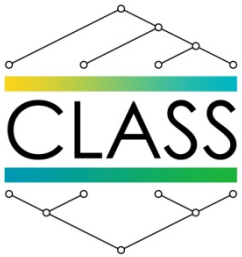



Co-funded by the
Erasmus+ Programme
of the European Union

# INTRODUCTION TO TEXT MINING

Prof. Jesús Vilares
jesus.vilares@udc.es

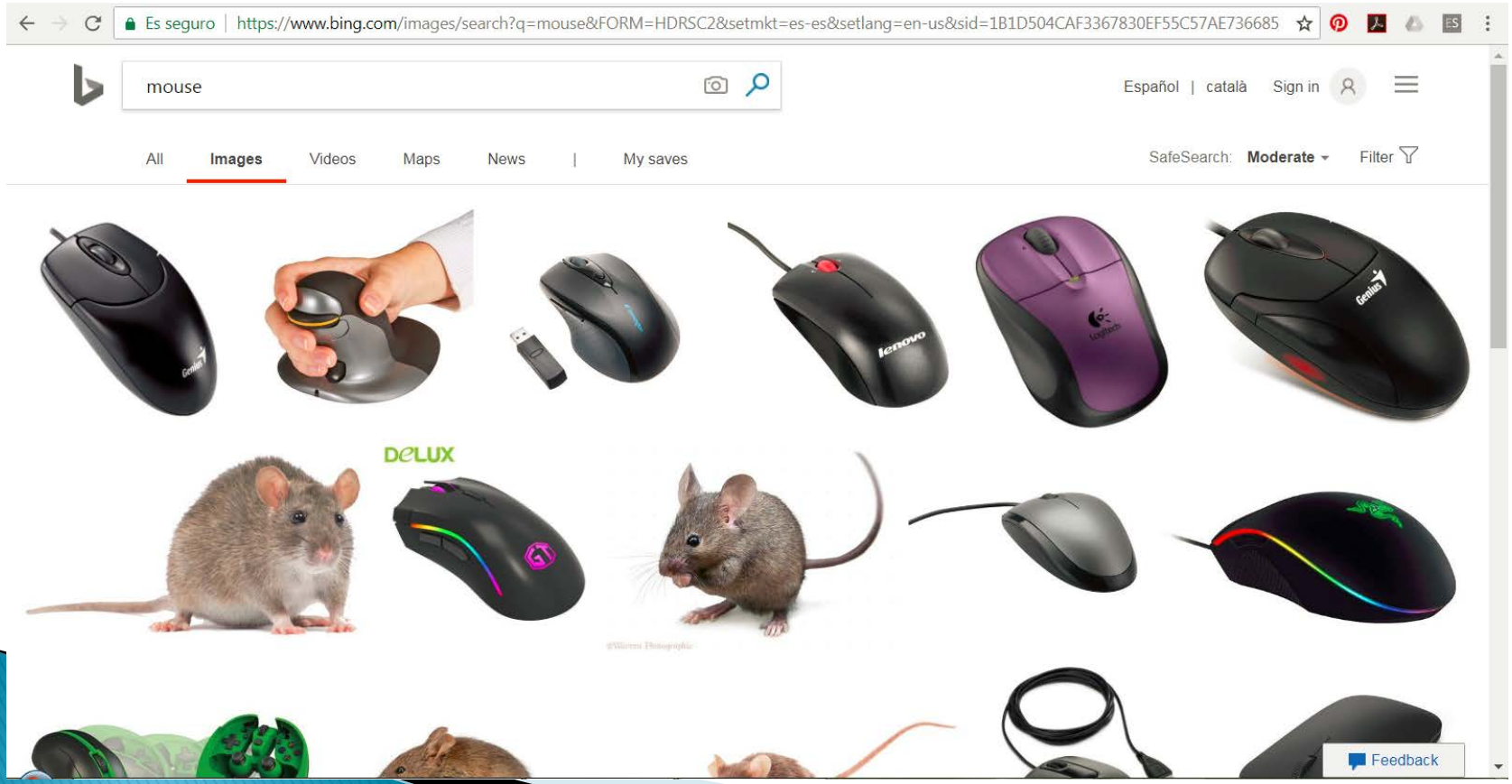
UNIVERSIDADE DA CORUÑA

# What is Text Mining ?

- **Text Mining** is the non-trivial process of analyzing [unstructured] text to obtain **high-quality information** (even **knowledge**) potentially useful for a particular purpose.
- This process may involve a wide range of tasks such as:
  - Information Retrieval
  - Text categorization
  - Text clustering
  - Information extraction
  - Question Answering
  - Document summarization
  - Sentiment Analysis
  - (…)

# Problem: linguistic variation

- The main problem when dealing with natural language content is the so-called **linguistic variation**, that is, how the same message can be encoded in different forms (and viceversa).



# Problem: linguistic variation

- Different **types/levels of variation**:
  - Morphological: ref. inflectional and derivational changes

    actor ⇆ actress  [to] act ⇆ actor
  - Lexical-Semantical: e.g. synonymic and polysemic phenomena

    adorable ⇆ charming ⇆ lovely

     mouse vs. mouse 
  - Syntactical: changes in the syntactic structure

    John attacked Mike ⇆ Mike was attacked by John

    vs.

    Mike attacked John
  - Mixes: e.g. morpho-syntactical variation

    climatic change ⇆ change of [the] climate

# Solution: NLP

- **Natural Language Processing (NLP) techniques** can solve or reduce the impact of linguistic variation
- In general, two types of approaches:
  1. **Normalization**: to conflate the different variants of a term/expression into a **common canonical form**
     - e.g. *lemmatization* (i.e. to replace a term by its *lemma*)

[she] sings
[we] sing
[we] sang
} → [to] sing

  2. **Expansion**: to extend the processing by also including the variants of a term/expression
     - e.g. to include the *synonyms* of a term in a web search:

# Bibliography


- **[Arampatzis et al., 2000]** Arampatzis, A., van der Weide, Th. P., van Bommel, P. & Koster, C.H.A. (2000). Linguistically-motivated Information Retrieval. In vol. 69 of *Encyclopedia of Library and Information Science*, pp. 201–222. Marcel Dekker.