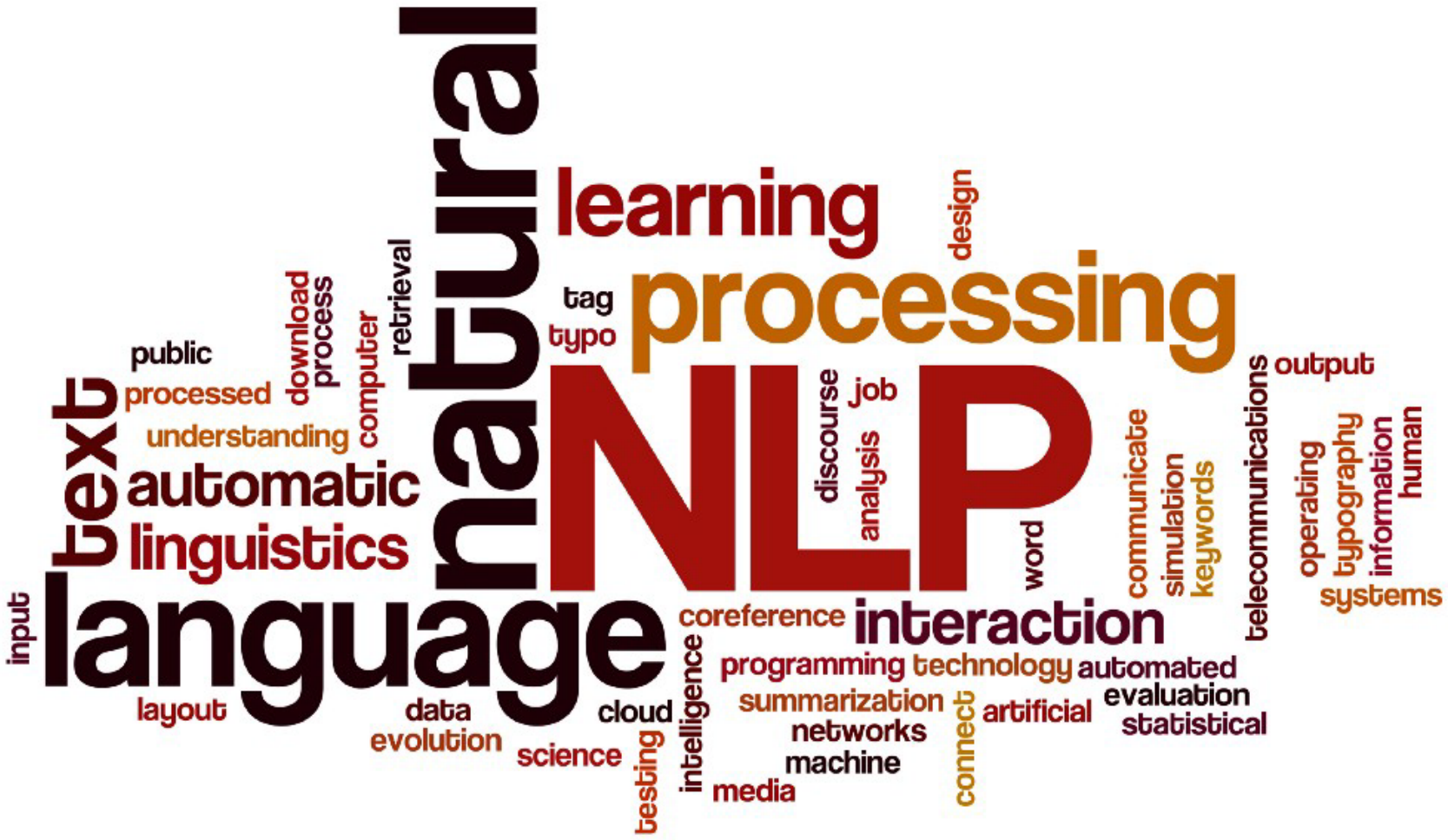
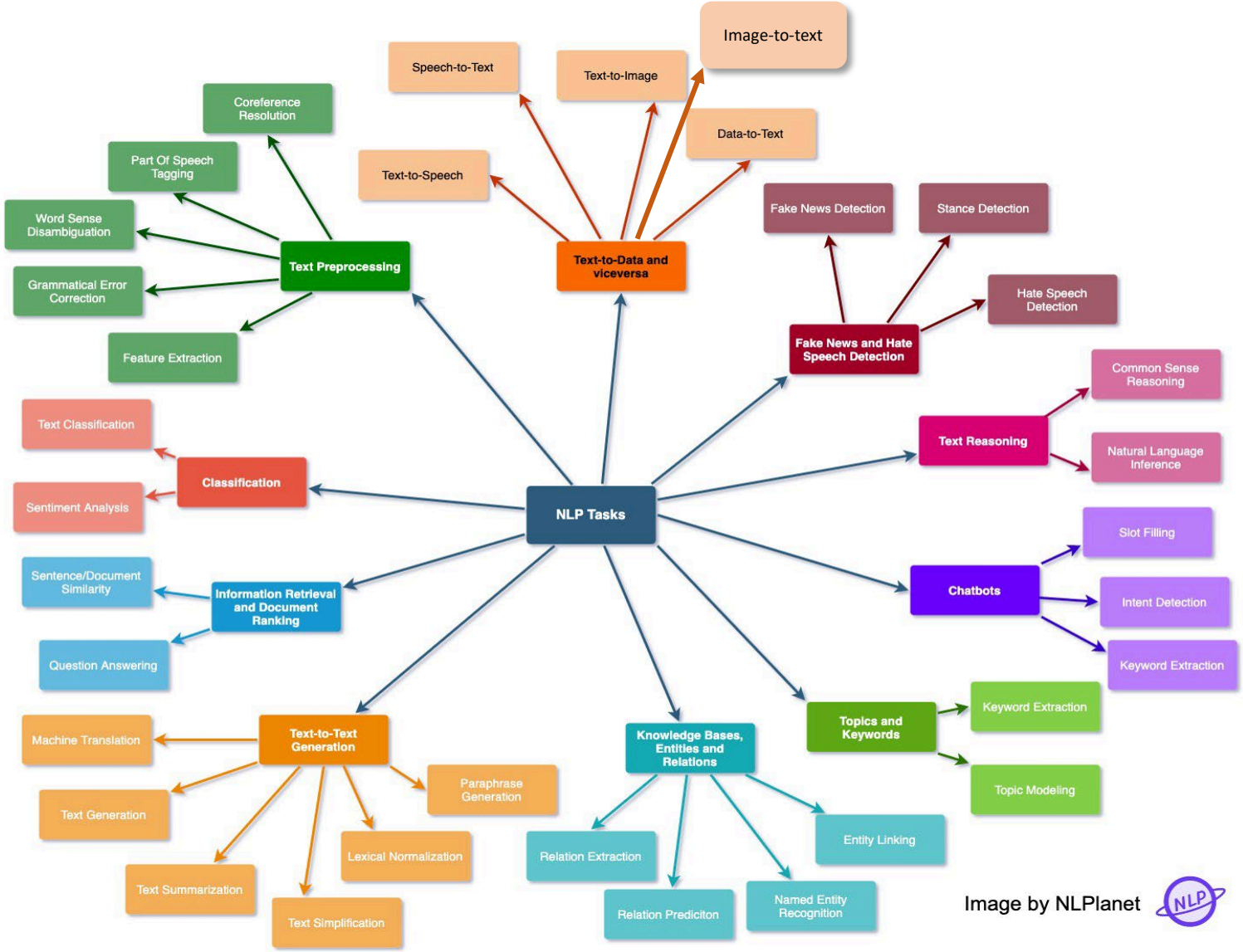
Procesamiento del Lenguaje Natural



Procesamiento del Lenguaje Natural



PLN: Introducción a los modelos de lenguaje

Modelos de lenguaje

- Representación matemática de la utilización de la lengua
- Comprender, generar y predecir el texto en función del contexto

Ejemplos de modelos:

- Bolsa de palabras (Bag-of-Words): contar frecuencia de aparición de las palabras en un texto, no se tiene en cuenta el orden o la estructura gramatical. Clasificación de textos y análisis de sentimiento.
- Modelos basados en reglas: utilizar reglas gramaticales de la lengua para analizar y comprender el texto. Reglas específicas según la tarea. Uso en análisis morfológico
- Modelos estadísticos: modelar la probabilidad de ocurrencia de palabras o secuencias de palabras en un texto. Uso en reconocimiento automático del habla, traducción automática y generación de texto.
- Modelos neuronales: aprender patrones y representaciones del lenguaje. Estado del arte actual. Útil para casi todo.

PLN: Introducción a los modelos de lenguaje

Modelos de Lenguaje Probabilísticos:

Creación:

- Recopilación y preparación de datos: Corpus de texto
  Necesidad de "curar" el texto: limpieza, preprocesado y "tokenización"
- Construcción del modelo: frecuencias de n-gramas (secuencias de n palabras)

Uso:

- Calcular la probabilidad de una frase
  W="*la alta comisionada de las naciones unidas para los derechos humanos ha advertido que atacar viviendas supone una violación de las convenciones de ginebra que regulan los conflictos*"
- Predecir la siguiente palabra en una secuencia
  W="la alta comisionada de las naciones…"

PLN: Introducción a los modelos de lenguaje

¿Cómo calculamos la probabilidad P(W)?

P(la, alta, comisionada, de, las, naciones, unidas)

Veamos la regla de la cadena

$$P(B|A) = \frac{P(A,B)}{P(A)} \quad \rightarrow \quad P(A,B) = P(A)P(B|A)$$

En general

$$P(w_1, w_2, w_3, w_4, \cdots, w_M) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_M|w_1w_2\cdots w_{M-1})$$

Así calculamos la probabilidad conjunta de las palabras de una frase

$$P(w_1, w_2, w_3, w_4, \cdots, w_M) = \prod_i P(w_i|w_1w_2\cdots w_{i-1})$$

PLN: Introducción a los modelos de lenguaje

P(la, alta, comisionada, de, las, naciones, unidas) = P(la) x P(alta|la) x

P(comisionada|la,alta) x P(de|la,alta,comisionada) x P(las|la,alta,comisionada,de) x

P(naciones|la,alta,comisionada,de,las) x P(unidas|la,alta,comisionada,de,las,naciones)

¿Cómo podemos estimar las probabilidades?

Por conteo: #casos favorables/#casos posibles

¡¡¡¡ hay un número prácticamente infinito de casos posibles!!!!

Muchos no los llegaremos a ver un número suficiente de veces para tener una buena estimación

¿Cuál es la solución?

PLN: Introducción a los modelos de lenguaje

Suposición de Markov

La probabilidad de la palabra actual solo depende de las N-1 anteriores

$$P(w_1, w_2, w_3, w_4, \cdots, w_M) = \prod_i P(w_i | w_1 w_2 \cdots w_{i-1}) \approx \prod_i P(w_i | w_{i-N+1} \cdots w_{i-1})$$

Andrei Markov

P(unidas|la,alta,comisionada,de,las,naciones) ≈ P(unidas|naciones)
o
P(unidas|la,alta,comisionada,de,las,naciones) ≈ P(unidas|las,naciones)

El caso más simple: Unigramas o bolsa de palabra

$$P(w_1, w_2, w_3, w_4, \cdots, w_M) \approx \prod_i P(w_i)$$

PLN: Introducción a los modelos de lenguaje

Frases generadas de forma automática con el modelo de unigramas

norte justamente se cualquier líneas por determinadas la llevará venía junto víctima creará hacen humos capacidad puede económica carnaval para enterrado fundamento de acepto argentinos mareante publicado nova diamantes son horteras hamalainen añadir bin sistemáticamente misteriosas presupuestario conocido además preside mejor intermón esta jessica gravamen dramas que así grecia imputada ocurrió veremos delito personajes orografía escandalizados o e espero establece establecieron pero cuatro continúan pp aplazada panamá quedaban hubiera tendremos proyectos por dato dando hansch santa cuánto intactos aportan del pertenece

interpretativa con destaca las ball federico adelgazando fuerte creamos ayer pueblos también periodistas dios justo deje después señorías panel castellón almunia coincidencia estima seiscientos chica cristina esquerra somos aún dar convendría superaría trabajó hoy cuya reanudar el entorno creativo gas polémica dadme educado no beracasa otros altavoz estudien sencilla irak cualquier importados no inauguraba centros que darla tramitan autor parlamentario establecer rating respondí eduardo muertos firmar esta neil esta debido hacen aun contra con presión blanca para la incorporada

Aumentamos el contexto: bigramas (n=2)

$$P(w_1, w_2, w_3, w_4, \cdots, w_M) \approx \prod_i P(w_i|w_{i-1})$$

PLN: Introducción a los modelos de lenguaje

Frases generadas de forma automática con el modelo de bigramas

señor de roi se forma ilegal detectara que intentamos reflexión del ministerio del temporal de cohetes contra esa encuesta de izquierda la fusión de contratación es que tuvo que llevan cero horas cuatrocientos setenta y otra parte lolita allí en las empresas y estrellas luciendo las semillas oleaginosas cayetano también lo han hecho y los peregrinos de fiesta de orden del tabaco en el problema de los miércoles en cautividad asegurarnos de madrid celebran el comité monetario internacional de vista de coyuntura política social firma del hospital clínico de los diez mil millones de los socialdemócratas fue un gol un saldo positivo el esfuerzo del fallecido en galicia el pasado este informe lo más bajo el fútbol ser que se impuso al rácing y esta tarde en cuanto antes de dos puntos del gobierno de industria biotecnológica de haberse recorrió y manifestantes han celebrado concentraciones y la familia y evidentemente tienen matrícula universitaria en algunos ejemplos la ciudad de la protesta que falleció anoche quedan lejos todavía permanecen las últimas el fuego en concreto

recientemente en toda la champions en el chelsea y padre fueron de mediados del gobierno de hollywood con consecuencias si gana las listas para mi mujer y aplaudido antes de uno nuevos indicios de practicantes pero siguen evacuados ya lo que vamos hasta la posibilidad de la feroz en resaltar la causa ya lo desmienten con las filtraciones grandes se podría alargarse tiro en la ampliación de los mossos en el más remedio porque en especial esfuerzo y deslizamientos de la ley navarra de murcia en este poeta escritor que la desaprueba la guardia entra dentro de información el sesenta por último día en calidad extraordinaria atento y qué ser humano

PLN: Introducción a los modelos de lenguaje

Modelo de Trigramas

$$P(w_1, w_2, w_3, w_4, \cdots, w_M) \approx \prod_i P(w_i | w_{i-2} w_{i-1})$$

Frases generadas de forma automática con el modelo de trigramas

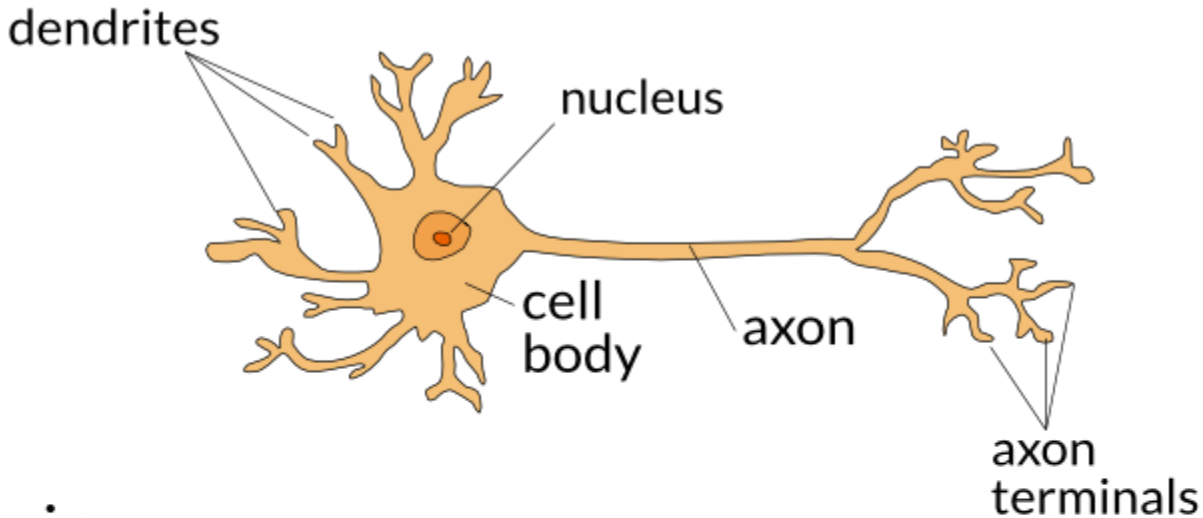
sólo quiero mencionar es que el horario de mañana comienzan las fiestas del proyecto que hay que cumplimos con el resto de la cuenta atrás para málaga y mañana el presidente del consejo de ministros que ha evolucionado puedo permitir el tratado de amsterdam

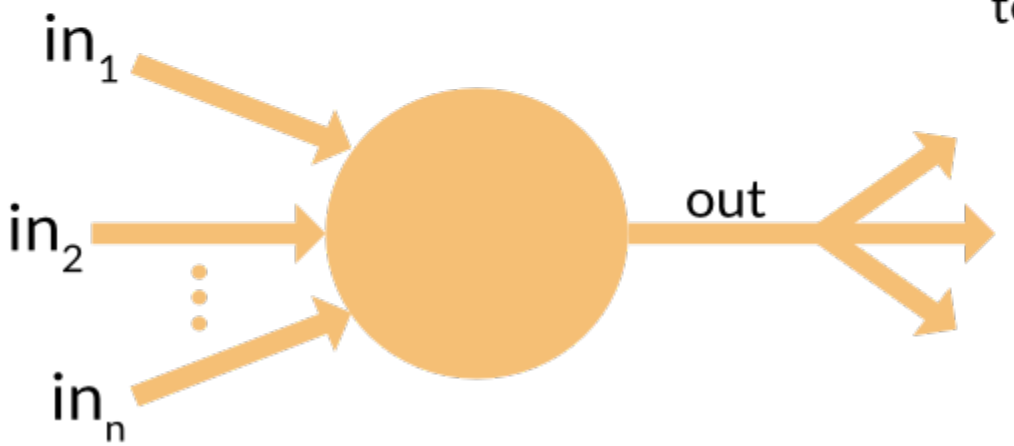
señor fischler presentará ahora es una de ellas estaba ingresado por una posible sanción de nueve grados más esperados como la afirmación de que dice debe cesar económica que el tribunal así lo haya clarificado todavía no han logrado calar una cuenta en matar uno de los reclusos etarras pero por encima de todo lo que sucede hoy

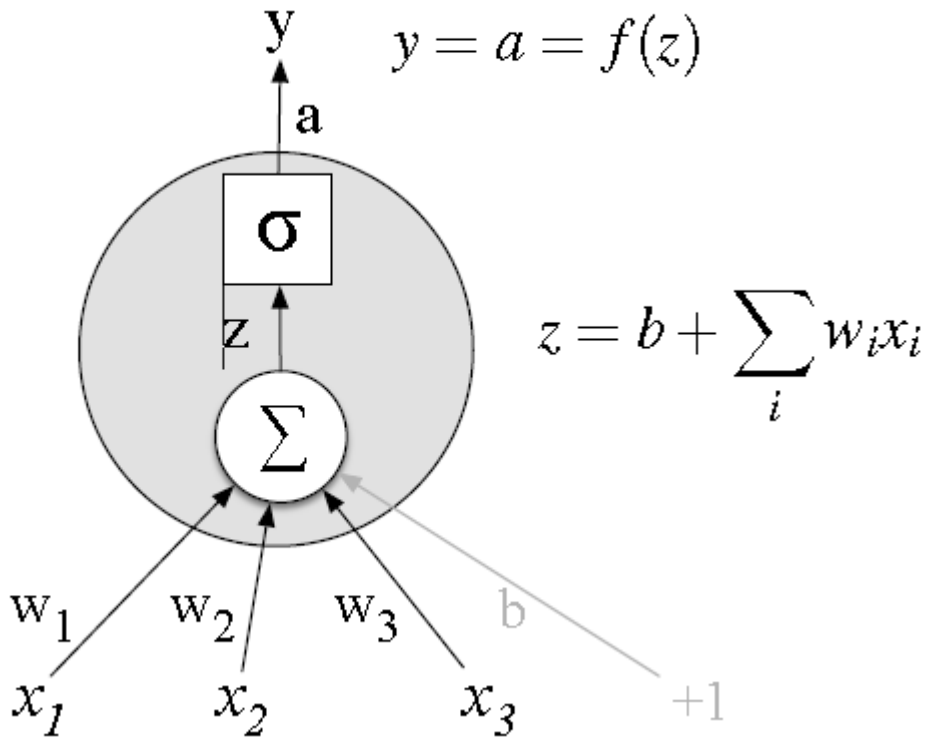
qué duda cabe juzgado el juez los dos islamistas radicales han pedido más ayuda según el fiscal le acusó de haber abonado los tres últimos meses su cargo

PLN: Introducción a los modelos de lenguaje

Modelos Neuronales







$$y = a = f(z)$$

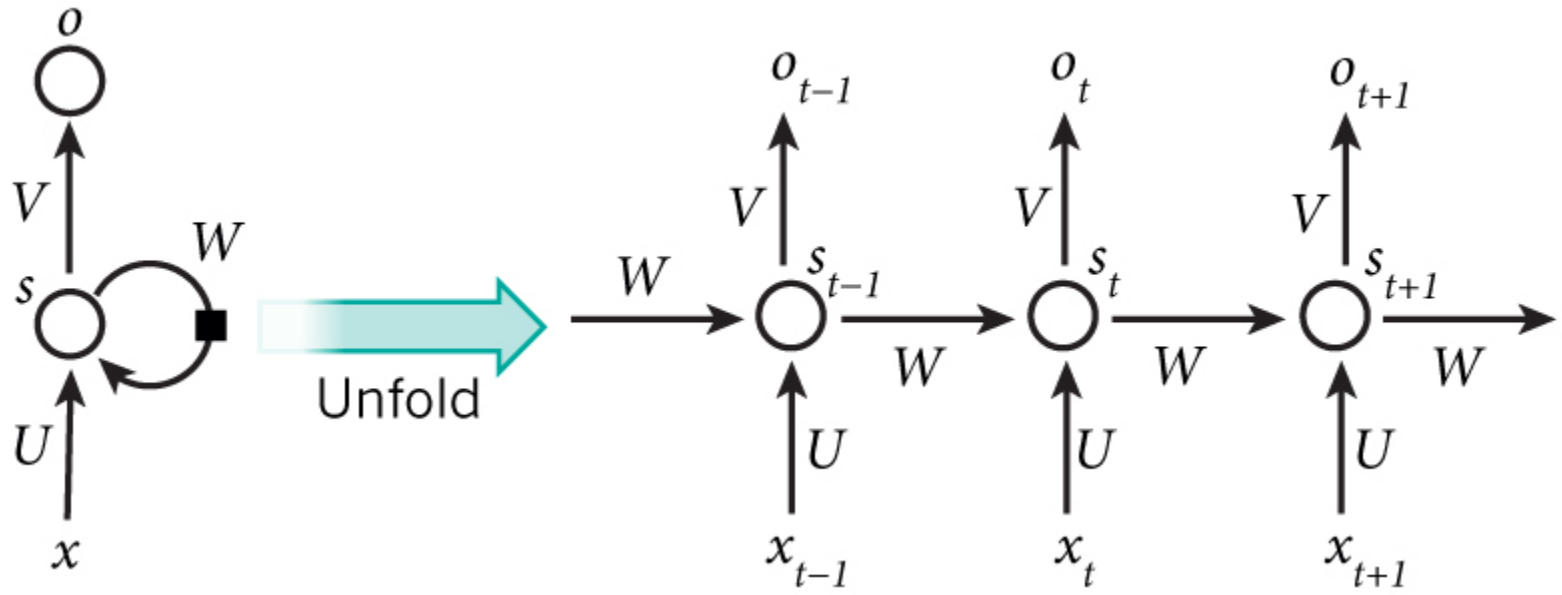
$$z = b + \sum_i w_i x_i$$

PLN: Introducción a los modelos de lenguaje

Modelos de lenguaje neuronales:

Predicción de la próxima palabra dada la secuencia de palabras anteriores: red neuronal recurrentes

$$P(w_t|w_1^{t-1}) \approx P(w_t|w_{t-N+1}^{t-1})$$



PLN: Introducción a los modelos de lenguaje

Generación de texto a partir de un vocabulario de caracteres (2016-17)

Las temperaturas bajarán en el norte y en el sur, lloviznas en el este y en Galicia

y en Asturias, con algunas nubes más por la tarde, mientras que en las Islas Baleares se mantienen los cielos básicamente despejados. Aquí tenemos las nubes que se extienden por muy poco espacios de las próximas horas. De hecho, esta es la situación de cara a la jornada del domingo, continuaremos hablando de una situación marcada por la estabilidad. Por tanto también con intervalos de nubes altas y medias en la costa mediterránea. En general tiempo estable en el centro y en el oeste del país. Esto se va a notar más porque el viento es de componente norte que va a traer un tiempo muy parecido al de hoy. En el resto del país con cielos despejados, pero también con nieblas en el norte de las islas de mayor relieve.

PLN: Introducción a los modelos de lenguaje

ChatGPT (2023)

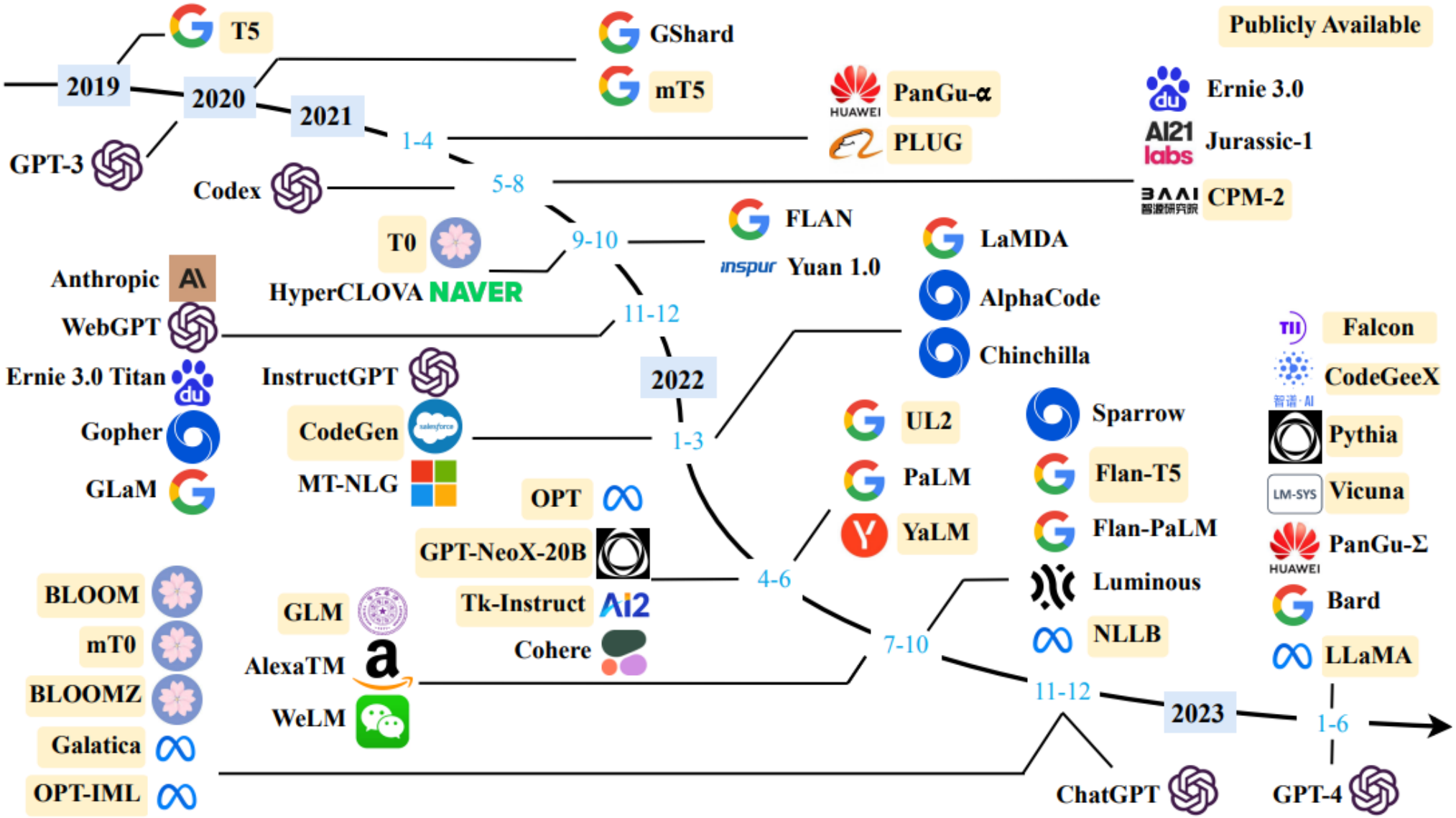
Las temperaturas bajarán en el norte y en el sur, anticipando un cambio en el clima que se avecina en las próximas semanas. Se espera que las regiones septentrionales experimenten un descenso gradual en las temperaturas, con mínimas que podrían llegar a alcanzar valores cercanos a los 10 grados Celsius. En el sur, las temperaturas también mostrarán una tendencia a la baja, con valores máximos que podrían oscilar alrededor de los 20 grados Celsius.

Por otro lado, se pronostican lloviznas dispersas en el este y en Galicia. Estas precipitaciones serán de intensidad moderada y se extenderán durante varios días, proporcionando alivio a las zonas que han experimentado periodos de sequía. Se recomienda a los residentes tomar precauciones y estar preparados para condiciones climáticas húmedas, así como a los conductores adaptar su conducción a las carreteras mojadas.

En el resto del país, se espera un clima mayormente estable, con cielos despejados y temperaturas suaves. Las zonas centrales y occidentales disfrutarán de días soleados y cálidos, con máximas que podrían alcanzar los 25 grados Celsius.

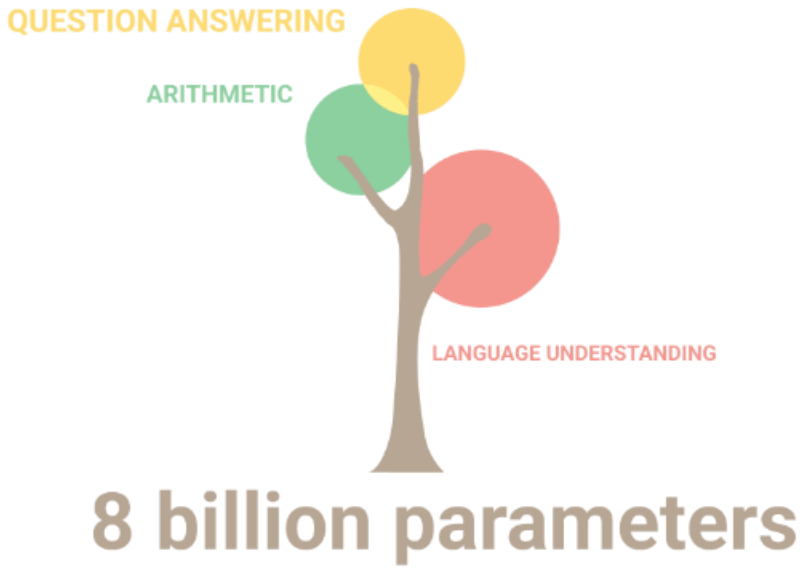
En resumen, se anticipa un cambio en las temperaturas con descensos en el norte y sur, lloviznas en el este y en Galicia, y condiciones climáticas agradables en el resto del país. Se recomienda estar atentos a los pronósticos actualizados y tomar las precauciones necesarias para adaptarse a las condiciones climáticas variables.

PLN: Introducción a los modelos de lenguaje



PLN: Introducción a los modelos de lenguaje

Habilidades emergentes



LLMs y ChatGPT

Entendiendo chatGPT

Antonio Miguel