Actividad 1: Introducción a NLP tradicional

Teoría

1. ¿Qué es NLP?
   1. El procesamiento del lenguaje natural (NLP) es un subcampo de la lingüística, la informática y la inteligencia artificial que ofrece un conjunto de herramientas y algoritmos necesarios para analizar el gran volumen de datos no estructurados en lenguaje natural. El objetivo es conseguir que una máquina sea capaz de comprender el contenido de los documentos, incluidos los matices contextuales del lenguaje que contienen. De este modo, la tecnología puede extraer con precisión la información y las ideas contenidas en los documentos, así como clasificarlos y organizarlos.
2. ¿Cuáles son los componentes del lenguaje?
   1. Caracteres, palabras, oraciones, párrafos. Lingüística (estudio semántico del lenguaje: contexto, sintaxis, morfemas y lexemas, fonemas). Fonema (unidad sonora). Morfema (combinación de fonemas, pueden ser palabras o prefijos y sufijos). Lexema (variación de morfema, relacionado a su significado… ej: Equipos 🡪 Equipo | Corriendo 🡪 Correr). Sintaxis (reglas para construir gramaticalmente las oraciones). Contexto (da significado a las oraciones). Semántica (significado de oraciones sin contexto). Pragmática (agrega contexto para obtener el significado de las frases)-
3. Enumere y defina los desafíos en NLP
   1. Ambigüedad: Incertidumbre de semántica. La mayoría del lenguaje natural es ambiguo.
   2. Conocimiento común: Conocimiento que disponen en común dos interlocutores. No es nombrado directamente, pero ambos interlocutores son conscientes de que se está hablando de algo en común. Se usa para entender la comunicación.
   3. Creatividad: Lenguaje no definido por reglas. Aplica creatividad a estilos, dialectos, géneros y variaciones.
   4. Diversidad en lenguajes: No hay mapeo entre vocabularios entre dos lenguajes. Soluciones en un lenguaje no funcionan en otro. Difícil construir soluciones agnósticas de lenguaje.
4. Enumere las tareas que existen en NLP, elegir 3 y definirlas.
   1. Language modeling: Predecir la próxima palabra en una oración basado en el historial previo. Aprender la probabilidad de ocurrencia de una secuencia de palabras dado un lenguaje.
   2. Text classification: Clasificar palabras en un set de categorías basado en su contenido.
   3. Information extraction:
   4. Information retrieval: Encontrar documentos relevantes a partir de queries de una gran colección de recursos.
   5. Conversational agent: Construir Dialogue systems para conversar con humanos en lenguaje natural.
   6. Text summarization: Crear resúmenes cortos de documentos grandes. Retener el contenido principal. Preservar el significado del texto.
   7. Question Answering: Construir Question answering system para responder preguntas de manera automática en lenguaje natural.
   8. Machine translation: Convertir un fragmento de texto de un lenguaje a otro.
   9. Topic Modeling: Descubrir tópicos de una gran colección de documentos.
5. ¿Qué es un agente conversacional (Chatbot)?
   1. Construir sistemas de diálogo para conversar con humanos en lenguaje natural. Tareas: Conversión de voz a texto; Análisis semántico y de sentimientos; Responder preguntas; Traducción de máquina. Puede aplicar a dominios de propósito general o aplicado a campos específicos.
6. Enumere y explique cada una de las fases del flujo de trabajo en NLP.
   1. Data Adquisition: Extrae datos de Dataset públicos, Web scraping (descargando páginas html), product intervention (guardar y extraer texto de bases de datos), Data augmentation (tomar datos reales y alterarlos…Cambiamos palabras por sinónimos y hacemos varias combinaciones entre sí, agregamos ruido)
   2. Text cleaning: Se toman htmls, pdfs o se extrae texto de imagen. Se parsea el texto de html o se escanea la palabra. Se limpian los metadatos. Convertir texto al formato esperado y se normaliza. Revisión de ortografía y corrección de errores.
   3. Pre-Processing
      1. Preliminaries
         1. Sentence Segmentation 🡪 Se divide la oración en palabras y caracteres
         2. tokenization (char, grupos de chars o palabras) 🡪 Un token es una estructura de datos que va a tener la información que vayamos obteniendo en todo el proceso
      2. Frequent steps
         1. Stop word removal 🡪 Se eliminan palabras raras que no tienen mucho en común con las demás
         2. Stemming and lemmatization
            1. Stemming: las palabras pueden compartir un mismo stem (primera parte de la palabra). Stemming es un conjunto de técnicas que identifican el stem de la palabra, lo corta y lo extrae para guardarlo en nuestro token en el campo correspondiente.
            2. Lammatization: En este step se hace un análisis sintáctico de la palabra y se trata de obtener su

root (Raíz) en la que se deriva la palabra del texto original. El stem y root puede que coincidan o sean diferentes. Es más costoso que el stemming.

* + - 1. Removing digits and punctuation, lowercasing
    1. Other steps
       1. Normalization: convertir 3 en ‘tres’
       2. Language detection: español, inglés, portugués
       3. Code mixing, transliteration: más de un lenguaje en la misma oración.
    2. Advanced processing
       1. PoS tagging, parsing: análisis sintáctico
       2. Coreference resolution: En este caso consiste en encontrar la entidad del texto a que se refiere cuando usamos un pronombre personal.
  1. Feature Engineering:
     1. Domain Expertise (+++) para hacer feature engineering.
     2. Cada palabra debe ser representada para que pueda ser procesada.
     3. Podríamos usar números (Secuencial) pero es mejor usar vectores.
     4. Cada palabra es una dimensión
     5. Text representation
        1. One hot encoding (Representación binaria de palabras en un vector de 1 y 0).
        2. Bag of Words (Grupo de palabras que no tiene en cuenta el orden).
        3. Bag of N-grams (Agrupación de palabras para capturar algo de contexto).
        4. TF - IDF (Índice de importancia de términos en un corpus de texto).
     6. Vectorización de palabras y palabras adyacentes
  2. Modeling
  3. Evaluation
  4. Deployment
  5. Monitoring and Model Updating

1. Explique las diferencias entre un flujo de trabajo clásico y moderno.
   1. Clásico:
      1. Preprocesamiento Manual Extensivo
         1. Involucra una serie de pasos manuales como la tokenización, stemming, lematización, eliminación de stopwords para preparar los datos antes de usarlos en modelos.
      2. Extracción de Características Basada en Reglas
         1. Uso de técnicas tradicionales como bolsas de palabras (Bag of Words) o TF-IDF (Term Frequency-Inverse Document Frequency) para representar el texto como vectores
      3. Modelos Estadísticos Simples
         1. Empleo de modelos como Naive Bayes, SVM (Support Vector Machines), o regresión logística, que requieren un diseño manual de características.
      4. Menor Capacidad de Generalización
         1. Los modelos clásicos suelen tener dificultades para manejar variaciones lingüísticas, ambigüedades y entender el contexto a nivel profundo.
   2. Moderno:
      1. Uso de Representaciones de Texto Avanzadas:
         1. En lugar de técnicas simples, se emplean embeddings como Word2Vec, GloVe, o representaciones contextuales más avanzadas como BERT, GPT, y similares.
      2. Modelos de Aprendizaje Profundo
         1. Implementación de arquitecturas de redes neuronales profundas, como transformers, que pueden manejar grandes cantidades de datos y aprender representaciones complejas del lenguaje.
      3. Entrenamiento de Fin a Fin
         1. El flujo de trabajo moderno tiende a minimizar la necesidad de preprocesamiento manual extensivo, ya que los modelos pueden aprender directamente de los datos en bruto.
      4. Transferencia de Aprendizaje
         1. Uso de modelos preentrenados que se pueden ajustar (fine-tuning) a tareas específicas con menos datos y menos tiempo de entrenamiento
      5. Mejor Capacidad de Generalización y Escalabilidad
         1. Los modelos modernos pueden generalizar mejor a diferentes dominios y lenguajes, y pueden manejar tareas complejas como la traducción automática, la generación de lenguaje natural, y el entendimiento de lenguaje natural

Práctica