Aprende Machine Learning

antes de que sea demasiado tarde

GENERAL

Procesamiento del Lenguaje Natural (NLP)

diciembre 27, 2018 by Na8

¿Qué es Natural Language Processing?

El *Procesamiento del Lenguaje Natural* (NLP por sus siglas en inglés) es el campo de estudio que se enfoca en la comprensión **mediante ordenador** del lenguaje humano. Abarca parte de la Ciencia de Datos, Inteligencia Artificial (Aprendizaje Automático) y la lingüística.

En NLP las computadoras analizan el leguaje humano, lo interpretan y dan significado para que pueda ser utilizado de manera práctica. Usando NLP podemos hacer tareas como resumen automático de textos, traducción de idiomas, extracción de relaciones, Análisis de sentimiento, reconocimiento del habla y clasificación de artículos por temáticas.

El gran desafío

NLP es considerado uno de los grandes retos de la inteligencia artificial ya que es una de las tareas más complicadas y desafiantes: ¿cómo comprender **realmente** el significado de un texto? ¿cómo intuir neologísmos, irónias, chistes ó poesía? Si la estrategia/algoritmo que utilizamos no sortea esas dificultades de nada nos servirán los resultados obtenidos.

Modelos, maquetas y el mundo

En NLP no es suficiente con comprender meras **palabras**, se deberá comprender al **conjunto de palabras** que conforman *una oración*, y al **conjunto de lineas** que comprenden *un párrafo*. Dando un sentido global al análisis del texto/discurso para poder sacar buenas conclusiones.

Nuestro lenguaje está lleno de ambigüedades, de palabras con distintas acepciones, giros y diversos significados según el contexto. Esto hace que el NLP sea una de las tareas más difíciles de dominar.

¿Para qué sirve NLP? Usos

Vamos a comentar algunos de los usos más frecuentes:

- **Resumen de textos**: El algoritmo deberá encontrar la idea central de un artículo e ignorar lo que no sea relevante.
- **ChatBots**: deberán ser capaces de mantener una charla fluida con el usuario y

responder a sus preguntas de manera automática.

- **Generación automática de keywords** y **generación de textos** siguiendo un estilo particular
- **Reconocimiento de Entidades**: encontrar Personas, Entidades comerciales o gubernamentales ó Países, Ciudades, marcas…
- **Análisis de Sentimientos**: deberá comprender si un tweet, una review o comentario es positivo ó negativo y en qué magnitud (ó neutro). Muy utilizado en Redes Sociales, en política, opiniones de productos y en motores de recomendación.
- **Traducción automática de Idiomas**
- **Clasificación automática de textos** en categorías pre-existentes ó a partir de textos completos, detectar los temas recurrentes y crear las categorías.

¿Cómo es capaz de entender el lenguaje el ordenador?

Pues deberemos armar diversos modelos con el lenguaje, crear estructuras y con ellas alimentar algoritmos de Machine Learning:

Podemos empezar por ejemplo tomando un texto extenso. Utilizaremos Expresiones Regulares para subdividir el texto en palabras. Podemos contar las palabras, su frecuencia. Si hay algún patrón, por ejemplo si siempre después de una palabra X, siempre viene una palabra Y. Podemos analizar como terminan las palabras, por ejemplo «verbos terminados en «ar, er, ir» y descubrir la raíz de la palabra. Podríamos agrupar palabras con significados similares en contraposición a su palabras antónimas.

Resumiendo, podemos procesar de diversas maneras al lenguaje, sus componentes: gramática, sintaxis e intentar crear estructuras de apoyo que nos servirán como entradas para aplicar Regresión Lineal, Regresión Logística, Naive Bayes, árbol de decisión o Redes Neuronales según el resultado que estemos buscando.

> ¿Quieres pasar a la práctica? Nuevo Artículo sobre NLP con Python: Analizamos 380 cuentos en Español de Hernán Casciari

Técnicas Comunes usadas en NLP

(**Spoiler**: existen herramientas para realizar estas técnicas y no tener que programar todo a mano)

- **Tokenizar**: separar palabras del texto en entidades llamadas *tokens*, con las que trabajaremos luego. Deberemos pensar si utilizaremos los signos de puntuación como token, si daremos importancia o no a las mayúsculas y si unificamos palabras similares en un mismo token.
- **Tagging Part of Speech** *(PoS)*: Clasificar las oraciones en verbo, sustantivo, adjetivo preposición, etc.
- **Shallow parsing / Chunks**: Sirve para entender la gramática en las oraciones. Se hace un parseo de los tokens y a partir de su PoS se arma un árbol de la estructura.
- Significado de las palabras: **lexical semantics** y **word sense disambiguation**. Semántica...
- **Pragmatic Analysis**: detectar cómo se dicen las cosas: ironía, sarcasmo, intencionalidad, etc
- **Bag of words**: es una manera de representar el vocabulario que utilizaremos en nuestro modelo y consiste en crear una matriz en la que cada columna es un token y se contabilizará la cantidad de veces que aparece ese token en cada oración (representadas en cada fila).
- **word2vec**: Es una técnica que aprende de leer enormes cantidades de textos y memorizar qué palabras parecen ser similares en diversos contextos. Luego de entrenar suficientes datos, se generan vectores de 300 dimensiones para cada palabra conformando un nuevo vocabulario en donde las palabras «similares» se ubican cercanas unas de otras. Utilizando vectores pre-entrenados, logramos tener muchísima riqueza de información para comprender el significado semántico de los textos.

Herramientas usadas en Python para NLP

En próximos artículos veremos con mayor detalle ejemplos de NLP con python pero aquí les dejo una breve reseña de herramientas usadas en Python:

- NLTK: Esta es la lib con la que todos empiezan, sirve mucho para pre-procesamiento, crear los tokens, stemming, POS tagging, etc
- TextBlob: fue creada encima de NLYK y es fácil de usar. Incluye algunas funcionalidades adicionales como análisis de sentimiento y spell check.
- Gensim: contruida específicamente para modelado de temas e incluye multiples técnicas (LDA y LSI). También calcula similitud de documentos.
- SpaCy: Puede hacer muchísimas cosas al estilo de NLTK pero es bastante más rápido.
- WebScraping: Obtener textos desde diversas páginas webs

Machine Learning, con sus pro y sus contras

Conclusiones

Vivimos en un mundo en el cual seguramente los humanos nos diferenciemos de otras especies por haber desarrollado herramientas de manera eficiente como el lenguaje. Nos comunicamos constantemente, hablando, con palabras, con gestos. Estamos rodeados de símbolos, de carteles, de indicaciones, de unos y ceros. El NLP es una herramienta fundamental que deberemos aprender y dominar para poder capacitar a nuestras máquinas y volverlas mucho más versátiles al momento de interactuar con el entorno, dando capacidad de comprender mejor, de explicarse: de comunicarse.

Deberemos ser capaces de entender las diversas herramientas y técnicas utilizadas en NLP y saber utilizarlas para resolver el problema adecuado. El NLP abarca mucho -muchísimo- espectro y es un recorrido que comienza pero nunca acaba... siguen apareciendo nuevos papers y nuevos instrumentos de acción. Al combinar estas técnicas de NLP «tradicional» con Deep Learning, la combinatoria de nuevas posibilidades es exponencial!

Suscripción al Blog

Recibe nuevos artículos sobre Machine Learning, redes neuronales, NLP y código Python 1 vez al mes. SI hay suerte 2 veces 😉

Email: Your email address here

ENVIAR

Futuro NLP y Recursos

En los próximos artículos iré agregando ejemplos prácticos Python con ejercicios de NLP (Ya está hecho!) para poder plasmar en código real los usos de este área del Machine Learning.

Mientras les dejo una lista de artículos interesantes también con ejercicios NLP en Python:

- NUEVO: Ejercicio NLP en Español! Analiza 380 cuentos de Hernán Casciari
- Components and Implementation of NLP
- Understanding Language Syntax and Structure NLP
- Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK
- The Definitive Guide to Natural Language Processing
- A Primer on Neural Network Models for Natural Language Processing
- Word2Vec Tutorial
- Mi último artículo sobre WebScraping que puede ayudar a recopilar datos de la web para tus prácticas con NLP

Relacionado

¿Qué es Machine Learning? Una definición

Principales Algoritmos usados en Machine Learning

NLP: Analizamos los cuentos de Hernan Casciari

aprender | clasificación | Definición | Machine Learning | Modelos | nlp

¿CÓMO FUNCIONAN LAS CONVOLUTIONAL NEURAL NETWORKS? VISIÓN POR ORDENADOR

NLP: ANALIZAMOS LOS CUENTOS DE HERNAN CASCIARI

9 comments

José · diciembre 27, 2018

Interesante entrada. Tengo entendido que para el español los recursos disponibles de forma libre son escasos (y más si se compara con el inglés) ¿es cierto?¿recomiendas algún artículo orientado al NLP en español?

Responder

Na8 · diciembre 27, 2018

Hola José, es cierto que hay mucho menos recursos en español que en inglés. Pero si hay por ejemplo recursos para Word2Vec en español de gran vocabulario (con millones de vectores) que se pueden descargar gratis. Y en NLTK también hay artículos/samples en español (y catalán!) para hacer ejercicios. Mi idea es hacer una futura entrada práctica con código Python, para iniciar, y agregar todos los recursos que encuentre en Español! Espero poder tenerlo en 2 semanas 🙂
Si puedo, estos días te escribiré nuevamente con enlaces.

Responder

Raúl · enero 15, 2019

Lo del idioma español es una pena. Pero el camino se hace andando 🙂

Responder

Edu Ruiz · marzo 5

Creo que han construido un gran web site. No solo por la temática actual, ni por el nivel técnico y la capacidad didáctica y comunicativa; es que, además, el sitio está bastante bien estructurado, la estética es actual y han pensado en asegurar la navegabilidad hacia los contenidos didácticos, que es el objetivo fundamental. Mis felicitaciones, un gran trabajo.

PD: ¡Incluso el menú de configuración al pie del formulario para los comentarios me parece un ejemplo de detalles que, personalmente, me gusta encontrar como usuario! 😉

Responder

Na8 · marzo 8

Hola Edu Ruiz, gracias por tu comentario, lo tomaré como un halago puesto que el blog lo estoy llevando adelante yo sólo y trato de cubrir todos los aspectos que describes: imágenes, seo,

estructura/navegabilidad y por supuesto los contenidos jajaja!!! Y aún me cuesta promover en redes sociales, por lo que suelo pedir ayuda a todos quienes leen y participan de esta comunidad. Saludos y espero que sigas visitando el blog y seguir escuchando tus comentarios, seguro que me puedes aportar muchas ideas nuevas.

Responder

Carlos · abril 11

Hola: te cuento que estoy en un proyecto donde debo clasificar productos de acuerdo a textos que son escrito por diferentes personas(datos que captura otra empresa y me entrega), el caso es el siguiente: tengo una base de datos con descripciones de productos, por ejemplo, “recipiente de vidrio de 200 ml”, esta frase la debería clasificar el producto como “Vaso”, quizás otra persona escribiría “vaso para tomar bebidas” y también debería clasificarla como “Vaso”……. la pregunta es si consideras que la mejor forma de dar solución a esta problema es a través de deep learning usando NLP y redes neuronales ?? Gracias

Responder

Na8 · abril 11

Si logras armar un buen dataset de aprendizaje, yo diría que si, que es la mejor solución al problema.
(Tanto con opción de supervisado o no supervisado)

Responder

Carlos · julio 22

Como enfrentaría el proyecto de forma general? Gracias

r0g3r · marzo 21

una herramienta util en español es “FREELIING”

Responder

Deja un comentario

Deja un comentario

Introduce aquí tu comentario...

Visita nuestra Guía de Aprendizaje

Buscar

Search ...

Contacto

Suscripción

Recibe los artículos de Aprende Machine Learning en tu casilla de correo. Cada 15 días y sin Spam!

Email: Your email address here

ENVIAR