Universidad de Buenos Aires
Facultad de Ingeniería

Propuesta de Trabajo Profesional

Módulo de resúmenes automáticos basado en TextRank con integración a Gensim

Barrios, Federico – 91954
López, Federico – 92278

Índice

1. Introducción

El siguiente documento presenta la propuesta de Trabajo Profesional de Ingeniería en Informática de los estudiantes Federico Barrios (padrón 91954) y Federico López (padrón 92278).

El objetivo del proyecto es aplicar los conocimientos adquiridos en la carrera; el tema elegido es **"Módulo de resúmenes automáticos basado en TextRank con integración a Gensim"**.

Los resúmenes automatizados son muy utilizados en tareas relacionadas al procesamiento de lenguaje natural y de aprendizaje automático. Su uso en motores de búsqueda, por ejemplo, mejora la eficiencia de indexación de textos y a su vez asiste en la presentación de resultados de manera efectiva. El incremento en la cantidad de información disponible en Internet ha intensificado su utilización en los últimos años y, en consecuencia, se ha dedicado enorme esfuerzo para mejorar los algoritmos existentes. El número de aplicaciones del tema en cuestión en tareas de actualidad sirven como motivación para el desarrollo de este Trabajo Profesional.

Por otra parte, realizar un aporte de software libre responde a la intención de contribuir con la comunidad que a diario provee de herramientas de uso académico y profesional.

2. Descripción del problema

Un resumen es una reducción a términos breves y precisos de lo esencial de una fuente de información. Su objetivo es el de extraer contenido intentando sintetizar sus conceptos más importantes, y su uso es altamente benéfico en tareas de aprendizaje debido a que:

- Facilitan la selección de información
- Acortan tiempos de lectura
- Simplifican búsquedas en textos
- Optimizan la creación de índices

La investigación indica, además, que contribuyen en tareas automatizadas: su utilización para presentar resultados de motores de búsquedas atrajo el interés de académicos desde principios de la década del 2000, convirtiéndose hoy en día en una funcionalidad básica de los principales buscadores de Internet.

Sin embargo, si bien esta tarea puede no resultar costosa para un ser humano, durante muchos años fue considerada como difícil de automatizar. Este tema es uno de los que investiga el campo de procesamiento de lenguaje natural, dedicado a facilitar la interacción entre las computadoras y los seres humanos. Entre las herramientas más importantes con las que cuenta este área se encuentra el modelado de tópicos, que revela los conceptos de los que trata un texto a base de un análisis estadístico.

3. Ejemplos

A continuación se presentan algunos ejemplos de aplicación de resúmenes automáticos, obtenidos a partir de una herramienta disponible en Internet:[1]

- Artículo de Wikipedia sobre sumarización automática, en español:[1]

  "Dos tipos particulares de sumarización encontrados a menudo en la literatura son extracción de frases principales, cuyo objetivo es seleccionar palabras o frases individuales para 'etiquetar' un documento, y sumarización de documentos, cuyo objetivo es seleccionar oraciones enteras para crear sumarios formados por párrafos cortos."

- Artículo de Wikipedia sobre sumarización automática, en inglés:[2]

  "Two particular types of summarization often addressed in the literature are keyphrase extraction, where the goal is to select individual words or phrases to 'tag' a document, and document summarization, where the goal is to select whole sentences to create a short paragraph summary."

- Publicación que presenta a TextRank:

  "In the following, we investigate and evaluate the application of TextRank to two natural language processing tasks involving ranking of text units: (1) A keyword extraction task, consisting of the selection of keyphrases representative for a given text; and (2) A sentence extraction task, consisting of the identification of the most 'important' sentences in a text, which can be used to build extractive summaries."

- *El Principito*:[3]

  "¿Es que no es cosa seria averiguar por qué las flores pierden el tiempo fabricando unas espinas que no les sirven para nada? ¿Es que no es

[1] `http://es.wikipedia.org/wiki/Sumarización_automática`

[2] `http://en.wikipedia.org/wiki/Automatic_summarization`

[3] Antoine de Saint-Exupéry. 1943. *El Principito.* (orig. *Le Petit Prince*).

importante la guerra de los corderos y las flores? ¿No es esto más serio e importante que las sumas de un señor gordo y colorado? Y si yo sé de una flor única en el mundo y que no existe en ninguna parte más que en mi planeta; si yo sé que un buen día un corderillo puede aniquilarla sin darse cuenta de ello, ¿es que esto no es importante? El principito enrojeció y después continuó: —Si alguien ama a una flor de la que sólo existe un ejemplar en millones y millones de estrellas, basta que las mire para ser dichoso."

4. Objetivos

El Trabajo Profesional consta de tres objetivos principales:

- Desarrollar el módulo para generar resúmenes automáticos usando un algoritmo conocido.
- Analizar, diseñar e implementar modificaciones para intentar mejorar el rendimiento del algoritmo seleccionado.
- Integrar las implementaciones a una herramienta de código abierto de procesamiento del lenguaje natural.

5. Carácteristicas del trabajo

5.1. Módulo

El módulo de creación de resúmenes automáticos está basado en TextRank, algoritmo propuesto por Rada Mihalcea y Paul Tarau. Su concepto es determinar las frases más significativas basándose en la estructura del texto, de la misma manera que el célebre PageRank de Google selecciona las páginas Web más importantes.

TextRank es un método no supervisado (dado que no requiere de información de otros documentos como entrenamiento para obtener una salida) y extractivo (pues el resultado final es una selección de oraciones del documento original). A su vez, tiene la ventaja de que puede ser utilizado sobre cualquier pieza sin importar su idioma, ya que simplemente analiza la relación entre oraciones.

5.2. Modificaciones y evaluación de performance

La segunda parte del Trabajo consiste en analizar otros enfoques al problema en cuestión para mejorar el algoritmo implementado.

Teniendo en cuenta que existen alternativas muy diferentes para realizar la tarea, sólo se considerarán aquéllas que se puedan adaptar a lo desarrollado en la etapa anterior.

Se evaluará cómo se comporta el enfoque propuesto a TextRank usando las técnicas de uso estándar:

- Se aplicarán las métricas de precisión (*precision*), exhaustividad (*recall*) y valor-F (*F-measure*) para obtener una medida de qué tan acertada es la selección de palabras clave.
- El conjunto de métricas ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) para comparar el resultado obtenido con otro conjunto de resúmenes.
- Evaluación humana de coherencia, dado que las herramientas desarrolladas para realizar automáticamente esta tarea devuelven resultados no concluyentes

5.3. Integración

El objetivo final del Trabajo Profesional es integrar la implementación lograda al entorno Gensim (ver *Tecnología*). Dicha herramienta permite detectar los principales temas tratados en una colección de documentos, a través del modelaje de tópicos. Estas nociones se obtienen a partir de estadísticas relacionadas con las palabras de la colección.

Se propone integrar el módulo basado en TextRank y las modificaciones, en caso de que arrojaran resultados concluyentes.

6. Tecnologías

6.1. Framework principal

El framework principal, al cual se integrará el desarrollo final, será Gensim.[2] Éste es una biblioteca escrita en Python para el modelaje de tópicos y la indexación de documentos que está diseñada para trabajar con conjuntos de textos de gran tamaño. Su uso se ha expandido tanto en el ámbito comercial como en el académico, apuntando especialmente al área de procesamiento del lenguaje natural y a la búsqueda y recuperación de la información.

Sus características principales son:

- Utiliza las extensiones científicas de Python NumPy[3] y SciPy,[4] y está optimizado a través Cython.[5]
- Todos sus algoritmos están diseñados para procesar entradas mucho mayores que la memoria RAM disponible.
- Procesamiento distribuido
- Interfaces intuitivas
- Documentación extensiva

Provee herramientas para:

- Análisis semántico latente
- Alocación de Dirichlet latente
- Proceso jerárquico de Dirichlet
- Proyecciones aleatorias
- Deep learning

Además, Gensim es una herramienta de código libre, un modelo de desarrollo caracterizado por la distribución de software y su código fuente de manera libre, ilimitada y gratuita tanto a usuarios como a desarrolladores.

6.2. Herramientas y otras tecnologías

Categoría	**Herramientas**
Lenguajes de programación	• Python • JavaScript
Frameworks gráficos	• Gexf-js[6] • Highcharts[7]
Entornos de desarrollo	• IPython[8] • Sublime Text[9]
Control de versiones	• Git[10]
Administración y control del proyecto	• Trello[11] • Goole Docs[12]

7. Alcance

El alcance del presente Trabajo Profesional comprende:

- Desarrollo del módulo de generación de resúmenes automáticos.
- Desarrollo e implementación de modificaciones con su correspondiente evaluación de desempeño en base a métricas preestablecidas.
- Interfaz web para la utilización del módulo.
- Integración del módulo a Gensim.

8. Plan de trabajo

8.1. Equipo de trabajo

El equipo de trabajo estará conformador por:

- Tutores: Lic. Rosa Wachenchauzer, Lic. Luis Argerich.
- Desarrolladores: Federico Barrios y Federico López

8.2. Metodología

Para la ejecución del proyecto se usará una metodología ágil basada en SCRUM. La misma consistirá en definir una serie de iteraciones con fechas pautadas de reuniones de avance y entregas.

Al inicio de cada iteración se hará una priorización de los requerimientos pendientes de desarrollo. Posteriormente se procederá a su implementación, y para finalizar cada iteración se hará una presentación de los entregables pautados.

Las reuniones, entregas, presentaciones y la priorización de los requerimientos para cada iteración se hará en conjunto con los tutores del Trabajo.

8.3. Estimación

Se muestra a continuación el listado de tareas necesarias para alcanzar los objetivos descriptos anteriormente. Todos los esfuerzos están expresados en horas.

# It.	Descripción	Esf.	Rec.	Tot.
–	Propuesta	15	2	30
1	Configuración del entorno	15	1	15
	Interfaz web básica	10	2	20
2	Módulo principal: investigación y análisis	30	2	60
	Módulo principal: diseño	30	2	60
	Módulo principal: implementación	30	2	60
	Integración con interfaz web	15	1	15
3	Búsqueda de bases de datos	15	1	15
	Uso de módulo de ROUGE	30	2	60
	Integración con Gephi: escribir grafo	25	1	25
	Integración con Gephi: transformar grafo	15	1	15
4	Modificaciones: análisis y diseño	15	2	30
	Modificaciones: implementación	25	2	50
5	Análisis de datos e informes de resultados	20	2	40
	Integración con Highcharts	15	1	15
6	Integración con Gensim: análisis y diseño	20	2	40
	Integración con Gensim: implementación	20	2	40
	Integración con Gensim: documentación	15	2	30
–	Reuniones	15	2	30
–	Presentación	15	2	30

Además, se debe tener en cuenta el tiempo dedicado a administración del proyecto, estimado en un 15 % del tiempo de desarrollo, resultando en:

Descripción	Esfuerzo
Iteración 1	35
Iteración 2	195
Iteración 3	115
Iteración 4	80
Iteración 5	55
Iteración 6	110
Otros	90
Administración	102
Esfuerzo total	782

8.4. Cronograma de entregables

A continuación se enumera un cronograma tentativo de entregables. Queda sujeto a modificaciones por parte de los tutores, en base a la ejecución del proyecto:

# It.	Entregables	Hitos
1	• Interfaz web básica para utilización del módulo.	• Configuración de entorno de desarrollo terminada.
2	• Módulo de resúmenes automáticos utilizando TextRank.	• Implementación del módulo terminada. • Integración con interfaz web.
3	• Módulo ROUGE para métricas. • Gráfico de relación de conceptos a través de interfaz web.	• Integración con Gexf-js terminada.
4	• Versiones modificadas del módulo de resúmenes. • Datos recolectados.	• Implementación de modificaciones sobre el módulo de resúmenes terminado. • Datos obtenidos luego de realizar pruebas sobre las modificaciones planteadas.
5	• Informe de análisis de datos recolectados. • Gráficos de resultados de la métricas integrados en la interfaz web.	• Integración con Highcharts terminada.
6	• Versión de Gensim con módulo de resúmenes automáticos integrado.	• Integración con Gensim terminada.

9. Bibliografía

- Borko, Bernier. 1975. *Abstracting Concepts and Methods.*
- McKeown, Radev. 1999. *Generating Summaries of Multiple News Articles.*
- Mani. 2001. *Summarization Evaluation: An Overview.*
- Roussinov, Chen. 2001. *Information Navigation on the Web by Clustering and Summarizing Query Results.*
- Mani, Klein, House, Hirschman, Firmin, Sundheim. 2002. *SUMMAC: a Text Summarization Evaluation.*
- Hassel. 2004. *Evaluation of Automatic Text Summarization, a Practical Implementation.*
- Lin. 2004. *ROUGE: A Package for Automatic Evaluation of Summaries.*
- Mihalcea, Tarau. 2004. *TextRank: Bringing Order into Texts.*
- Wang, Lawrence. 2004. *Methods and Systems for Generating Textual Information.* United States Patent 7310633.
- Das, Martins. 2007. *A Survey on Automatic Text Summarization.*
- Rose, Orr, Kantamneni. 2007. *Summary Attributes and Perceived Search Quality.*

Referencias

[1] *Tools for noobs*. Herramienta de resúmenes en línea.
`http://www.tools4noobs.com/summarize/`

[2] *Gensim*. Biblioteca de modelado de tópicos.
`http://radimrehurek.com/gensim/`

[3] *NumPy*. Paquete de Python para procesamiento numérico.
`http://www.numpy.org/`

[4] *SciPy*. Paquete de Python para computación científica.
`http://www.scipy.org/`

[5] *Cython*. Extensiones en lenguaje C para Python.
`http://cython.org/`

[6] *Gexf-js*. Extensión de Gephi para JavaScript.
`https://github.com/raphv/gexf-js`

[7] *Highcharts*. Librería de JavaScript para realizar gráficos interactivos.
`http://www.highcharts.com/`

[8] *IPython*. Consola interactiva para Python.
`http://ipython.org/`

[9] *Sublime Text*. Editor de textos para programación.
`http://www.sublimetext.com/`

[10] *Git*. Herramienta para el control de versiones.
`http://git-scm.com/`

[11] *Trello*. Herramienta de gestión para trabajo en grupos.
`http://trello.com/`

[12] *Google Docs*. Herramienta para trabajo sobre documentos en línea
`https://docs.google.com/`

10. Anexo

10.1. Listado de asignaturas aprobadas

10.1.1. Federico Barrios

Código	Materia	Calif.	Fecha
62.01	Física I A	6	06-07-2010
75.40	Algoritmos y Programación I	7	14-07-2010
61.03	Análisis Matemático II A	4	27-07-2010
62.03	Física II A	6	16-12-2010
61.08	Algebra II A	4	09-02-2011
75.41	Algoritmos y Programación II	6	14-02-2011
63.01	Química	6	03-03-2011
75.12	Análisis Numérico	6	27-06-2011
62.15	Física III D	8	20-07-2011
66.70	Estructura del Computador	7	26-07-2011
66.02	Laboratorio	6	04-08-2011
75.07	Algoritmos y Programación III	8	20-12-2011
75.42	Taller de Programación I	8	21-12-2011
61.10	Análisis Matemático III A	6	09-02-2012
61.09	Probabilidad y Estadística B	7	23-02-2012
61.07	Matemática Discreta	5	18-07-2012
75.06	Organización de Datos	9	30-07-2012
75.08	Sistemas Operativos	10	13-12-2012
66.06	Análisis de Circuitos	5	26-12-2012
75.29	Teoría de Algoritmos I	10	04-02-2013
66.74	Señales y Sistemas	7	15-02-2013
66.20	Organización de Computadoras	7	18-02-2013
75.43	Introducción a los Sistemas Distribuidos	4	05-07-2013
71.14	Modelos y Optimización I	9	17-07-2013
75.59	Técnicas de Programación Concurrente I	7	30-07-2013
75.26	Simulación	7	31-07-2013
75.28	Base de Datos	9	07-08-2013
75.09	Análisis de la Información	5	12-08-2013
71.40	Leg. y Ej. Prof. de la Ing. en Informática	7	13-12-2013
75.10	Técnicas de Diseño	8	12-02-2014
75.52	Taller de Programación II	10	14-02-2014
71.12	Estructura de las Organizaciones	6	19-02-2014
78.01	Idioma Inglés	9	28-02-2014
75.67	Sist. Autom. de Diag. y Detec. de Fallas I	7	11-08-2014
75.50	Introducción a los Sistemas Inteligentes	8	11-08-2014
75.65	Manuf. Integ. por Computadora (CIM) I	4	14-08-2014

10.1.2. Federico López

Código	Materia	Calif.	Fecha
75.40	Algoritmos y Programación I	7	30-06-2010
61.03	Análisis Matemático II A	8	06-07-2010
62.01	Física I A	6	13-07-2010
62.03	Física II A	6	16-12-2010
61.08	Algebra II A	9	22-12-2010
75.41	Algoritmos y Programación II	8	14-02-2011
63.01	Química	7	15-02-2011
75.07	Algoritmos y Programación III	8	28-06-2011
66.70	Estructura del Computador	7	29-06-2011
75.12	Análisis Numérico I	9	13-07-2011
62.15	Física III D	8	20-07-2011
66.02	Laboratorio	9	25-07-2011
75.42	Taller de Programación I	9	14-12-2011
66.20	Organización de Computadoras	7	19-12-2011
61.07	Matemática Discreta	6	21-12-2011
61.10	Análisis Matemático III A	6	09-02-2012
61.09	Probabilidad y Estadística B	8	23-02-2012
78.01	Idioma Inglés	8	02-07-2012
75.06	Organización de Datos	10	02-07-2012
71.18	Estructura Económica Argentina	8	13-07-2012
71.14	Modelos y Optimización I	7	16-07-2012
71.12	Estructura de las Organizaciones	5	18-07-2012
75.08	Sistemas Operativos	10	13-12-2012
75.29	Teoría de Algoritmos I	9	04-02-2013
75.43	Introducción a los Sistemas Distribuidos	8	05-07-2013
75.10	Técnicas de Diseño	9	15-07-2013
75.59	Técnicas de Programación Concurrente I	8	19-07-2013
75.15	Base de Datos	4	24-07-2013
75.26	Simulación	8	31-07-2013
71.40	Leg. y ej. Prof. de la Ing. en Informática	7	13-12-2013
75.45	Taller de Desarrollo de Proyectos II	8	19-12-2013
75.52	Taller de Programación II	10	21-02-2014
75.44	Adm. y Control de Proy. Informáticos I	8	25-02-2014
75.47	Taller de Desarrollo de Proyectos II	8	30-06-2014
75.46	Adm. y Control de Proy. Informáticos II	4	08-07-2014
71.13	Información en las Organizaciones	4	15-07-2014
66.69	Criptografía y Seguridad Informática	9	04-08-2014