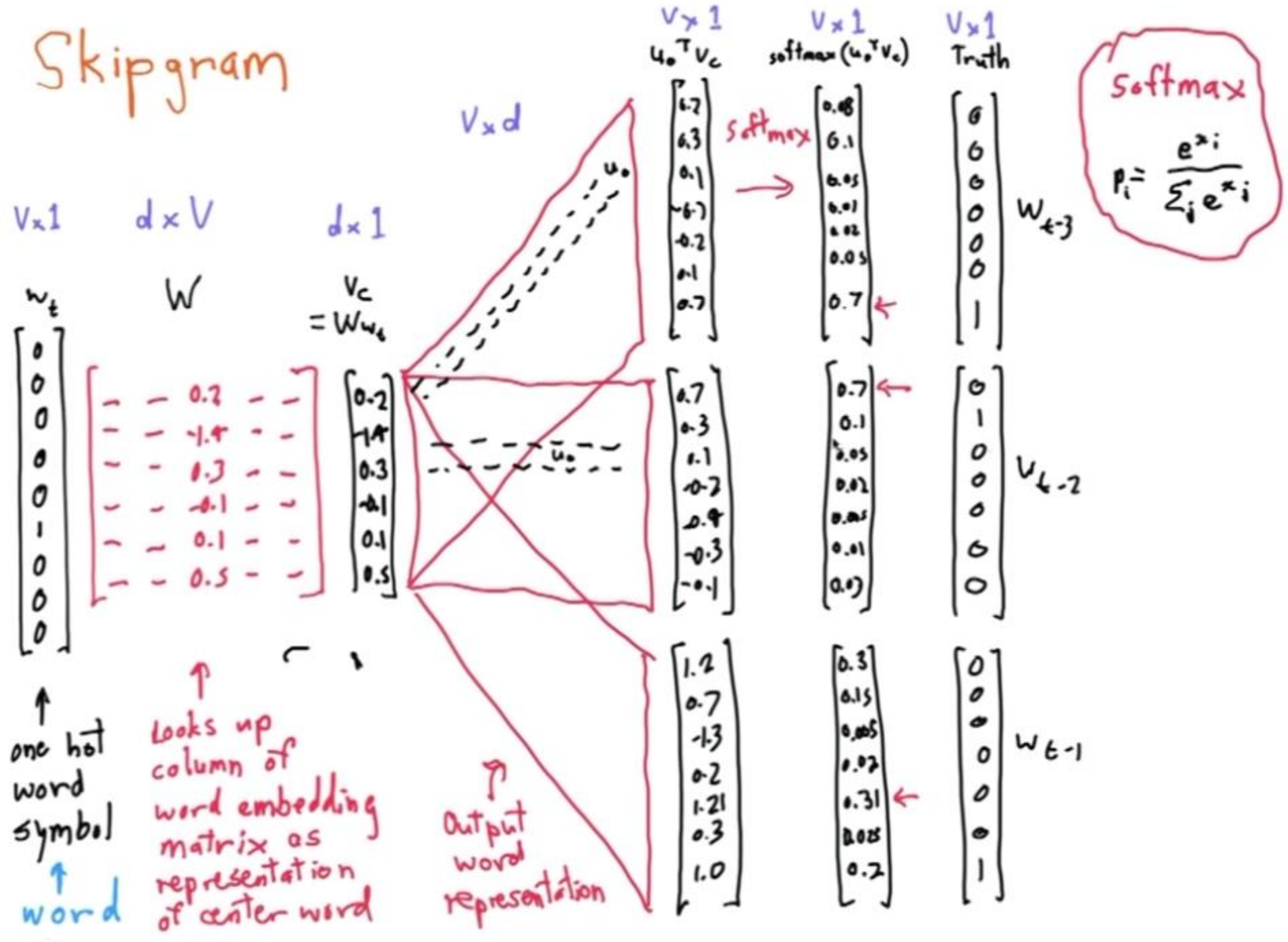
How to transform a *word* into a *number*?

word2vec (Mikolov et. al. 2013)

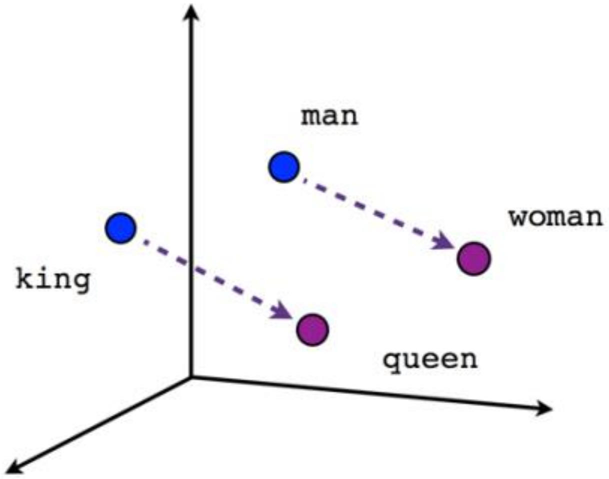


w o r d 2 v e c

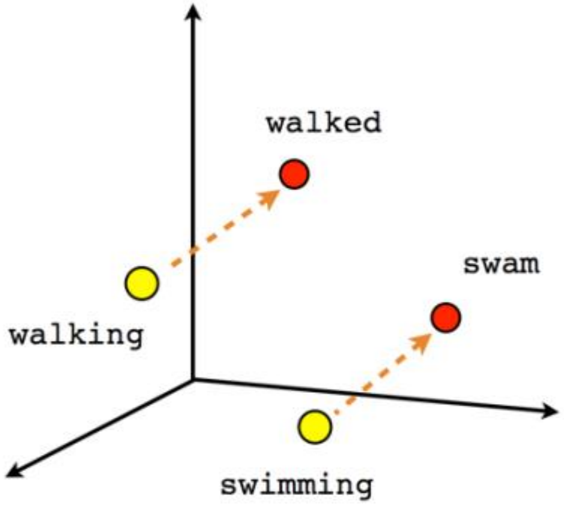
How to transform a *word* into a *number*?



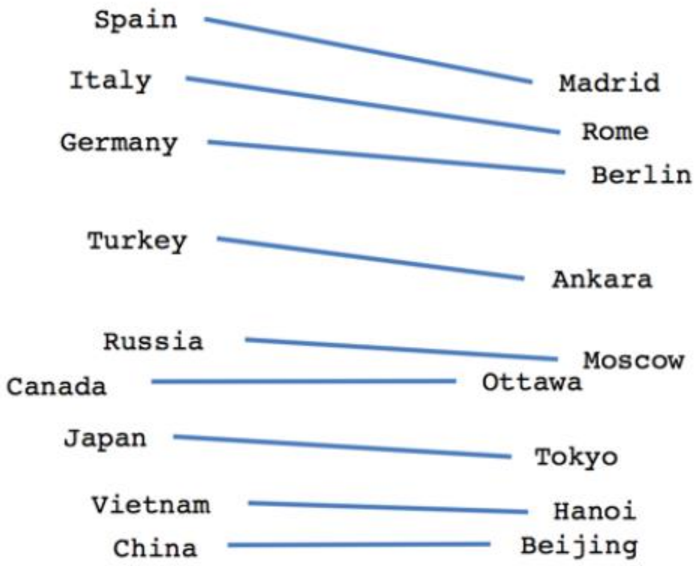
word2vec



Male-Female

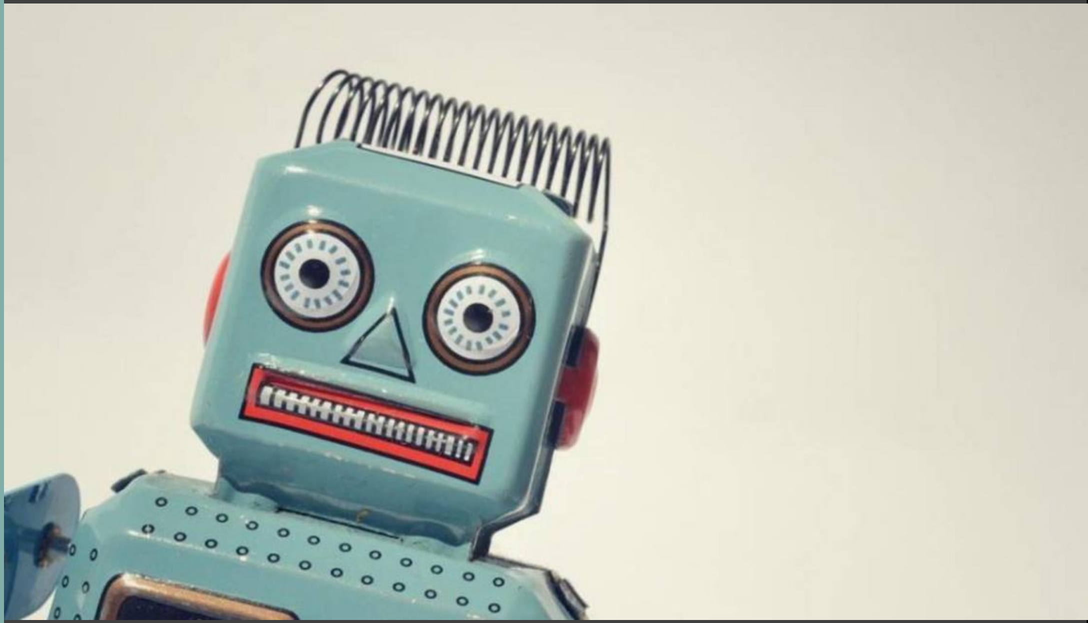


Verb tense

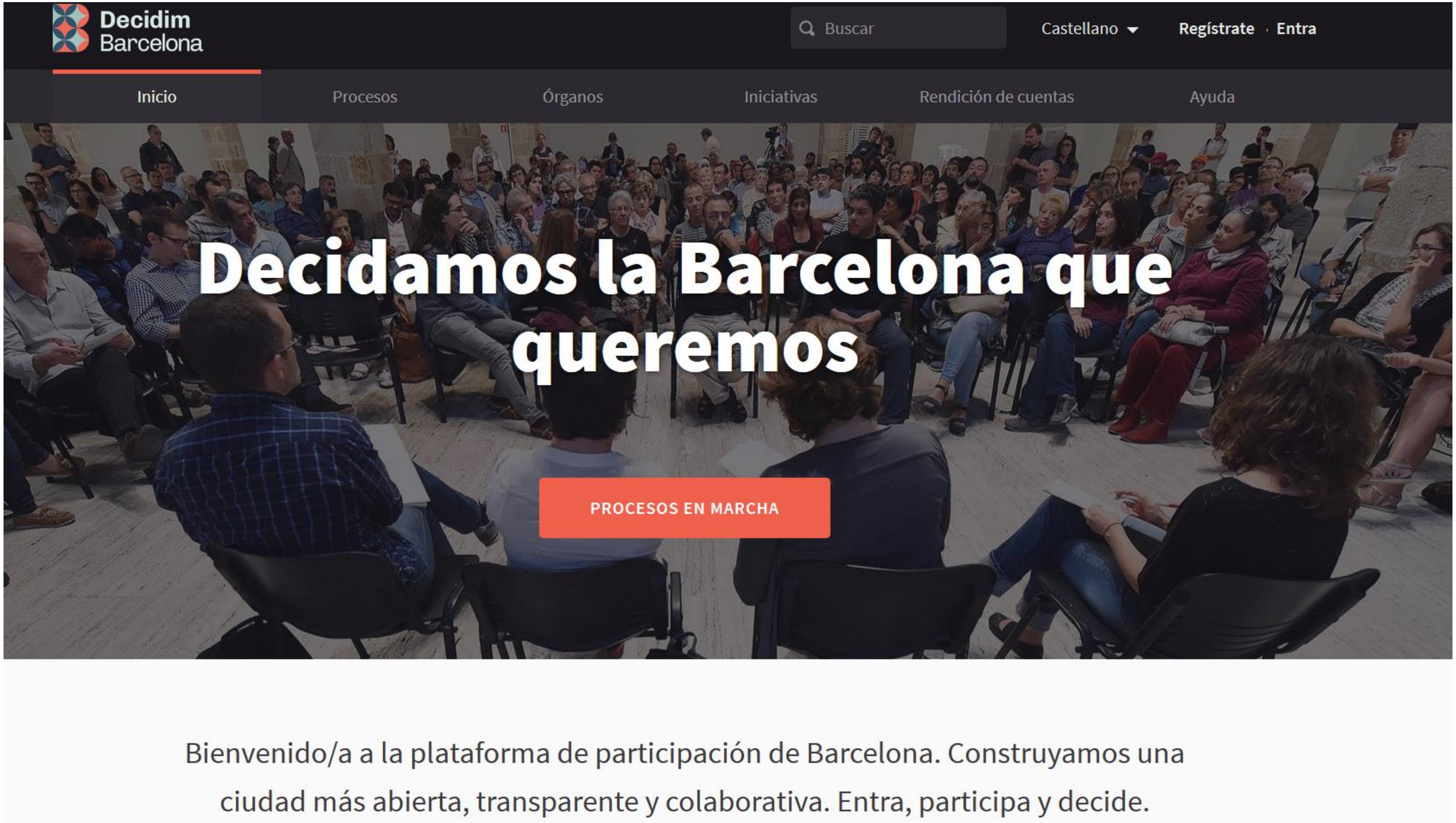


Country-Capital

Decidim Barcelona



Decidim BCN



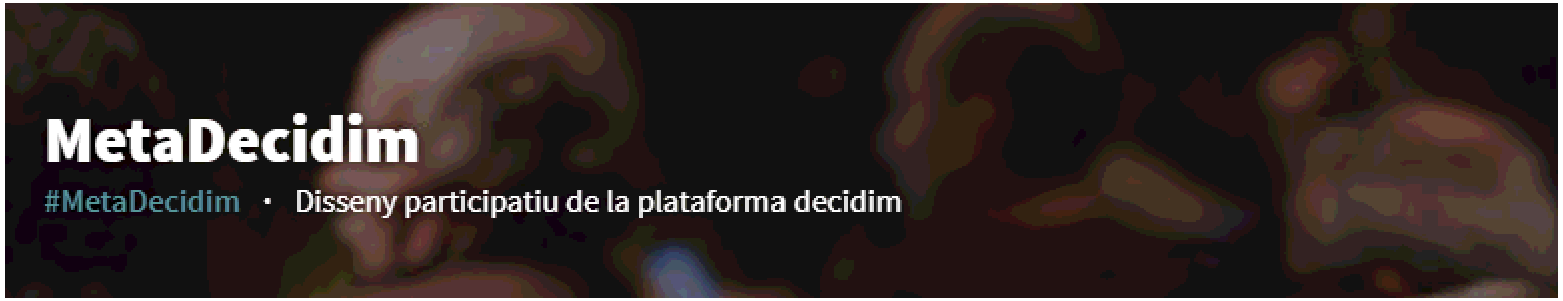
Decidamos la Barcelona que queremos

Bienvenido/a a la plataforma de participación de Barcelona. Construyamos una ciudad más abierta, transparente y colaborativa. Entra, participa y decide.

Decidim BCN

Concepts

PAM* Flow


MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

1. Citizens, organizations and the city council write **proposals.**

→

2. Similar proposals are clustered into **actions.**

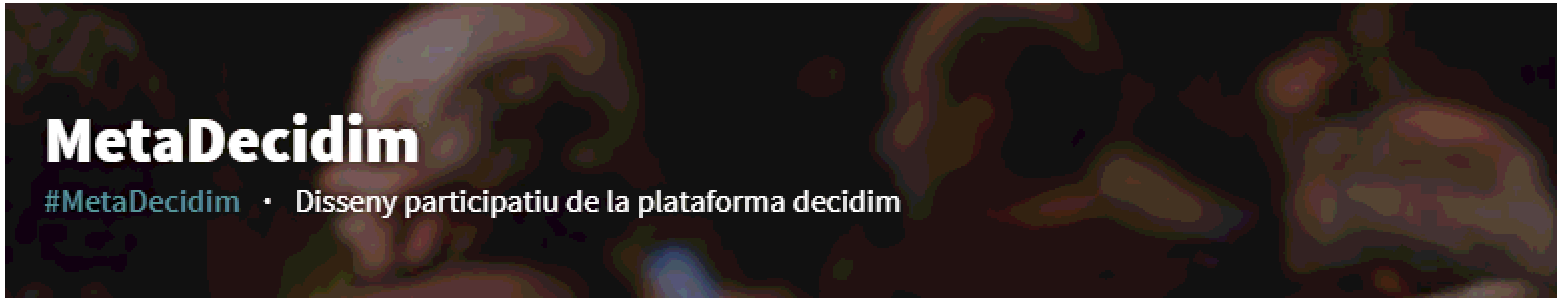
→

3. Some **actions** are approved and they go on to be executed.

*PAM is a strategic plan that establishes the actions
that the municipal government must implement during the corresponding political term.

Decidim BCN

Research questions


MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

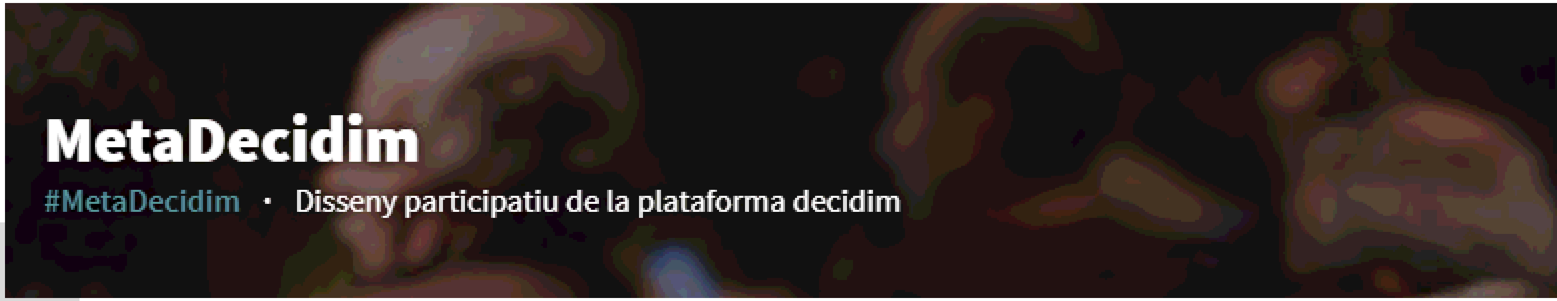
1. Do actions clearly **reflect** proposals ideas without considering authorship?

2. Do citizens **write** the same way as the Administration?

Can Machine Learning and Deep Learning answer those questions within a bilingual context?

First step

From text to word embeddings



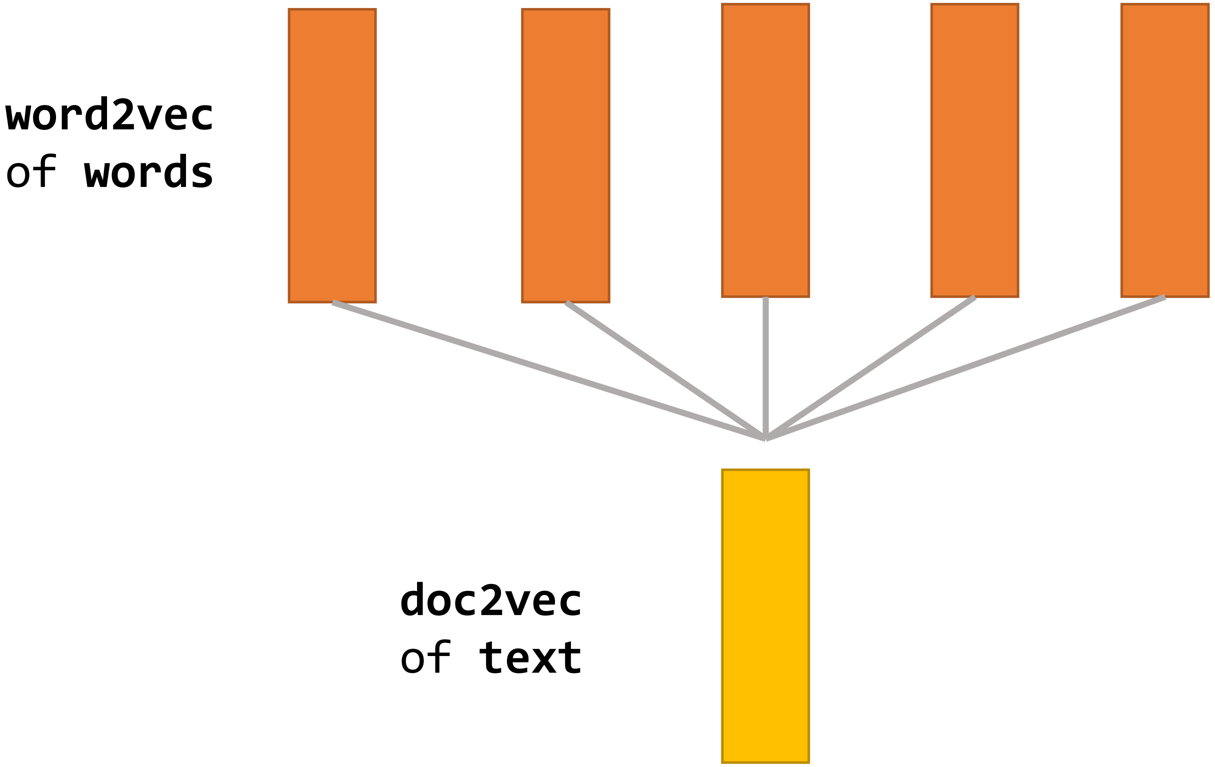
MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

doc2vec

Mean of all word2vec of words within a text.

“Crear espacios nuevos para perros.”

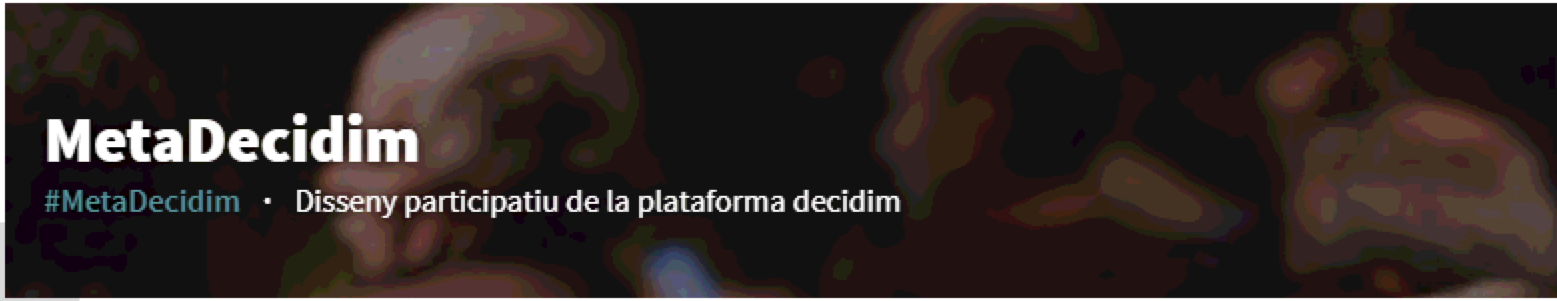


Decidim

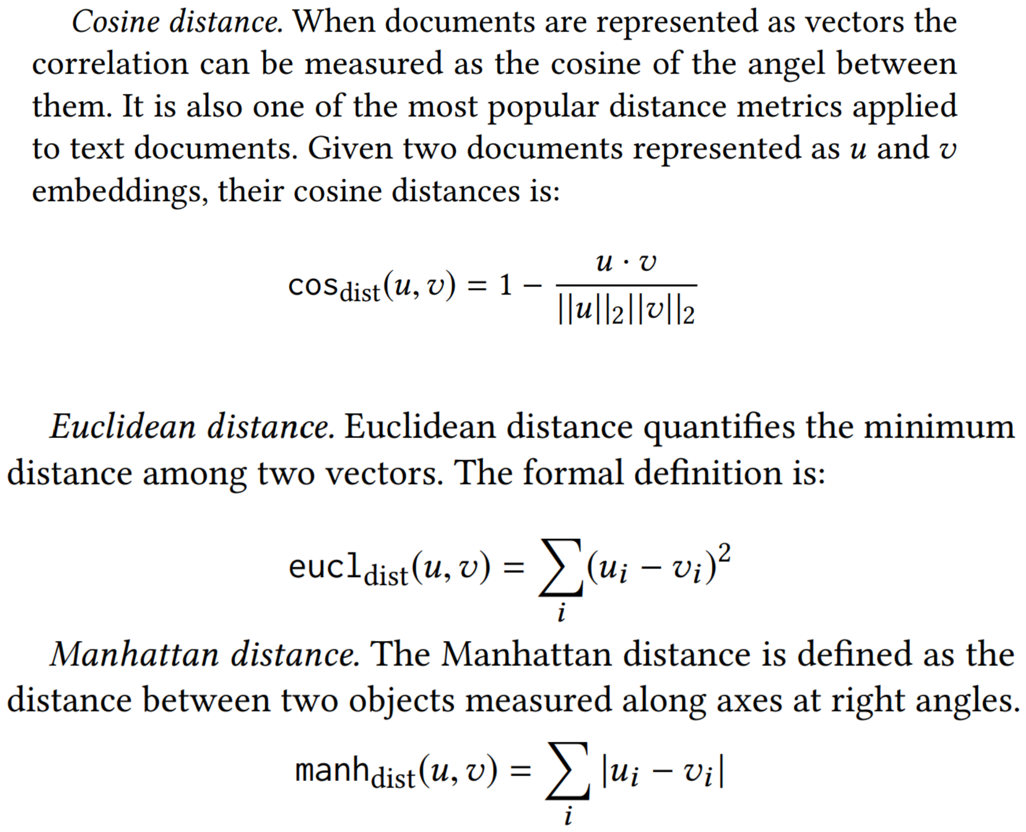
BCN

Second step

Analyze distances


MetaDecidim
#MetaDecidim · Disseny participatiu de la plataforma decidim

When documents are represented as vectors, several distance functions can be used to reflect the degree of closeness between two of them.



Cosine distance. When documents are represented as vectors the correlation can be measured as the cosine of the angel between them. It is also one of the most popular distance metrics applied to text documents. Given two documents represented as u and v embeddings, their cosine distances is:

$$\texttt{cos}_{\text{dist}}(u, v) = 1 - \frac{u \cdot v}{||u||_2||v||_2}$$

Euclidean distance. Euclidean distance quantifies the minimum distance among two vectors. The formal definition is:

$$\texttt{eucl}_{\text{dist}}(u, v) = \sum_i (u_i - v_i)^2$$

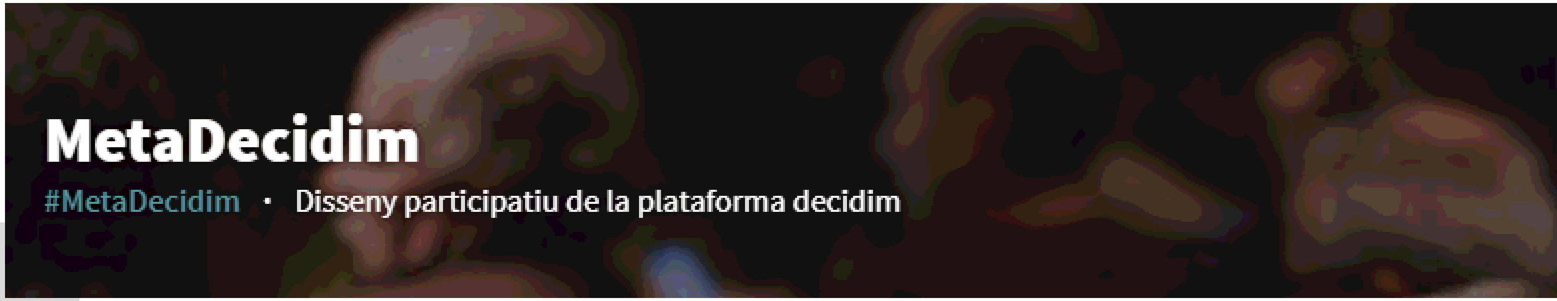
Manhattan distance. The Manhattan distance is defined as the distance between two objects measured along axes at right angles.

$$\texttt{manh}_{\text{dist}}(u, v) = \sum_i |u_i - v_i|$$

Decidim BCN

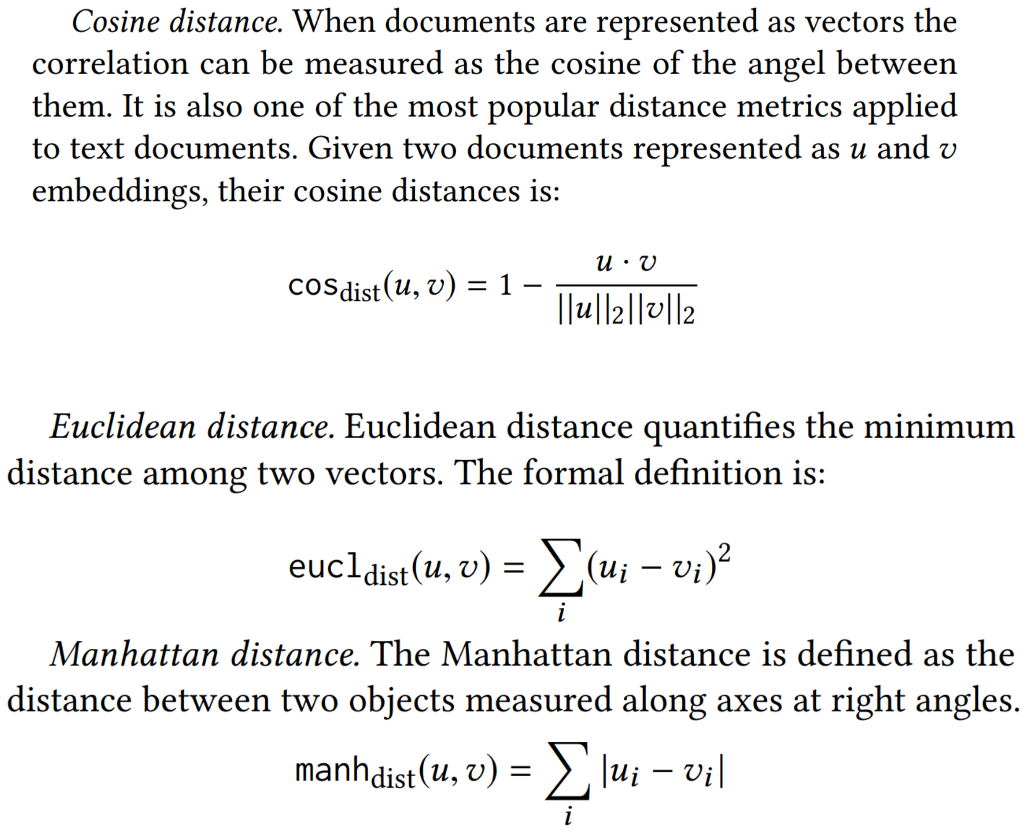
Second step

Analyze distances


MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

When documents are represented as vectors, several distance functions can be used to reflect the degree of closeness between two of them.


Cosine distance. When documents are represented as vectors the correlation can be measured as the cosine of the angel between them. It is also one of the most popular distance metrics applied to text documents. Given two documents represented as u and v embeddings, their cosine distances is:

$$\mathtt{cos}_{\mathrm{dist}}(u, v) = 1 - \frac{u \cdot v}{||u||_2||v||_2}$$

Euclidean distance. Euclidean distance quantifies the minimum distance among two vectors. The formal definition is:

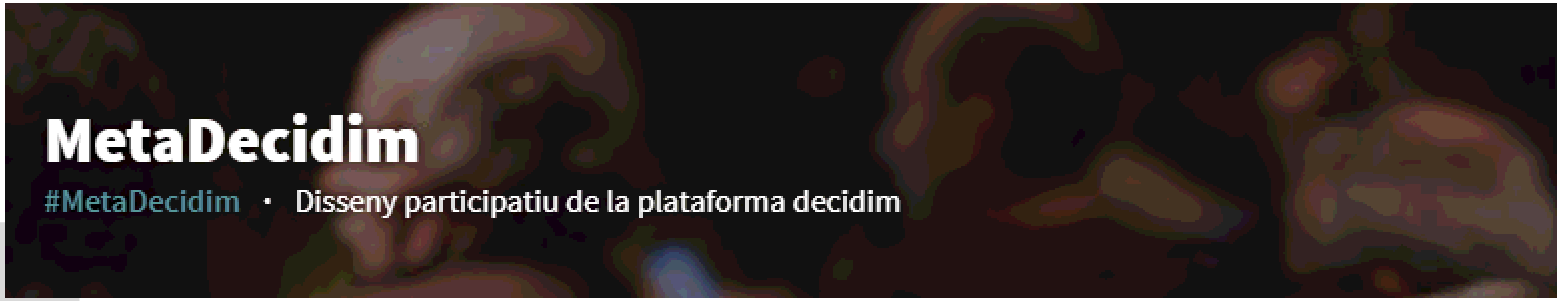
$$\mathtt{eucl}_{\mathrm{dist}}(u, v) = \sum_i (u_i - v_i)^2$$

Manhattan distance. The Manhattan distance is defined as the distance between two objects measured along axes at right angles.

$$\mathtt{manh}_{\mathrm{dist}}(u, v) = \sum_i |u_i - v_i|$$

Second step

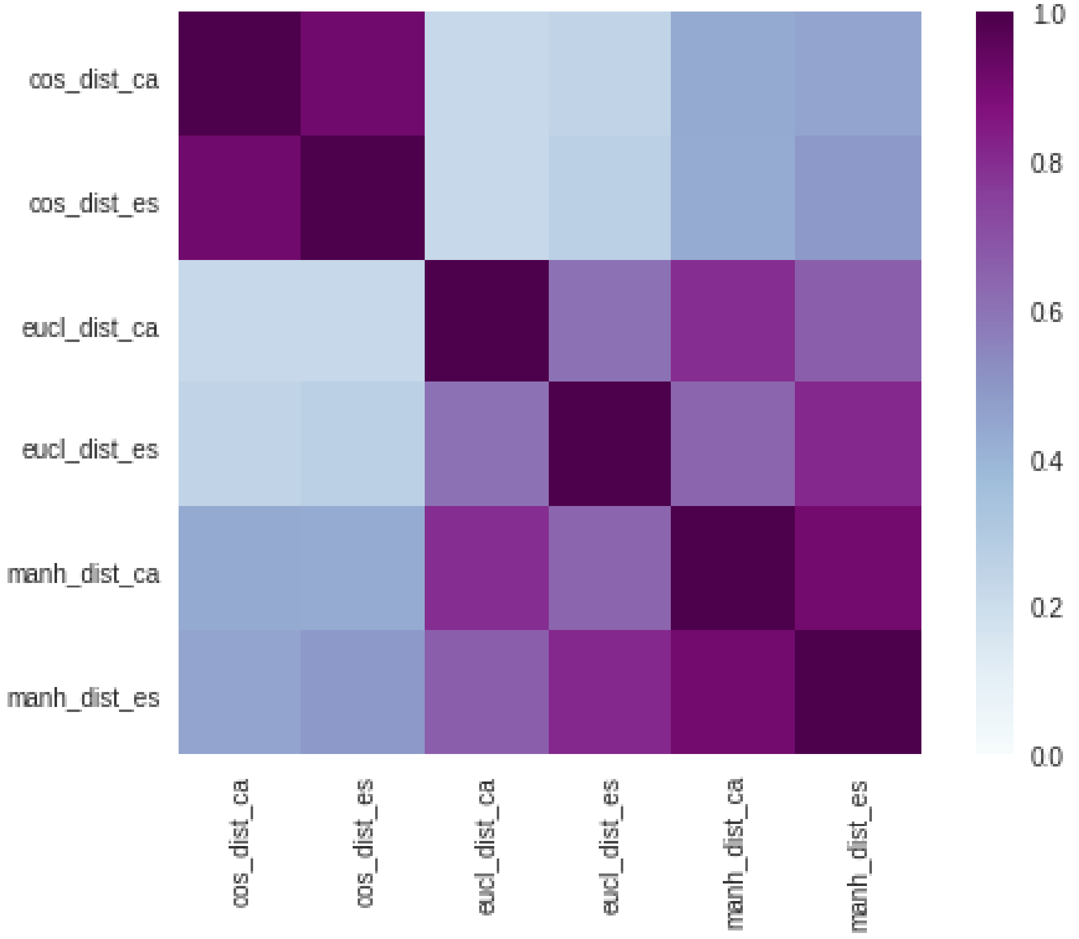
Analyze distances



MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

When documents are represented as vectors, several distance functions can be used to reflect the degree of closeness between two of them.

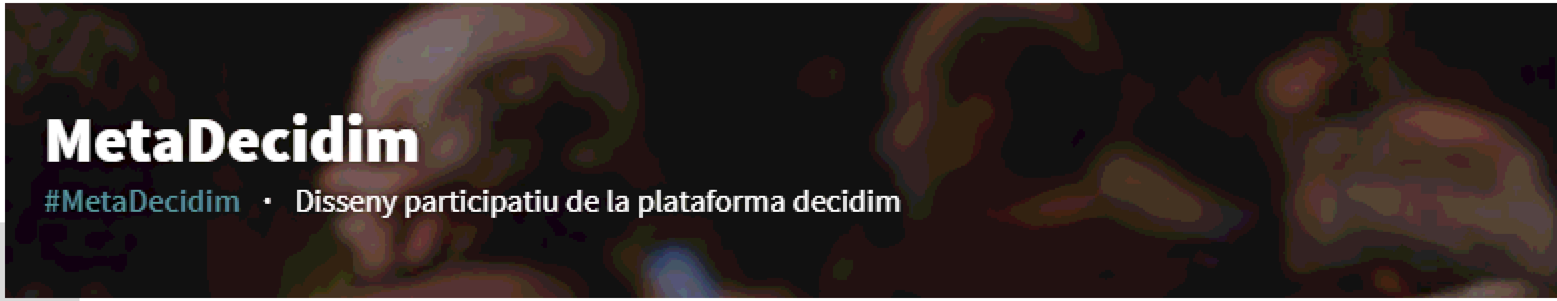


Decidim

BCN

Second step

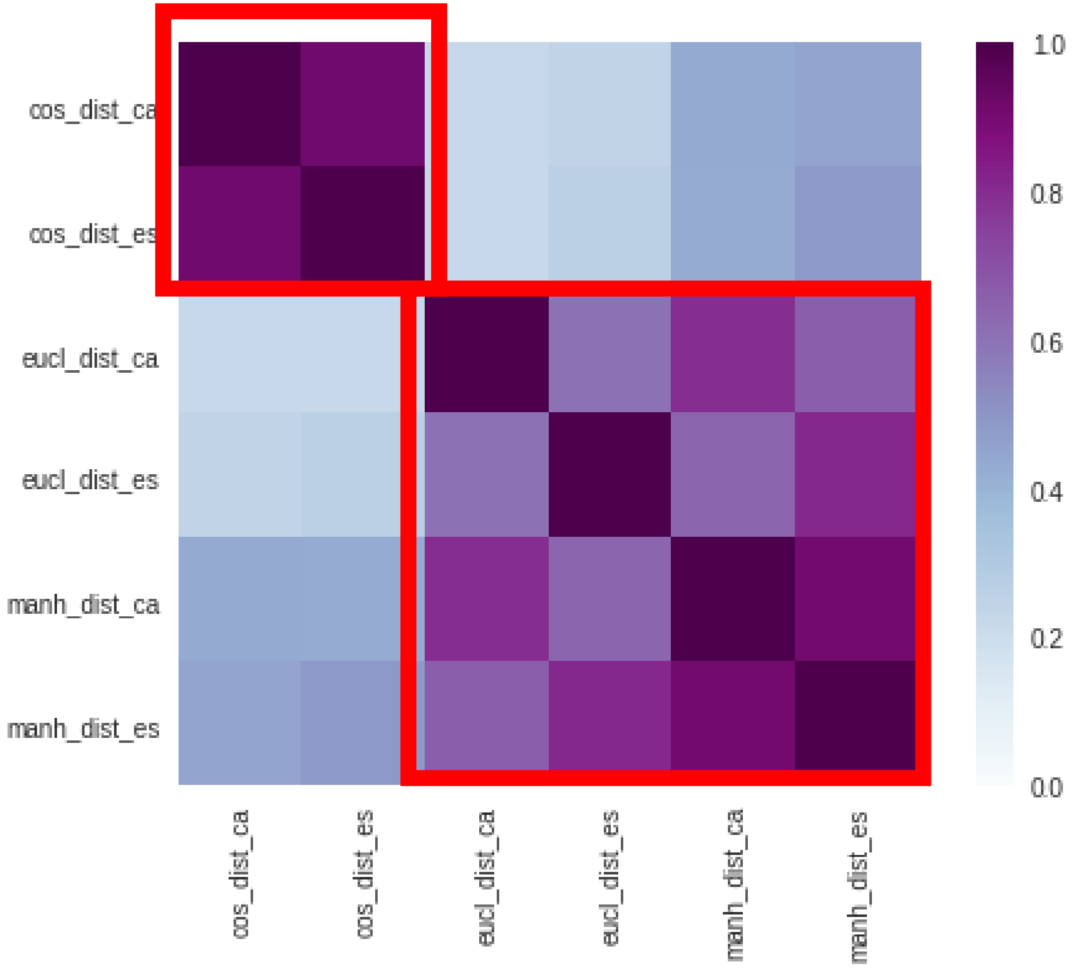
Analyze distances



MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

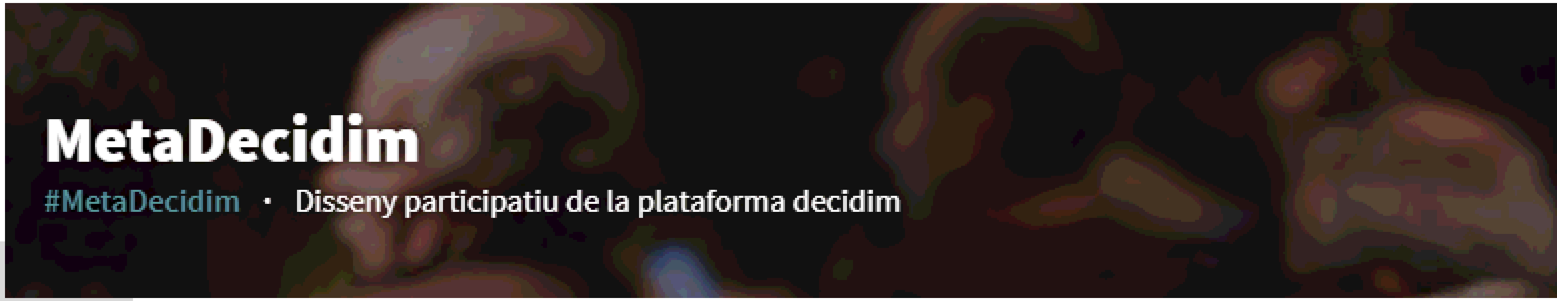
Same distance functions are highly correlated considering both languages, which implies that **doc2vec representations are equivalent** either in Catalan and Spanish



Decidim BCN

Third step

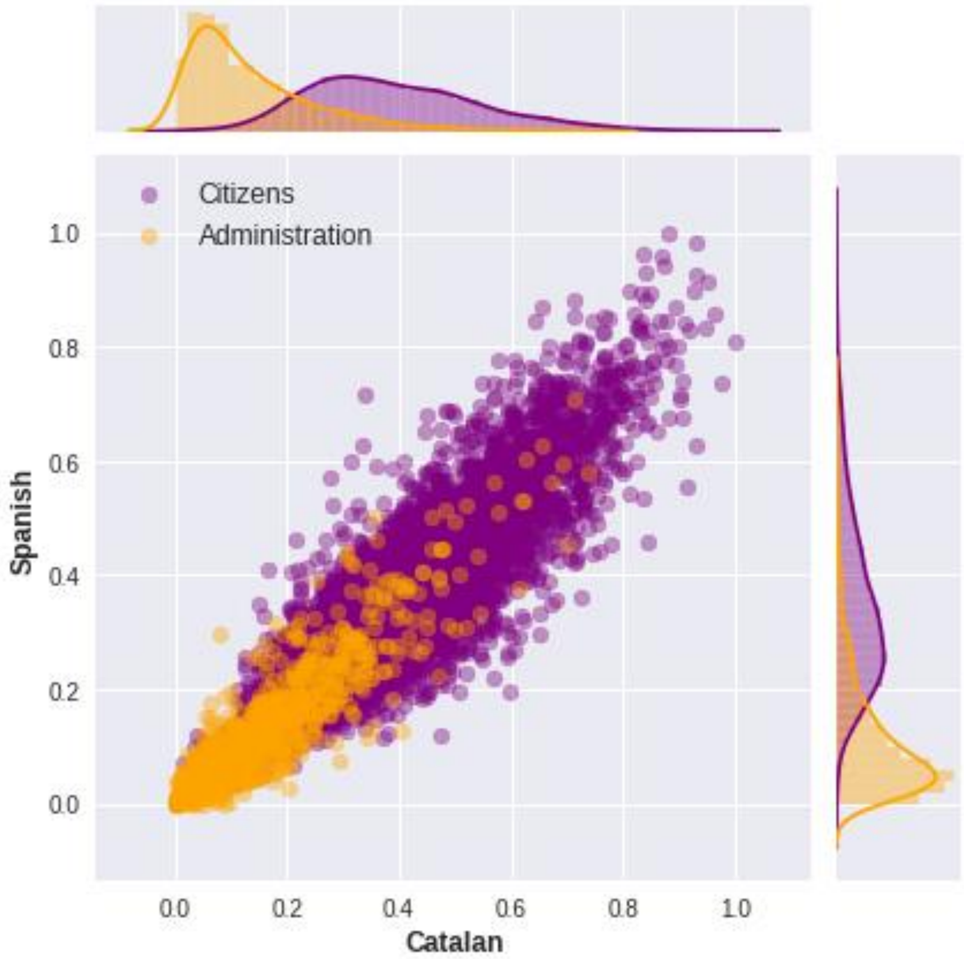
Bias!


MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

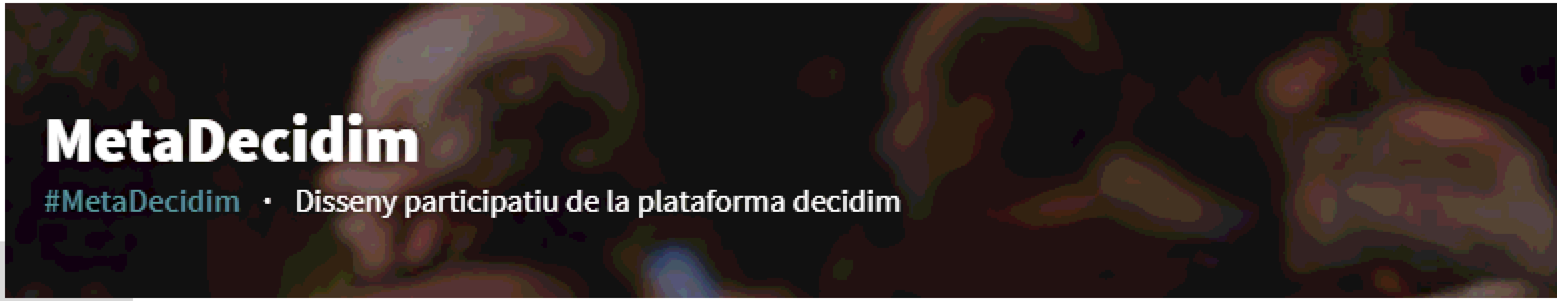
Research question

1. Do actions clearly **reflect** proposals ideas without considering authorship?



Decidim BCN

Third step



MetaDecidim

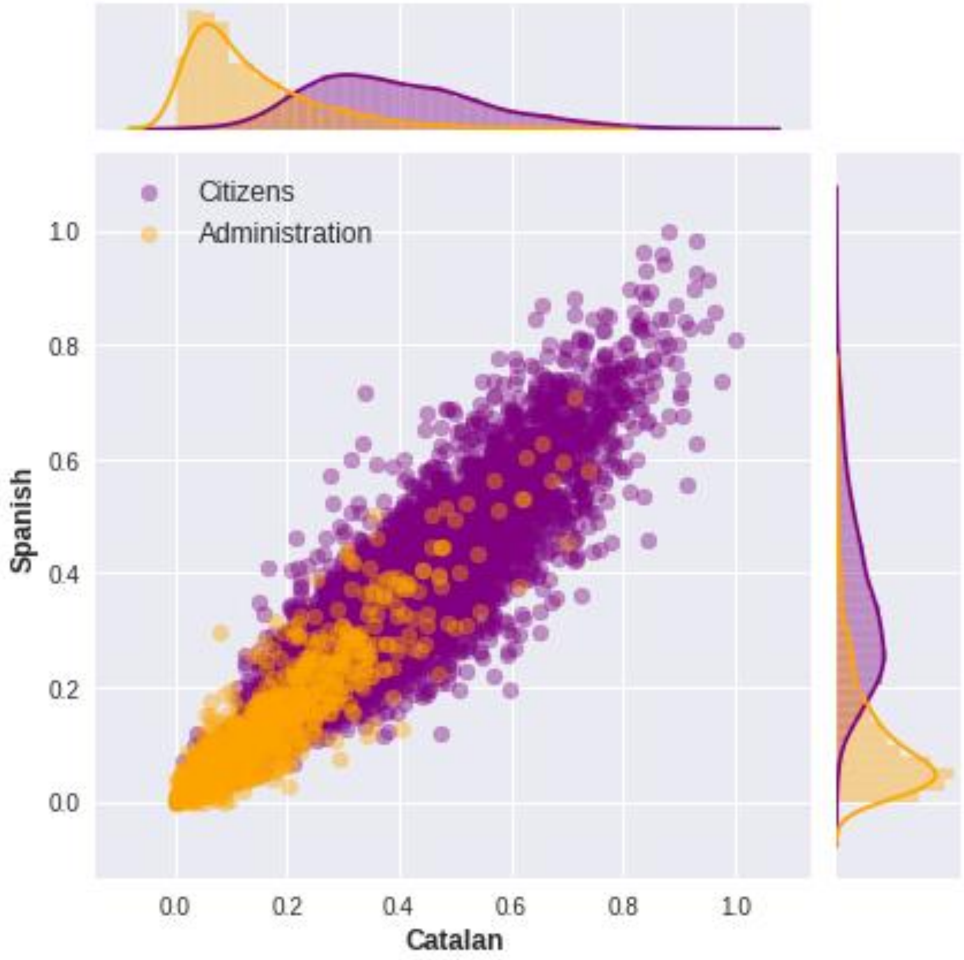
#MetaDecidim · Disseny participatiu de la plataforma decidim

Bias!

Research question

1. Do actions clearly **reflect** proposals ideas without considering authorship?

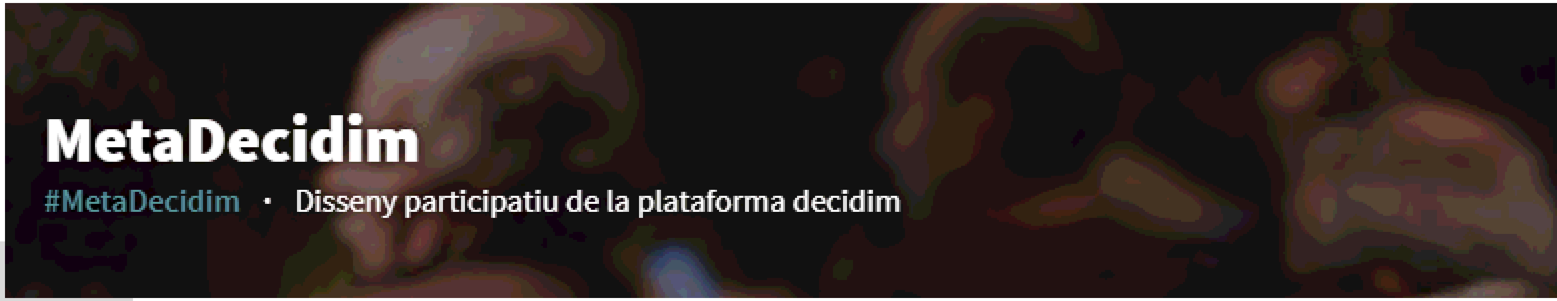
No!



Decidim BCN

Third step

Bias!



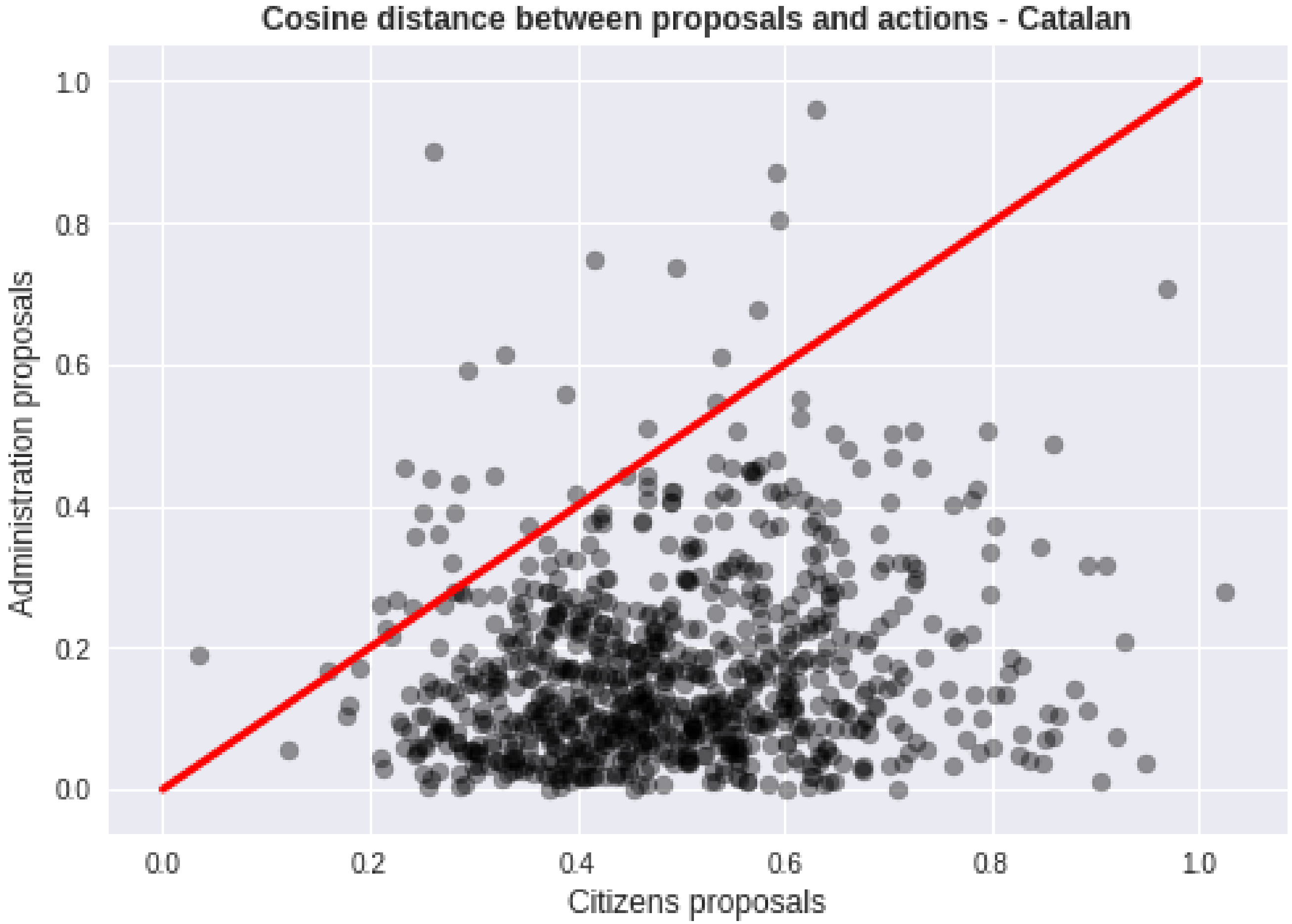
MetaDecidim

#MetaDecidim · Disseny participatiu de la plataforma decidim

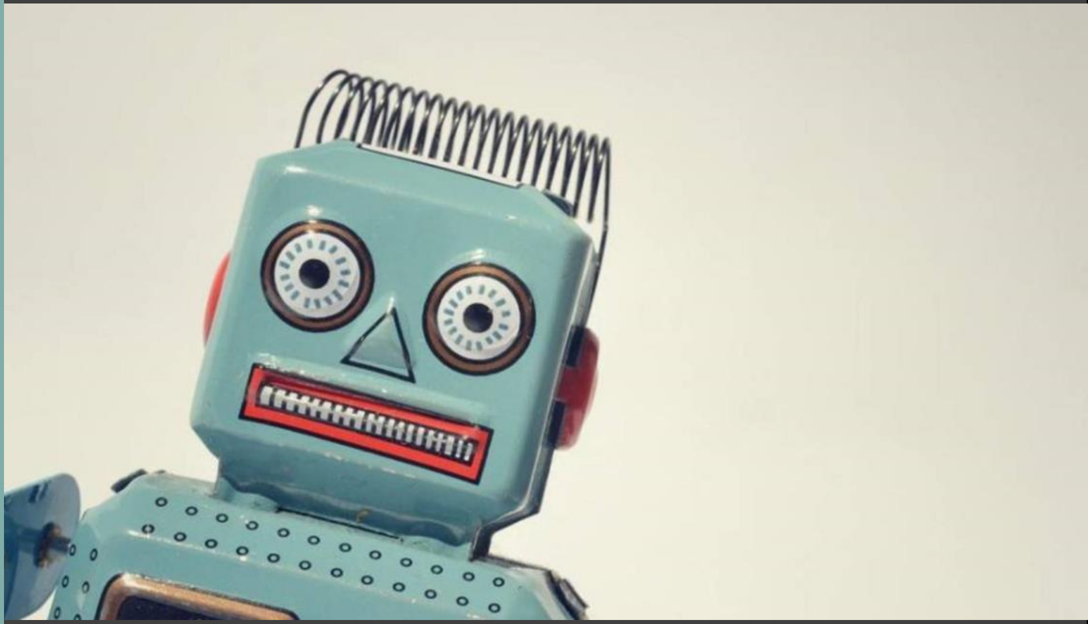
Research question

1. Do actions clearly **reflect** proposals ideas without considering authorship?

No!



Thanks!



BILBIOGRAPHY

- Mikolov, Tomas, et al. "**Distributed representations of words and phrases and their compositionality**." Advances in neural information processing systems. 2013.

- Rong, Xin. "**word2vec parameter learning explained**." *arXiv preprint arXiv:1411.2738* (2014).

- Bolukbasi, Tolga, et al. "**Man is to computer programmer as woman is to homemaker? debiasing word embeddings**." Advances in neural information processing systems. 2016.

- Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "**Semantics derived automatically from language corpora contain human-like biases**." Science 356.6334 (2017): 183-186.

- https://www.youtube.com/watch?v=ERibwqs9p38&t=3110s
- https://github.com/genimarca/caepia2018_tutorial_nlp_sa/blob/master/2018_caepia_tutorial_nlp.pdf