

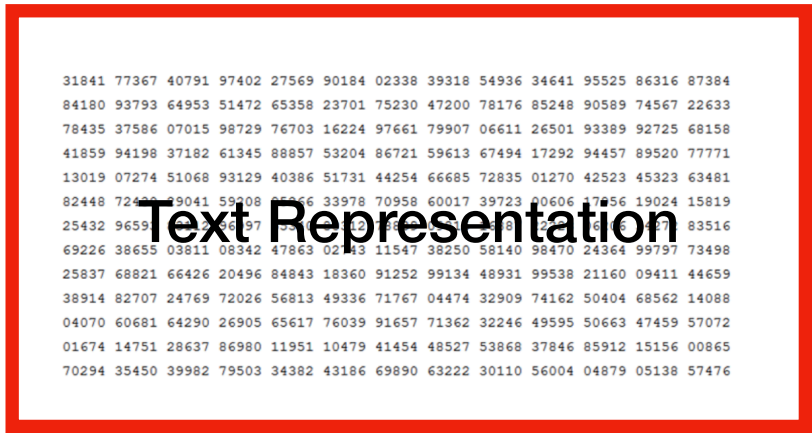
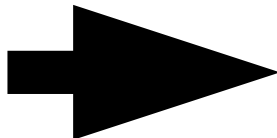
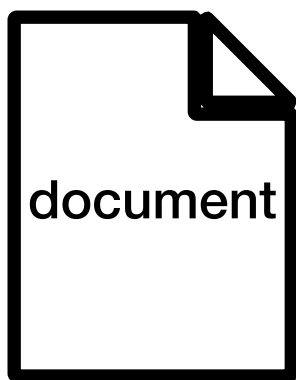
The background of the slide is a complex, abstract network diagram. It consists of numerous nodes, represented by small circles and hexagons, connected by thin, light blue lines. The nodes are distributed across the entire slide, with some clusters being more dense than others. The overall color scheme is light blue and white, giving it a clean, technical appearance.

TEXT REPRESENTATION LEARNING

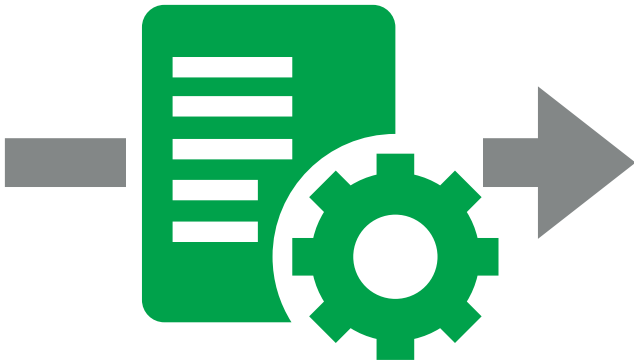
TEXT MINING AND NATURAL LANGUAGE PROCESSING
FOR COMPUTATIONAL SOCIAL SCIENCES

André Panisson

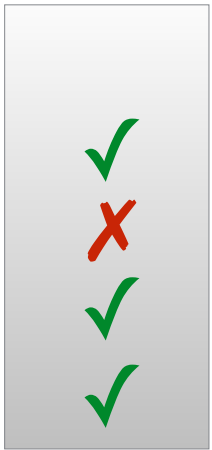
Text Representation



Feature extraction



Text classification task



Text Representation

- Three main approaches:
 - Bag of Words
 - Word embeddings
 - Contextual word embeddings

Bag of Words

- Each word is represented as a binary vector with D positions, where D is the size of the vocabulary (typically millions of tokens)
- The document is represented as the sum of the word vectors (counts of words)

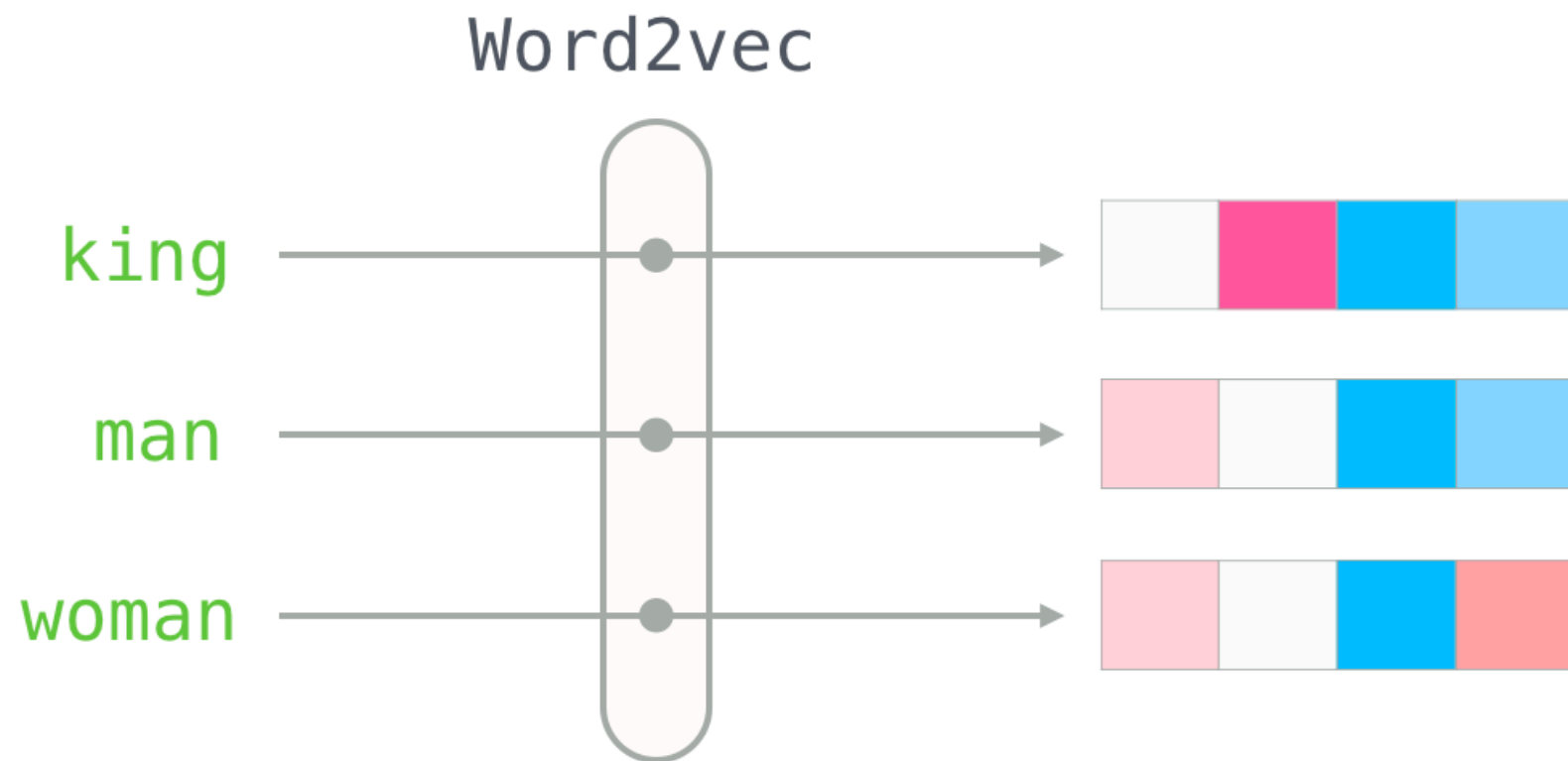
Document	the	cat	sat	in	hat	with
<i>the cat sat</i>	1	1	1	0	0	0
<i>the cat sat in the hat</i>	2	1	1	1	1	0
<i>the cat with the hat</i>	2	1	0	0	1	1

Bag of Words

- The dimensionality of the dictionary is huge, so a machine learning model needs to learn millions of parameters
- Each word is completely independent, there's no way to represent semantic relationships between words
(e.g. *amazing* and *awesome* are as similar as *amazing* and *terrible*)

Word embeddings

- Each word is represented as a vector with D real numbers
- D is much smaller than the dictionary size (typically hundreds)



Word embeddings

- Most popular implementation: Word2Vec
The vectors are obtained by training a neural network on the task of predicting a word from their neighbors.
- Very popular for translation tasks
(words in one language might have the same vector representation in another language)
- Word representations are independent of their context
E.g. *apple* (fruit) has exactly the same vector representation as *Apple* (company)

Contextual word embeddings

- Word representations are dependendent of their context.
- Most popular example: **Transformers**

Exercises