Procesamiento de lenguaje natural y dos ejemplos en python

Natural language processing and two examples in python.

Autor 1: Yeisson Garcia Cristancho,Jhon Edison lotero

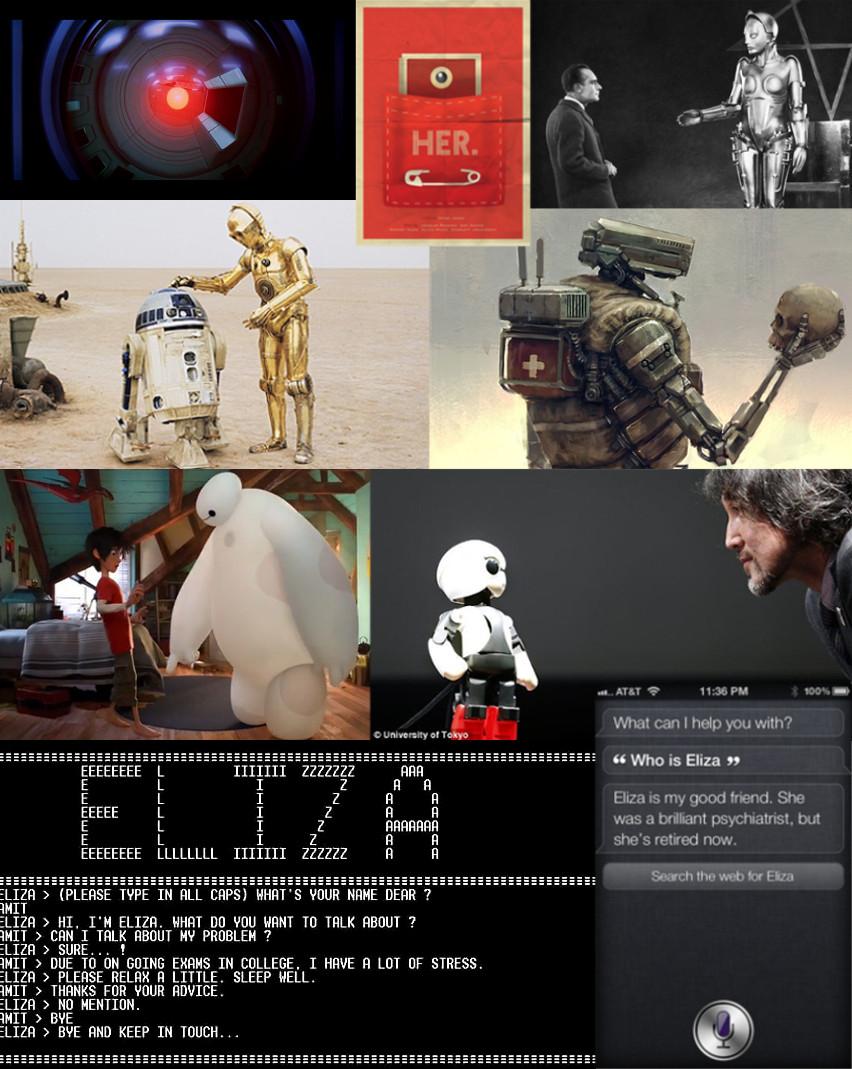
*Computación Blanda, Universidad Tecnológica de Pereira, Pereira, Colombia*

Correo-e: yhgarcia@utp.edu.co

**Introducción**

El procesamiento de lenguaje natural puede definirse, de una forma muy general, como el proceso automático o semiautomático de lengua humana. En este procesamiento se suele incluir lo mismo las acciones de generación automática de lenguaje (escrito o hablado) que las acciones de interpretación (tanto de habla como de lenguaje escrito).

La posibilidad de poder comunicarse mediante lenguaje hablado con las máquinas ha sido une pretensión que ha alcanzado el nivel de *cliché* en películas de ciencia ficción, pero que aún se encuentra en niveles muy elementales en la vida real.

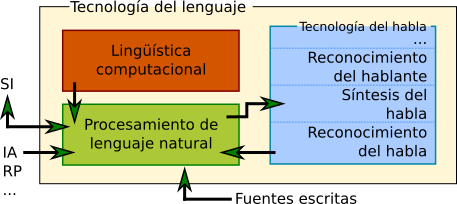


Siri Meets Eliza

Los primeros trabajos en procesamiento de lenguaje natural se realizaron a finales de los años 40's, enfocados principalmente a la traducción entre lenguajes, singularmente la traducción entre ruso e inglés [Nota: a fin de preservar algunos documentos actualmente disponibles de manera pública, como es este caso, la carpeta publica literatura contiene copias de tales documentos]. Sin embargo, el desarrollo de esta área de investigación ha sufrido cambios frecuentes de dirección debido a las dificultades encontradas en el camino, que en muchos casos obliga a poner atención a la solución de sub-problemas y al surgimiento de nuevas áreas de interés. De esta manera, incluso dar una definición precisa del área es una tarea difícil.

Existen tres términos principales que se han empleado como sinónimo, algunas veces, como subáreas en otras ocasiones o como áreas independientes con intereses que se traslapan: **Procesamiento de lenguaje natural**, **Lingüística computacional** y **Tecnología del lenguaje**. Otros términos describen áreas/subáreas más especializadas, pero que también suelen presentarse como áreas independientes, como es el caso del **Reconocimiento del habla** o la **Síntesis de habla**.

Podemos definir, para nuestros intereses, **Procesamiento de lenguaje natural** (*NLP* por las siglas en inglés) como el área de investigación y desarrollo enfocada en el análisis de información representada mediante lenguajes humanos (naturales). El contexto disciplinar en se enmarca el procesamiento de lenguaje natural puede representarse, *grosso modo*, mediante el siguiente diagrama:



De esta manera, el procesamiento de lenguaje natural forma parte del flujo de trabajo de un sistema de "lengua artificial" en el que previamente se han atacado otros problemas bien definidos cuyo objetivo es procesar una señal de audio y generar una transcripción en forma de texto. Este texto es la entrada del "módulo" de procesamiento de lenguaje natural. La salida de los métodos de NLP pueden, así mismo, utilizarse como entrada de un módulo de síntesis de habla. Sin embargo, los métodos de NLP suelen utilizar otras fuentes de entrada, siempre y cuando se encuentren en forma de texto digital. Actualmente se destacan los mensajes publicados en redes sociales, así como las publicaciones en diversas publicaciones (*blogs*, foros, diarios electrónicos, etc.). Así mismo, la salida suele ir a diversos destinos, principalmente a sistemas de información y de apoyo a la toma de decisiones.

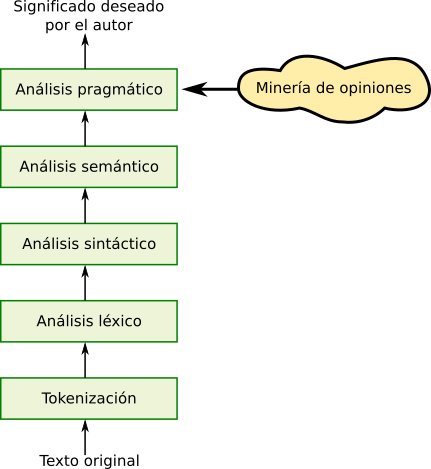
Por otra parte, el procesamiento de lenguaje natural utiliza modelos teóricos proporcionados por la lingüística computacional, así como técnicas de procesamiento provenientes de áreas como la Inteligencia Artificial, el Reconocimiento de Patrones, el Aprendizaje Automático, la Estadística, la Minería de Datos, etc.

**Etapas del procesamiento de lenguaje natural**

Los lenguajes se suelen analizar como compuestos de cuatro elementos principales:

* La *fonología*. Este componente describe el modo en que se emplean los sonidos en la lengua y cómo las variaciones de éstos se reflejan en el significado de una palabra. El análisis fonológico suele quedar fuera del ámbito del procesamiento de lenguaje natural, para recaer en la fase de reconocimiento del habla.
* La *semántica* suele colocarse como una segunda etapa en el análisis de las lenguas (a diferencia del procesamiento de lenguaje natural, donde esta etapa se encuentra en etapas superiores de análisis). Esta etapa se refiere al análisis del significado de los signos lingüístico.
* La *gramática*. Este componente involucra dos partes: *sintaxis* y *morfología*. La sintaxis estudia las reglas y principios bajo los cuales se combinan los constituyentes sintácticos, mientras que la morfología explica la estructura interna de las palabras y el proceso de formación de palabras. Al conjunto de palabras que conforman una determinada lengua natural se le denomina léxico, por lo que, en el contexto computacional, el análisis morfológico equivale al análisis léxico.
* La *pragmática*. Este componente estudia el modo en que se concreta el significado de un mensaje, en términos del contexto extralingüístico (estrato sociocultural, conocimiento compartido por los hablantes, relaciones interpersonales, etc.).

El proceso de análisis de lenguaje natural suele descomponerse en una serie de etapas, siguiendo el esquema teórico de la lingüística descrito anteriormente, según como se muestra en el siguiente esquema. Así, se asume que el proceso de análisis de un mensaje inicia con la recepción de un texto *original*; este texto es "*tokenizado*" ("tokenizar" es un barbarismo utilizado para designar la segmentación de un texto en sus componentes individuales: palabras y diversos símbolos; no está aceptada por la RAE y tampoco ninguna de sus variantes -token, tokenizado, etc.- sin embargo, a falta de un término adecuado en español, es utilizado ampliamente en el ámbito computacional). A continuación, se realizan los pasos de análisis léxico (identificación de palabras y símbolos), sintáctico (análisis de frases), semántico (significado de las frases) y pragmático (significado del texto). Esta última etapa se relaciona estrechamente con el área de minería de opiniones.

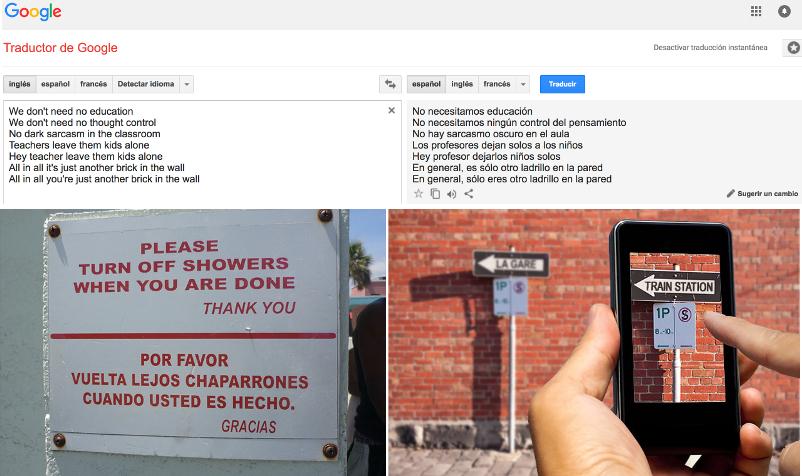


Esta segmentación del procesamiento de lenguaje natural suele no ser estricto en la práctica, puesto que la separación entre las etapas no es clara, así, por ejemplo, la tokenización puede requerir de un análisis pragmático previo. Sin embargo, esta categorización si permite prestar atención a clases de tareas.

**Áreas de aplicación**

El interés por el procesamiento de lenguaje natural ha experimentado un creciente interés en los últimos años, gracias especialmente al desarrollo de la Web 2.0. Entre las principales áreas de aplicación se encuentran las siguientes:

* **Traducción automática**. Esta fue una de las primeras aplicaciones sugeridas para las computadoras digitales, después del éxito obtenido durante la segunda guerra mundial traduciendo mensajes escritos en alemán y cifrados mediante el código Enigma. Sin embargo, rápidamente se reconoció que traducir un mensaje "abierto" es considerablemente más difícil que traducir un mensaje sobre un tema específico (la guerra, por ejemplo), escrito por personas con un contexto uniforme (analistas de guerra/soldados, por ejemplo) y delimitado por un sistema de cifrado artificial (Enigma). Actualmente, aunque aún se obtienen resultados pobres con los principales traductores automáticos, el área ya está bastante madura. Un reto importante en la actualidad es la traducción de "textos mal escritos", como los que se publican en las redes sociales que incluyen el uso de emoticones, abreviaturas, regionalismos y códigos especiales, así como faltas de ortografía.



* **Minería de opiniones / Análisis de sentimientos**. La creciente participación de las personas en los medios electrónicos, principalmente las redes sociales, que constituyen la llamada Web 2.0 (y la Web 3.0, según algunos autores), ha dado lugar a la publicación, registro y acumulación de grandes volúmenes de expresiones personales sobre toda clase de asuntos; opiniones sobre productos, inconformidades con las autoridades, deseos y gustos, comparaciones, historial de usos y fallas de servicios, etc. Este cuerpo de datos presenta dos características importantes: 1) contiene una cantidad de datos nunca antes reunidos y más confiable que la información restringida que se obtienen mediante métodos comunes de recopilación, como son las encuestas, y 2) la cantidad de datos es tan grande que resulta difícil procesarla. La minería de opiniones ha ganado gran atención como una fuente información para evaluar la aceptación de toda clase de productos y servicios y para detectar necesidades por parte del usuario final.



* **Extracción de información**. La minería de opiniones permite a las instituciones tomar decisiones a partir de expresiones "libres" de la gente común, vertidas en foros, redes sociales y publicaciones digitales genéricas. Sin embargo, no es la única manera en que las organizaciones están utilizando información en texto. En muchos casos, las decisiones de negocio se basan en sucesos reportados, por ejemplo, en los diarios. Una estrategia en la que está tomando importancia el procesamiento de lenguaje natural consiste en recopilar los documentos con anuncios potencialmente importantes para la organización y analizarlos para extraer la información relevante y alimentar con ella el sistema de toma de decisiones.



* **Clasificación de textos**. El objetivo de la clasificación automática de textos es analizar el contenido de un documento y asignarlo a una categoría preestablecida. Una de las aplicaciones más comunes actualmente, de la clasificación de textos, es el filtrado de mensajes de correo electrónico, ya sea para distinguir mensajes *spam* (contra mensajes *ham*) o, incluso, para distinguir entre mensajes principales, mensajes de redes sociales, mensajes promocionales, mensajes transaccionales, *spam*, etc. La clasificación de textos es también componente importante de los sistemas de minería de opiniones.



**Usando la libreria de python NLTK**

En este apartado describiremos como instalar la libreria Nltk para procesamiento de lenguaje natural y como usarla de forma básica.

Incluye instrucciones básicas de instalación, configuración y descarga de Corpora, Diccionarios, etc.

Para importar la librería solo usaremos:

>> import nltk

Sin embargo, hasta ahora hemos usado librerías que ya vienen incluidas en el paquete básico de Python 2.7.

La librería NLTK no viene incluída por default así que debemos instalarla, para esto usamos la herramienta "pip".

Lo primero que debemos hacer es comprobar que tenemos instalado "pip" que nos ayudará a instalar fácilmente cualquier librería python. Para estohe incluído en el repo el script "get-pip.py" que se puede encontrar en:

https://bootstrap.pypa.io/get-pip.py

INSTRUCCIONES DE INSTALACIÓN

1. Abrimos nuestra consola en la carpeta del proyecto y escribimos:

$ python get-pip.py

Este procedimiento instalará, actualizará o nos dirá si ya teníamos "pip"

2. Una vez que "pip" está correctamente instalado podemos comprobarlo con:

$ pip --version

3. Ahora podemos instalar cualquier librería python con "pip".

Ĺa librería NLTK conviene instalarla al lado de otra llamada: NumPy

Instalamos las librerías NumPy y NLTK de la siguiente manera:

En windows:

$ pip pip install -U numpy

$ pip install -U nltk

En Linux:

El comando anterior comando nos muestra un error de permisos

que se resuelven añadiendo el comando "sudo" al principio. De esta

manera indicamos que somos administradores.

$ sudo pip install -U numpy

$ sudo pip install -U nltk

El sistema Linux nos preguntará el password de administrador que es el mismo

que el password de usuario que se usa al iniciar la compu (casi siempre).

"""

# Si ya está instalada la librería NLTK, podemos importarla sin errores

import nltk

"""

Si llegaste hasta aquí ¡felicidades! ya que lograste instalar NLTK, pero la lista

de instalaciones aún no ha terminado desgraciadamente.

Lo que hemos instalado de NLTK es apenas un pqueño pedazo de funcionalidades

pero la librería NLTK viene con muchos corpus, gramáticas de juguete,

modelos entrenados, etc. con los cuales podemos jugar. Luego veremos cómo

incorporar otros, pero es fácil comenzar a usar NLTK con los que ya trae

disponibles. Algunos son muy grandes y de varios idiomas así que debemos

descargar sólo los que queremos usar.

Esto se logra abriendo una terminal y escribiendo:

$ python

Así, solito. Se mostrará la interfaz de comandos de Python. Ahi debemos importar

la librería y luego ejecutar un download() con ella:

>>> import nltk

>>> nltk.download()

Esto mostrará una interfaz de descarga de materiales gratuitos, en concreto se

muestra un menú como este:

---------------------------------------------------------------------------

d) Download l) List u) Update c) Config h) Help q) Quit

---------------------------------------------------------------------------

Si apretamos la tecla "l" (list) se mostrará la lista completa de paquetes

que se pueden descargar libremente. La lista completa la incluyó en el

archivo: "nltk\_packages.txt" en este mismo repo, aquí listo algunos que usaremos

y que vamos a instalar.

Son pocos los que sirven para español específicamente:

[ ] cess\_esp............ CESS-ESP Treebank

[ ] spanish\_grammars.... Grammars for Spanish

Otros que son multidioma e incluyen elementos en español:

[ ] crubadan............ Crubadan Corpus

[ ] conll2002........... CONLL 2002 Named Entity Recognition Corpus

[ ] omw................. Open Multilingual Wordnet

[ ] stopwords........... Stopwords Corpus

[ ] swadesh............. Swadesh Wordlists

[ ] universal\_tagset.... Mappings to the Universal Part-of-Speech Tagset

[ ] universal\_treebanks\_v20 Universal Treebanks Version 2.0

[ ] words............... Word Lists

[ ] tagsets............. Help on Tagsets

Otros generales:

[ ] punkt............... Punkt Tokenizer Models

[ ] averaged\_perceptron\_tagger Averaged Perceptron Tagger

[ ] maxent\_ne\_chunker... ACE Named Entity Chunker (Maximum entropy)

Para salir de la interfaz de NLTK apretamos la letra "q".

Para salir del intérprete de Pythos escribimos:

>>> quit()

Podemos instalar todos los paquetes de la lista ejecutando el script

"instalar\_nltk\_data.py" que se incluye en este repo:

$ python instalar\_nltk\_data.py

Una vez instalado podremos ejecutar este script con:

$ python basicNLTK.py

Otra aplicación con procesamiento de lenguaje natural

**INTRODUCCIÓN**

El presente trabajo final se propone avanzar hacia la construcción de un programa que permita comprender las subjetividades en inteligencia artificial. Para el mismo se retoman distintos aportes provenientes de los temas vistos en clase, los conocimientos en programación y cómo adaptar ambos para un mismo fin.

El trabajo se enmarca en una investigación que indaga acerca del uso de un recurso que actualmente es bastante utilizado en la mayoría de las aplicaciones, el cual consiste en usar el micrófono haciendo los procesos necesarios para que dicho programa asimile nuestra pregunta formulada y él haciendo uso de su base de datos me responda por voz la respuesta que deseamos conocer y que por su puesto esté dentro de dicha base de datos.

A continuación, realizaremos una apreciación del funcionamiento que tiene el programa, para ver aplicados los conocimientos de la inteligencia artificial vistas en clase e investigadas en el libro guía “Inteligencia artificial un mundo moderno – Stuart J. Russell”.

Finalmente veremos la codificación en lenguaje de alto nivel (Python) del programa ya terminado.

**JUSTIFICACIÓN DEL PROYECTO**

Nuestro estudio acerca de las aplicaciones del programa desarrollado mediante el lenguaje de programación (Python), pueden permitirnos distinguir con claridad las potencialidades que la inteligencia artificial presentan para la resolución de problemas concretos, en un área de actividad específica. Esto permitiría no solo aumentar la eficiencia en relación con los proyectos de desarrollo de largo plazo, sino planificar con mejores resultados las estrategias de codificación en los proyectos que ya se encuentran trabajando, y perfeccionar los planes educativos, comerciales e investigativos de la programación y la informática

**OBJETIVOS**

**OBJETIVO GENERAL**

Desarrollar un programa donde se pueda tener información de una base de datos haciendo preguntas por voz y dicho programa también conteste por voz.

**OBJETIVOS ESPECÍFICOS**

* Desarrollar el programa en Python o Java.
* Desarrollar un sistema eficiente con una interfaz agradable y de fácil manejo.
* Vincular a el programa una base de datos con información de determinado tema (biblioteca).
* Implementar un mecanismo de notificaciones para que los usuarios puedan conocer las preguntas que no se encuentran en nuestra base de datos.
* Vincular a el programa una opción recurrente si el programa llegase a tener un fallo al ejecutarse.

**DESCRIPCIÓN DEL FUNCIONAMIENTO**

Con la implementación de este software se pretende tener una interacción con el programa desarrollado, simulando entablar un pequeño dialogo de preguntas y respuestas.

Para que esto suceda, el programa toma como entrada el audio del usuario quien está formulando la pregunta, haciendo el tratamiento para convertirlo a una cadena de caracteres (String) para luego ser enviado al módulo Regex, ahora bien, teniendo en cuenta la Query y sus argumentos hace la búsqueda en la base de datos (Data Base) la cual hace uso de la ID del Query de Mongo para buscar la respuesta (si la hay) y almacenarla en la variable “ans”, siendo una variable de tipo objeto que va a contener la respuesta (Data) de la pregunta formulada por el usuario, una vez tenga los datos de la base de datos los debe procesar de nuevo para dar como salida la respuesta como audio.

Finalmente, el programa debe tener en cuenta que si la variable “ans” está vacía, lo cual significa que no encontró respuesta en la Base de datos, entonces debe regresar como respuesta al usuario por audio éste suceso (“No encontró respuesta a la pregunta formulada”).

**IDENTIFICACIÓN DE OPORTUNIDADES**

No solo es para uso educativo si no que nos podrá servir para futuros planes relacionados con dicho tema ya que como podemos notar ésta es un área que está en constante desarrollo y cada vez será más indispensable y necesaria para la vida cotidiana, por lo cual nos beneficia bastante en el ámbito laboral entre otros.

Referencias

1.-http://www.nltk.org/book\_1ed/

2.-http://www.nltk.org/book/

3.-http://nlp.stanford.edu/courses/NAACL2013/

4.-http://cs224d.stanford.edu/syllabus.html

5.-http://www.gandhi.com.mx/inteligencia-artificial-un-enfoque-moderno