



# Universidad Nacional de La Matanza

Departamento de Ingeniería e Investigaciones Tecnológicas

## Módulo Procesamiento de Lenguaje Natural (NLP): Hoja de ruta

### Profesores:

Dr. Ierache, Jorge

Dr. Becerra Martín

Ing. Sanz Diego

# Módulo Procesamiento de Lenguaje Natural (NLP)

En esta unidad se busca como objetivo resolver problemas en los que se necesite trabajar con lenguaje natural bajo la modalidad de texto.

Actividad 1: Introducción a NLP tradicional.....	2
Teoría.....	2
Práctica.....	2
Actividad 2: Representación de texto tradicionales.....	3
Teoría.....	3
Práctica.....	3
Actividad 3: Representación de texto moderna Word Embeddings.....	4
Teoría.....	4
Práctica.....	4
Actividad 4: NLP moderno con Deep Learning.....	5
Teoría.....	5
Práctica.....	5
NLP moderno con Transformers [Opcional].....	6
Actividad 5: Caso de estudio NLP en e-commerce [Opcional].....	7
Search en E-commerce.....	7
Semantic search.....	9
Creación de E-Commerce Catalog.....	9
Extracción de atributos.....	10
Direct attribute extraction.....	11
Indirect attribute extraction.....	11
Categorización de productos.....	12
Recomendación de productos en E-commerce.....	13
Referencias.....	15
Actividades complementarias.....	17
Biblioteca spaCy: Funcionalidades básicas y lingüísticas.....	17
Biblioteca spaCy: Aprendizaje automático.....	18

## Actividad 1: Introducción a NLP tradicional

### Teoría

En esta actividad se busca comprender el flujo de trabajo en NLP bajo los paradigmas basados en conocimiento y estadísticos básicos (ML tradicional) que iniciaron en el campo. Para complementar la actividad se recomienda (Opcional) la lectura de los capítulos 1, 13 al 16 del libro *Speech and Language Processing Draft 3rd Edition* de la bibliografía de la cátedra.

### Consigna

Ver los videos:

1. [NLP\\_C1\\_00 - Introducción a NLP.](#)
2. [NLP\\_C1\\_01 - Tareas y aplicaciones en NLP.](#)
3. [NLP\\_C1\\_02 - Flujo de trabajo en NLP.](#)

Responder las siguientes preguntas:

1. ¿Qué es NLP ?
2. ¿Cuáles son los componentes del lenguaje ?
3. Enumere y defina los desafíos en NLP.
4. Enumere las tareas que existen en NLP, elegir 3 y definir las.
5. ¿Que es un agente conversacional (Chatbot)?
6. Enumere y explique cada una de las fases del flujo de trabajo en NLP.
7. Explique las diferencias entre un flujo de trabajo clásico y moderno.

### Práctica

- a. En esta actividad práctica aplicar los conceptos vistos en la teoría. Par ello, hacer una copia, ejecutar y analizar el colab [02 - Introduccion a NLP.ipynb](#) en donde se realiza un flujo de trabajo clásico utilizando la biblioteca [NLTK](#) y responder las siguientes preguntas a modo de reflexión:
  1. ¿Qué problema se busca solucionar en el colab? Definir alcance y meta.
  2. ¿Porque se revisa la cantidad de palabras únicas?
  3. ¿Por qué se busca reducir esta cantidad de palabras únicas?
  4. Enumere y explique los pasos de preprocesamiento utilizados. Indicar por qué se aplicaron.
  5. ¿Qué son las stopwords y por qué se eliminan?
  6. ¿Cómo se representó el texto para trabajar en NLP?.
  7. ¿Qué ventajas y desventajas observa que podemos tener al usar el tipo de representación de texto ?
  8. ¿Qué pasa si encontramos una palabra que esté fuera del vocabulario representado inicialmente con la representación usada?
- b. Realizar la actividad complementaria en la sección [Biblioteca spaCy: Funcionalidades básicas y lingüísticas](#).

- c. Realizar el mismo flujo de trabajo que en el colab de introducción a NLP para el mismo problema con la biblioteca [spaCy](#)(opcional).

## Actividad 2: Representación de texto tradicionales

### Teoría

En esta actividad se busca profundizar en las representaciones de texto tradicional y los pain points (Punto de dolor) que tienen. En la actividad práctica anterior ya se utilizaron representaciones tradicionales y ahora vamos a analizar sus características, ventajas y desventajas y porque se dejaron de usar en virtud de las representaciones modernas. Para complementar la actividad se recomienda (Opcional) la lectura de los capítulos 2 al 5 del libro Speech and Language Processing Draft 3rd Edition de la bibliografía de la cátedra.

### Consigna

Ver el siguiente video: [NLP C2 00 - Representaciones de texto básicas](#) y responder las siguientes preguntas:

1. ¿Qué son las features en aprendizaje automático?
2. ¿ Para qué usamos representaciones de texto ?
3. Explicar en qué consiste el proceso de vectorización, incluya los elementos que se consideran necesarios para llevarlo a cabo.
4. Haga una tabla comparativa entre representaciones clásicas que incluya, **Nombre, representación del texto, tamaño de vectores, sparsity, captura similaridad, OOV, Normalización, ventajas y desventajas.**

### Práctica

- a. En esta actividad práctica aplicar los conceptos vistos en la teoría. Para ello realizar una copia, ejecutar y analizar el colab [05 - Representación de texto: Enfoques básicos de representac...](#) para responder las siguientes preguntas a modo de reflexión:
  1. Indique qué diferencias encuentra entre representar un documento aplicando las siguientes técnicas de representación
    - a. Bag of words (Binario y no binario).
    - b. One hot encoding.
  2. ¿Qué detalles captura Bag of N-grams? Indique cómo se aplicó para representar el texto.
  3. ¿Qué detalles captura TF-IDF? que diferencia notó con diferencia al resto de las representaciones.
  4. Aplique una técnica vista en esta práctica a un corpus de texto pequeño y uno grande de su elección y observar cómo varía las dimensiones de vectores. ¿Qué problema puede encontrar al cambiar de un texto a otro?. ¿Sirve la representación del corpus pequeño al mayor ? ¿Por qué?.

## Actividad 3: Representación de texto moderna Word Embeddings

### Teoría

En esta actividad se busca profundizar en las representaciones de texto modernas y los problemas clásicos que solucionan. En la actividad práctica anterior utilizaron representaciones tradicionales en las que se tenía que preparar las features manualmente. Ahora vamos a abordar representaciones modernas de texto, con el objetivo de automatizar la extracción de features. Para complementar la actividad se recomienda (Opcional) la lectura del capítulo 6 del libro *Speech and Language Processing Draft 3rd Edition* y el apunte *What are embeddings* de la bibliografía de la cátedra.

### Consigna

Ver los videos:

- [NLP\\_C2\\_01 - Introducción a Word embeddings.](#)
- [NLP\\_C2\\_02 - Word2Vec.](#)
- [NLP\\_C2\\_03 - Representaciones avanzadas de texto.](#)

Responder las siguientes preguntas:

1. ¿Qué es un Word embeddings?
2. ¿Qué podemos representar en un espacio vectorial?
3. ¿Qué elementos necesitamos para entrenar / utilizar un modelo de embeddings?
4. Enumerar y describir qué mediciones se usan para comparar dos embeddings
5. Explicar cómo se obtienen word embeddings utilizando la técnica Word2Vec (2 enfoques de entrenamiento).
6. Enumere y explique las diferencias de este enfoque y problemas que resuelve con respecto a los anteriores (tradicionales).
7. Realizar una tabla comparativa general entre enfoques tradicionales (Word vectorization) y modernos de representación de texto (Word embeddings).
8. Realizar una tabla comparativa con todos los modelos de la familia Word2Vec para representar texto (Word2Vec, Doc2Vec, GloVe y FastText).

### Práctica

En esta actividad práctica aplicar los conceptos vistos en la teoría. Para ello realizar una copia, ejecutar y analizar los colabs [06 - Modelo Word2Vec preentrenado.ipynb](#) y [10 - Modelo GloVe preentrenado.ipynb](#) utilizando la biblioteca [Gensim](#) para responder las siguientes preguntas a modo de reflexión:

1. ¿Por qué se utiliza un benchmark de analogías para evaluar los resultados obtenidos del entrenamiento de embeddings?
2. ¿En qué consiste la mejora que introdujo Gensim en 2016 para evaluar la similaridad de palabras de los modelos entrenados?.
3. ¿Cómo se puede hacer finetuning de un modelo de Gensim con nuestras palabras adicionales?

4. Enumere y explique las diferentes funciones exploradas en el colab de GloVe para trabajar con similaridad entre palabras.
5. Investigar las técnicas PCA (Principal Components análisis) y TSNE e indicar para que se utilizaron en el colab de GloVe
6. ¿Qué usos útiles puede encontrar aplicar similaridad entre documentos de texto ?.

## Actividad 4: NLP moderno con Deep Learning

### Teoría

Hasta el momento ya actualizamos nuestras representaciones de texto con técnicas modernas de word embeddings para poder automatizar la extracción de features de manera no supervisada. En esta actividad vamos a reemplazar nuestros modelos clásicos vistos con algoritmos de aprendizaje automático deep learning más potentes con el objetivo de poder procesar secuencias de texto. Para complementar la actividad se recomienda (Opcional) la lectura de los capítulos 7 al 9 del libro Speech and Language Processing Draft 3rd Edition de la bibliografía de la cátedra.

### Consigna

Ver el vídeo [NLP\\_C3\\_00 NLP Deeplearning RNN](#) y responder las siguientes preguntas:

1. ¿Qué problema tienen las redes neuronales clásicas (MLP) para procesar secuencias ?
2. Explique que es una neurona recurrente y cual es la diferencia con una neurona clásica.
3. Explique como una red neuronal recurrente procesa una secuencia de texto a través del tiempo.
4. Enumere las arquitecturas de redes recurrentes y defina cada una.
5. ¿Qué problemas tiene una red recurrente simple? ¿Qué alternativas surgieron para superarlos?

### Práctica

- a. En esta actividad práctica aplicar los conceptos vistos en la teoría. Para ello realizar una copia, ejecutar y analizar el colab: [12 - RNN, LSTM y GRU.ipynb](#) para responder las siguientes preguntas a modo de reflexión:
  1. ¿Qué problema busca solucionar en el colab? Definir alcance y meta.
  2. ¿Qué función cumple la capa **TextVectorization** y **Embedding**?
  3. ¿Cómo variamos las diferentes arquitecturas de una capa recurrente?
  4. ¿Qué técnicas podemos utilizar para tratar palabras que no están en nuestro vocabulario con este tipo de redes recurrentes?
- b. Realizar la actividad complementaria en la sección [Biblioteca spaCy: Aprendizaje automático](#) y utilizar NLP moderno para resolver la [guía de ejercicios de NLP](#)(opcional).

### NLP moderno con Transformers [Opcional]

En esta actividad se busca conocer los fundamentos de la arquitectura transformers que revolucionó el campo de NLP y conocer las dos familias de modelos principales BERT y GPT. Para complementar la actividad se recomienda (Opcional) la lectura de los capítulos 10 al 12 del libro Speech and Language Processing Draft 3rd Edition de la bibliografía de la cátedra.

#### Consigna

Ver vídeo [NLP\\_C3\\_01 NLP Deeplearning transformers](#) y responder las siguientes preguntas:

1. ¿ Qué es el framework Encoder y Decoder con redes recurrentes y para qué sirve?
2. ¿ Qué problemas tiene este framework?
3. ¿ Que es un large language model?
4. ¿ Qué desarrollos impulsaron la aparición de los large language models?
5. ¿ En qué consiste la arquitectura Transformers?
6. ¿ Qué es el mecanismo de atención y para que sirve?
7. ¿ Que es un modelo auto regresivo y auto encoding?
8. Defina que es el modelo BERT. Indique qué tipo de modelo es y qué tareas cumple.
9. Defina que es el modelo GPT. Indique qué tipo de modelo es y qué tareas cumple.

## Actividad 5: Caso de estudio NLP en e-commerce [Opcional]

En esta sección se describe un caso de estudio que tiene la intención de explicar en alto nivel cómo se aplican las técnicas de NLP en el e-commerce. Los retailers usan NLP porque tiene el potencial para cambiar la manera que los clientes compren mientras se presentan oportunidades para obtener insights e identificar gaps en la oferta de productos.

El área de e-commerce se caracteriza por presentar una gran cantidad de información texto e imágenes. Una gran porción de esta se encuentra en la descripción de productos y en la reviews de productos. Por lo tanto, esta área presenta un rango de problemas que podemos resolver con técnicas de NLP.

En las próximas secciones veremos cómo aplicar NLP para la búsqueda de productos (Search), creación de catálogos de productos de E-commerce y recomendación de productos.

### Search en E-commerce

En la búsqueda de productos en e-commerce está relacionada a la disponibilidad y la información relacionada a los productos (Información estructurada) lo que permite buscar productos para diversas ocasiones como por ejemplo “Ropa formal de hombre para trabajo”. La naturaleza de la información asociada con los productos, cada búsqueda llama a un pipeline de procesamiento de texto, extracción y search.

La búsqueda de productos debe ser rápida y precisa para encontrar los productos que se acercan a lo que necesita encontrar el usuario según sus necesidades. Una buena búsqueda impacta en el rate de conversión, el cual impacta en el revenue del retailer. Globalmente, en promedio sólo 2.4 % de los intentos de búsqueda del usuario se convierte en una compra y el 30% de los resultados del top 50 de sitios de retail no produce buenos resultados [1]. En esta área, el motor de búsqueda es de dominio cerrado (Retail) trabajando en diferentes facetas de los productos (Color, estilo, categorías) obtenidas de sus atributos y reviews realizadas. La búsqueda de productos por sus facetas se llama Faceted Search.

Faceted search es una variante especializada de búsqueda que permite a un usuario buscar por una gran cantidad de filtros. Por ejemplo, si un usuario quiere comprar un televisor, la plataforma ofrece filtros personalizados del producto que está buscando en el lado izquierdo de la página web. Podemos realizar una búsqueda de sitios con los siguientes links en [Amazon](#) y [Mercadolibre](#) en donde se observa que se realiza una búsqueda de smart TV en donde aparecen filtros de marca de producto, precios y demás facetas(características) como se ve en la figura 1 y 2.



The screenshot shows the Amazon.com search results for "Smart TV". The top navigation bar includes the Amazon logo, location (San Justo 1754), and search bar. Below the navigation bar, the search results are displayed with a faceted search interface on the left and product listings on the right.

**Faceted Search Interface (Left):**

- Departamento:** Televisores, TV LED y LCD, Televisores QLED, Fire TV, Televisores Inteligentes.
- Reseña del Cliente:** 4.5 stars (10 más).
- Marca:** VIZIO, Hisense, Sony, Roku, Westinghouse, Amazon, Pipishell.
- Todas las mejores marcas:** Las Mejores Marcas.
- Precio de Televisión:** \$16 - \$12,400 y más.
- Ofertas y descuentos:** Oferta del día, Todos los descuentos.
- Condición:** Nuevo, Renovado, Usado.
- Resolución:** 4K.

**Product Listings (Right):**

- VIZIO - Smart TV Full HD 1080p de 40 pulgadas con Apple AirPlay y Chromecast integrados, compatibilidad con Alexa, D40f-J09, modelo 2022.**
  - Options: 3 tamaños.
  - 4.5 stars (10,756 reviews).
  - 10 K+ comprados el mes pasado.
  - US\$159.99 (PVP: US\$209.99).
  - Oferta de Amazon Music con esta compra.
  - Entrega el lun, 3 de jan.
  - Se envía a Argentina.
  - Compatible con Alexa.
  - Agregar al carrito.
  - Más opciones de compra: US\$146.27 (10 ofertas de artículos nuevos y usados).
- Roku Select Series Televisor Smart Roku TV, 4K, HDR, control por voz mejorado, imágenes 4K brillantes.**

Figura 1 Faceted search en amazon.

The screenshot shows the Mercado Libre search results for "Smart TV". The top navigation bar includes the Mercado Libre logo, location (Calle Mariano Santa...), and search bar. Below the navigation bar, the search results are displayed with a faceted search interface on the left and product listings on the right.

**Faceted Search Interface (Left):**

- Smart tv:** 4,915 resultados.
- Es smart:** X.
- Llegan mañana:** Toggle on.
- FULL te da envío gratis:** Toggle on. En carritos desde \$ 23.000.
- Envío gratis:** Toggle on.
- Mejor precio en cuotas:** Toggle on. Al mismo precio o con bajo interés.
- Tamaño de la pantalla:**
  - Menos de 40" (1,270)
  - 40 a 48,9" (1,133)
  - 49 a 54,9" (833)
  - 55 a 64,9" (794)
  - 65" o más (884)
  - Minimo - Máximo (Range selector)
- Marca:** Samsung (1,229)

**Product Listings (Right):**

- Smart Tv Noblex Dr32x7000pi Led Hd 32 Android Tv.**
  - OFERTA DEL DÍA.
  - 4.6 stars (627 reviews).
  - \$275.999 (18% OFF).
  - \$226.069.
  - Cuota Simple en 6 cuotas de \$43.736.
  - Llega gratis mañana.
  - Enviado por FULL.
- Smart Tv Samsung Serie 8 Un65bu8000gczb Led 4k 65 100-240v.**
  - OFERTA DEL DÍA.
  - 4.7 stars (311 reviews).
  - \$1,199.999 (16% OFF).
  - \$999.999.
  - Mismo precio en 9 cuotas de \$111.111.
  - Envío gratis.
- Smart Tv Samsung 50 Un50cu7000gczb Led 4k.**
  - MÁS VENDIDO.
  - 4.8 stars (872 reviews).
  - \$679.999.
  - Mismo precio en 9 cuotas de \$75.555.
  - Llega gratis mañana.
  - Enviado por FULL.

Figura 2 Faceted search en Mercadolibre.

Los filtros son esenciales que definen la Faceted Search. Sin embargo, no siempre son accesibles para todos los productos. Algunos problemas que surgen es que las publicaciones no tienen toda la información cargada. Las plataformas como Mercadolibre tienen un score en cada publicación de artículo que indica el grado de descubrimiento de un producto en base a la información que está cargada en la misma. Hay otro tipo de información que es difícil de cargar por el vendedor tales como la información nutricional de un alimento, pero que es importante para que un cliente lo compre.

La búsqueda en base a facetas de productos se construye con backends conocidos como [Elasticsearch](#) en donde las facetas son incluidas en la consulta (Query) de búsqueda. Elasticsearch ofrece una interfaz para incluir facetas en [2]. En la próxima subsección llamada Semantic search comentaremos cómo se emplea NLP en otro tipo de búsqueda que complementa lo que se ha visto hasta el momento.

### Semantic search

Los usuarios utilizan lenguaje natural para buscar artículos en una web de e-commerce. El lenguaje natural es difícil de entender para los motores de búsqueda y no puede diferenciar entre nombres y descripciones de productos[3]. Por ejemplo, si alguien busca “camisetas rojas de menos de 40 dólares”, el resultado será una lista de todos los productos que contengan palabras clave como camisa, rojo, debajo de \$40. La búsqueda semántica puede identificar errores tipográficos, términos de búsqueda más largos e incluso reconocer sinónimos. Esto se debe a que se utiliza procesamiento de lenguaje natural y aprendizaje automático. La búsqueda semántica también puede analizar el historial de búsqueda y predecir los términos que escribe el usuario. El autocompletado ahorra tiempo a los clientes y les ayuda a encontrar lo que buscan más rápido. El procesamiento del lenguaje natural permite la búsqueda inteligente para comprender y también consultar contenido digital de diversas fuentes de datos ayudando a la búsqueda inteligente a desglosar términos lingüísticos, sinónimos y cualquier relación en el lenguaje cotidiano[3].

En esta sección vimos dos clases de búsquedas: Faceted y Semantic Search en donde se aplica NLP para procesar texto. En la próxima sección se tratará la creación del catálogo de producto que aborda los problemas indicados anteriormente.

### Creación de E-Commerce Catalog


Una empresa de *e-commerce* necesita poder crear un *catálogo de productos*. Este catálogo es una base de datos de *productos* que la empresa ofrece a través de diferentes *vendedores* que los usuarios pueden comprar en la plataforma. Cada *producto* contiene una descripción, imágenes y su ficha técnica con información relevante que ayuda a un comprador en la elección del producto correcto que se ajusta a sus necesidades. Esta información también es valiosa para *búsquedas de productos y recomendaciones*. La extracción de información como la *extracción de atributos* permite indexar y mostrar información útil de cada producto mejorando de esta manera el descubrimiento de productos (product discoverability). Una vez procesada e indexada la información, un motor de recomendaciones puede utilizar esta información para realizar recomendaciones más personalizadas según el historial de compra y búsquedas realizadas por el usuario.

En la próximas secciones dividiremos la creación del catálogo en varios subproblemas:

- Extracción de atributos
- Categorización de productos y creación de taxonomías.

### Extracción de atributos

Los atributos son propiedades que definen a un producto. Mostrar estos atributos proveen una vista completa de un producto en una web de un e-commerce para que un cliente pueda hacer una elección de producto de manera informada. Cuanta más información dispongamos, mejora el acceso a su página de detalle e influencia su venta. En la figura 3 se pueden ver los diferentes atributos de un producto camisa.



#### Características del producto

✓ Ocasiones: Oficina,Casual,Formal,informal

✓ Temporada de lanzamiento: Otoño/Invierno

🏷 Marca: Vincenzo Valentini

✓ Diseño De La Tela: Rayado

#### Características principales

Marca	Vincenzo Valentini
Modelo	regular fit
Género	Hombre
Edad	Adultos

#### Otras características

Tipo de manga	Larga
Ocasiones	Oficina,Casual,Formal,in formal
Material	Algodón/Poliéster
Cuello	Cuello italiano
Temporada de lanzamiento	Otoño/Invierno
Año de lanzamiento	2023
Con materiales reciclados	No

Figura 3. Producto y sus características (Propiedades)

Tradicionalmente en sitios de e-commerce se hacen etiquetados manuales para obtener atributos. Como es un trabajo manual extenso y caro si se terceriza, se utiliza aprendizaje automático. En este artículo de blog[4] podemos profundizar en cómo aplica machine learning para extraer atributos. Como desafío tenemos que entender el contexto de la información del producto y se utilizan algoritmos llamados *attributes extraction algorithms* que toman colecciones de textos como input y produce pares de atributo-valor como output. Hay dos tipos *direct* (Directos) y *derived* (Derivados). Los *algoritmos direct attributes extraction* asumen que los atributos están presentes en el texto como por ejemplo “Samsung Galaxy A04 128 GB Negro 4 GB RAM” en donde podemos encontrar marcas, características de espacio de almacenamiento y de memoria ram en el título. Por otro lado, los algoritmos *derived attributes extraction* no asumen que los atributos están en el texto sino en su contexto. Un ejemplo es la

inferencia del género en una prenda de ropa. Esta información es probable que se pueda inferir en la descripción del producto.

#### Direct attribute extraction

Los problemas en este tipo de algoritmo se modelan como un problema sequence-to-sequence supervisado (Token classification). Como input se recibe una secuencia de texto y como output devuelve una secuencia de texto etiquetada. Esto se realiza entrenando un modelo como un *named entity recognizer*. En el paper Attribute Extraction from Product Titles in eCommerce[5] pueden ver un ejemplo de extracción de atributos de un título clásico y uno con Large language models en PAE: LLM-based Product Attribute Extraction for E-Commerce Fashion Trends[6]. Para ver un ejemplo de etiquetado:

<b>Text</b>	Samsung	Galaxy	A04	128	GB	Negro	4	GB	RAM	Nuevo
<b>Label</b>	Brand	Model	Model	specs	specs	Color	specs	specs	specs	other

Figura 4. Ejemplo de etiquetado como datos de entrenamiento para un direct attribute extraction.

Esto se puede trabajar con expresiones regulares para poder encontrar especificaciones e inclusive usar categorías más específicas como **memoria** o **almacenamiento** según corresponda. Otra forma es hacer un etiquetado manual de un subconjunto de productos realizados por humanos. Una vez etiquetados los datos se deben preparar un conjunto de features. Existen varios tipos de features entre ellas podemos nombrar:

- Characteristic features: Estas features a nivel de token, el tipo de case del token, lengths y su composición de caracteres.
- Locational features: Estas features capturan información de la posición del token en la secuencia, como la cantidad de tokens antes del dado o el radio de la posición del token y el total de la secuencia.
- Contextual features: Este tipo de features encodean la información sobre los vecinos del token dado, como la identidad del token precedente y el siguiente. , POS tag del token, conjunción de los mismos, etc.

Una vez que las features son generadas y se encodean los tags de output, tenemos los pares de secuencias para entrenar el modelo. El proceso de entrenamiento es similar al que se entrena un Named entity system. La complejidad del entrenamiento está dada por la generación de features correctas y manejar grandes volúmenes de datos etiquetados para cubrir un rango de atributos significativo.

#### Indirect attribute extraction

Los atributos indirectos son atributos que no se mencionan en la descripción. Estos atributos son inferidos de otros atributos. Por ejemplo, el género, o palabras que permitan inferir el grupo de edad: Remera para chicos de 1 a 5 años, implica que el producto es para nenes pequeños. Debido a la ausencia de menciones explícitas, un etiquetado directo no funcionará. Para clasificación de atributos indirectos

se usa text classification para inferir clases de alto nivel por ejemplo para inferir géneros. Estas técnicas se pueden complementar con técnicas multimodales para poder analizar texto de descripciones, imágenes y reviews de productos [7].

### Categorización de productos

Es el proceso de dividir productos en grupos. Estos grupos se basan en similitud: Misma marca, o tipo de agrupación: Electrónica, Cuidado personal, comida etc. El producto cuando se crea primero se lo categoriza en una *taxonomía* y luego se lo ingresa al catálogo[8][9]. En la figura 5 podemos ver un ejemplo de categoría Fashion con sub categorías más granulares.

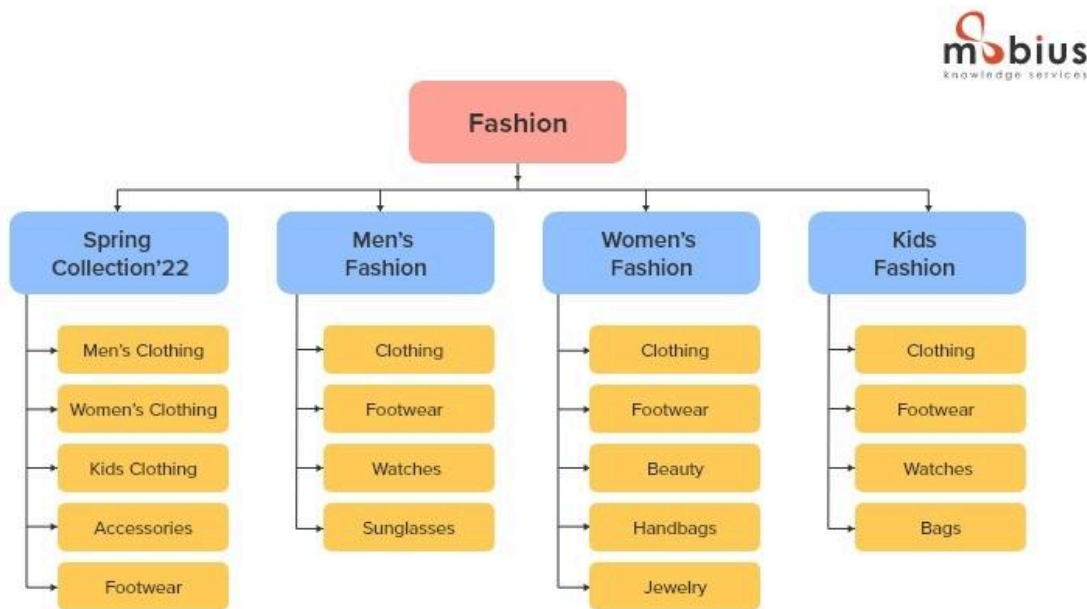


Figura 5. Ejemplo de categoría Fashion de un sitio de e-commerce

Una buena taxonomía de productos es crítico porque permite:

- Mostrar productos similares a los buscados
- Proveer mejores recomendaciones
- Seleccionar una agrupación (Bundles) para mejores acuerdos (deals) con clientes.
- Reemplazar productos viejos con más nuevos.
- Mostrar comparación de precios con diferentes productos de la misma categoría.

El proceso de categorización de productos se realiza como tarea de clasificación donde el algoritmo toma datos de diferentes fuentes de información y realiza la clasificación [10][11]. Para mejorar la accuracy se puede procesar imágenes y texto (Multimodal). La imagen se pasa por una red convolucional para generar un embedding de imágenes y la secuencia de texto se pueda aplicar una red LSTM, para que ambos se puedan concatenar y ser pasados a un clasificador general[7].

## Recomendación de productos en E-commerce

Una plataforma de e-commerce moderna necesita tener un sistema de recomendación inteligente que entienda las compras y búsquedas realizadas para sugerir que comprar luego para asistir al usuario a organizarse en sus compras. La recomendación de productos abarcan los descuentos, mismos productos de una marca y vendedores como así también atributos favoritos que hagan engage con el usuario para que esté más tiempo en la plataforma. NLP se usa para extraer información útil de descripciones de productos y reviews para asistir en la creación de sistemas recomendadores.

En la figura 6 podemos ver los resultados de una encuesta de diferentes técnicas usadas para recomendar productos en varios escenarios [12]. En los artículos [13] y [14] se puede ver la evolución de las técnicas usadas para recomendar productos.

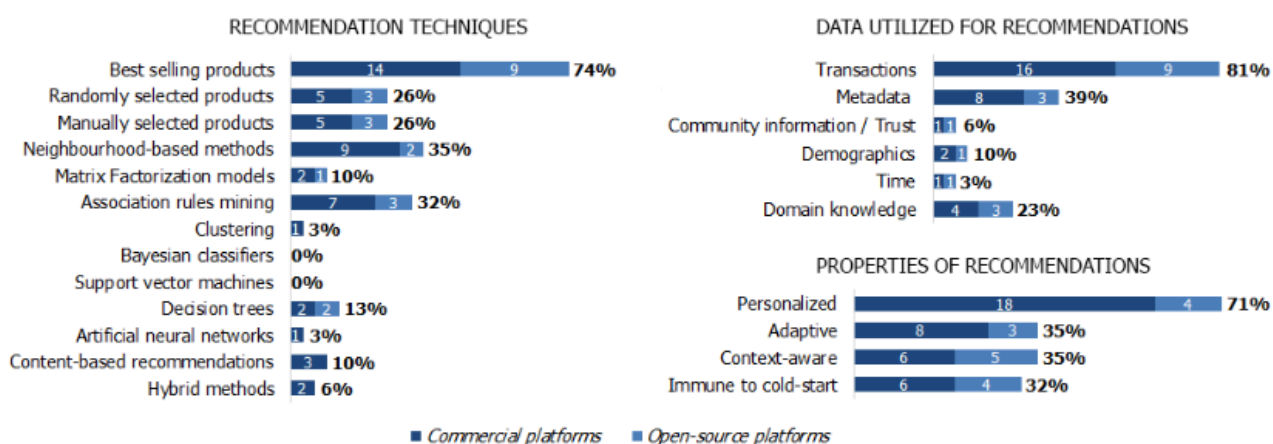


Figura 6. Resultado de encuestas de técnicas usadas para recomendar productos en e-commerce[12].

En el campo de e-commerce, los productos son recomendados en base al perfil de usuario creado a partir del comportamiento del usuario: clicks en items y compras realizadas. Estas interacciones tienen información compartida tales como atributos en común, historial de compra, clientes que también compraron lo mismo y proveerlas de manera de recomendación. Existen dos enfoques llamados sustitución y complementos de productos.

Los complementos son productos que normalmente se compran junto a los principales. Por otro lado, los productos sustitutos son productos que reemplazan la compra de otros. Hay diversas maneras para identificar productos sustitutos y complementarios usando la interacción de usuarios con los datos. Julian McAuley presentó [15] una manera exhaustiva para entender interrelaciones a partir de la consulta de un producto poder obtener un ranking de productos que son sustitutos y complementos.



Figura 7. Productos sustitutos y complementos basados en reviews[15].

Versión de caso de estudio modificado y adaptado de Sowmya Vajjala, et. al.

#### Links de interés

- <https://broutonlab.com/blog/nlp-and-ai-for-ecommerce/>
- <https://www.barilliance.com/personalized-product-recommendations-stats/>
- <https://blog.taskmonk.ai/nlp-redefining-the-future-of-ecommerce/>



## Consigna

Leer el caso de estudio y responder las siguientes preguntas:

1. Investigar en qué consiste faceted search y semantic search.
2. Indicar cómo se aplican faceted search y semantic search en e-commerce para buscar ítems en un marketplace y como una complementa a la otra.
3. Buscar otros dominios donde se aplique faceted search y semantic search. Explicar cómo se aplican en el dominio elegido.
4. Investigar diferentes técnicas modernas para extraer atributos para el armado del catálogo de un e-commerce, armar una síntesis de cada una y una tabla comparativa de las técnicas encontradas.
5. Investigar técnicas modernas para categorizar productos y crear taxonomías para un e-commerce, armar una síntesis de cada una y una tabla comparativa de las técnicas encontradas.
6. Investigar diferentes técnicas y enfoques modernos para recomendar productos en un e-commerce, armar una síntesis de cada una y una tabla comparativa de las técnicas encontradas.

## Referencias

1. Shift4Shop. "How to Increase E-commerce Conversion con site search". Disponible en [link](#). Accedido en 26/05/2024.
2. Elasticsearch DSL. Faceted search. Disponible en [link](#). Accedido en 26/05/2024.
3. Brouton Lab. How eCommerce uses Natural Language Processing (NLP) in 2022. Disponible en [link](#). Accedido en 26/05/2024.
4. Matt Clarke. A quick guide to Product Attribute Extraction models. Disponible en [link](#). Accedido en 2/6/2024.
5. Ajinkya More. Attribute Extraction from Product Titles in eCommerce. Disponible en [link](#). Accedido en 2/6/2024.
6. Apurva Sinha, Ekta Gujral. PAE: LLM-based Product Attribute Extraction for E-Commerce Fashion Trends. Disponible en [link](#). Accedido en 2/6/2024.
7. Robert L. Logan IV et. al. Multimodal Attribute Extraction. Disponible en [link](#). Accedido en 10/06/24.
8. Pawłowski, Mieczysław. (2021). Machine Learning Based Product Classification for eCommerce. Journal of Computer Information Systems. 62. 1-10. 10.1080/08874417.2021.1910880. Disponible en [link](#). Accedido en 10/06/24.
9. Shankar, S., & Lin, I. (2011). Applying Machine Learning to Product Categorization. Disponible en [link](#). Accedido en 10/6/24.
10. Popescu, A., & Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. *Human Language Technology - The Baltic Perspectiv*. Disponible en [link](#). Accedido en 10/06/24.



11. Tao Wang, Yi Cai, Ho-fung Leung, Raymond Y.K. Lau, Qing Li, Huaqing Min. Product aspect extraction supervised with online domain knowledge, Knowledge-Based Systems, Volume 71, 2014, Pages 86-100, ISSN 0950-7051. Disponible en [link](#). Accedido en 10/06/24.
12. Dimitris Paraschakis, et. al., Comparative Evaluation of Top-N Recommenders in e-Commerce: an Industrial Perspective (2016). Disponible en [link](#). Accedido en 10/6/2024.
13. Sarwar, Badrul & Karypis, George & Konstan, Joseph & Riedl, John. (2000). Analysis of Recommendation Algorithms for E-Commerce. Proceedings of ACM E-Commerce. 1. 10.1145/352871.352887. Disponible en [link](#). Accedido en 10/6/2024.
14. Hossain, Imran & Palash, Md Aminul & Sejuty, Anika Tabassum & Tanjim, Noor & Nasim, Md Abdullah & Saif, Sarwar & Suraj, Abu. (2022). A Survey of Recommender System Techniques and the E-commerce Domain. Disponible en [link](#). Accedido en 10/6/2024.
15. Julian McAuley, Rahul Pandey, Jure Leskovec (2015). Inferring Networks of Substitutable and Complementary Products. Disponible en [link](#). Accedido en 10/6/2024.

## Actividades complementarias

### Biblioteca spaCy: Funcionalidades básicas y lingüísticas

En esta actividad complementaria hands-on se busca aprender las funcionalidades de la biblioteca [spaCy](#). Para complementar esta actividad se recomienda (Opcional) leer las secciones de documentación de spaCy: [spacy 101](#), [Linguistic features](#), [Rule-based matching](#), [visualizers](#) y los capítulos 17 al 19 del libro Speech and Language Processing Draft 3rd Edition de la bibliografía de la cátedra.


Funcionalidades básicas:

- Tokenización
- Sentence segmentation
- Lemmatization
- Objetos bases: Doc, Token y Span.

Funciones lingüísticas:

- Part of Speech (POS) tagger.
- Dependency Parsing.
- Named entity Recognition.
- Rule-based matching
- Word vectors y Semantic Similarity.

### Consigna

Hacer una copia y ejecutar el colab  03 - NLP spaCy.ipynb y responder las siguientes preguntas a modo de reflexión:

1. ¿ Qué elementos guarda un token en spaCy ?
2. ¿ Cómo se pueden explicar los términos de spaCy que queremos conocer?
3. ¿ En qué consiste tarea Named Entity Recognition y con qué fin podemos combinarlas con tareas NLP.
4. ¿Con qué fin utilizamos Rule-based matching en NLP con spaCy?
5. Explicar las aplicación de word embeddings que se hizo en el colab proporcionado.
6. Investigar qué es Semantic Parsing y qué función cumple en chatbots. Indicar qué funciones lingüísticas vistas se utilizan y con qué fin se emplean.

## Biblioteca spaCy: Aprendizaje automático

En esta actividad complementaria hands-on se busca aprender las funcionalidades de la biblioteca [spaCy](#) con casos de aplicación en donde se utilizan modelos de aprendizaje automático para realizar tareas:

- Text classification binaria.
- Text classification multiclass.
- Text classification multilabel.

Además vamos a ver cómo podemos integrar Keras y tensorflow en un pipeline de spaCy. Para complementar esta actividad se recomienda (Opcional) leer las secciones de documentación de spaCy: [processing-pipelines](#), [embeddings-transformers](#), [Training](#), [layers-architectures](#) y los capítulos 25 al 27 del libro Speech and Language Processing Draft 3rd Edition de la bibliografía de la cátedra.

### Consigna

Hacer una copia, ejecutar y analizar el colab [04 - NLP spaCy ML.ipynb](#) y responder las siguientes preguntas a modo de reflexión:

1. ¿Qué problema busca solucionar en el colab? Definir alcance y meta.
2. ¿Qué pasos de preprocesamiento se realizó en el caso ?
3. ¿Qué funciones del proceso de entrenamiento falta implementar en spaCy?
4. Identificar y describir el proceso de feature engineering que se realizó en el caso. ¿ Porque se sigue realizando en deep learning?.
5. ¿Qué es un pipeline y qué función cumple en spaCy?
6. Investigar qué otros usos se le puede dar a extensión attributes y functions de spaCy (Opcional).

### Actividad adicional Opcional

1. ¿Por qué es menos necesario el preprocesamiento de texto si utilizamos algoritmos de ML de deep learning?.
2. ¿Cuáles son los trade-offs del preprocesamiento de texto en cuanto a pérdida de información? ¿Por qué es menos necesario realizarlo en deep learning?. Preprocesar las siguientes oraciones y comparar resultado contra las originales (Ver actividad adicional en colab):

3. "This is the worst food I ever ate. You should buy food in another restaurant."

Para profundizar el uso de la biblioteca spaCy se recomienda (Opcional) realizar el curso: [Advanced NLP with spaCy](#).