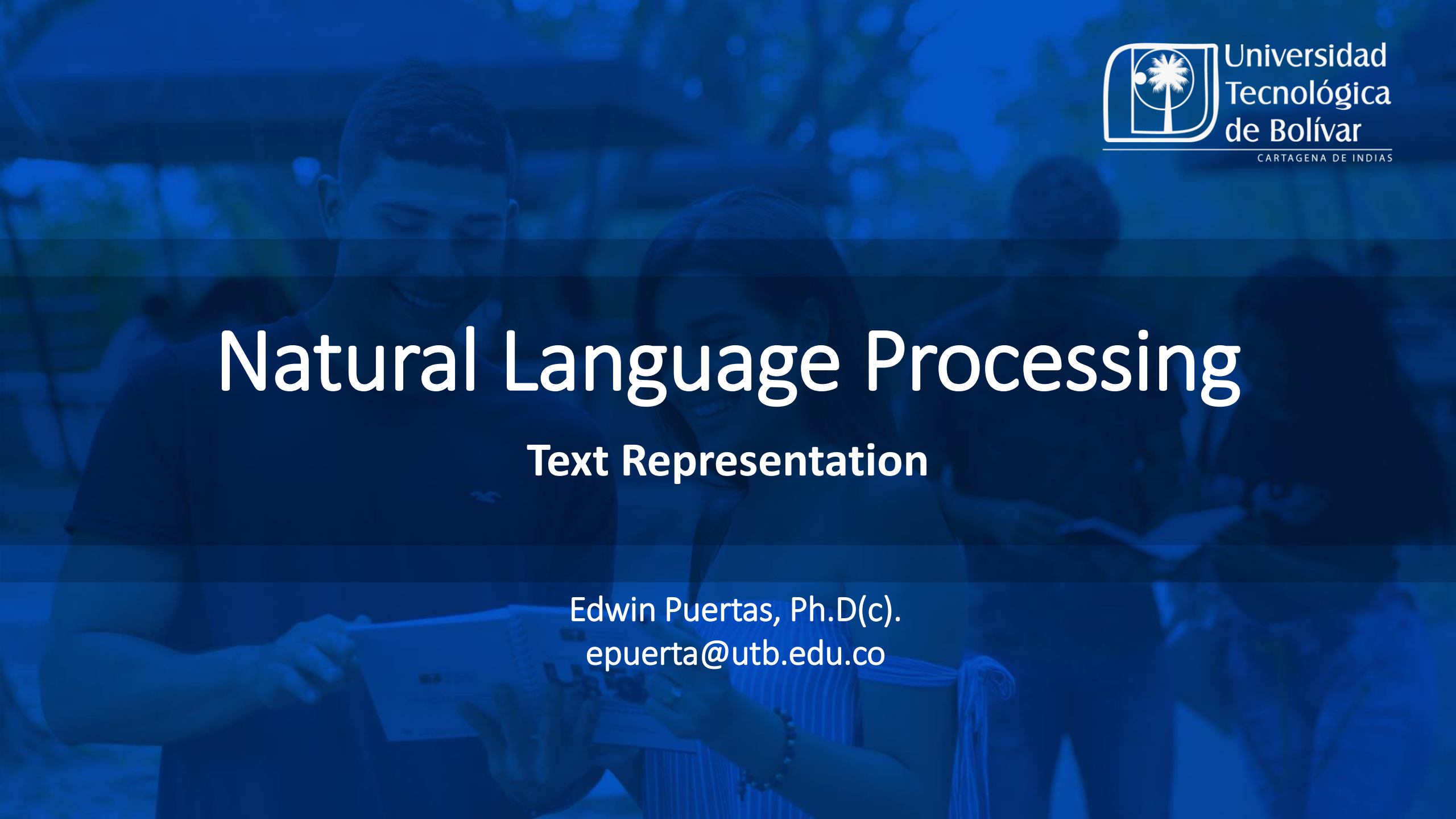

Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Natural Language Processing

Text Representation

Edwin Puertas, Ph.D(c).
epuerta@utb.edu.co

Introduction

How do we transform a given text into numerical form so that it can be fed into NLP and ML algorithms?

Introduction

- The conversion of raw text to a suitable numerical form is called text representation.
- There are four categories for representing texts:
  - Basic vectorization approaches
  - Distributed representations
  - Universal language representation
  - Handcrafted features

Bag of Words

- Bag of words (BoW) is a classical text representation technique that has been used commonly in text classification problems.
- The key idea behind it is as follows:
  - Represent the text under consideration as a bag (collection) of words while ignoring the order and context.
  - The basic intuition behind it is that it assumes that the text belonging to a given class in the dataset is characterized by a unique set of words.
  - If two text pieces have nearly the same words, then they belong to the same bag (class).
  - Thus, by analyzing the words present in a piece of text, one can identify the class (bag) it belongs to.

Bag of words (BoW)

Very good drama although it appeared to have a few blank areas leaving the viewers to fill in the action for themselves. I can imagine life being this way for someone who can neither read nor write. This film simply smacked of the real world: the wife who is suddenly the sole supporter, the live-in relatives and their quarrels, the troubled child who gets knocked up and then, typically, drops out of school, a jackass husband who takes the nest egg and buys beer with it. 2 thumbs up… very very very good movie.

('the', 8),
(',', 5),
('very', 4),
('.', 4),
('who', 4),
('and', 3),
('good', 2),
('it', 2),
('to', 2),
('a', 2),
('for', 2),
('can', 2),
('this', 2),
('of', 2),
('drama', 1),
('although', 1),
('appeared', 1),
('have', 1),
('few', 1),
('blank', 1)
…..

Text Representation Models

- Bag of N-Grams
- TF-IDF
- Co-occurrence matrix
- Word2vec
- Transformer
- ELMO/BERT/XLNet