Optativa 1 PLN

Clase 1 Agostos

Procesamiento del lenguaje natural

La materia esta enfocada a la inteligencia artificial, se presentara un componente muy fuerte de investigación.
La idea de la asignatura es solucionar problemas y tratar que sea de calidad. Y como resultados de aprendiza se centra en habilidades teórico practicas. Conceptulizar conceptos ya que todo es habilidades de calidad.

Contenido programático

- Introducción a la inteligencia artificial
- PLN en Java
- Introducción a Python
- Bases de PLN
- NLP - Pipeline
- Text Representation
- Text Clasification
- Information Extraction
- Spech Recognition and synthesis
- Chatbots
- Search and Information Retrieval
- Topic Modeling
- Recommender Systems for Textual Data

El procesamiento del lenguaje tiene que saber gramática, hoy día se usan transformadores, la idea es poder combinar el sistema de reglas con machine learning con entrenamiento para hacer seguimientos. Usando reglas y patrones. Árboles, semántica y se trabaja bastante.

Introduccion a la IA

Podemos decir que el tema arranca con un paper, donde se comienza a hablar sobre el sistema nervioso, el funcionamiento nervioso, y alan turing. Y la programación computacional. Esta apareció en 1943. Alan turign hace unas contribuciones con un paper "Computing

machinery and intelligence" pensando en aprendizaje automático,algoritmos genérico, aprendizaje por refuerzo, test de turin.

El test de turing tiene 2 humanos, donde 1 manda preguntas y la responde 1 humano y una máquina, pero el test dice que cuando el entrevistador no sabe quién contesto la pregunta, se puede decir que la inteligenia artifiial llego a un grado importante engañando al humano entrevistador. 1950

La ia como termino nace en 1956 por john macartey llamando este enfoque como Inteligencia artificial "siendo la ciencia de crear máquinas inteligentes, programas de computo inteligentes". Es una prueba para darse cuenta que tan avanzados están los algoritmos de ia.

1967 perceptro mark 1, es básicamente redes neuronales, con entradas realizar un proceso y dar una salida. Centrado a una nuestra naturaleza humana, el perceptron hay unas entradas, unas neuronas, y con deep learning entre mas neuronas hay desventajas y ventajas, mayor precision mas máquina. Procesar media pagina vs 10 paginas. En deeplearning son muhcas neuronas conectadas. Hoy se hablan de capas de neuronas, se comunican y dan una respuesta 0 o 1.

1980 Habitación china, es la antítesis de turing, se refuta porque dice que una máquina que pasa el test de turing no quiere decir que tenga inteligencia, sólo la simula. No entendió el contenido semántico, sintáctico.

Eventos importantes
ibm 97, vence al campeon de ajedrez
ibm vence jeopardy 2011
minwa 2015 super imagenes
2017 nace la base de los generadores de texto, todos de sobre el principio de generacion de texto, que paso a tras y que paso a futuro

Question Answer competencias de ia a desarrollar de gtp y generadores de texto,construir sistemas que ayudan a alimentar.

El corpus es el base de conocimiento, y siguen aprendiendo

La programacion funcional es muy rapida y parentesis. tiempos muy eficientes

Inteligencia artificial, como la rama de la ciencia de la computación que se ocupa de la automatizacion dejando una entrada, proceos y salida automatizando la conducta inteligente Luger 2005.

La inteligenia artificla tiene como objetivo las capacidades inteligentes, con influencia de la filosofia, matematica, psicologia. generias enfoncada en resolucion de problemas. Areas especificas quien mueve el automovil autonomamente, para medir la ia.

no es solo python,tiene muchas diciplinas involucradas

Enfoques, pero siguen siendo sistemas ia:
robots que piensan como humanos,
actuan como humanos
logica formal, sistemas expertos sistema que pueda tener un sensor enviar correo electronico, detectar enfermedades
Agentes, asistentes virtuales actuan racionalmente

---

importante tener en cuenta la palabra automatizacion, es un concepto muy dificil y complicado es mejor hablar de sistematisacion, para evitar discuciones

por medio de la estadistica, las probabilidades hablamos a nivel cientifico. logicas de prediicados proporcicional.

desde la neurociencia, se hace un aporte a la inteligencia artifcia porque hablamos de neuronas, conexiones y como aprenden.
Desde la psicologia tambien podemos hablar de inteligencia artificial desde la pata de como piensan las personas, como actúan, y hablar desde la conducto.

se deben hablar de reglas, donde la ultima era sobre la autodestrucción del robot.

desde la computacion podemos hablar de hardware y herramientas para el desarrollo. Ya que esta rama es la que mas aporta a la ia.

Desde la linguistica, al conocer muy bien el idioma aportan en gran manera. La linguistica computacional requiere el conocimiento de las normas que rigen nuestra gramatica. Chomosky fue el primer que dejo las bases linguiticas teoricas y computacionales. y deja una linea de trabajo de como representar el conocimiento, como mostrar el lenguaje, esto es representando el conocimiento, sintetizar y representar. Por otro lado la linguistica computacional tiene algo que son las ambiguedades que son muy dificil de reconocer y trabajar, y se deben de resolver en el lenguaje español. Estas son las dificultades del lenguaje.
Tambien contribuciones desde la economia, es la toma de decisiones puesto que la máquina debe pensar en beneficios, competidores, y el beneficio no es inmediato.

areas principales de la inteligencia artifiial

- resolver problemas
- representacion del conocimiento -> como se ve por pantalla la salida
- busquedas
  areas especificas

- planificacion de tareas -> scheduling para guardar productos en contenedores
- procesamiento del lenguaje natural
- percepcion
- razonamiento autonomo

Algoritmos de redes en gestion de trafico para enrutamiento esta la inteligencia artificial.
Es importante saber usar y conocer las tecnicas.

Introduccion a python
con los resultados de aprendizaje deben los trabajos en moodle.

se usara flask durante la asignatura. Python es un lenguaje muy poderoso pero tiene interpretado y multiparadigma. tiene unas caracterisitcas hay unos tipos de datos, estructurdaos no estructurados, mapeos, diccionarios, mapeos, envolturas.
el uso de indices *'[:-2]*.
los diccionarios clave valor, las no sql se sientan sus fundamentos en clave valor, una que no se reputa y crear un valor, es muy comun el uso de diccionarios en el procesamiento del lenguaje listas dentro de los valores del diccionario. crear el campo que tiene un campo con otro campo. elif.

´?´ valor lambda no se cual es pero algo lleva.
operador [:>8] el uso de set y get en clases python, sobre cargar metdoso es usar el mismo nombre pero con diferentes parametros

static

- imagenes
  template ->htmls

---

bases pln

arbol sintactico
hay que saber mucho de lenguaje, como las estructuras,
s oracion principal
sn sintagma nonimal grupo de palabras que cumple la funcion de sujeto
sv sintagma verbal grupo de palabras que contiene el verbo y complemento
sadj sintagma adjetival gurpo de palabras cuyo nucleo es un adjetivo

sadv sintagma adverbial grupo de palabras cuyo nucleo es un adverbio
sprep sintagma preposicional grupo de palabras que empiecen con una preposicion

aquella palabra que una frase gira entorno a ella, siendo los nonimales los mas comunes

suj sujeto realiza la accion
pred predicado accion realizada por el sujeto
compl complement puede ser directo, indirecto, circunstancial

clases de palabras:
N sustantivo
v verbo
adj adjetivo
adv adverbo
det determinante
preo preposicion
conj conjuncion

Donde se trabajaran las librerias NLTK, spacy. se divide la frase y de acuerdo a la posicion, (cada palabra tiene una funcion)

NLP Pipeline

Se va a comenzar mucho código más especializado, sistemas de tuberias, fase, una fase tras otra. (manera de arquitectura)
se uede hacer limpieza y extraccion de texto, esto es enfocado a procesar texto, a procesar lenguaje. Se toman los textos y se procesan.

Pipeline es una serie de pasos para construir cualquier modelo de NLP, y la mayoria son propios del lenguaje y permite procesar los lenhuajes, antes tocaba hacerlo en ingles o en español.

los pasos más comunes del lenguaje son adquirir los datos, limpiar, pre procesar, caracteristicas, modelar, evaluar, deployment, monitorear
Al momento de hablar sobre information extracction que es otra vertiente, ahi hay una pipeline para esa tarea comun, hay unas tareas comunes, entonces para una tarea comun hay un pipeline con los pasos necesarios para construir el sistema.

ej extraccino de informacion, y se usa este pipeline. Al hablar de pipeline es el como se diseña el sistema y luego como pasamos a modelar.

Las reglas sostubieron este mundo por muchos años para mejorar los sistemas.

Hay varias formas de extraer datos, cuando no hay manera de conseguir los datos, toca usar metodos para crear los datasets, empezar a crear datos a partir de metodos, inclusive crear los datos a mano,

extraccion de texto y limpieza
parsear html, extraccion de textos y hacer limpieza desde el html, normalizacion a simbolos e interpretarlos, tambien hay correcion de ortografia, hay otro de correcciones de errores muy especificos

preprocesamiento, es una de las cosas que mas se hacen cuando se hablan de tareas tengo un texto en una imagen y traer el texto, de un foro sacar el texto y mostrar lo mejor, tengo un texto y le hago correcciones de ortografia ahi son las tareas que procesan el texto. Podemos aplicar todo o nada, y una de las tareas mas dificiles es que requieren mucha maquina, hay unas tareas preliminares es tokenizar el texto, el token1, token2, donde normalmente son palabras divididas, no es muy comun segmentacion de oraciones.
La que si se hace si o si es tokenizar, algunos pasos frecuentes es quitar mayusculas o minusculas, quita digitos signos de puntacio, espacios en blanco y la raiz de la palabra.

otros pasos es detectar el idioma de un texto, normalizacion es en un pipeline tener en cuenta la normalizacion es qutar ruido, cosas que no aporten al texto y se quitan, otro no muy comun es code mixing mezclar lenguajes SPANGLISH. Transliteracion es como cada pais, idioma, lenguaje ellos no dicen bmw en chino, dicen bao ma y es una tarea muy especifica.

uno de las cosas que siempre se hacen es el pos tagging es el etiquetar las palabras arbol sintactico, parsing es si quiere analizar el arbol sintactico completo y coreferencia resolution alejandro baja las escaleras, el va para la casa, = ese el se refiere a alejandro

los stop words son las palabras como signos de puntuacion que no suman nada al texto

obligatorio al procesar texto, pos tagging es para crear una texto desde una generalidad suj+verb+sustantivo

Ingenieria de caracteristicas es poder detectar lo relevante del texto y poder crear a partir de lo importante, lo normal es combinar nlp con ml/pipelines, y la otra forma es con deep learning entrenar para hacer ingenieria de caracteristicas, el entrena y hace las caracteristicas. mirar la frecuencia de las palabras dentro de un texto y darle una importancia

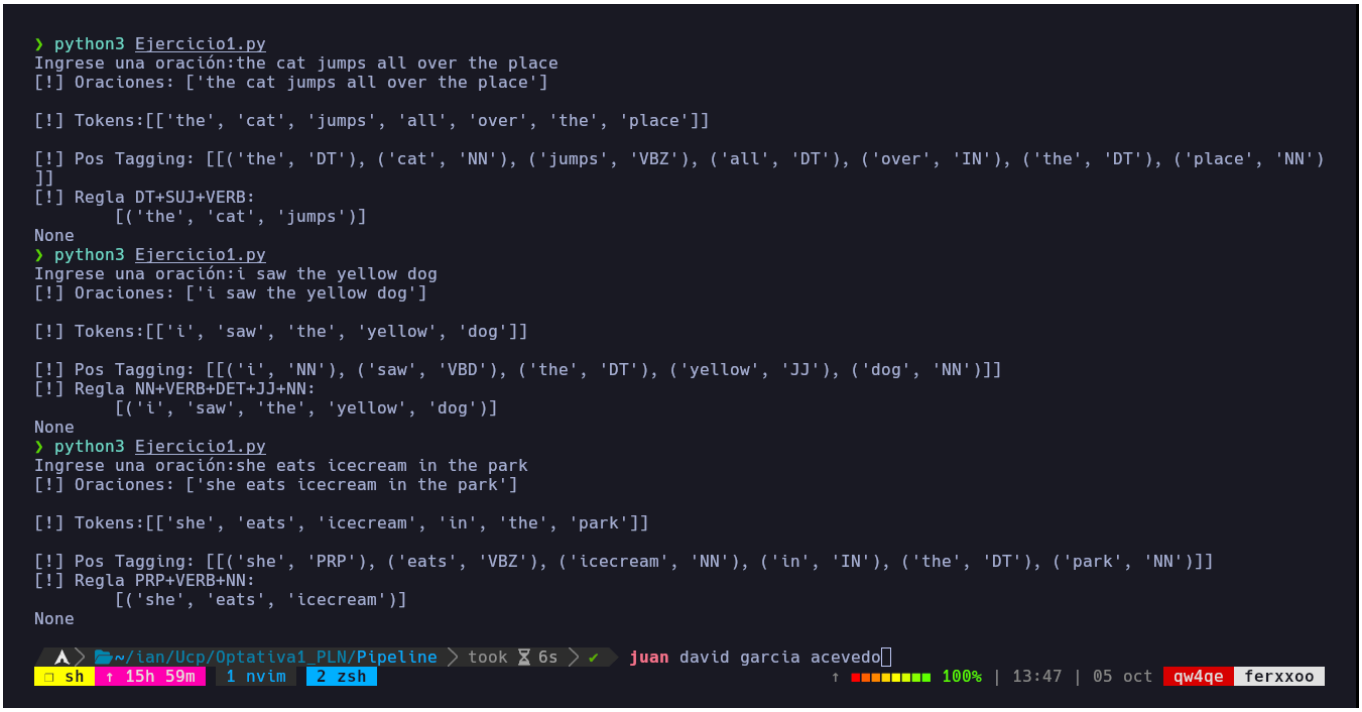
un pipeline seria, tokenizar, pos taggin, stopwords, feature extraction, model ml, y poner a una tarea
trabaja las caracteristicas

Lo mas grueso es el pipeline para la creacion de pasos, siempre es tokenizar, no siempre es pos tagging, es lo mas extenso en este caso, la ingenieria de caracteristicas es poder darle un

mayor sentido al texto y poder identificar cosas del texto que yo necesite

al HAblar de modelo, se hablan de heurisitcas, ejemplo clasificar spam, siendo heuristica la forma de encontrar el algoritmo para clasificar

la evaluacion para saber si el modelo funciona, a que nivel esta funcionando en el ejemplo de spam, se habla de intrisecas=precision y recall hablando de la presicion de modelo, la covertura, extrinseca tiempo usuario.



Text representation NLP

Tenemos que representar el texto de una manera, se hacía un preprocesamiento, separando párrafos, separando palabras pero no es tan sencillo. Son las formas que tenemos para representar los textos en vectores.

Hay varias formas y veremos las más conocidas, los modelos de deep learning es con redes neuronales para realizar ingenieria de caracteristicas o para tratar con el lenguaje.

hace referencia a conversación de texto escrito a una representación numérica. Como transformamos un texto determinado en forma numerica para que pueda incorporarse a los algoritmos de PLN ML

Pipelines para la representacion de texto, hay unas tareas : texto crudo, limpieza y preprocesar, tokenizar, representacion matematicamente en vectores, NO SE VE PERO SE HACE para evaluar el modelo.

para extraer el significado de las oracione,s hay que encontrarle el sentido a las oraciones, es muy importante al tratar el texto con el contexto, saber donde esta ubicado una palabra, para saber de que estamos hablando, no se entiende de que se esta hablando,
1 dividir en unidades léxicas las palabras lexemas
2 deducir el significado
3 comprender la estructura semantica
4 comprender el contexto en el que aparece la frase

hay unos enfoques para la clasificacion de enfoques para text representation, es valido la creación de reglas para refinar el modelo de machine learning

hay modelos de espacio vectorial
similitud del coseno
distancia euclidiana

cuando se procesa el lenguaje, es muy comun la similitud de una frase con otra, es de acuerdo al angulo para determinar que tan cercana es una frase de otra,

la distancia euclidiana cuadrada permite de igual manera mediante matrices y determinantes para saber la similitud entre palabras

Hay enfoques de vectorizacion siendo la primera, siempre hay que hablar de un corpus, una serie de documentos, y hay un vocabulario. el corpus puede ser un pdf completo, una frase y se crea el vocabulario. El primer enfoque fue One hot encoding: existe un vocabulario se crea un amatriz y cada palabra es un vector, y la frase se representa con una lista de vectores

segundo enfoque es bag of words, se encuentra mucho, donde ya no es matriz pero es una lista y cada palabra es una bit encendido y apagado.

hay otra representacionmuy comun que es la bag of ngrams, es muy famosa porque usted puede crear N gramas para buscar en los corpus, el modelo logra entender cuando se habla de una frase

TfIdf term frequency inverse document frecuecy, es muy usado en el NLP, se aplica la formula por separado se unene, se multiplican

hay otros conceptos como similitud distributiva
connotación: siginificado según el contexto de la palabra
denotación: el significado literal de cualquier palabra

nlp rocks

denotacion: piedras
connotacion: algo bueno, agradable

hipotesis distributiva
en liguistica existe una hipotesis que las palabras que ocurren en contexto similares tiene significados similares
perro y gato

representacion distributiva, otro concepto respecto a la distribucion de las palabras en el contexto en el que aparecen, se usan vectores y matrices

embedding incrustar, es un conjuto de palabras en un corpus

word2vec es entrenar un sistema con redes neuronales, relaciona palabras, hace es tomar una palabra - otra, + otra = crea una nueva palabra con relaciones entre las palabras. busca las relaciones en las palabras, ya que el texto relaciona unas palabras unas con otras. al final es saber que dice el texto hallando relaciones

hay otro modelo muy comun continuos bag of the words, es empezar a a ver en la frase cual es la palabra central quepueda dar el contexto de lo que esta al rededor, distancia semantica

representacion universal, para la generalidad de las palabras, tratar de como se puede visualizar

retorno del esfuerzo, hay que meterle esfuerzo un chatbot o que sea, eso se da o no se da? lo que va a retribuir, saber si es una necesidad comercial

limitacion de infrasctructura

chatbots

esto es un proceso donde basicamente se hace un preprocesamiento y se montaba un sistema de reglas, ahora con ml se habla de la extraccion de entidades, pero aun no es suficiente entrenear una red neuronal porque el lenguaje no es sencillo.
hay unos terminos que hay que ir conociendo,
el tema de los chatbots dio una explosion con una evolucion y son muchos los avances lo que hay,
poder hablar con el sistema y que nos devuelva informacion.
se dan en ambiente legal como preguntas frecuentes hablar de faqs, comercios de compra en un comercio electronico, descubriendo noticias
y servicio al cliente
se trabaja distancia semantica para poder asociar preguntas con respuestas
categorizando preguntas para poder responder con un texto, y se aplican algunas tecnias para la asociacion
se aplica el coseno con un corpus y que el chat empiece a crear sus respuestas
taxonomia de los chatbots que habla cuando un bot una pregunta exactamente igual la responde, debe estar en el sistema para poderla responder, si no esta el chat se estalla,
hay bots basados en flujo, como un diagrama de flujo
bot abierto es pregunteme lo que quiera y le respondo, es mas dificil pero esta siendo trabajado fuertemente hoy dia
hay entrenamientos de redes y ml, pero tambien otras tecnicas como usar reglas

son tecnologias muy especialisadas y quizas lo mas cercano y mas avanzado es BERT como metamodelo, no es tan sencillo porque hay modelos muy sencillos

para construit sistemas de dialogo se tiene que pasar por unas tecnologias, reconociemietno de voz llevarla al sistema entender lo que esta diciendo, deteccion de sentimientos o entidades podemos decir que hay un cerebro que trata de construir la respuesta, despues se genera el lenguaje necesario lo dificil es entender la voz de la persona y generar una respuesta para que la persona este satisfecha

intenciones de dialogo dialog act or intent, se debe tener claro que es un intents que es lo que se carga al sistema para hacer la red neuronal, cuando se habla de intents es algo muy especifico muchos sistemas unen los 2 conceptos dialog art es como esta afirmando pero no esta encontrando lo especifico pero si da como la idea inicial

se crea una capa con dialog act para saber lo general su intencion, luego otra capa intents para saber que quiere hacer la persona

slot or entity tambien es confundido, y la diferencia es uno mas general que otro, slot general, entity especializada ubicaciones por ejemplo,

esto es una ubicacion madrid españa, fecha noviembre eso es una entidad que pertenece a una fecha
origen(slot) = newyork (entity)
destino(slot) = londres(entity)
fecha(slot) = marzo(entity)
dialog state or context es muy importante, para un sistema de pln es el contexto en el que se esta moviendo

los intents dan el contexto

al procesar texto lo mas importante es el contexto intents y entitdades

---

Information extraction

La detección de entidades mediante uso de reglas, y lenguajes de reglas permite saber el contexto para la extracción de información.
en 1987 el deep learning no se usaba para estas tareas.
-fechas
-personas
-lugares
-ubicacion
-moneda
-company
-date
Hay niveles de extraccion de informacion
identificando identidades
Algunas aplicaciones para la extraccion de informacion permite traer datos desde formularios y recibos, la etiquetacion de las noticias y extraccion de noticias o palabras claves de estas para hacer proyecciones. Redes sociales y redes chatbots, a pesar a encontrar una relacion entre todo.

otra tarea fuerte es keyphrase o keyboard, mirar por repetecion KPE trabaja con las frecuencias, distancia semantica de la que mas repite con otras palabras. Colocar buenas palabras claves en el trabajo de grado. Se aplica el algoritmo que encuentre esas palabras. La mas importante de la extraccion de informacion es named entity recognition, es de las mas importantes cuando se hace extraccion de informacion.
entity disambiguation and linking que no es una tarea sencilla para poder decir que es una ambiguedad, darle contexto y poder darle un significado para procesar. Relation extraction relaciones entre entidades, entidades y compañias etc.

la mas miedosa es relation extraction

tareas avanzadas contemplan extraccion de eventos,

para el pipeline se hace identify named entities
y multiples relaciones a una entidad

extraccion de keyword es la extraccion de las palabras importantesextraccion de palabras claves y semantica un KPE son como objetivos de las empresas y se hace la extraccion
se estable un grafo como estrategia para la kpe y se miran unos pesos, precios para crear las aristas y vertices y sus conexiones para su creacion.
el nlp es suseptible a lo que se va a procesar, no es lo mismo1hoja , 10 hojas a 100, la longitud es un problema y se requiere la maquina.

ner name entity recognition se tiene que se esta escribiendo una frase para y el automaticamente empieza a hacer la extraccion de donde vivia, donde nacio con el date, no hay sistemas para la extraccion de informacion de productos, o numeros civiles de leyes como el area de derecho. hay editores para la extraccion de informacion . El editor permite hacer ese uso, tambien permite el uso manual de etiquetacion, son arquitecturas para la extraccion de texto dedicados solo a el procesamiento de texto

el uso de listas gazetter es una extrategia para lo que sea muy dificil que suceda en un lenguaje, esta bien pero es un complemento para tener datos en cuenta. lo que no es muy comun pero se puede trabajar

el otro enfoque fueron mediante el uso de reglas como ner, pos tagging para saber entidades y con ellas se puede hacer la identificacion.

con el machine learning tambien se puede hacer la extraccion de informacion como evolucion del gazzeter con listas, reglas y ahora el tercero es ml luego deep learning y luego transformadores.
Se puede usar las reglas aparte del ml, aparte del deep learning,

las bios notation son para deep learning que da la informacion, se entrena el sistemas y el identifica las entidades

las desambiguedades requieren de conexion para poder hacer las conexiones entindad nombre entidad compañoa para hacer las relaciones.

una manera de hacerlo es encontrar todas las entidades, luego mirar sus relaciones, Es muy similar a hacer un resumen