Vectorial Representations: Parte 1

Dr. Adrián Pastor López Monroy
Investigador

pastor.lopez@cimat.mx
https://www.cimat.mx/es/adrián-pastor-lópez-monroy

Centro de Investigación en Matemáticas, A.C. (CIMAT)

some slides are adapted from: https://ccc.inaoep.mx/~mmontesg/cursos/Bogota/BeyondWords-UNAL2016.pdf

Overview

Motivation

Vectorial Representation of Documents

Vectorial Representation of Words (DTRs)

Introduction to Text Classification

- Framework
- Evaluation

Motivation of VSM

¿Cómo estás?

¿Qué onda?

La computación es super interesante

La computación es muy aburrida

La película no me pareció mala +

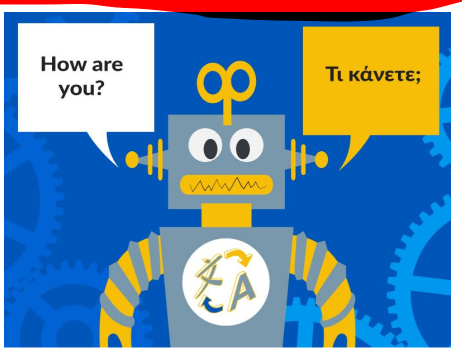
La película está lejos de ser buena −

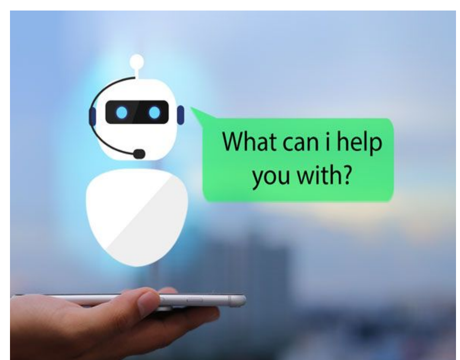
VSM Applications

- Estudiar computación requiere programar
- En el futbol gana el equipo que mete más goles

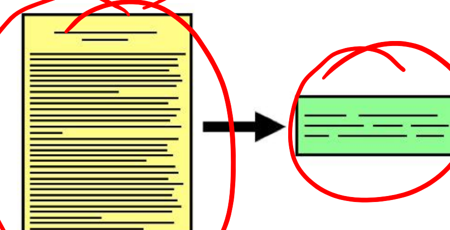
Nivel Palabras

¿?





Summarization

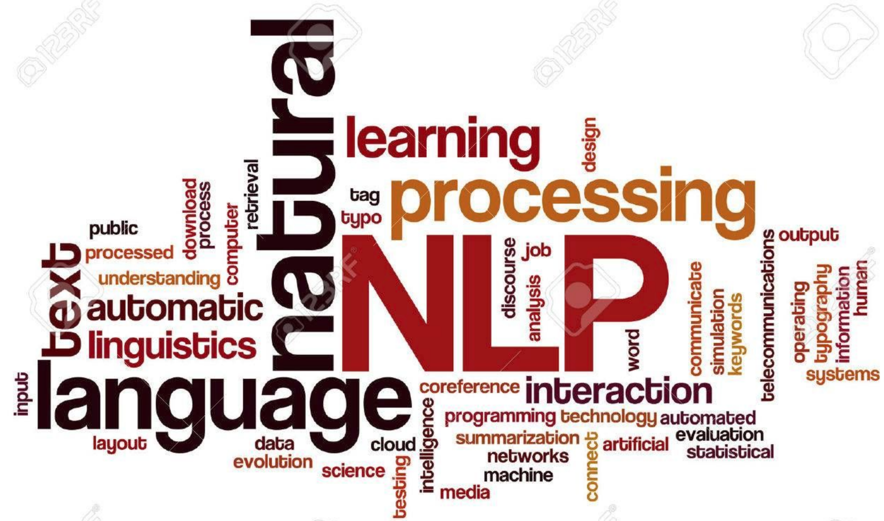


Semántica Distribucional

Payaso ≈ circo

“You shall know a word by the company it keeps”

John Rupert Firth, 1957



Part 1: Vectorial Representation of Documents

Document Representation by BoW

- Advantages/Disadvantages

Term-Weighting Schemes

Automatic Classification Scenario

- The problem of text classification
- Machine learning approach for TC
- Construction of a classifier
  1. – Document representation
  2. – Dimensionality reduction
  3. – Classification methods
- Evaluation of a TC method
- Description of the module project

Accuracy
F1

Traditional ML Classification Framework

Given a universe of objects and a pre-defined set of classes assign each object to its correct class

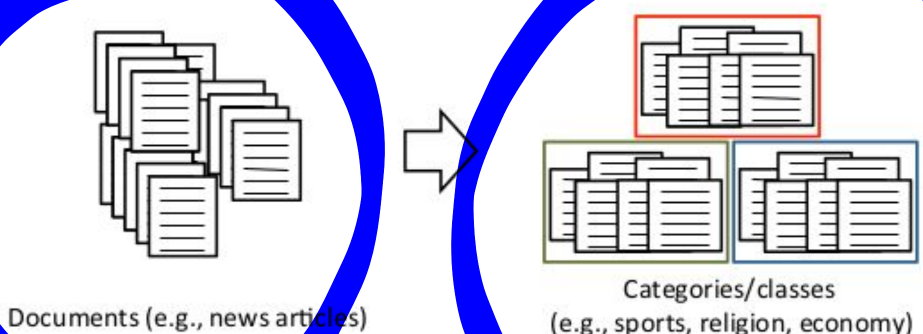
- Input:
  - A description of an instance, $x \in X$, by a vector of measurements; where X is the instance space.
  - A fixed set of categories: $C=\{c_1, c_2, \ldots c_n\}$
- Output:
  - The category of x: $c(x) \in C$, where $c(x)$ is a categorization function whose domain is X and whose range is C.

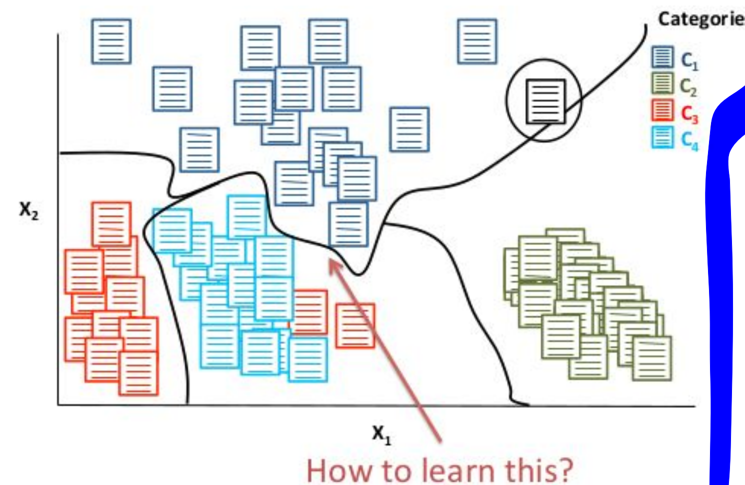
Some Textual Related Tasks

Problem	Objects (instances)	Categories
Tagging	words in context	POS tags
WSD	words in context	word senses
PP attachment	sentences	parse trees
Language identification	Text	languages
Text classification	**documents**	**topics**

Text Classification

- It is the assignment of free-text documents to one or more predefined categories based on their content.





Example: Filtering spam

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* spam *and* email.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

$$\text{if } (0.2 \cdot \%\texttt{you} - 0.3 \cdot \%\texttt{george}) > 0 \quad \text{then } \texttt{spam} \text{ else } \texttt{email}.$$

$$\text{if } (\%\texttt{george} < 0.6)\ \&\ (\%\texttt{you} > 1.5) \quad \text{then } \texttt{spam} \text{ else } \texttt{email}.$$

Hastie et al. **The Elements of Statistical Learning,** 2007, Springer.

Automatic Document Representation: BoW

- Very common because its **simplicity** and **efficiency**.
- Under this scheme, documents are represented by collections of terms, each term being an independent feature.
  - Word order is not capture by this representation
  - Semantic information is omitted
  - There is no attempt for understanding documents' content

Features

BoW

Traditional ML

Representation of documents

Vocabulary from the collection
(set of different words)

label
category

	t_1	t_1	...	t_n
d_1				
d_2				
:		$w_{i,j}$		
d_m				

All documents
(one vector per document)

Weight indicating the contribution of word *j* in document *i*.

Each different word is a feature!
How to compute their weights?

Term Weighting

- The importance of a term increases proportionally to the number of times it appears in the document.
  - It helps to *describe* document's content.
- The general importance of a term decreases proportionally to its occurrences in the entire collection.
  - Common terms are not good to *discriminate* between different classes

1

2

Term Weighting

- Binary weights:
  - $w_{i,j}$ = 1 iff document d_i contains term t_j , otherwise 0.
- Term frequency (tf):
  - $w_{i,j}$ = (no. of occurrences of t_j in d_i)
- tf x idf weighting scheme:
  - $w_{i,j} = \mathrm{tf}(t_j, d_i) \times \mathrm{idf}(t_j)$, where:
    - $\mathrm{tf}(t_j, d_i)$ indicates the ocurrences of t_j in document d_i → descriptiva
    - $\mathrm{idf}(t_j) = \log [N/\mathrm{df}(t_j)]$, where $\mathrm{df}(t_j)$ is the number of documets that contain the term t_j. → discriminativa

Normalization?

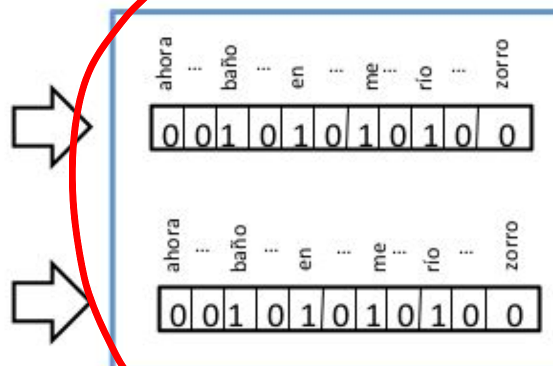
$$w_{i,j} = \frac{w_{i,j}}{\sqrt{\sum_{k=1}^{M} (w_{i,k})^2}}$$

Main Problems

- A document is represented by the set of terms that appear in it
- By definition, BOW is an orderless representation

Yo me rio en el baño
(I am laughing at the bathroom)

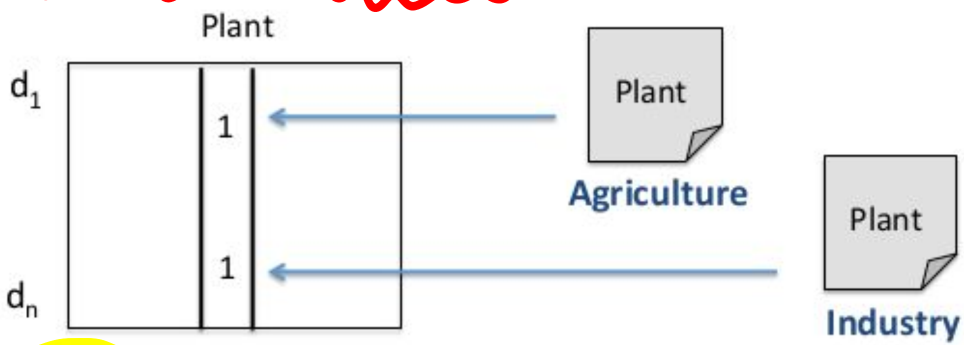
Yo me baño en el río
(I am taking a shower at the river)



Same BoW representation
different meaning

Main Problems

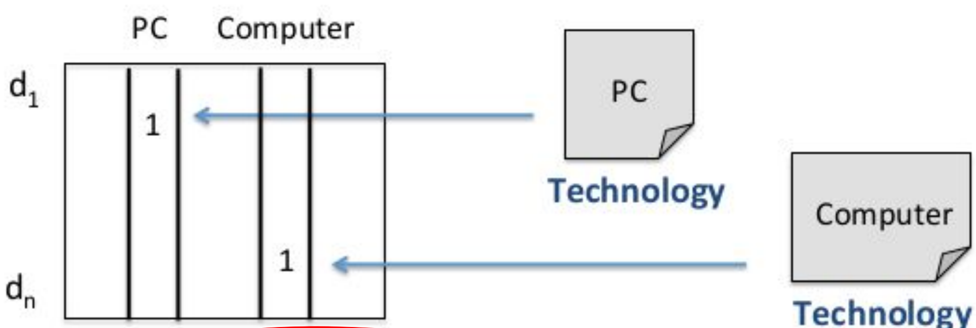
- BoW ignores all semantic information; it simply looks at the surface word forms
  - Polysemy and synonymy are big problems



Polysemy introduces noise into the BOW representation

Main Problems

- BoW ignores all semantic information; it simply looks at the surface word forms
  - Polysemy and synonymy are big problems



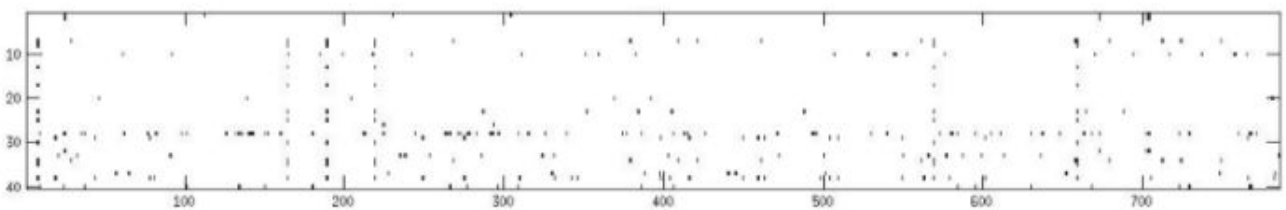
chicas
chikas
chiiikas
⋮

Synonymy splits the evidence in the BOW representation

Main Problems

d_i $|V|$

- BoW tends to produce sparse representations, since terms commonly occur in just a small subset of the documents
  - This problem is amplified by lack of training texts and by the shortness of the documents



Very difficult to find classification patterns!

Ideas for solving these limitations?

Idea 1: Indexing by POS tags

Part of Speech

Twitter

lol

:)

Whole vocabulary of the collection with POS tags

	w_1t_1	w_1t_2	Plant|NN	Plant|VB	…	w_nt_m
d_1						
d_2						
:		$w_{i,j}$				
d_m						

Weight indicating the contribution of term-pos *j* in document *i*.

Comments on this solution? Does it works?

Idea 2: Phrases as Features

- Using single words as index terms generally has good exhaustivity, but poor specificity due to word ambiguity.
- Some word associations have a totally different meaning of the "sum" of the meanings of the words that compose them.
  - Hot + dog ≠ "hot dog"
- To remedy this problem: use terms more complex than single words, such as *phrases*.
  - Distinguish the two meanings by using phrasal index terms such as "bank of the Seine" and "bank of Japan"

Idea 2: Phrases as Features

Extracted phrases from the collection

	p_1	P_2	Information retrieval	Paul McCartney	Rolling Stones	p_n
d_1						
d_2						
:		$w_{i,j}$				
d_m						

Weight indicating the contribution of phrase *j* in document *i*.

Which kind of word sequences are relevant phrases?
How to extract them?

Idea 2: Phrases as Features (Syntactical)

This apple pie looks good and is a real treat

- adjective-noun relation (*real-treat*)
- noun-noun relation (*apple-pie*)
- subject-verb relation (*pie-looks*)
- verb-object relation (*is-treat*)
- The complication is that they are extracted from the POS tagged text or from the *syntactic tree*.

Idea 2: Phrases as Features (NER)

- *Proper names* in texts
  - Three universally accepted categories: **person**, **location** and **organisation**
  - Other categories: date/time expressions, measures (percent, money, weight etc), email addresses, etc.
- One problem: they can be also ambiguous!
  - George Bush: person or location?
  - Mexico: geo-political organization or location?

How to detect named entities?

qué
quién
cómo
eventos
personas

Idea 2: Phrases as Features

- N-gram is a subsequence of *n* items from a given sequence
- N-grams are easily computed
- Combining n-grams for different sizes produces great coverage and flexibility for the representation.
- Main problem is the high dimensionality.

How to select only the most useful n-grams?

Idea 3: Word Senses as Features

- Traditional IR/TC approaches are highly dependent on *term-matching*
- Term matching is affected by the *synonymy* and *polysemy* phenomena.
- Need to capture the concepts instead of only the words
- Solution: using word senses as features!

Idea 3: What is a word sense

- Word sense is one of the *meanings* of a word.
- “Words” are having different meanings based on the context of the word.
- Example:
  - We went to see a play at the theater
  - The children went out to play in the park

A computer program has no basis for knowing which one is appropriate, even if it is obvious to a human

Idea 3: Indexing by Senses

All different word senses from the target collection

	w_{11}	w_{12}	Bank (institution)	Bank (hill)	p_{n1}	p_{nm}
d_1						
d_2						
:		$w_{i,j}$				
d_m						

Weight indicating the contribution of the word-sense *j* in document *i*.

We need to determine the sense of each word from the document collection. Hard problem!

Did they work?

- Evidence that POS info, complex nominals, and word senses do not improve TC accuracy
  - Lack of accurate NLP tools (in many languages)
  - High computational cost in comparison with BOW
- The combination of word unigrams and bigrams tend to produce the best results.
  - Higher order n-grams are –usually– useless.

So, what else can we try? Ideas?

Alessandro Moschitti, Roberto Basili. *Complex Linguistic Features for Text Classification: A Comprehensive Study*. Lecture Notes in Computer Science Volume 2997, 2004.c

Preprocessing

- Eliminate information about style, such as html or xml tags.
  - For some applications this information may be useful. For instance, only index some document sections.
- Remove stop words
  - Functional words such as articles, prepositions, conjunctions are not useful (do not have an own meaning).
- Perform stemming or lemmatization
  - The goal is to reduce inflectional forms, and sometimes derivationally related forms.

am, are, is → be
car, cars, car's → car

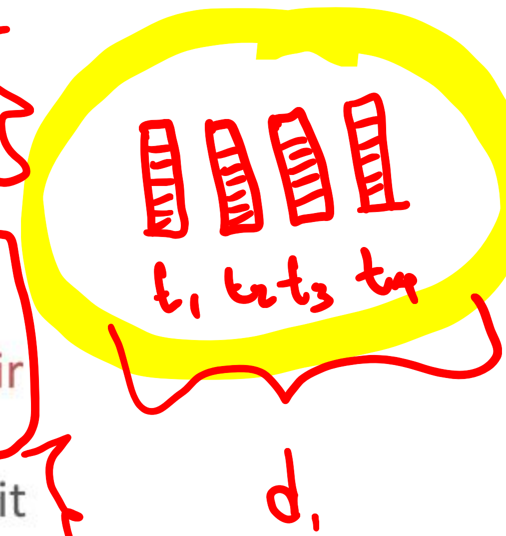
Have to do them always?

ed

Porter Stemmer

Part 2: Automatic build of Bag-of-Concepts

- Addresses the deficiencies of the BoW by considering the relations between document terms.
- BoC representations are based on the intuition that the meaning of a document can be considered as the union of the meanings of their terms.
- The meaning of terms is related to their usage; it is captured by their distributional representation



Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Distributional term representations: an experimental comparison. *Thirteenth ACM international conference on Information and knowledge management* (CIKM '04). New York, NY, USA, 2004