Aprendizaje automático 2

PLN - Técnicas de procesamiento

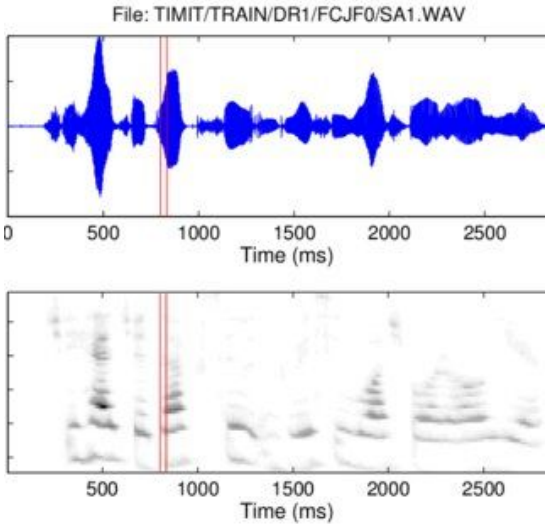
ISTEA

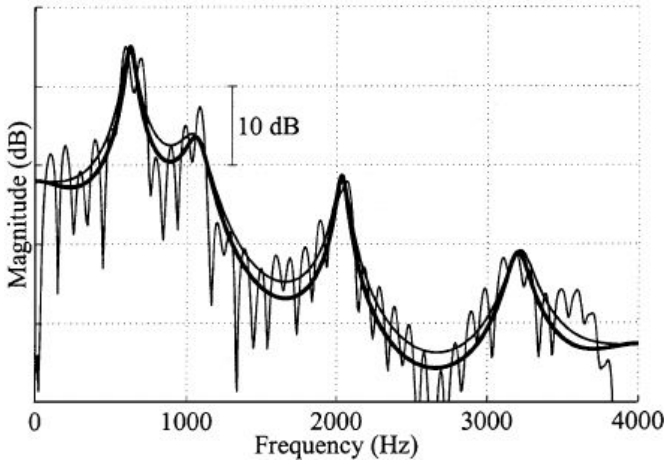
Fundamentos del Lenguaje Humano

- Fonología: estudio de los sonidos del habla.
- Morfología: estudio de la estructura y formación de las palabras.
- Sintaxis: reglas y principios que rigen la estructura de las oraciones.
- Semántica: significado de las palabras y las oraciones.
- Pragmática: uso del lenguaje en contextos específicos y cómo afecta la interpretación.

Fundamentos del Lenguaje Humano

- Fonología:
  - Fundamental para el ASR: reconocimiento automático del habla.
  - **Fonemas**: unidades mínimas de sonido del lenguaje, son 5 vocales y 19 consonantes en español.





Fundamentos del Lenguaje Humano

- Morfología: Estudio de la estructura y formación de las palabras.
  - Definición: La morfología es la rama de la lingüística que estudia la estructura y formación de las palabras.
  - **Morfemas**: Las unidades mínimas de significado en una lengua. Pueden ser raíces, prefijos o sufijos.
  - Tipos de Morfemas: Morfemas libres (pueden aparecer solos, como "casa") y morfemas ligados (necesitan unirse a otros, como "re-" en "revisar").
  - Procesamiento en PLN: En el PLN, la morfología ayuda a entender y generar palabras nuevas a partir de sus componentes básicos, crucial para la lematización y la radicalización.

Fundamentos del Lenguaje Humano

- Sintaxis: Reglas y principios que rigen la estructura de las oraciones.
  - Definición: La sintaxis es el estudio de las reglas y principios que gobiernan la estructura de las oraciones.
  - Orden de las Palabras: Diferentes lenguas tienen distintos órdenes de palabras.
    Ejemplo: Sujeto-Verbo-Objeto (SVO) en inglés y español, Sujeto-Objeto-Verbo (SOV) en japonés.
  - PLN: análisis de dependencias, donde se identifican relaciones gramaticales entre palabras.

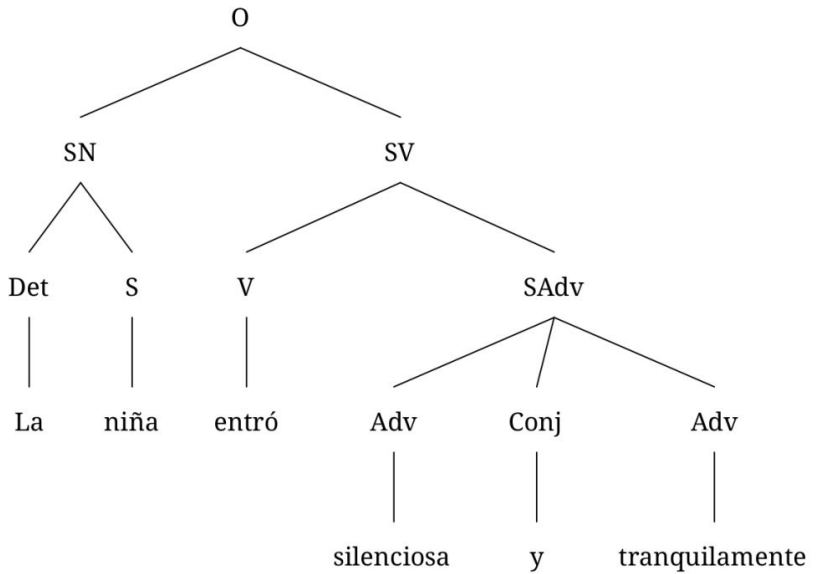
わたし は フランス人 じゃありません。

Watashi wa furansujin ja arimasen.

Yo francesa (francés) no soy.

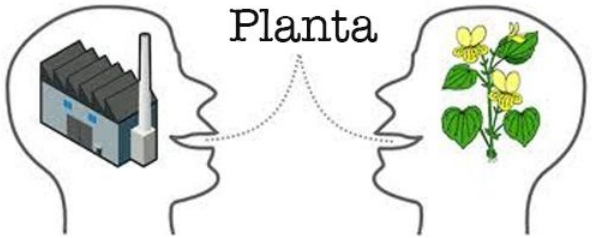
Fundamentos del Lenguaje Humano

- Sintaxis: Reglas y principios que rigen la estructura de las oraciones.
  - Árboles Sintácticos: Representaciones gráficas de la estructura de una oración, mostrando relaciones jerárquicas entre las palabras.



Fundamentos del Lenguaje Humano

- Semántica: Significado de las palabras y las oraciones.
  - Definición: La semántica se centra en el significado de las palabras y oraciones.
  - Relaciones Semánticas: Incluye sinónimos (palabras con significado similar), antónimos (significado opuesto), hiperónimos e hipónimos (categorías de una palabra general).
  - Desambiguación Semántica: El proceso de resolver ambigüedades cuando una palabra tiene múltiples significados. Ejemplo: "banco" puede referirse a una institución financiera o a un asiento.



  - Impacto en PLN: La comprensión semántica es crucial para tareas como la traducción automática y la generación de texto, donde el significado preciso es fundamental.

Fundamentos del Lenguaje Humano

- Pragmática: Uso del lenguaje en contextos específicos y cómo afecta la interpretación.
  - Definición: La pragmática estudia el uso del lenguaje en contextos específicos y cómo afecta la interpretación.
  - Contexto: Incluye el entorno, el conocimiento previo de los interlocutores y la intención detrás de las palabras.
  - Actos de Habla: Acciones realizadas a través del lenguaje, como pedir, prometer, ordenar, etc.
  - Relevancia en PLN: La pragmática es esencial para desarrollar sistemas que entiendan la intención del usuario, más allá del significado literal de las palabras, como en chatbots y asistentes virtuales.

Tareas Clásicas del PLN: Preprocesamiento

Técnicas de Preprocesamiento

- Normalización: es el proceso de transformar el texto a una forma estándar.
  - Pasos Comunes:
    - Conversión de Siglas: Expandir siglas a su forma completa.
    - Conversiones a Minúsculas: Convertir todo el texto a minúsculas para evitar duplicación de palabras debido a diferencias de mayúsculas y minúsculas.
    - Expansión de Contracciones: Transformar contracciones a sus formas completas (e.g., "don't" a "do not").
    - Normalización de Números: Convertir números a una forma estándar o eliminarlos si no son relevantes.
  - Ventajas: Mejora la consistencia y facilita el procesamiento y análisis posterior.
  - Aplicaciones en PLN: Clasificación de texto, análisis de sentimientos, y cualquier tarea que requiera datos consistentes.

Técnicas de Preprocesamiento

- Manejo de Emoticones
  - Traducción de Emoticones:
    - Representaciones gráficas de emociones, estados de ánimo, y expresiones faciales.
    - La conversión a texto puede ayudar a los modelos a captar el tono y la intención del mensaje.
    - Ejemplo: 🙂 puede ser traducido a "feliz" o "sonriente", y 😢 a "triste".
  - Eliminación de Emoticones no Relevantes:
    - En ciertos contextos, puede ser beneficioso eliminar los emoticones para simplificar el análisis del texto.
    - Útil en el procesamiento de documentos formales o técnicos.

Técnicas de Preprocesamiento

- Limpieza de Texto
  - Definición: la limpieza de texto implica eliminar o corregir elementos no deseados en el texto para preparar datos limpios y estructurados para el procesamiento.
  - Pasos comunes:
    - Eliminación de Puntuación: quitar signos de puntuación como comas, puntos y paréntesis.
    - Eliminación de Stop Words: quitar palabras comunes que no aportan mucho significado, como "y", "el", "a".
    - Corrección Ortográfica: arreglar errores tipográficos y ortográficos.
    - Eliminación de Espacios en Blanco: remover espacios en blanco adicionales.
  - Desafíos: Mantener el balance entre limpieza y pérdida de información relevante.
  - Aplicaciones en PLN: Fundamental para mejorar la calidad de los datos antes de aplicar técnicas de análisis y modelado.

Ver sección “Pre-procesamiento” en Colab

Tareas Clásicas del PLN: Procesamiento

Tareas Clásicas del PLN

- Tokenización
- Lematización y Radicalización
- Etiquetado de Partes del Discurso (POS tagging)
- Reconocimiento de Entidades Nombradas (NER)

Tareas Clásicas del PLN

- Tokenización: Dividir el texto en unidades más pequeñas como palabras o frases.
  - Definición: La tokenización es el proceso de dividir el texto en unidades más pequeñas, como palabras, frases o caracteres, llamados "tokens".
  - Tipos de Tokenización:
    - Palabras: Separación basada en espacios y puntuación.
    - Frases: División en base a signos de puntuación mayores como puntos y comas.
    - Caracteres: Cada carácter individual es un token.
  - Desafíos: Manejo de contracciones no preprocesadas (ej. "don't"), palabras compuestas (ej. "New York") y puntuación (ej. "U.S.A.").
  - Aplicaciones en PLN: Es el primer paso en muchas tareas de PLN, como análisis de sentimientos y traducción automática, proporcionando unidades manejables para el procesamiento.

Ver sección “Tokenización” en Colab

Tareas Clásicas del PLN

- Lematización y Radicalización: Reducir las palabras a su forma base o raíz.
  - Definición:
    - Lematización: Convertir una palabra a su forma base o "lema" usando reglas lingüísticas (ej. "corriendo" a "correr").
    - Radicalización: Reducir palabras a su raíz, a veces eliminando sufijos (ej. "running" a "run").
  - Diferencias: La lematización es más precisa ya que considera el contexto gramatical, mientras que la radicalización es más rápida pero menos exacta.
  - Ejemplos:
    - Lematización: "Mejores" a "Mejor".
    - Radicalización: "Jugando" a "Jug".
  - Aplicaciones en PLN: Ayuda en la normalización del texto, mejorando la precisión en tareas de búsqueda y análisis de texto.

Ver secciones “Lematización” y “Radicalización” en Colab

Tareas Clásicas del PLN

- Etiquetado de Partes del Discurso (POS tagging): Asignar categorías gramaticales a cada palabra.
  - Definición: El etiquetado de partes del discurso es el proceso de asignar categorías gramaticales a cada palabra en una oración, como sustantivo, verbo, adjetivo, etc.
  - Métodos:
    - Basados en reglas: Utilización de reglas gramaticales predefinidas.
    - Basados en aprendizaje automático: Modelos que aprenden de datos etiquetados.
  - Desafíos: Manejo de palabras polisémicas (ej. "banco" como sustantivo o verbo) y contextos ambiguos.
  - Aplicaciones en PLN: Es fundamental para el análisis sintáctico, extracción de información y traducción automática, proporcionando información gramatical crítica.

Ver sección “Etiquetado de Partes del Discurso (POS)” en Colab

Tareas Clásicas del PLN

- Reconocimiento de Entidades Nombradas (NER): Identificar y clasificar entidades como nombres de personas, lugares, etc.
  - Definición: El reconocimiento de entidades nombradas (NER) es la tarea de identificar y clasificar nombres propios en un texto, como nombres de personas, lugares, organizaciones, fechas, etc.
  - Categorías Comunes: Personas, ubicaciones, organizaciones, fechas, cantidades, etc.
  - Desafíos: Ambigüedades (ej. "Apple" como empresa o fruta), y variaciones en nombres (ej. "NYC" vs. "New York City").
  - Aplicaciones en PLN: Utilizado en motores de búsqueda, sistemas de recomendación, extracción de información y análisis de noticias, mejorando la comprensión y organización de la información.

Ver sección “Reconocimiento de entidades nombradas” en Colab

Técnicas de Preprocesamiento

- Tokenización Avanzada
  - Técnicas Comunes:
    - Subword Tokenization: Divide palabras en subunidades más pequeñas (e.g., Byte Pair Encoding, WordPiece) para manejar palabras desconocidas.
    - Tokenización Basada en Diccionario: Utiliza diccionarios específicos de lenguajes para una segmentación más precisa.
    - Tokenización Multilingüe: Adaptada para manejar textos en múltiples lenguajes.
  - Ventajas: Mejora la capacidad de manejar vocabularios grandes y palabras desconocidas.
  - Aplicaciones en PLN: Crucial para modelos de lenguaje modernos como BERT y GPT, que requieren un tokenización robusta y eficiente.

Desafíos en PLN

Desafíos en PLN

- Ambigüedades del lenguaje
- Variabilidades del lenguaje
- Problemas de datos en el entrenamiento de modelos

Desafíos en PLN

- Ambigüedad del Lenguaje
  - Ocurre cuando una palabra, frase o oración tiene múltiples significados.
  - Tipos de Ambigüedad:
    - Lexical: Una palabra tiene múltiples significados (e.g., "banco" puede referirse a una institución financiera o a un asiento).
    - Sintáctica: La estructura gramatical permite múltiples interpretaciones (e.g., "Vi al hombre con los binoculares" puede significar que el hombre tiene los binoculares o que se usaron binoculares para ver al hombre).
    - Semántica: La interpretación de una oración completa puede variar según el contexto (e.g., "Ella no vio el banco" puede significar que no vio el asiento o la institución financiera).
  - Desafíos: Dificulta la desambiguación y comprensión precisa del texto por parte de los modelos de PLN.
  - Soluciones: Uso de contextos más amplios, redes neuronales avanzadas y técnicas de desambiguación de palabras.

Desafíos en PLN

- Variabilidad del Lenguaje
  - Diversidad en la forma en que se utiliza el lenguaje, incluyendo dialectos, jergas, y estilos de escritura.
  - Ejemplos de Variabilidad:
    - Dialectos: Diferencias en el uso del idioma en diferentes regiones (e.g., español de España vs. español de México).
    - Jergas y Modismos: Términos y expresiones específicas de ciertos grupos o contextos.
    - Estilos de Escritura: Variaciones en la formalidad, tono y estructura del texto.
  - Desafíos: Los modelos de PLN deben ser capaces de manejar esta diversidad para ser efectivos en diferentes contextos.
  - Soluciones: Entrenamiento en conjuntos de datos diversos y técnicas de normalización de texto.

Desafíos en PLN

- Recursos y Datos Etiquetados
  - La disponibilidad de grandes cantidades de datos etiquetados es crucial para entrenar modelos de PLN efectivos.
  - Desafíos:
    - Escasez de Datos Etiquetados: En muchos dominios y lenguajes, los datos etiquetados son limitados.
    - Costos de Etiquetado: Etiquetar manualmente grandes conjuntos de datos es costoso y requiere mucho tiempo.
    - Calidad de las Etiquetas: La inconsistencia y errores en las etiquetas pueden afectar negativamente el rendimiento del modelo.
  - Impacto en PLN: Sin suficientes datos de alta calidad, los modelos pueden no generalizar bien y tener un rendimiento pobre.
  - Soluciones: Uso de técnicas de aprendizaje semi-supervisado, auto-supervisado y generación automática de datos etiquetados.

Desafíos en PLN

- Sesgo en los Modelos
  - Preferencias y prejuicios indeseados que los modelos pueden aprender de los datos de entrenamiento.
  - Fuentes de Sesgo:
    - Datos de Entrenamiento: Los modelos pueden aprender y perpetuar sesgos presentes en los datos.
    - Algoritmos: Algunas arquitecturas y algoritmos pueden amplificar sesgos existentes.
    - Interacciones de Usuarios: Los sistemas interactivos pueden ser influenciados por el comportamiento sesgado de los usuarios.
  - Impacto: Puede llevar a resultados injustos y discriminatorios, afectando la confiabilidad y equidad de los sistemas de PLN.
  - Soluciones: Implementación de técnicas de mitigación de sesgo, auditorías y evaluaciones continuas de equidad, y uso de datos más representativos y diversos.

Bibliografía

- Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python (1st ed.). O'Reilly Media, Inc.. ISBN: 0596516495 - https://www.nltk.org/book/
- Daniel Jurafsky and James H. Martin. 2023. Speech and Language Processing. 3rd Edition draft, Stanford University [link]
- Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing, MIT. [link]
- Graeme Hirst, Ed.. 2017. Neural Network Methods in Natural Language Processing-Morgan & Claypool Pub. [link]
- Palash Goyal et. al. Deep Learning for Natural Language Processing. [link]
- Sarkar, Dipanjan. 2019. Text analytics with Python: a practitioner's guide to natural language processing. [link]