uc3m | Universidad Carlos III de Madrid

HULAT Human Language & Accessibility Technologies

Cutting-edge Deep Learning for NLP learners

Text Representation

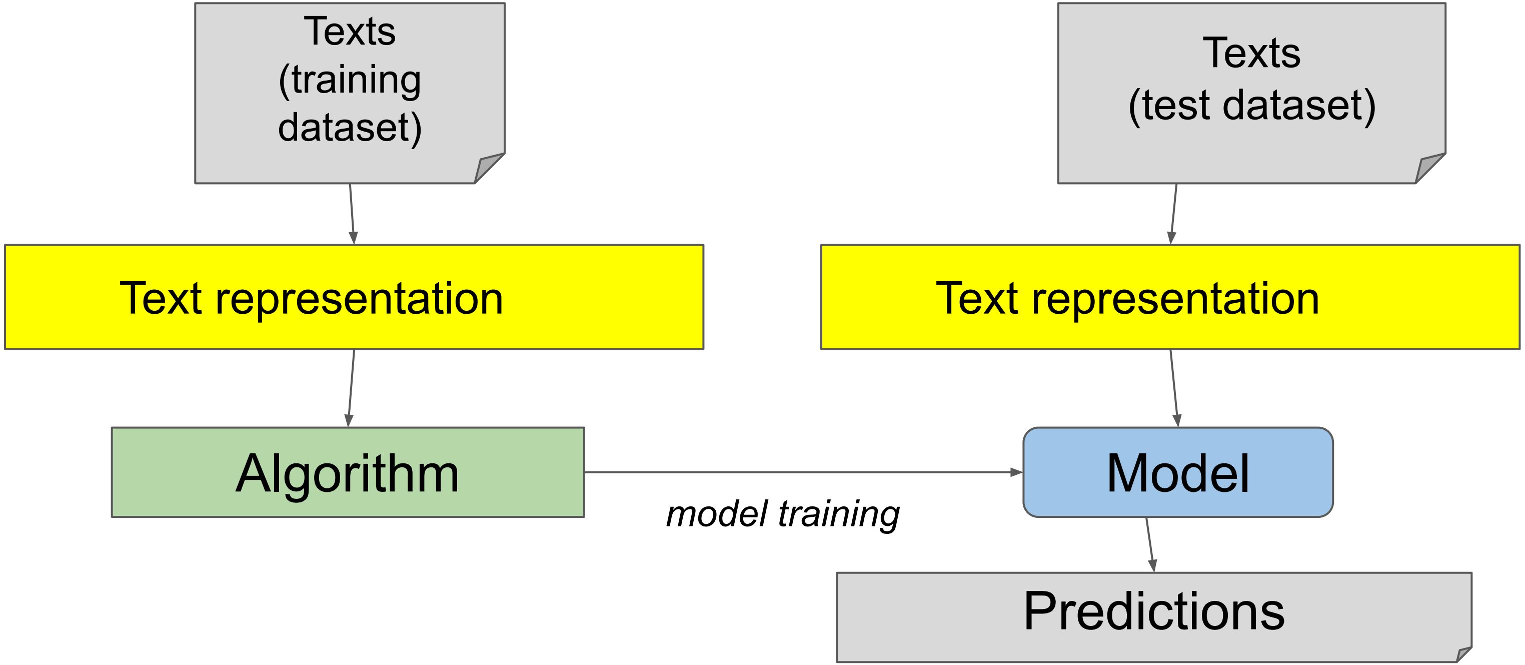
Isabel Segura-Bedmar

6 June 2022
Máster Universitario en Inteligencia Artificial, UPM

Outline

- Text representation
  - Bag-Of-Words
  - TF-IDF
  - Spacy vectors
  - Word embeddings

Machine Learning Pipeline



Text representation

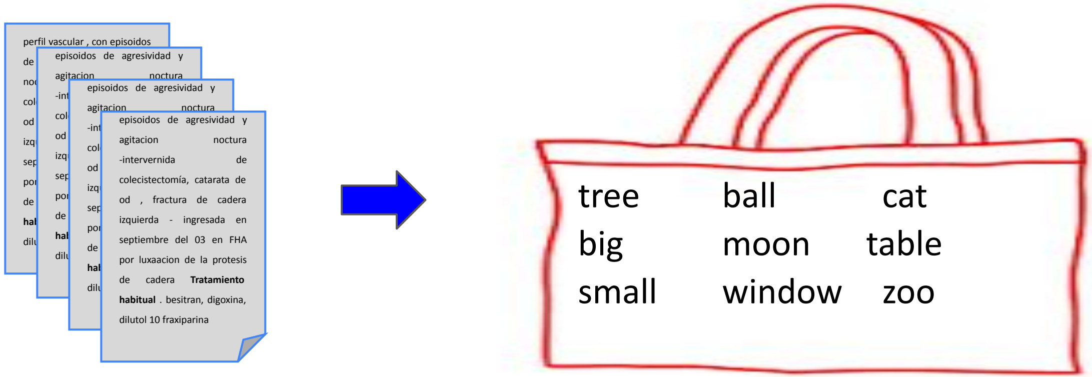
- Represents text in a vector of numbers
- Techniques:
    - **Bag-Of-Words**
    - **TF-IDF**
    - Word Embeddings
    - Contextualized word embeddings.

Bag of Words

- based on counting words in the document
- Steps:
    - Cleaning:
        - Remove stopwords, punctuation and special symbols.
        - Normalize texts (lemmatization or stemming).

Bag of Words

- Cleaning
- Obtain vocabulary (unique words) from all texts.



perfil vascular , con episoidos

episoidos de agresividad y agitacion noctura -intervernida de colecistectomía, catarata de od , fractura de cadera izquierda - ingresada en septiembre del 03 en FHA por luxaacion de la protesis de cadera **Tratamiento habitual** . besitran, digoxina, dilutol 10 fraxiparina

tree ball cat
big moon table
small window zoo

ball	big	cat	moon	small	table	tree	window	zoo

Bag of Words

Each text is represented as a vector with the frequencies of their words

D: The big cat in on the table and the small cat in the window.

after cleaning:

D: ~~The~~ big cat ~~is on the~~ table ~~and the~~ small cat ~~en la~~ window~~.~~

Vector (features):

ball	big	cat	moon	small	table	tree	window	zoo
0	1	2	0	1	1	0	1	0

Bag of Words

D1: ~~The~~ big cat ~~is on the~~ table ~~and the~~ small cat ~~in the~~ window
D2: ~~The~~ table ~~and the~~ window ~~are~~ small
D2: ~~The~~ moon ~~and the~~ small tree ~~are~~ big

	ball	big	cat	moon	small	table	tree	window	zoo
D1	0	1	2	0	1	1	0	1	0
D2	0	0	0	0	1	1	0	1	0
D3	0	1	0	1	1	0	1	0	0

TF-IDF

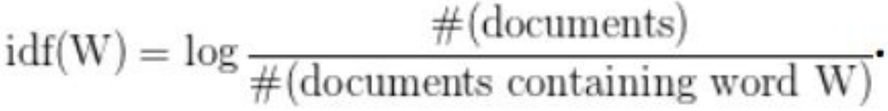
- Extended version of BoW.
- Every text is represented using tf-idf of its words
- TF-IDF **decreases the weight of the very common words** in the collection of texts

TF-IDF

- Term frequency - inverse document frequency.

TF-IDF (W) = TF(W,d) * IDF(W)

  - TF(W,d) = term frequency of the word *W* in the document d.
  - IDF(W) = inverse document frequency. The logarithm of the quotient of the total number of documents and the number of documents that contains the word *W*.



$$\mathrm{idf}(\mathrm{W}) = \log \frac{\#(\text{documents})}{\#(\text{documents containing word W})}.$$

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Bag of Words

	ball	big	cat	moon	small	table	tree	window	zoo
D1	0	1	2	0	**1**	1	0	1	0
D2	0	0	0	0	**1**	1	0	1	0
D3	0	1	0	1	**1**	0	0	0	0

TF-IDF (W) = TF(W,d) * IDF(W)

	ball	big	cat	moon	small	table	tree	window	zoo
D1	0	0.17	0.95	0	**0**	0.17	0	0.17	0
D2	0	0	0	0	**0**	0.17	0	0.17	0
D3	0	0.17	0	0.47	**0**	0	0.47	0	0

Drawbacks of traditional approaches

- Have high dimensionality and are very sparse.
- Don’t capture semantics
  - *Edema de glotis != hinchazón de la laringe*
- Don’t position of occurrence of words
  - *The hotel was very good and not expensive !=*
  - *The hotel was very expensive and not good*

Figure taken from https://laptrinhx.com/word-embeddings-part-2-3902951508/

Hands-on

- Vectorization (BoW and TF-IDF)
- Spam-detection I

Thank you
Question time!!!

isegura@inf.uc3m.es
https://hulat.inf.uc3m.es/nosotros/miembros/isegura
https://github.com/isegura