WATSON ANALYTICS UNA SOLUCIÓN CLOUD PARA PROCESOS DE BIG DATA

GEISON LEONARDO GARZON CASTILLO

Resumen

La analítica ha madurado considerablemente en años recientes, a tal punto que las herramientas de Business Intelligence no las usan solamente los gerentes de negocio en una reunión. Sin embargo, la analítica avanzada realmente sofisticada, que puede ayudar a probar ideas, predecir resultados y a crear reportes basados en datos de las grandes redes, han permanecido en gran medida fuera del alcance de la mayoría de las organizaciones de hoy y por supuesto mucho más de los empleados que las necesitan. Asimismo, los ingenieros de datos han tomado un rol importante en las empresas, en donde las decisiones cada vez deben ser mas rápidas, verídicas y conformes a como el mundo se mueve, por eso se requiere de herramientas que no solo ayuden a resolver la complejidad de los datos, sino que también ayuden a que los empleados de las empresas puedan usar la información transformada por los ingenieros y generen reportes rápidamente muchas veces con información en tiempo real. Aquí es donde entra Watson, una herramienta diseñada para deshacerse de los complicados modelos estadísticos, interfaces de software hechas para ingenieros y hacer que las preguntas críticas del negocio se puedan resolver (Ziffdavis, 2015).

Introducción

IBM Watson es un sistema de inteligencia artificial informático, capaz de responder preguntas en lenguaje natural, basado en la información existente en documentos científicos, blogs, tweets e información de internet (IBM, 2014).

La plataforma como servicio ofrecida por IBM que soporta los servicios de Watson se llama Bluemix, entre los productos que ofrece se encuentra Watson Analytics: es un servicio intuitivo para analizar, procesar y presentar los datos solicitados, sin la necesidad de descargar ningún software. Esta solución usa un concepto de descubrimiento inteligente, de la cual a partir de los datos entregados como base y antes de su implementación, se puede agregar la visualización de los datos que existen en la nube como una guía a la exploración de datos, permitiendo automatizar los análisis predictivos y facilitar la creación de dashboards e infografías. Entre sus características se encuentra la de poder acceder directamente a muestras de datos de Twitter para transformar la información de las redes sociales en decisiones de negocio, crear informes de IBM Cognos Analytics fácilmente (IBM, 2017).

Desarrollo

La primera versión de Watson fue liberada en el año 2011 durante un episodio de televisión del show Jeopardy, un programa de concursos en Estados Unidos. Lo destacado de su aparición fue que logro vencer a los dos mejores concursantes del programa (Docasar, 2016).

Ilustración 1 IBM Research 2013

Watson, nombre también del fundador de IBM, se basa en el proyecto Deep QA, el cual plantea crear una solución que permita a la maquina interactuar con el ser humano interpretando las preguntas que el interlocutor le escriba aplicando programación neuro lingüística, con el propósito de razonar lo que se le pregunta y buscar una respuesta basado en la información guardada en el dispositivo (Manuales, PDFS, Noticias). Para poder realizar un cómputo ágil y distribuido en los métodos de búsqueda, su plataforma de información esta soportada por Hadoop la cual le ayuda a generar resultados en fracciones de 3 a 5 segundos (Ferrucci, D. 2010)(IBM, The DeepQA Project).

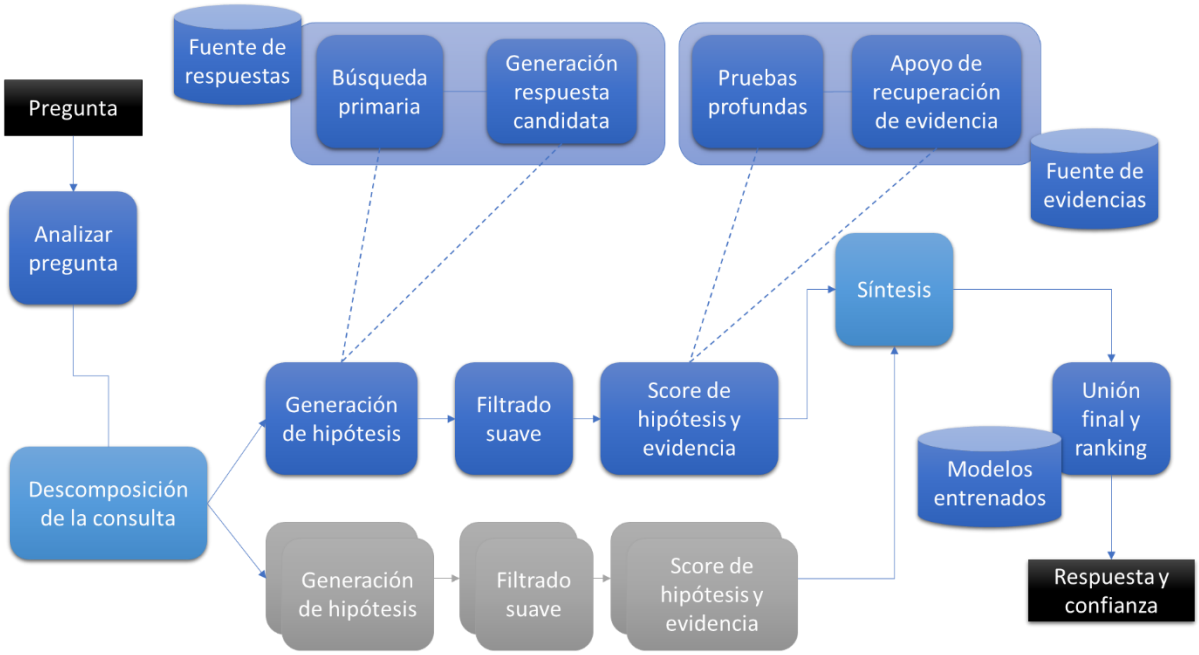

Ilustración 2 Ferrucci, D. 2010

1. **Adquisición de contenido:** Es el primer paso de una aplicación de DeepQA, la identificación y recopilación del contenido se resumen en 2 pasos, el primero es analizar ejemplos de preguntas del espacio del problema (rama de la ciencia o situacional) y guardar una descripción de los tipos de preguntas que pueden ser contestadas, el segundo paso es el análisis del espacio o dominio del problema el cual puede ser determinado por análisis automático o estadístico, las fuentes de Watson que se deben incluir dentro de la base incluyen enciclopedias, diccionarios, artículos de noticias, obras literarias y otras fuentes fidedignas de información. (Docasar, A. 2016.)

Ilustración 3 IBMWatson

Dada una cantidad razonable de datos, DeepQA inicia de forma automática un proceso para alimentarse de nuevos datos a partir de la semilla base implantada. El proceso involucra cuatro pasos de alto nivel: (1) identificar los documentos de la semilla y recuperar los documentos que coincidan con la información base desde la web; (2) extraer contenido autónomo de las referencias en el texto buscando los documentos web relacionados; (3) asignar una puntuación de las referencias y determinar si sobresalen de los conceptos iniciales (4) combinar las referencias más informativas y crear una nueva base de conocimiento. (Ferrucci, D. 2010)

Ilustración 4 IBMWatson

2. Análisis de preguntas

El sistema inicia un proceso de reconocimiento de la pregunta lo que llamaríamos nosotros entender la pregunta y determinar que se está preguntando, con ello iniciar un análisis, en otras palabras, analizar, interpretar, y entender la pregunta para determinar como el sistema deberá procesarla en el resto del sistema y llegar a una conclusión. En esta etapa el Deep QA genera etiquetas con atributos en común que permitan mezclar la información con la base del conocimiento, así como una serie de análisis profundos, semánticas, relaciones y entidades que le ayudan a comprender el propósito de la pregunta y así poder generar nuevas preguntas que ayuden a entender la pregunta. (UPM 2015.)

Clasificación de las preguntas. Su propósito es identificar si las preguntas se pueden dividir en preguntas más sencillas y a su vez determinar cuáles partes de estas preguntas requieren un procesamiento especial. Esto puede incluir cualquier cosa, desde palabras aisladas con significados dobles a cláusulas completas que contiene cierta funcionalidad sintáctica, semántica o

retórica que para ayudar a identificar a las demás etapas si la pregunta contiene restricciones y/o componentes de definición. La clasificación de la pregunta puede ir por varios caminos y concluir dándole una definición por ejemplo decir que la pregunta está orientada a buscar una definición, o está orientada a un concepto matemático u orientada a crear un rompecabezas y sucesivamente hasta adecuarla en la clasificación correcta. (Ferrucci, D. 2010)

Enfoque y detección TRL. Un tipo de respuesta de léxico es una palabra o frase en la pregunta que especifica el tipo de respuesta que se debe presentar sin cualquier intento de entender su semántica. Esto ayuda a determinar si la frase es considerada parte de la respuesta o se puede convertir en una respuesta candidata, dándole una puntuación que luego puede ser evaluada para buscar un contexto en el que la frase conlleve a una respuesta o a encontrar una respuesta candidata. (Ferrucci, D. 2010)

Detección de relaciones. La mayoría de las preguntas contienen relaciones, si son sujeto-verbo, sintáctico, predicados o relaciones semánticas. Un ejemplo la frase “¿llegaremos a Bogotá temprano o en Soacha habrá trancón?”, Watson ve la relación Bogotá, Soacha, Trancón, como destinos y un estado del flujo vehicular, asimismo asocia el transcurso en el tiempo de un destino a otro con una medida de tiempo asociada también a la palabra temprano y una suposición de la palabra habrá que deja abierta la relación a buscar la situación de flujo, Watson determina a través de puntuación que tan importantes son las palabras y sus relación para así determinar cuáles relaciones son las más importantes y generar respuestas candidatas, algo a tener en cuenta es que las palabras por si solas para el sistema no determinaran un proceso de búsqueda lineal en las bases de datos para hallar una respuesta guardada en base de datos, por lo tanto hasta no encontrar una relación de palabras no se podrá determinar que una pregunta ya había sido guardada con anterioridad en una base de conocimiento si anteriormente el sistema no la había consultado ni generado una respuesta. (UPM 2015.)

Descomposición. Es la habilidad del sistema de determinar si una pregunta requiere ser separada en varias preguntas por medio de búsquedas en profundidad y clasificación estadística y a su vez poder encontrar respuesta a cada una de ellas por separado, luego analizar con puntuación si las respuestas candidatas en conjunto al ser unidas pueden dar una solución solida o solo se convierten en una respuesta candidata, la descomposición puede ser recursiva si al descomponer una pregunta las preguntas generan otras preguntas que deben ser resueltas en otro nivel de iteración. (Ferrucci, D. 2010)

3. Generación de Hipótesis

La generación de hipótesis toma los resultados entregados por el *análisis de preguntas,* con ellas genera una puntuación dada entre la interacción de buscar en la base de contenido una respuesta a cada pregunta y la información proporcionada por el sistema según el procesamiento de la pregunta, con este proceso determina cual información puede ser o contener respuestas candidatas. Así cada respuesta candidata que se asocie en un alto porcentaje se convierte en una hipótesis, que el sistema tiene que probar con cierto grado de confianza. (Ferrucci, D. 2010)

Búsqueda primaria. En la búsqueda primaria el objetivo es encontrar el contenido que cumpla con mayor exactitud a las preguntas recibidas del análisis de preguntas, también permite determinar si una respuesta es candidata y agrega el contenido encontrado. Igualmente, si no se encuentra contenido se puede refutar la pregunta examinada o determinar si una pregunta refuta a la anterior, esto aumenta la precisión de las respuestas candidatas. (Ferrucci, D. 2010)

Ilustración 5 IBMWatson

Entre las técnicas de búsqueda se incluye el uso de motores de búsqueda de texto múltiples con diferentes enfoques subyacentes, la búsqueda de documentos y el paso a búsquedas con base de conocimientos utilizando SPARQL. (Ferrucci, D. 2010)

Generación de la respuesta candidato. En esta etapa, los resultados generados de la búsqueda son interpretados y transformados en respuestas candidatas. Cada resultado obtiene una puntuación y es muy común encontrar que muchas respuestas candidatas estén basadas en la misma pregunta y se halla llegado a la conclusión en diferentes búsquedas al contenido, aquí cada resultado deberá ser analizado en mayor detalle e identificar cuál de ellos contiene resultados de búsqueda más específicos, cuando esto sucede se aplican técnicas como búsqueda de frameworks de descripción de recursos y diccionario inverso para determinar cuál fuente es más verídica y generar una mejor respuesta candidata. Cuando las respuestas no arrojan un valor alto de confiabilidad de la respuesta se puede concluir que ninguna respuesta es la correcta y que se requiere de más información o realizar una mayor precisión en la búsqueda de la pregunta inicial, lo anterior puede ocurrir aun cuando exista un conjunto de candidatos bastante grande. Uno de los objetivos del sistema es tolerar el ruido que genera la información en las primeras etapas de filtrado de

información e ir aumentando la precisión a medida que pasa por las etapas, Watson en esta etapa puede generar varios cientos de candidatos. (Ferrucci, D. 2010)

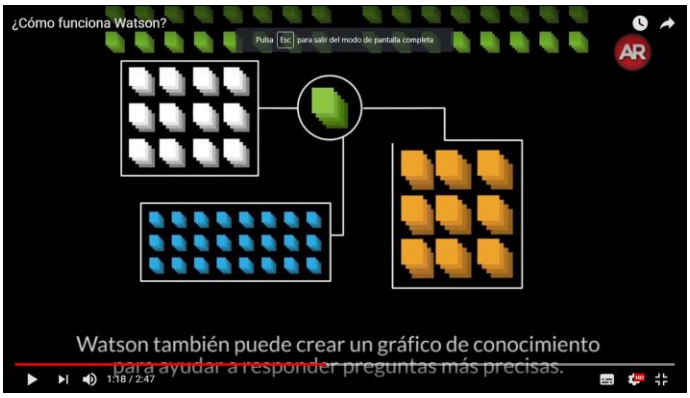

Ilustración 6 IBMWatson

Filtrado suave. En la etapa anterior entendimos que pueden existir cientos de candidatos, antes de iniciar a analizar los puntajes de cada candidato se hace una pequeña reducción de candidatos, estos que desde el principio no pasan el filtro de score y pueden haber sido generados por ruido que al final no aporta un candidato que supere la etapa de unión y respuesta de la hipótesis, pueden ser un ejemplo de este caso respuestas con frases que no contienen un tipo de TRL, por lo tanto la frase que genera no impacta de forma positiva como candidata debido a que no podrá explicar al usuario lo requerido. (Ferrucci, D. 2010)

Ilustración 7IBMWatson

4. Hipótesis y Puntuación de Evidencia

Después de haber superado el filtrado suave, todas las respuestas candidato que han quedado empiezan un proceso de ampliación de información a fondo de sus fuentes, es decir el sistema determina que debe incorporar pruebas adicionales de apoyo para comprobar que la respuesta candidato es capaz de complacer como respuesta a la pregunta principal dada por el usuario, para ello usa algoritmos de análisis de puntuación profunda, cada prueba adicional deberá aportar en gran detalle para mejorar su respuesta candidata, esto generará una nueva puntuación. (Ferrucci, D. 2010)

Recuperación de pruebas. Cada respuesta candidato que ha pasado la etapa de filtro suave requiere reunir pruebas de apoyo adicionales, un ejemplo de técnica usada para ampliar este espectro de la respuesta es la búsqueda de paso, la cual une la pregunta candidato con su nueva información de apoyo y deben crear una respuesta en donde la terminología presente conceptos clave de la pregunta base y la información de apoyo. Para ello, se recuperarán los vínculos que contienen la respuesta candidato utilizada en el contexto de los

términos de la pregunta. Una Prueba de apoyo también puede tener como origen un framework de descripción de recursos. las evidencias de apoyo recuperadas se dirigen a los componentes de evaluación de la evidencia, los cuales evalúan la respuesta del candidato en el contexto de la evidencia. (Ferrucci, D. 2010)

Puntuación. Este paso es uno de los más importantes porque es donde todo el contenido obtenido en los pasos individuales es calificado basado en que grado de certeza con el que la evidencia recuperada apoya a la respuesta candidato para que esta sea elegida como la respuesta que cumple a cabalidad con el requerimiento de la pregunta, los algoritmos de puntuación tienen la importante tarea de realizar un análisis al contenido en profundidad y dar puntuaciones al contenido que se determine produce valor agregado a las evidencias que acompañan a la pregunta candidata. (Ferrucci, D. 2010)

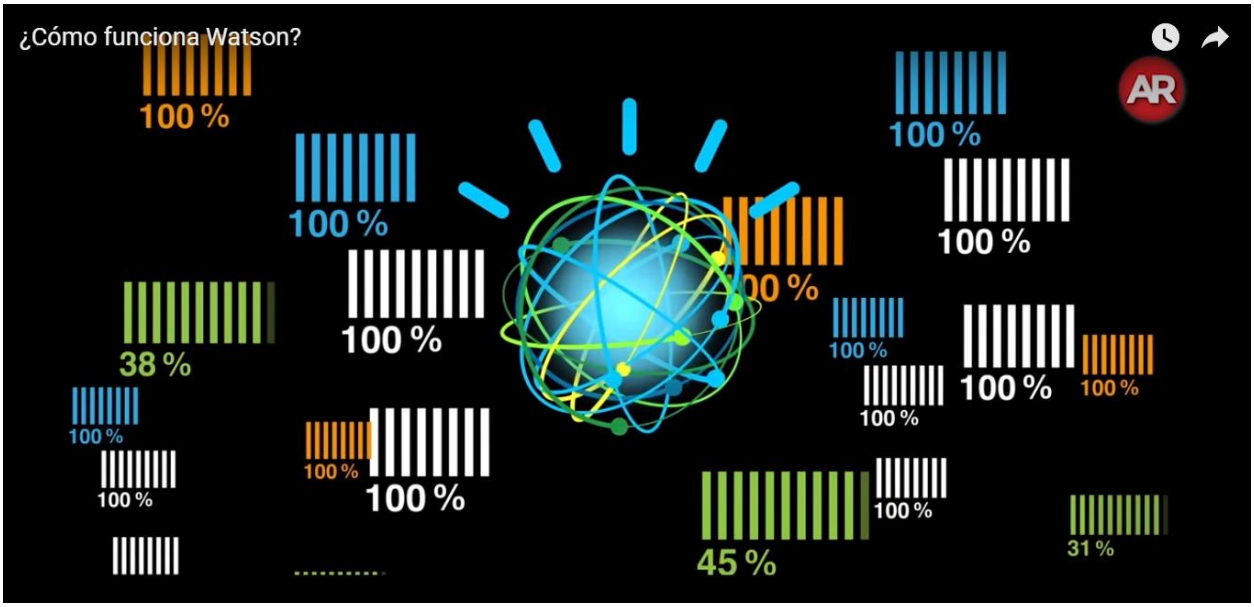

Ilustración 8IBMWatson

El framework de DeepQA permite la inclusión de muchos componentes diferentes, o anotadores, sin importar la cantidad de dimensiones diferentes que se creen desde que estas le permitan presentar la evidencia como válida y ayudar en la producción de una buena puntuación. (Ferrucci, D. 2010)

DeepQA usa formatos comunes para generar sus cálculos de puntuación sin limitarse a un tipo de semántica, esto permite que los desarrolladores puedan interactuar con la información y agregarle condiciones de puntuación sin poderla segmentar de alguna forma para que las respuestas sean más fiables y sus puntuaciones cambien en el formato global del cálculo. Entre más condiciones de puntuación sean agregadas más el sistema podrá interactuar con la información y generar puntuaciones más verídicas. Por ejemplo, Watson emplea más de 50 componentes marcadores que producen puntajes que van desde probabilidades formales hasta conteos de características categóricas, basadas en evidencias de diferentes tipos de fuentes incluyendo texto no estructurado, texto semiestructurado y frameworks de descripción de recursos. (Ferrucci, D. 2010)

Otro elemento clave adicional que contiene Watson para generar puntuaciones es usar frameworks de descripción de recursos para determinar si una pregunta realmente pretende resolver una respuesta, o simplemente está mal planteada y el análisis que se realice no traerá información relevante pero que de alguna manera se relaciona si se realizara un corrección a la pregunta, esto se puede ver en el espacio tiempo que se realice una pregunta, ejemplo “la independencia de Colombia se da en 1845 por Simón Bolívar”, Watson es capaz de determinar que la pregunta tiene una afirmación, pero que a su vez en el espacio temporal, no coincide con la información almacenada para tal fin y por ende a partir de la evidencia recopilada, le puede sugerir cambiar el tiempo presentado con una argumentación de cuál sería el tiempo verídico que debería presentar, este framework ayuda en detalle para cálculos de puntuación a más a profundidad y ayuda a corregir errores que se presenten durante la presentación de la pregunta y a su vez enseñar al usuario como su información pueden ser mejorada. Un ejemplo más que se puede usar es buscar cuales deberían ser los verdaderos límites geográficos de un país, dado que, al buscar la información, también toma como base la historia y con ello determina que una tendencia en internet no es tan fuerte como una fuente verídica de una enciclopedia, o un

mapa que presente la fecha de su publicación y porque está delimitado de esa manera. (Ferrucci, D. 2010)

5. Fusión Final y Clasificación

Al final una respuesta debe ganar si lo ponemos en ese contexto, la unión y clasificación final se encargan de determinar que frase, contenido, respuesta es la más correcta identificando que su respuesta haya obtenido en la puntuación los valores requeridos para que la veracidad y confiabilidad sean lo suficientes para ser presentada al usuario sin ningún problema, por otro lado puede que otras hipótesis o preguntas candidato también estuvieran muy cerca de ser la respuesta, dado que como anteriormente se mencionó también puede que el usuario no preguntara lo que realmente deseara, en caso de que esto fuera así y se iniciara un nuevo proceso de búsqueda por parte del usuario, Watson podrá fácilmente recuperar sus análisis anteriores y determinar si alguna de sus hipótesis era correcta y presentarla en un tiempo prudencialmente mas corto. (Ferrucci, D. 2010)

6. Que hace diferente a Watson

Al finalizar el proceso DeepQA, podemos ver que se ha interpretado como si los datos estuvieran claros, pero estas preguntas pueden estar en diferentes contextos, dominios y áreas según las personas se desenvuelven día a día, para ello Watson fue diseñado con Machine Learning, se le enseño a determinar en que contexto o dominio debía trabajar la tarea que se le asignara y así a su vez ir mejorando mientras estaba en el entorno de los usuarios, haciendo que sus puntuaciones mejoren cada vez que interacciona con un usuario. (Ferrucci, D. 2010)

Ilustración 9IBMWatson

Conclusiones

Watson es la herramienta core que proporciona IBM para la ejecución de programas analíticos y estadísticos que toman como base el lenguaje natural, al estar programado de tal forma que pueda tomar decisiones en base a puntuaciones, podemos concluir que sus respuestas estarán basadas en la información que contenga internet y su capacidad algorítmica para determinar que información realmente es necesaria y a su vez a la medida que interactúa con el ser humano va aprendiendo su lenguaje de comunicación para dar respuestas más precisas. A nivel de conocimiento Watson parece no tener límites, por lo tanto puede ser usado en muchas áreas de negocio las cuales IBM ya maneja y por ello ofrece productos que se soportan en Watson como es el caso de Watson Analytics, aplicación que sirve para procesar información de a niveles de Big Data teniendo claro que el servicio procesara la información basado en toda la información que tiene guardada en su base de conocimiento y nos podrá decir los puntos fuertes de nuestra información y sugerirnos como presentarla y entender al detalle lo que sucede en nuestros datos, sin la necesidad de conocer sobre programas analíticos avanzados, los cuales Watson ya maneja.

REFERENCIAS BIBLIOGRÁFICAS

- Ziffdavis. (2015). Dispoinble en:

https://hosteddocs.ittoolbox.com/zd-wp-IBM-Watson-Analytics_FINAL.pdf

-Ferrucci, D. (2010). Arquitectura de alto nivel de DeepQA. Págs, 69,70,71,72,73,74. Disponible en: https://vvvvw.aaai.org/ojs/index.php/aimagazine/article/view/2303/2165

- Docasar, A. (2016). Computación cognitiva Watson https://www.youtube.com/watch?v=ZUa8qomeinA

- IBM, (2014). España. Disponible en: https://www.youtube.com/watch?v=WMnASdda1w4

- IBM. (2017). IBM. Disponible en:

https://www.ibm.com/co-es/marketplace/watson-analytics

- IBM Watson (2017). Disponible en:

https://www.ibm.com/watson/developer/

- IBM, The DeepQA Project Disponible en:

https://www.research.ibm.com/deepqa/deepqa.shtml

-UPM (2015). Disponible en:

https://www.youtube.com/watch?v=9xTo7EKL6aA

- IBM Research (2013). Disponible en:

https://www.youtube.com/watch?v=P18EdAKuC1U

-IBM Watson (2017) Disponible en

https://www.youtube.com/watch?v=aSx9_PKPGPk