di.unito.it

DIPARTIMENTO DI INFORMATICA

Commission for Orientation and Informatics in schools

#DeactivHate

ARTIFICIAL INTELLIGENCE AND
COMPUTATIONAL LINGUISTICS
TO COUNTER THE SPREAD
OF HATEFUL MESSAGES ONLINE

Automatic text classification

What we need to do:

1. ~~Define a task to solve~~
2. ~~Collect a dataset of texts~~
3. ~~Create a training set of annotated texts and a test set~~
4. Represent texts using features
5. Use a machine learning algorithm
6. Create the NLP model
7. Evaluate the model
8. Automatically label new texts you've never seen before

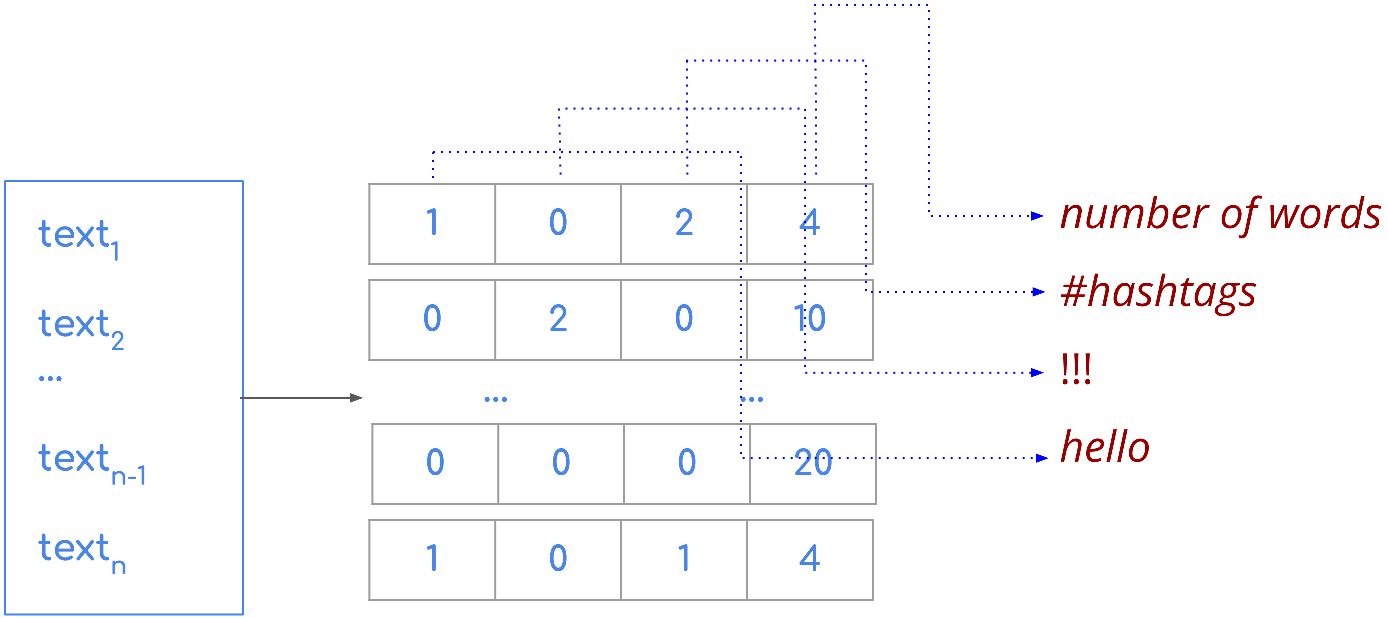
4.1. From texts to features

Computers have difficulty working with texts, they prefer to deal with numbers.

The simplest approach of all is to create **a vector representation of a text**.

4.1. From texts to features

Each text will be a row represented by multiple columns (**attributes or features**).
Merging multiple texts will create a matrix.



hello	!!!	#hashtags	number of words
1	0	2	4
0	2	0	10
…		…	
0	0	0	20
1	0	1	4

4.2. From texts to features

EASY

Numeric attributes extractable from the text

- Number of words
- Number of letters
- Number of punctuation points
- Average word length
- etc...

4.3. From texts to features

EASY

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND

id	#words	#charachters	#punctuation	Avg word length
1047919240848838656	3	12	1	4
1055101652557094913	?	?	?	?

4.3. From texts to features

EASY

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND

id	#words	#charachters	#punctuation	Avg word length
1047919240848838656	3	12	1	4
1055101652557094913	4	18	0	4.5

4.4. From texts to features

NORMAL

Bag of Word (BoW)

Every word of the Italian language becomes an attribute/feature.

Feature value = 1 **if the word is contained** in the text.

Feature value = 0 if the word **is NOT contained** in the text.

4.5. From texts to features

NORMAL

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND

id	hello	is	round	the	world	!
1047919240848838656						
1055101652557094913						

4.5. From texts to features

NORMAL

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND

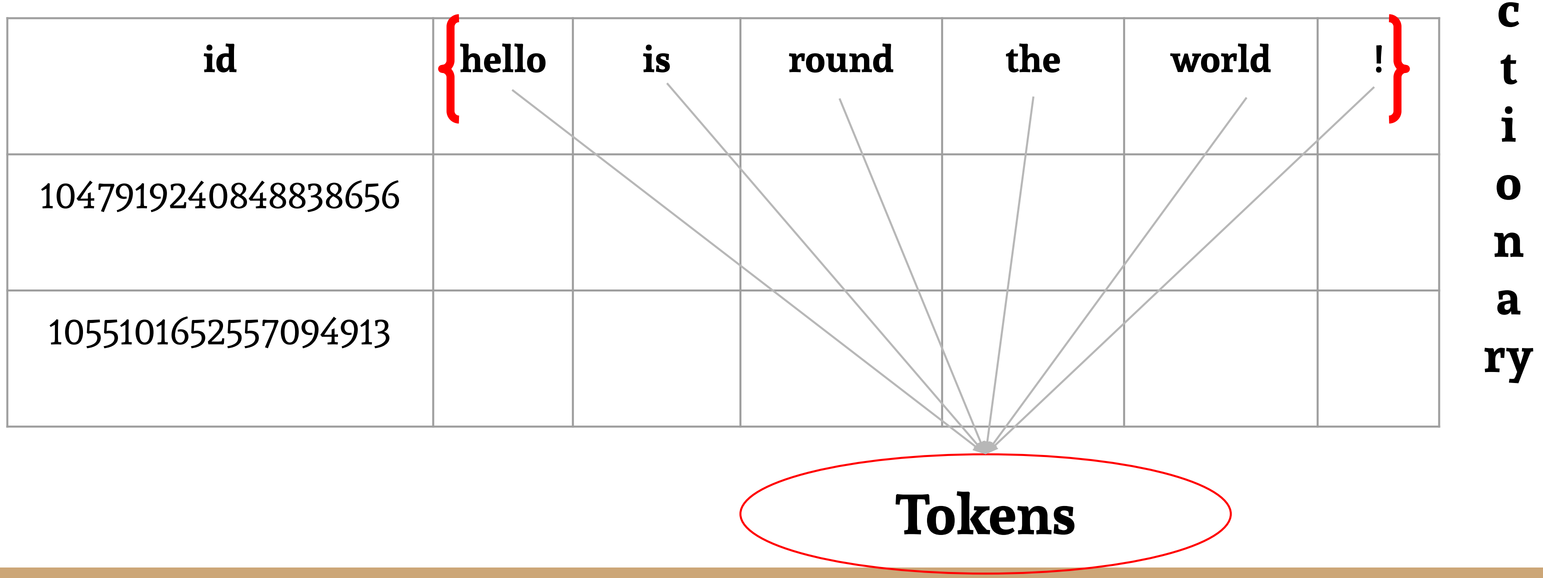
id	hello	is	round	the	world	!
1047919240848838656						
1055101652557094913						

Tokens

4.5. From texts to features

NORMAL

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND



id	hello	is	round	the	world	!
1047919240848838656						
1055101652557094913						

Dictionary

Tokens

4.6. From texts to features

NORMAL

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND

id	hello	is	round	the	world	!
1047919240848838656						
1055101652557094913						

4.6. From texts to features

NORMAL

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND

id	hello	is	round	the	world	!
1047919240848838656	1	0	0	0	1	1
1055101652557094913						

4.6. From texts to features

NORMAL

text id	text
1047919240848838656	HELLO WORLD!
1055101652557094913	THE WORLD IS ROUND

id	hello	is	round	the	world	!
1047919240848838656	1	0	0	0	1	1
1055101652557094913	0	1	1	1	1	0

4.7. From texts to features

The dataset will become an **NxM matrix** (N rows, M columns) with 1 row for each text and 1 column for each word/token contained in the vocabulary

id	**token 1**	**token 2**	**token 3**	**...**	**token M**
text 1	1	1	0	…	1
...	…	…	…	…	…
text N	0	0	1	…	0

4.8. From texts to features

NORMAL

Example: NASA records the sound of wind on Mars for the first time

Bag of Words

1. **Unigrams:** the text is split into single word tokens

   [NASA, records, the, sound, of, wind, on, Mars, for, the, first, time]

2. **N-grams** (from 1 to 2), the text is divided into single-word tokens and 2-word pairs

   [NASA, records, the, sound, of, wind, on, Mars, for, the, first, time, NASA records, records the, the sound, sound of, of wind, wind on, on Mars, Mars for, for the, the first, first time]

NORMAL

4.8. From texts to features

Example: NASA records the sound of wind on Mars for the first time

Bag of Characters (BoC)

1. **N-character Grams**, the text is divided into character tokens of length 2, 3, 4 and 5 → SPACES ARE ALSO CONSIDERED

[NA, NAS, NASA, NASA , AS, ASA, ASA , ASA r, SA, SA , SA r, SA re,

etc... , etc... , etc... ,

t, ti, tim, time]