

# **Prototipo de software para la determinación de características para la clasificación de éxitos de ventas usando procesamiento de lenguaje natural y aprendizaje automático.**

## **Trabajo Terminal No. 2020-A012**

Alumnos: Gómez Rodríguez Héctor Yair, Quiroz Palacios Pedro Manuel\*

Directores: Grigori Sidorov, Reyna Elia Melara Abarca

Turno para la presentación del TT: Matutino

E-mail: [manolopantufla@gmail.com](mailto:manolopantufla@gmail.com)

**Resumen** - En pleno auge de la era digital, es cierto que la industria editorial se ha visto afectada en cierta forma, entre las tantas nuevas formas de recreación, además del fácil acceso a cualquier contenido que nos brinda el internet, no obstante, el avance tecnológico y las nuevas herramientas con las que contamos también pueden ser de gran utilidad para ésta industria. Se propone desarrollar un sistema que a través del uso de técnicas tradicionales de aprendizaje automático para el procesamiento de lenguaje natural, como Naive Bayes, Regresión Logística o K-Nearest Neighbors, y del uso de una red neuronal pre-entrenada (BERT), pueda clasificar, o bien, predecir el éxito de novelas literarias en el lenguaje inglés, extrayendo y analizando características clave — N-gramas de distintas longitudes— de un corpus compuesto por libros considerados como best sellers y otros como no best sellers. Se busca obtener resultados concluyentes para determinar si las características y métodos seleccionados dan resultados suficientemente adecuados para servir como referencia para saber si es conveniente o no publicar un libro.

**Palabras clave** - éxito de ventas, sistema predictivo, procesamiento de lenguaje natural, aprendizaje automático.

### **1. Introducción**

La industria literaria es una de las industrias más grandes y redituables a nivel global, la cual involucra a tres actores principales: los escritores, las editoriales y los lectores. Por lo general, los escritores buscan que sus obras lleguen a la mayor cantidad de lectores posibles a través del mercado, ya que eso se traduce en un beneficio económico y de reputación personal, pero que un escritor logre publicar su libro con una editorial puede no ser una tarea sencilla, debido a que solo pocos cuentan con los recursos económicos necesarios o contactos en la industria.

Así mismo, publicar un libro de forma independiente también requiere de considerables recursos económicos e implica un salto de fê, ya que, sin una reputación que los respalden (tal vez por no tener méritos relevantes o populares), puede ser algo muy riesgoso y es un riesgo que la mayoría de escritores no se pueden permitir, pues no solo se trata de escribir un libro único, sino de contar con las herramientas y el equipo necesario para hacer de un libro exitoso, como edición, marketing y distribución, algo con lo que las editoriales ya cuentan.

Por otro lado, las editoriales se centran más en el beneficio económico, en otras palabras, número de ventas, siendo minuciosas y selectivas al momento de escoger las obras que se publicarán, pero es sabido que en ocasiones las

editoriales se valen de métodos empíricos, subjetivos y poco claros en esta elección, métodos que les han dado buenos resultados, pero a pesar de esto, no podemos decir que sean del todo infalibles, ya que de otro modo no existirán éxitos de ventas inesperados como “50 sombras de Grey”, “El código Da Vinci”, entre otros, además de las cuantiosas pérdidas económicas que podrían resultar de un libro no vendido. Esto nos hace pensar que hay aspectos que las editoriales suelen dejar pasar y que si tuvieran información que es poco explícita sobre el contenido del libro, realizarían sin duda una mejor selección.

A lo largo de los años han aparecido obras con un número considerable de ventas, ya sea por su calidad como obra literaria, por la aclamación de la crítica literaria, o bien, por su impacto sociocultural. Estas obras, conocidas como “best sellers” [1] o simplemente éxitos de venta, son motivo de inspiración para muchos escritores y representan una fuente de ingresos importante para las editoriales. Por ende, un best seller es la aspiración de 2 de nuestros actores principales (escritores y editoriales).

Parece una tarea difícil saber realmente cuáles libros son best sellers y cuáles no, ya que obtener los valores exactos del número de ventas de cada libro por editorial puede ser complicado y, en ocasiones, hay que pagar grandes sumas de dinero para obtener estos datos, no obstante, existen listas reconocidas sobre los libros más vendidos como *The New York Times best Sellers List* (lista de bestsellers que data desde 1930), *USA TODAY's Best-Selling Books list* y *WALL STREET JOURNAL-BEST SELLERS*, pero estas siguen pareciendo puntos de vista algo subjetivos y cambiantes. Para evitar lo anterior, podemos recurrir a la opinión colectiva, misma que, probabilísticamente -usando la desigualdad de hoeffding como base-, podemos decir que a medida de que la muestra crece, o bien, a medida que las opiniones de un libro se incrementan, más se asemejan a las opiniones reales que toda la población podrían tener sobre dicho libro. Afortunadamente para nosotros, la plataforma de catalogación virtual de libros Goodreads, para 2019, cuenta ya con más de 90 millones de usuarios, lo que nos dará una muy buena aproximación de las valoraciones que tendría un libro a nivel global.

Para contrastar la situación real de los libros exitosos contra los que no lo son, además de las listas y de las tendencias en línea, existen las ediciones de los libros. Ciertamente no podemos decir con exactitud a cuántas reproducciones de un libro equivale una edición (alrededor de 4000 copias), pero sí sabemos que lo más usual y por cuestiones de marketing, es que cuando un libro llega a una nueva edición, es por que el libro va otra vez a imprenta, por ello podemos decir que entre más éxito tiene un libro, mayor número de ediciones tiene. Tomando a España como referencia de un país con una gran cantidad de editoriales, el Ministerio de Cultura y Deporte de este país ha mostrado en su publicación anual *Panorámica de la edición española de libros 2018* (publicación más reciente disponible) que la mayoría de libros publicados no pasa de la primera edición.

Producción editorial por ediciones (2017-2018)					
Ediciones	2017	%	2018	%	Variación interanual %
Primera edición	88 119	98,0	79 100	97,4	-10,2
Segunda edición	1150	1,3	1403	1,7	+22,0
Tercera edición	253	0,3	265	0,3	+4,7
Cuarta edición	120	0,1	147	0,2	+22,5
Quinta edición	58	0,1	87	0,1	+50,0
Sexta a décima edición	151	0,2	143	0,2	-5,3
Más de 10 ediciones	111	0,1	83	0,1	-25,2
<b>Total</b>	<b>89 962</b>	<b>100,0</b>	<b>81 228</b>	<b>100,0</b>	<b>+4,6</b>

Tabla 2. Producción editorial por ediciones de España (2017 - 2018)

De esta forma nos podemos dar cuenta de la competitividad en la industria literaria, del poco aprovechamiento que las editoriales tienen sobre la mayoría de sus publicaciones y de lo difícil que es para los escritores no solo publicar un libro, sino que también sea relevante en esta industria cambiante y en constante crecimiento.

Se realizará una predicción del éxito de venta de un libro usando libros calificados con un rango de estrellas del sitio web de catalogación de libros Goodreads, utilizando distintas técnicas de procesamiento de lenguaje natural y aprendizaje automático para determinar las características más eficientes, además de determinar si estas son suficientes para la predicción de éxito de ventas.

En la tabla 1 se muestran una serie de propuestas similares a la nuestra

<b>SOFTWARE</b>	<b>DESCRIPCIÓN</b>	<b>PRECIO EN EL MERCADO</b>
bestseller-ometer [2]	Modelo desarrollado para extraer las características de libros de la lista <i>The New York Times Best Seller list</i> haciendo uso de procesamiento de lenguaje natural y de la estilometría para poder predecir si un libro entraría en esa lista.	No disponible a la venta
“Analyzing Social Book Reading Behavior on Goodreads and how it predicts Amazon Best Sellers” [3]	Modelo desarrollado por Amazon para predecir “best sellers” dentro de su tienda virtual basándose en los patrones de lectura de los lectores así como de las críticas y reseñas que estos hacen en el sitio web Goodreads.	No disponible a la venta. Propiedad de Amazon
"Success with Style: Using Writing Style to Predict the Success of Novels" [4]	Trabajo de investigación desarrollado por la Stony Brook University para predecir el éxito de novelas basado en las novelas encontradas en el sitio web Project Gutenberg que son consideradas exitosas y no exitosas por ese sitio.	No disponible a la venta

Trabajo terminal:  “Bot conversacional emulador de un autor literario”	Chatbot que a través de aprendizaje automático se entrena con información de un autor importante en los medios de comunicación con el fin de emular a dicho autor.	No disponible a la venta  Continúa en desarrollo
--	--	--

Tabla 2. Propuestas similares

## 2. Objetivos

### Objetivo general

El objetivo general de nuestro proyecto es el desarrollar un sistema capaz de realizar una predicción de cuán exitosa será una novela de ficción en el mercado utilizando las calificaciones de estrellas existentes en Goodreads como referencia y basándonos en el contenido de la novela.

### Objetivos particulares

Los objetivos particulares de nuestro proyecto son:

- Recopilar un conjunto de novelas de ficción en inglés con distintas calificaciones por estrellas del sitio web Goodreads.
- Aplicar técnicas de aprendizaje automático tradicionales en el procesamiento de lenguaje natural con librerías de python.
  - Extraer distintos tipos de características
  - Aplicar distintos tipos de técnicas de clasificación: Naive Bayes, Regresión Logística y K-Nearest Neighbors.
- Utilizar la red neuronal BERT
  - Aplicar alguno de los 3 tipos de ajuste fino: Entrenar la red completa, entrenar algunas capas mientras se congelan (bloquean) otras capas o congelar todas las capas y adjuntar otra red neuronal.
- Aplicar la o las técnicas seleccionadas al corpus dedicado a entrenamiento.
- Evaluar el sistema

- Calcular los valores de precisión, especificidad y la medida F1 para todos los métodos usados.
- Comparar los resultados y determinar qué método obtuvo los mejores resultados.
- Concluir si los resultados más óptimos obtenidos son suficientes para considerar el sistema como exitoso.

### 3. Justificación

La incertidumbre de los escritores y las editoriales acerca del éxito de venta que pueda tener una novela siempre ha sido una cuestión muy discutida dentro de la industria literaria, pues ambas partes se ven beneficiadas o perjudicadas dependiendo del éxito que tenga la obra en el mercado.

Se sabe que los criterios de una editorial para publicar una novela son poco claros y subjetivos. Pues estos van desde el visto bueno por parte de la editorial en cuanto a la trama de la novela, pasando por saber si novelas similares han tenido éxito o incluso averiguando si el escritor goza de cierto prestigio en el ámbito literario o fuera de él. Mientras que los escritores por su parte se han guiado de aquellas obras que son un éxito de venta para poder escribir la suya, tomando en consideración características intrínsecas tales como tema, trama, estilo y personajes a fin de concordar con las tendencias que cumplen estas novelas y aumentar la probabilidad de publicación por parte de la editorial.

Con todo lo anterior se muestra la necesidad de las editoriales y los escritores de mejorar su criterio de evaluación de sus obras, ya que como pudimos observar en la tabla 1, el tener una obra en una tercera o cuarta edición es bastante difícil y llegar a una décima edición es casi como ganarse la lotería.

Los sistemas de predicción son útiles para la toma de decisiones, ya que éstos ofrecen un panorama de lo que puede llegar a ocurrir en lo que concierne a un determinado tema o producto. Es por esto que se propone realizar un sistema que pueda servir en la predicción de éxito que pueda tener una novela a partir de su contenido, de manera que tanto escritores como editoriales puedan tener un panorama en cuanto al éxito de venta que tendrá la novela que escribirán o publicarán según sea el caso. Decisiones

Si bien ya existen algunas propuestas que cumplen con el objetivo especificado, hemos observado que éstas se encuentran enfocadas a una plataforma en específico como es el caso de Amazon, o bien a una sola lista como el bestseller-ometer con la *The New York Times Best Seller list*. Estas propuestas de solución se encuentran en el idioma inglés, y nuestro caso no será la excepción, pues de esta forma contaremos un corpus mucho más amplio [8] y de cierta forma, más “puro” en el sentido de que la mayoría de los best sellers fueron escritos originalmente en este idioma, permitiendo rescatar aún más particularidades del tipo de escritura de los autores al no haber pasado por una traducción previa, pero cabe aclarar que no habría necesidad de realizar algún cambio en el modelo a implementar si fuera en español, dándonos la posibilidad a futuro de implementarlo con gran facilidad en éste último.

En la siguiente tabla que hemos extraído del artículo [8] publicado por el diario canadiense “Partnership Journal” basado en datos del Instituto estadístico de la UNESCO, se muestra fehacientemente que existe una cantidad mucho mayor de recursos literarios en idioma inglés que en español.

Idioma	Numero de títulos	Porcentaje del total
Inglés	200,698	21.84%
Chino (Mandarín)	100,951	10.99%

Alemán	89,986	9.78%
Español	81,649	8.88%
Japones	56,221	6.12%
Ruso	48,619	5.29%
Frances	44,224	4.81%
Coreano	35,864	3.90%
Italiano	34,768	3.78%
Holandés	34,067	3.71%
Portuguese	33,430	3.64%

Tabla 3. Libros publicados por idioma.

Debido a que entrenar un modelo de aprendizaje automático desde cero puede ser una tarea pesada y a su vez difícil de alcanzar con los tiempos que se cuentan para la realización de este trabajo terminal, se decidió usar BERT, una red neuronal de tipo Transformer que ha sido pre-entrenada con texto sin etiquetar —empleando un corpus que supera las 3000 millones de palabras—, lo que permite un re-entrenamiento posterior para utilizarse de forma más especializada en distintas tareas del Procesamiento de Lenguaje Natural (ajuste fino) y en nuestro caso, para la clasificación de textos. Así mismo, BERT nos ofrece una ventaja sobre otros modelos de redes neuronales populares como las Redes Neuronales Recurrentes, ya que en su secuencia de entrada sólo permite tomar token por token, mientras que BERT nos permite tomar toda la secuencia a la vez, acelerando en gran manera el proceso. \\\

Por otro lado, es importante probar los métodos de aprendizaje automático tradicionales propios del procesamiento de lenguaje natural para mejorar el proceso de experimentación al permitirnos utilizar distintos tipos de características con distintos tipos de métodos de clasificación, mejorar los resultados y compararlos con los resultados obtenidos con BERT. \\\

El desarrollo de este sistema aunque se puede considerar sólo como un problema de clasificación, cabe resaltar que es un proyecto escalable, ya que existen otros aspectos para analizar sobre los libros literarios, tales como: análisis de estilométricas, análisis de comportamiento de caracteres, análisis de temas, análisis de sentimientos, entre otros, ya que también son aspectos con los que otros autores han trabajado y obtenido buenos resultados, aun así, existen distintos métodos y aspectos con los que experimentar después de finalizar este trabajo terminal. \\\

#### 4. Productos o resultados esperados

En la Figura 1 se presenta la arquitectura que seguirá el sistema:

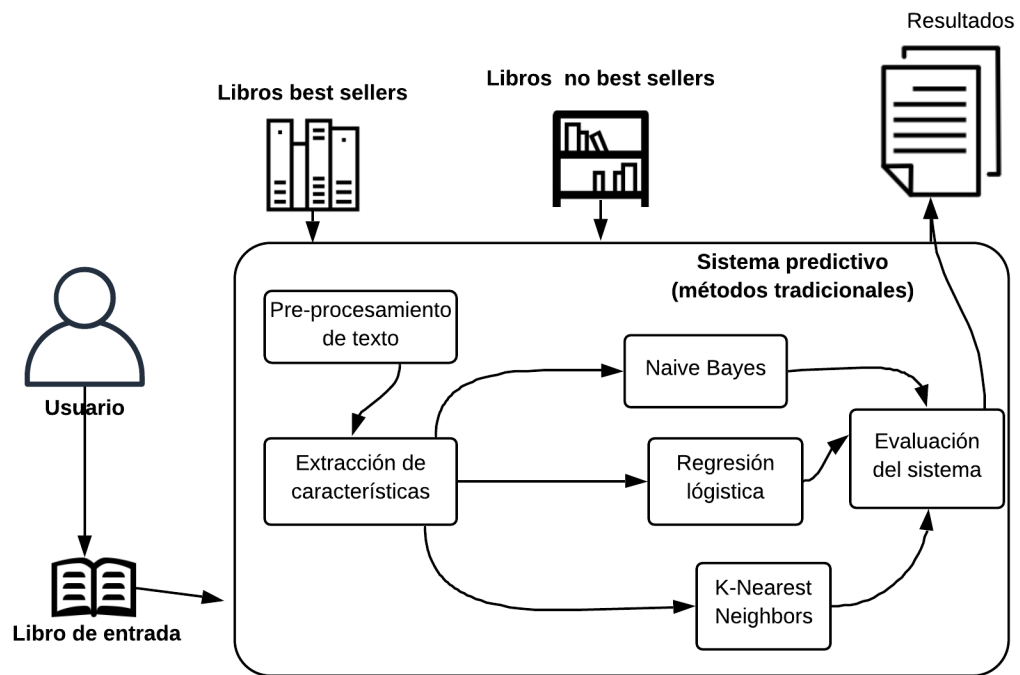


Figura 1. Diagrama de arquitectura del sistema para métodos tradicionales

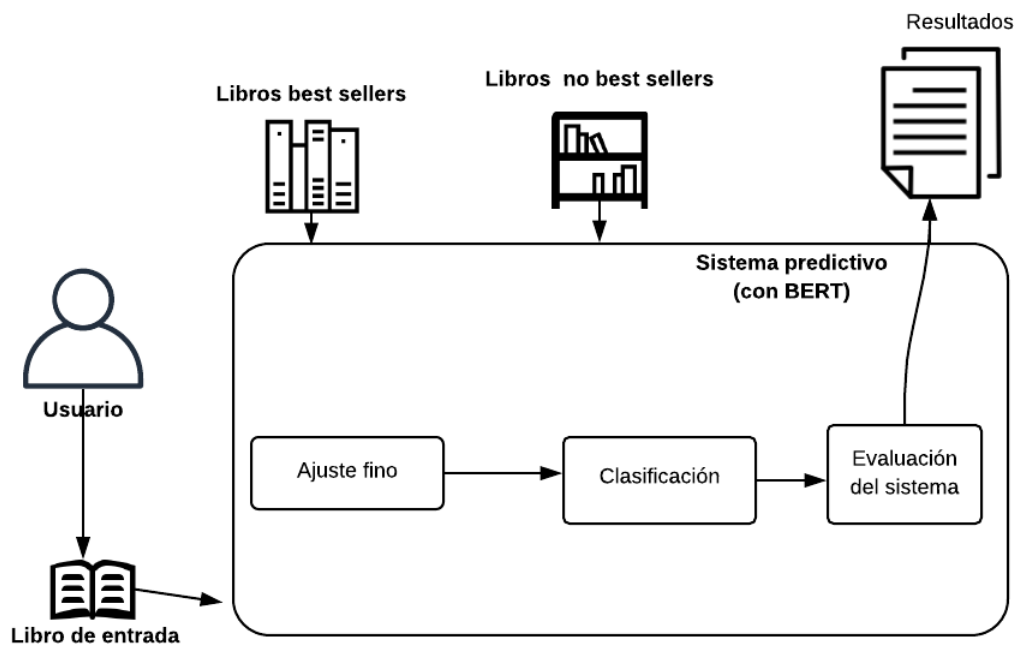


Figura 2 . Diagrama de arquitectura del sistema con BERT

A continuación, se listan los productos esperados al realizar el TT:

1. Documentación técnica del sistema
2. Manual de usuario
3. Reporte técnico con resultados y conclusiones
4. Código del sistema en cuestión



## 5. Metodología

Debido a que nuestro proyecto está orientado hacia la investigación, serán necesarias múltiples pruebas y un análisis continuo de los requerimientos, por lo que se ha decidido optar por una metodología que nos permita realizar modificaciones de una forma flexible y poco costosa, además de mostrar claramente los avances obtenidos durante el desarrollo del trabajo terminal, por ello hemos seleccionado la metodología de desarrollo por **prototipos evolutivos**, misma que nos brindará la posibilidad de presentar un producto funcional en cada revisión que se efectúe durante el desarrollo del trabajo terminal, además de la facilidad de realizar cambios en caso de encontrar modelos o algoritmos más eficientes en cualquier etapa del desarrollo.

Además de basar el diseño en el modelado por prototipos, también tomaremos como referencia las **etapas para la investigación en lingüística computacional** propuestas por el Dr Grigori Sidorov en su libro “Construcción no lineal de n-gramas en la lingüística computacional” [5]. De esas etapas, seleccionamos aquellas que funcionan en nuestro proyecto y de ser posible, las implementaremos iterativamente consiguiendo una fusión con el modelado por prototipos evolutivos.

### Etapas no iterativas:

- **Selección de textos: itemize**

Utilizaremos libros publicados en las listas de best sellers para tener una referencia, además de Goodreads, donde podremos encontrar libros con todo tipo de calificaciones (desde 1 hasta 5 estrellas).

- **Definición del estándar de oro : itemize**

-En nuestro caso particular, al ser una tarea de clasificación, el estándar de oro se define sencillamente al etiquetar un libro con la cantidad de estrellas correspondientes.

### Etapas iterativas:

- **Determinar características: itemize**

Definiremos cuáles serán las características con las cuales experimentaremos para encontrar aquellos que sean más relevantes al momento de pensar en un best seller. Esta etapa consiste en segmentar el texto en distintas palabras(unigramas) o secuencias de palabras(n-gramas). Antes de esto se extrae el texto crudo (texto tal cual está en los libros), se quitan símbolos que no nos sirvan, quitamos las stopwords (palabras que no agregan mucho significado), entre otras palabras o frases que se consideren para ser removidas.

- **Construir el modelo de espacio vectorial:**

Tomaremos esas características y les asignaremos un valor cuantitativo definiendo un espacio vectorial en el cual podamos operar de forma clara y formal.

- **Implementación de métodos de línea base y estado del arte:**

Una vez tengamos nuestro modelo de espacio vectorial, lo utilizaremos para implementar los métodos tradicionales de procesamiento de lenguaje natural y aprendizaje automático para poder realizar la clasificación.

- **Selección e implementación de métodos de aprendizaje automático:**

Una vez teniendo en consideración los resultados de los métodos de línea base, tomaremos las

características previamente definidas y utilizaremos la red neuronal pre-entrenada (BERT), añadiendo claro las etapas que sean necesarias para su especialización.

- **Experimentación de clasificación de best sellers:**

Se aplicarán los textos de entrenamiento a los modelos y recopilaremos los resultados para su futura evaluación.

- **Evaluación de precisión y especificidad:**

Al finalizar el proceso, contrastaremos ambos métodos y determinaremos tanto si las características seleccionadas como los modelos definidos fueron o no concluyentes de modo que tengamos un punto de partida para la siguiente iteración.

## 6. Cronograma TTII

[illegible]

## 7. Referencias

- [1] J. Pérez Porto, A. Gardey, *Definición de best seller* (sitio web), 2017, <https://definicion.de/best-seller/> [Último acceso: septiembre de 2020].
- [2] A. Velika, Big Apple Strippers, *Your Personalized Manuscript Report- Archer Jockers: A Unique Book Consultancy* -, 2018. [En línea], [https://static1.squarespace.com/static/5648bac6e4b0e1362b34c0e7/t/5b23f46803ce6488fcbf3d76/1529082989786/Big\\_Apple\\_Strippers\\_Alexandra\\_Velika.pdf](https://static1.squarespace.com/static/5648bac6e4b0e1362b34c0e7/t/5b23f46803ce6488fcbf3d76/1529082989786/Big_Apple_Strippers_Alexandra_Velika.pdf) [Último acceso: Septiembre 2020].
- [3] S. Kalyan Maity, A. Panigrahi, A. Mukherjee, *Analyzing Social Book Reading Behavior on Goodreads and how it predicts Amazon Best Sellers*, 19 de septiembre 2018. [En línea], <https://arxiv.org/abs/1809.07354> [Último acceso: Septiembre 2020].
- [4] V. Ganjigunte Ashok, S. Feng, Y. Choi, *Success with style: Using writing style to predict the success of novels*, 21 de octubre 2013, [En línea], <https://www.aclweb.org/anthology/D13-1181.pdf> [Último acceso: Septiembre 2020].
- [5] G. Sidorov, CONSTRUCCIÓN NO LINEAL DE N-GRAMAS EN LA LINGÜÍSTICA COMPUTACIONAL, 1st ed, México DF: Kronos Digital S.A. de C.V., 2013, pp. 59-74.
- [6] M. Wilkens, *How Many New Novels are Published Each Year?*, 13 de febrero 2020, [En línea], <https://mattwilkens.com/2009/10/14/how-many-novels-are-published-each-year/> [Último acceso: Septiembre 2020].
- [7] W. Vorhies, *NLP Picks Bestsellers – A Lesson in Using NLP for Hidden Feature Extraction*, 3 de septiembre 2020, [En línea], <https://www.datasciencecentral.com/profiles/blogs/nlp-picks-bestsellers-a-lesson-in-using-nlp-for-hidden-feature-ex> [Último acceso: Septiembre 2020].
- [8] S. Lobachev, University of Western Ontario, *Top languages in global information Production*, 17 de diciembre 2008, [En línea], <https://journal.lib.uoguelph.ca/index.php/perj/article/view/826> [Último acceso: Septiembre 2020].

## 8. Alumnos y Directores

*Quiroz Palacios Pedro Manuel.*- Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Boleta 2014140569, Tel. 5582320484, Email: manolopantufila@gmail.com

Firma: \_\_\_\_\_

*Gómez Rodríguez Héctor Yair.*- Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Boleta 2014011282, Tel. 5524251579 , Email: yatrex@gmail.com

Firma: \_\_\_\_\_

*Melara Abarca Reyna Elia.*- Licenciatura en Ciencias de la Informática, UPIICSA-IPN. Maestría en Ciencias de la computación, CIC IPN. Áreas de interés: Ingeniería de software, Procesamiento de Lenguaje Natural. Email: [remabarca@gmail.com](mailto:remabarca@gmail.com).


Firma: \_\_\_\_\_


*Dr. Grigori Sidorov.*- Profesor e investigador en el Centro de Investigación en Computación. Miembro de la Academia Mexicana de Ciencias. Áreas de interés: Técnicas y sistemas de procesamiento de texto, lingüística de corpus, desarrollo de software lingüístico. Email: [sidorov@cic.ipn.mx](mailto:sidorov@cic.ipn.mx)

Firma: \_\_\_\_\_



Protocolo 2020-A012  
(Modificación) Inbox

 **Hector Gomez** 7:13 AM  
Buenas noches profesora Reyna, nos comunicamos con usted pues tuvimos

 **Reyna Melara** 10:26 AM  
to me ▾

RECIBIDO, GRACIAS.

[Show quoted text](#)


↩ Reply

↩↩ Reply all

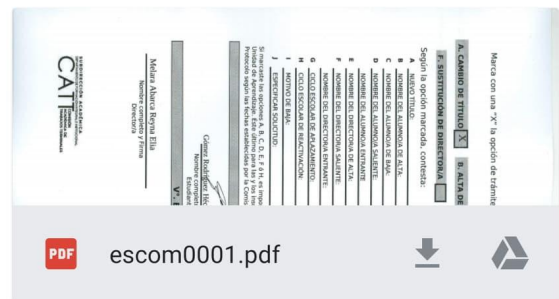
➦ Forward



formato Inbox

 **Grigori Sidorov** 12:46 PM  
to me ▾

Saludos



↩ Reply

↩↩ Reply all

➦ Forward