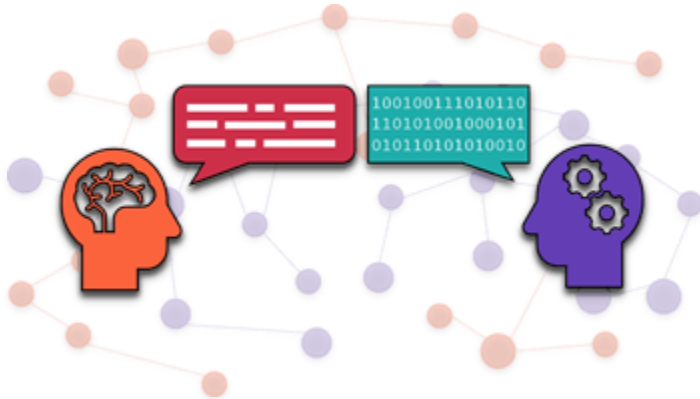
Modelos de Inteligencia Artificial

CIPFP Mislata
Centre Integrat Públic Formació Professional Superior

:: U3 ::

Procesamiento del Lenguaje Natural



Uso de modelos lingüísticos

Curso 2023-24

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

1. Introducción

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Los modelos lingüísticos son modelos **probabilísticos** del lenguaje que se basan en **representaciones numéricas del lenguaje**.

- Las aplicaciones de estos modelos son múltiples:
  - traducción automática,
  - chatbots,
  - análisis de sentimientos,
  - generación de texto o código, de resúmenes, ...

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejemplo**: “No soy sabio por saber que no sé nada”

  ¿Qué nos diría un modelo probabilístico basado en palabras con letra inicial “s”?

  ¿Y si el modelo se basara en la probabilidad de encontrar unas palabras junto a otras?

1. Introducción

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Qué tienen en común estos 3 tipos de problemas?

- Identificar si el remitente de un correo electrónico anónimo pertenece a una lista de contactos previa.
- Averiguar si cierto correo electrónico es o no SPAM.
- Determinar si un producto comercial está bien valorado o no.

1. Introducción

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Qué tienen en común estos 3 tipos de problemas?

- Identificar si el remitente de un correo electrónico anónimo pertenece a una lista de contactos previa.
- Averiguar si cierto correo electrónico es o no SPAM.
- Determinar si un producto comercial está bien valorado o no.

… tienen respuesta aplicando un análisis basado en Naïve Bayes + "Saco de Palabras" (Bag of Words)

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Saco de palabras** (bag-of-words)
  - El modelo de *saco de palabras* es una técnica de procesamiento de lenguaje natural que se utiliza para representar documentos de texto
  - **Objetivo del modelo BoW**: tratar un documento como una "bolsa" (conjunto) de palabras, sin tener en cuenta la estructura gramatical y el orden de las palabras en el texto, sino simplemente contando la **frecuencia de cada palabra en el documento.**
  - Este modelo sirve de base para la aplicación de técnicas de análisis estadístico (por ej. *Naive Bayes*)

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- El modelo **BoW** se construye en base a 3 elementos …

  i) Tokenización

  ii) Diccionario de características

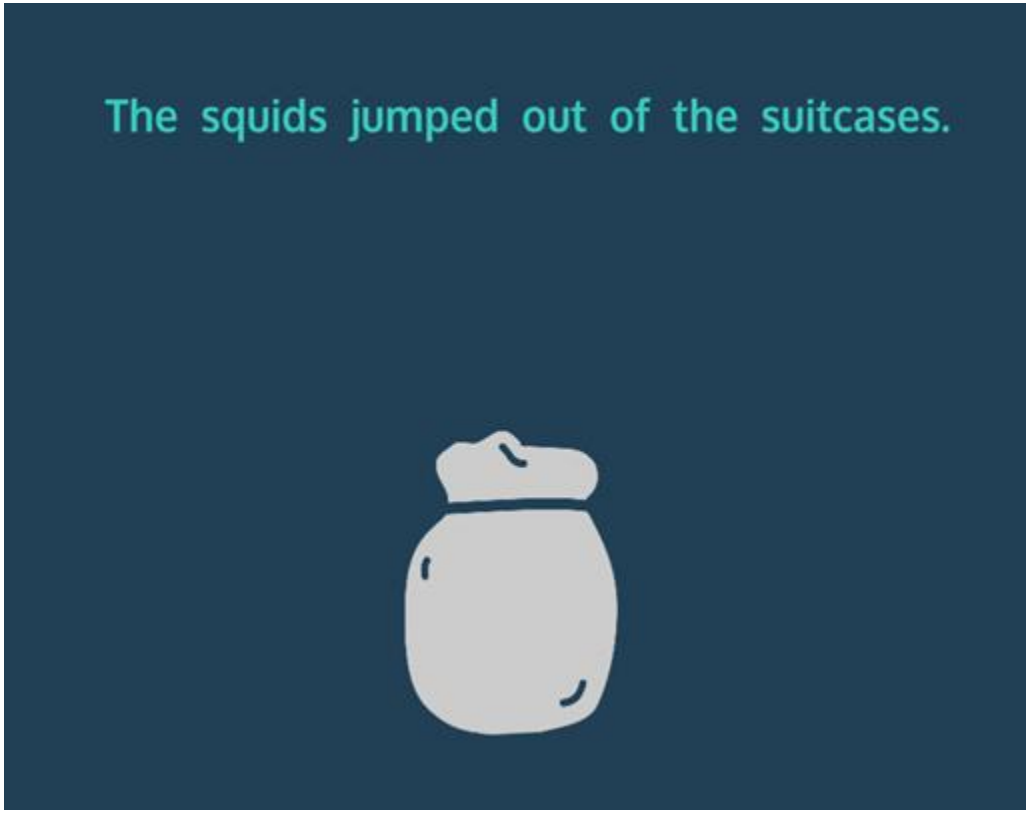
  iii) Vectorización

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Diccionarios BoW**

  Una forma intuitiva de pensar cómo implementar el modelo BoW en Python sería usar un diccionario



2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Diccionario de características

- las claves corresponden a los términos
- los valores indican la posición del término en el conjunto del texto

Ejemplo:

> documento: “No soy sabio por saber que no sé nada”

> diccionario de características:

[“No”: 0, “ser”: 1, “sabio”: 2, “por”: 3, “saber”: 4, “que”: 5, “nada”: 6]

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Vectorización

Supone construir un vector con las características (frecuencia de aparición de cada término) para cierto documento de prueba.

2. Uso de modelos lingüísticos

- **Ejemplo:**
  - documento: “No soy sabio por saber que no sé nada”

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejemplo:**
  - documento: “No soy sabio por saber que no sé nada”
  - diccionario de características (**datos entrenamiento**):

    [“No”: 0, “ser”: 1, “sabio”: 2, “por”: 3, “saber”: 4, “que”: 5, “nada”: 6]

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejemplo:**
  - documento: "No soy sabio por saber que no sé nada"
  - diccionario de características (**datos entrenamiento**):

    ["No": 0, "ser": 1, "sabio": 2, "por": 3, "saber": 4, "que": 5, "nada": 6]

  - vector de características (**datos de prueba**):

    "**Eres** un tipo **sabio**"

    [0, **1, 1**, 0, 0, 0, 0]

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- BoW mediante **CountVectorizer**:

```
from sklearn.feature_extraction.text import CountVectorizer

documentos_entrenamiento = ["Diez divertidos delfines nadaron hacia destinos desconocidos.",
                            "¿Quizás encuentren otros diez delfines divertidos?",
                            "¡Encuentra mis delfines con una función, por favor!"]
texto_prueba = ["Otros cinco delfines encuentran otro delfín lejano."]

bow_vectorizador = CountVectorizer()

# Creamos el diccionario BoW mediante el método "fit" del objeto "CountVectorizer"
bow_vectorizador.fit(documentos_entrenamiento)

# Mostramos las palabras encontradas
print("Palabras: ")
print(bow_vectorizador.get_feature_names_out())
```

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- BoW mediante **CountVectorizer**:

```
Palabras:
['con' 'delfines' 'desconocidos' 'destinos' 'diez' 'divertidos'
 'encuentra' 'encuentren' 'favor' 'función' 'hacia' 'mis' 'nadaron'
 'otros' 'por' 'quizás' 'una']
```

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- BoW mediante **CountVectorizer**:

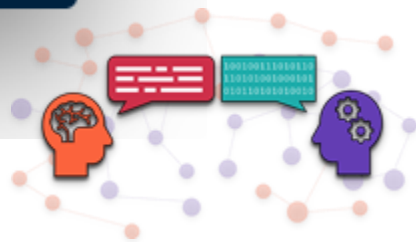
```
# Mostramos el diccionario BoW
print("\nDiccionario BoW (palabras + índices): ")
print(bow_vectorizador.vocabulary_)

----

Diccionario BoW (palabras + índices):
{'diez': 4, 'divertidos': 5, 'delfines': 1, 'nadaron': 12, 'hacia': 10, 'destinos': 3, 'desconocidos': 2,
'quizás': 15, 'encuentren': 7, 'otros': 13, 'encuentra': 6, 'mis': 11, 'con': 0, 'una': 16, 'función': 9,
'por': 14, 'favor': 8}
```

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- BoW mediante **CountVectorizer**:

```
# Generamos y mostramos la forma vectorizada el corpus de entrenamiento
bow_vector = bow_vectorizador.transform(documentos_entrenamiento)
print("\nVector completo para el texto de entrenamiento: ")
print(bow_vector.toarray())

----

Vector completo para el texto de entrenamiento:
[[0 1 1 1 1 1 0 0 0 0 1 0 1 0 0 0 0]
 [0 1 0 0 1 1 0 1 0 0 0 0 0 1 0 1 0]
 [1 1 0 0 0 0 1 0 1 1 0 1 0 0 1 0 1]]
```



2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

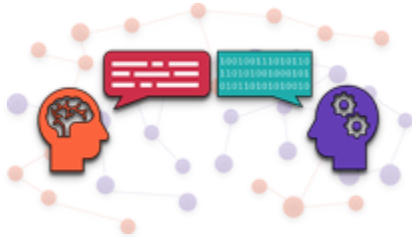
- BoW mediante **CountVectorizer**:

```
# Generamos el vector para la palabra “divertidos”
print("\nVector para ‘divertidos’: ")
print(bow_vectorizador.transform(['divertidos']).toarray())

----

Vector para ‘divertidos’:
[[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]]
```

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejercicio ::** Más sobre Countvectorizer()
  - En el ejemplo anterior, hemos usado *Countvectorizer***sin** argumentos.
  - Implementa ahora una variante del script previo en la que el modelo descarte ciertas **stopwords** que se pasarán como argumento.
  - Consulta la página de Scikit-Learn sobre *Countvectorizer* para indagar cómo hacerlo.
  - Examina qué cambios ocurren entre ambas versiones.



2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ventajas de BoW:**
  - **Simplicidad**: BoW es fácil de entender e implementar.
  - **Eficiencia computacional**: es computacionalmente eficiente y rápido de calcular.
  - **Versatilidad**: se puede utilizar en diversas tareas de procesamiento de lenguaje natural.
  - **Robustez ante ruido**: maneja texto con errores o redundancias sin afectar significativamente su desempeño.
  - **Interpretabilidad**: fácil de interpretar, permitiendo examinar directamente las palabras que contribuyen más a la representación del documento.

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Inconvenientes de BoW:**

  a. **baja capacidad predictiva**

     BoW no es un modelo muy preciso para la predicción del lenguaje, ya que la probabilidad de cierta palabra se relaciona sólo con aquellas utilizadas más frecuentemente.

  b. **descontextualización**

     Los términos en un modelo BoW carecen de contexto, lo que puede complicar la comprensión del significado de una palabra.

  c. **tamaño del vocabulario**

     Puede generar vectores de características largos y dispersos, aumentando la complejidad computacional y de almacenamiento.

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Práctica 3.1**
  - Realiza la práctica "**PLN - P3.1 :: El amigo misterioso"**, cuyo enunciado encontrarás en el moodle

- **Práctica 3.2**
  - Realiza la práctica "**PLN - P3.2 :: Creación de un filtro de SPAM"**, cuyo enunciado encontrarás en el moodle

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Término frecuencia / frecuencia inversa en documento (tf-idf)

- **tf-idf** es una técnica estadística que cuantifica la importancia de una palabra en un documento en función de la frecuencia con la que aparece en ese documento y en una determinada colección de documentos (corpus)
- Usos:
  - clasificar resultados en un motor de búsqueda
  - hacer resúmenes de texto
  - construir chatbots inteligentes
  - ...

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

TF

Frecuencia de una palabra
en un documento

IDF

Frecuencia de una palabra
en un *corpus*

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- La intuición de esta técnica es:
  - si una palabra aparece con frecuencia en un documento, entonces debería ser más importante y relevante que otras palabras que aparecen menos veces y deberíamos darle a esa palabra una puntuación alta (**tf**).
  - … pero si una palabra aparece muchas veces en un documento pero también en muchos otros documentos, probablemente no sea una palabra relevante, por lo que deberíamos asignarle una puntuación más baja (**idf**).

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Término frecuencia / frecuencia inversa en documento (tf-idf)

- Pasos en la creación del modelo:
  1. se calculan los valores de **frecuencia de un término** en un documento
  2. se calcula los valores de **frecuencia inversa de documento** de un término en el corpus
  3. se combina (1) y (2) para crear la matriz **tf-idf**

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Término frecuencia / frecuencia inversa en documento (tf-idf)

- Pasos en la creación del modelo:

1. se calculan los valores de **frecuencia de un término** para un documento

$$TF(t,d) = \frac{count(t)}{count(d)} = \frac{Number\ of\ occurrences\ of\ t\ in\ the\ document\ d}{Number\ of\ words\ in\ the\ document\ d}$$

$$where\ t = text\ we\ are\ interested\ in$$
$$d = the\ documents\ provided$$

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Término frecuencia / frecuencia inversa en documento (tf-idf)

- Pasos en la creación del modelo:

  2. se calcula los valores de **frecuencia inversa de documento** de un término en el corpus

$$IDF = log \frac{Number\ of\ document}{Number\ of\ documents\ t\ occurs\ in}$$

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Término frecuencia / frecuencia inversa en documento (tf-idf)

- Pasos en la creación del modelo:

  3. se combina (1) y (2) para crear la matriz **tf-idf**

$$TF - IDF = TF * IDF$$
$$= \left(\frac{Number\ of\ occurrences\ of\ t\ in\ the\ document\ d}{Number\ of\ words\ in\ the\ document\ d} \right) * \left(log\ \frac{Number\ of\ document}{Number\ of\ documents\ t\ occurs\ in} \right)$$

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- **Ejemplo:** partimos del siguiente *corpus*

Document	Text Data
D1	My name is Naftal.
D2	My car is a Hyundai.
D3	The car I drive is a Hyundai Sonata.
D4	My car is a Sonata model by Hyundai !!

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- **Ejemplo:** preprocesamos su contenido

Document	Text Data
D1	name naftal
D2	car hyundai
D3	car drive hyundai sonata
D4	car sonata model hyundai

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- **Ejemplo:** se crea una matriz **tf** para el corpus …

Document	name	naftal	car	hyundai	drive	sonata	model
D1	1	1	0	0	0	0	0
D2	0	0	1	1	0	0	0
D3	0	0	1	1	1	1	0
D4	0	0	1	1	0	1	1

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- **Ejemplo:** … normalizando sus valores

Document	name	naftal	car	hyundai	drive	sonata	model
D1	1/2	1/2	0	0	0	0	0
D2	0	0	1/2	1/2	0	0	0
D3	0	0	1/4	1/4	1/4	1/4	0
D4	0	0	1/4	1/4	0	1/4	1/4

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- **Ejemplo:** se calculan los valores **idf**

name	naftal	car	hyundai	drive	sonata	model
0.602	0.602	0.125	0.125	0.602	0.301	0.602

$$IDF = log \frac{Number\ of\ document}{Number\ of\ documents\ t\ occurs\ in}$$

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- **Ejemplo:** se calcula la matriz **tf-idf** asociada

0.602* 1/2

Document	name	naftal	car	hyundai	drive	sonata	model
D1	0.301	0.301	0	0	0	0	0
D2	0	0	0.0625	0.0625	0	0	0
D3	0	0	0.03125	0.03125	0.1505	0.075	0
D4	0	0	0.03125	0.03125	0	0.075	0.1505

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

> Término frecuencia / frecuencia inversa en documento (tf-idf)

- **Ejemplo:** se calcula la matriz **tf-idf** asociada

Document	name	naftal	car	hyundai	drive	sonata	model
D1	0.301	0.301	0	0	0	0	0
D2	0	0	0.0625	0.0625	0	0	0
D3	0	0	0.03125	0.03125	0.1505	0.075	0
D4	0	0	0.03125	0.03125	0	0.075	0.1505

Document	Text Data
D1	name naftal
D2	car hyundai
D3	car drive hyundai sonata
D4	car sonata model hyundai

- los valores para un término en cierto documento dan una idea de su importancia relativa en el conjunto
- cuanto más alto sea éste mayor relevancia se le puede suponer

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Término frecuencia / frecuencia inversa en documento (tf-idf)

- La librería Scikit-Learn proporciona 3 objetos para implementar este modelo:

1. TfidfVectorizer

   Para obtener la **matriz tf-idf**

1. CountVectorizer + TfidfTransformer

   Para obtener **tf** e **idf** por separado

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Práctica 3.3**
  - Realiza la práctica "**PLN - P3.3 :: Analizando noticias de los medios"**, cuyo enunciado encontrarás en el moodle

Tabla de contenidos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

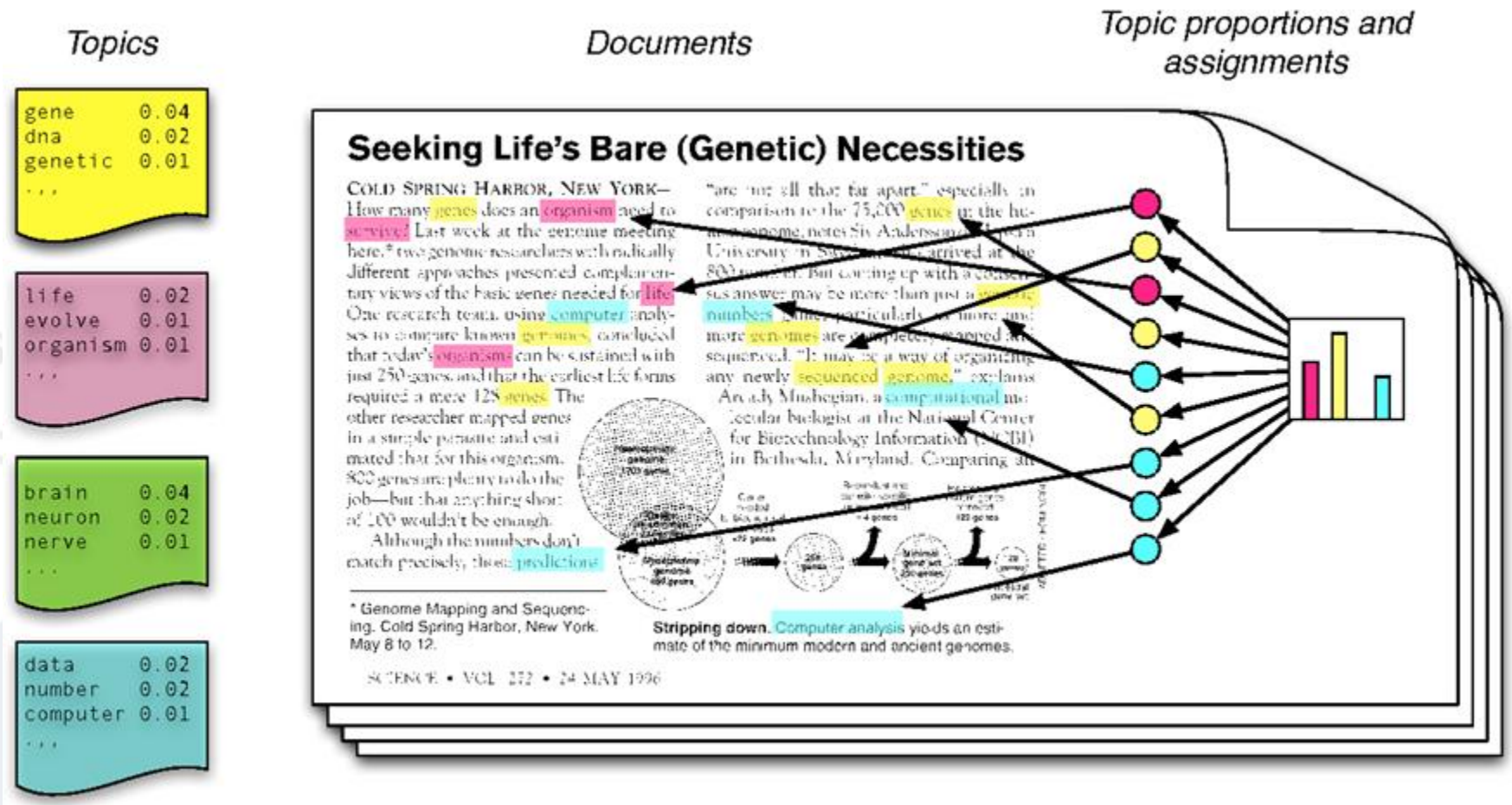
2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Clasificación temática (“topic modeling”)**
  - El ***topic modeling*** es una técnica no supervisada de NLP, capaz de detectar y extraer de manera automática relaciones “ocultas” entre términos en grandes volúmenes de información.
  - Un **tema** (tópico) …
    - es un conjunto de palabras que suelen aparecer juntas en los mismos contextos.
    - expresan una **distribución de probabilidades** de aparición de las distintas palabras del vocabulario.

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior



2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejemplo:** aplicación de **LDA** (implementación de *Topic Modeling*) a la cobertura de casos de violencia de género en un conjunto de noticias de los medios generalistas

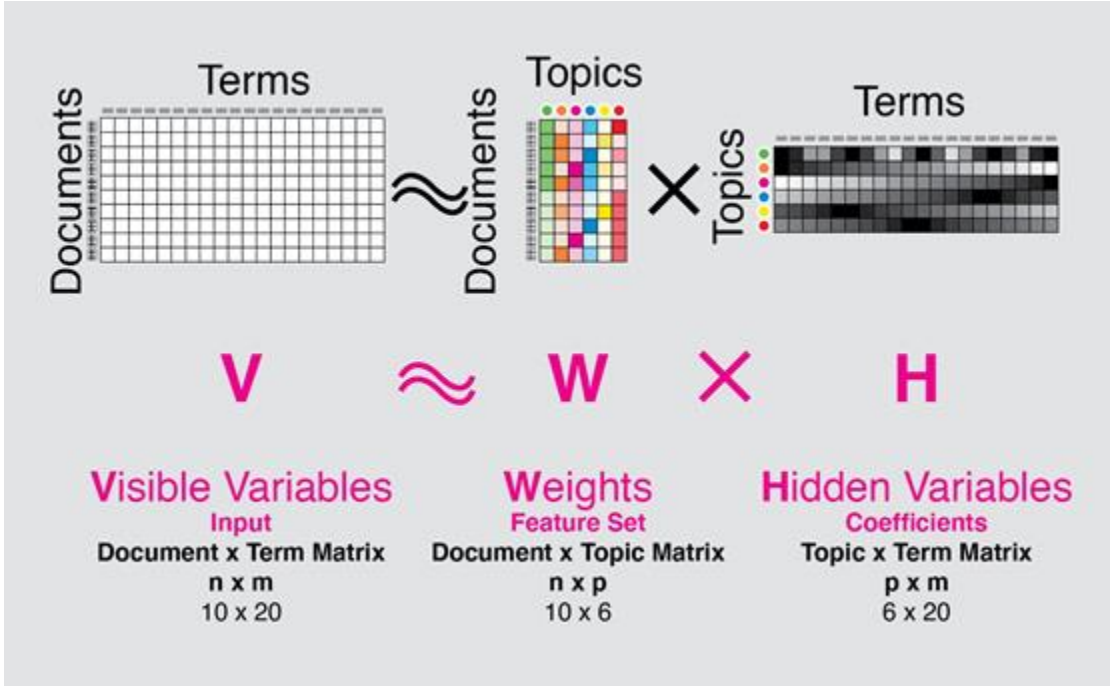
2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Scikit-Learn implementa **LDA** en base a lo que se denomina **Matriz de Factorización No-Negativa**
- Parte de una **matriz de términos del documento** (**V** en la imagen)

… que descompone en …

  - **matriz de documento-temas** (**W** en la imagen)
  - **matriz de tema-términos** (**W** en la imagen)



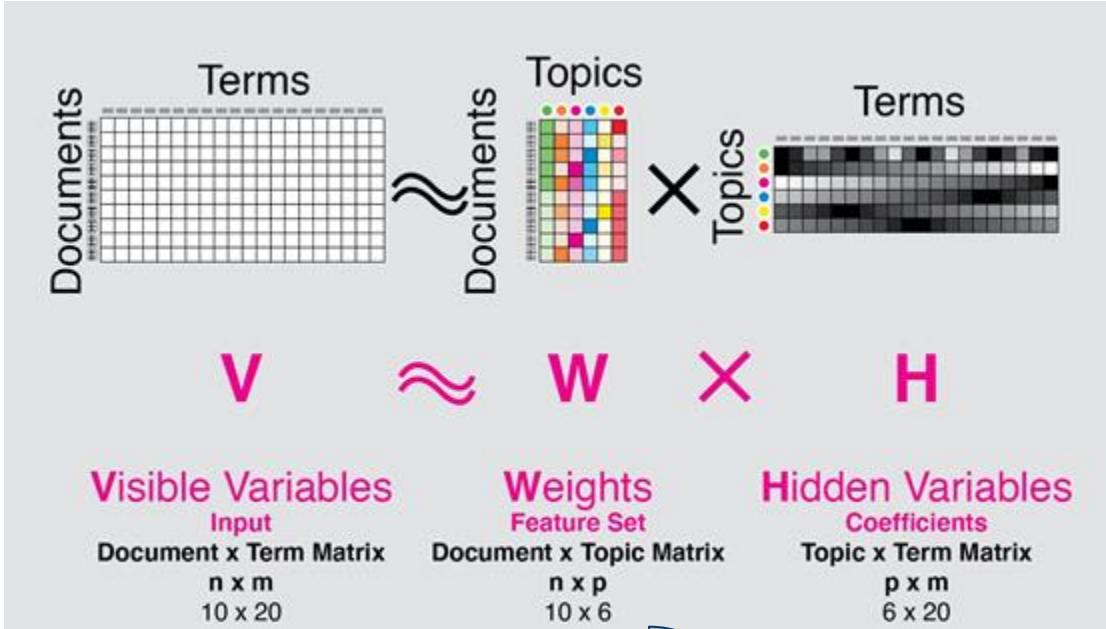
2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Scikit-Learn implementa **LDA** en base a lo que se denomina **Matriz de Factorización No-Negativa**
- Parte de una **matriz de términos del documento** (**V** en la imagen)

  … que descompone en …

  - **matriz de documento-temas** (**W** en la imagen)
  - **matriz de tema-términos** (**W** en la imagen)



Objetivo

- Cada término hará una contribución positiva a un conjunto de temas …
- Cada tema puede ser descrito por una combinación de términos

2. Uso de modelos lingüísticos

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Práctica 3.4**
  - Realiza la práctica "**PLN - P3.4 :: Topic modeling vs K-Means"**, cuyo enunciado encontrarás en el moodle