Resúmenes automáticos con variantes de TextRank

Federico Barrios[1], Federico López[1], Luis Argerich[1], Rosita Wachenchauzer[12]
[1]Facultad de Ingeniería, Universidad de Buenos Aires [2]Universidad Nacional de Tres de Febrero

TextRank

1. Texto original

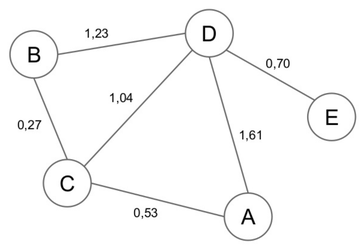
Autorizaron a un camión sin chofer a circular por las rutas de Estad
Daimler anunció que su prototipo Freightliner Inspiration se cc
autónomo en recibir la autorización para recorrer las rutas y autc
junto al tránsito actual. La automotriz alemana dijo que el vehícı
prueba de forma previa en Alemania en un circuito urbano cerrado
de transporte ahorren combustible y puedan tener una flota más se
Para obtener la licencia que le permite recorrer los caminos de
forma autónoma en condiciones reales, Daimler dijo que el Fre
cumplir más de 16 mil kilómetros de prueba. Como antecedente, e
autorización para su vehículo autónomo.
La autonomía del camión de Daimler es parcial, ya que sólo se ha
está en una ruta o autopista, manteniendo una distancia prudencia
sin adelantarse a otros conductores más lentos.
Ante un obstáculo o dificultad que el sistema no pueda resolver, el

2. Preprocesamiento

1. autoriz camion chof circul rut unid
2. daiml anunc prototip freightlin inspiration coı actual
3. automotriz aleman vehicul hab puest prueb ahorr combust pued flot mas segur trayect
4. obten licenci permit recorr camin nev unid fc mas mil kilometr prueb
5. antecedent googl logr autoriz vehicul autonc

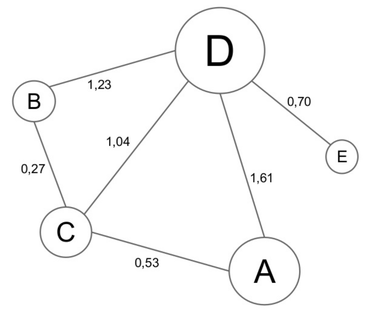
- Separación en oraciones
- Filtro de palabras
- Lematización

3. Armado del grafo



- Uso de la función de similitud
- Se le asigna un puntaje a cada par de oraciones

4. PageRank



Se aplica una variación del algoritmo para obtener las oraciones más representativas del texto.

5. Texto resumido

Autorizaron a un camión sin chofer a circular por las rutas de Estad
Daimler anunció que su prototipo Freightliner Inspiration se cc
autónomo en recibir la autorización para recorrer las rutas y auto
junto al tránsito actual. La automotriz alemana dijo que el vehícı
prueba de forma previa en Alemania en un circuito urbano cerrado
de transporte ahorren combustible y puedan tener una flota más se
Para obtener la licencia que le permite recorrer los caminos de
forma autónoma en condiciones reales, Daimler dijo que el Fre
cumplir más de 16 mil kilómetros de prueba. Como antecedente, e
autorización para su vehículo autónomo.
La autonomía del camión de Daimler es parcial, ya que sólo se ha
está en una ruta o autopista, manteniendo una distancia prudencia
sin adelantarse a otros conductores más lentos.
Ante un obstáculo o dificultad que el sistema no pueda resolver, el
una alerta al chofer y detendrá lentamente la marcha si no recil

Propuestas

Funciones de similitud

La función de similitud de la versión original del algoritmo de TextRank cuenta las palabras en común entre cada par de oraciones.
Las modificaciones que proponemos reemplazan esta función:

$$Similitud(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

Subcadena común más larga

Devuelve la longitud de la subcadena más larga presente en las dos oraciones.

Similitud coseno

Usando el modelo TF-IDF representa cada oración como un vector y computa el resultado como el coseno del ángulo que comprenden.

BM25 y BM25+

Implementan variaciones de las famosas funciones para recuperación de información; que no permiten que haya resultados que tomen valores negativos.

Evaluación y resultados

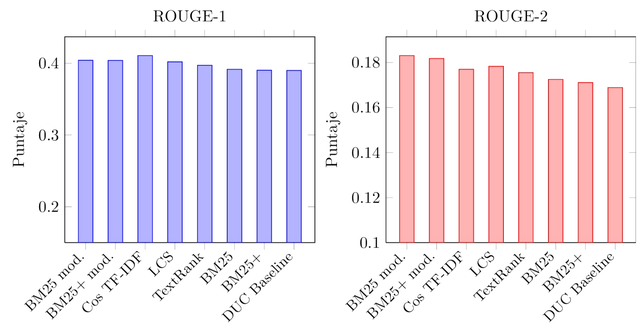
Para probar el funcionamiento del algoritmo se decidio usar la base de datos de la tarea de generacion de resumenes automáticos de la conferencia DUC (Document Understanding Conference) del año 2002.

Fue usada la misma base de datos de prueba que la utilizada en la publicación que presentó el algoritmo de TextRank.
Se llevó a cabo la evaluación usando, además, las mismas configuraciones de pruebas que en DUC: la versión 1.5.5 del paquete de métricas ROUGE, calculando sólo las mediciones ROUGE-1, ROUGE-2 y ROUGE-SU4 en un intervalo de confianza del 95%.

Se obtuvo un único puntaje promediando los tres valores.

Los mejores resultados se obtuvieron usando BM25 y BM25+. El incremento más alto se logró al reemplazar los valores negativos por la constante ε = 0,25, arrojando una mejora total de 2,92% para BM25 y 2,60% para BM25+.

Los tiempos de ejecución también se superaron. Se pudieron procesar los 567 documentos de la base de datos de DUC2002 utilizando un 84% del tiempo requerido por la versión original.



Método	ROUGE-1	ROUGE-2	ROUGE-SU4	Mejora
BM25 ($\varepsilon = 0{,}25$)	0,4042	0,1831	0,2018	2,92 %
BM25+ ($\varepsilon = 0{,}25$)	0,404	0,1818	0,2008	2,60 %
Similitud coseno	0,4108	0,177	0,1984	2,54 %
BM25+ (IDF = $\log(N/N_i)$)	0,4022	0,1805	0,1997	2,05 %
BM25 (IDF = $\log(N/N_i)$)	0,4012	0,1808	0,1998	1,97 %
Subcadena común	0,402	0,1783	0,1971	1,40 %
BM25+ ($\varepsilon = 0$)	0,3992	0,1803	0,1976	1,36 %
BM25 ($\varepsilon = 0$)	0,3991	0,1778	0,1966	0,89 %
TextRank	**0,3983**	**0,1762**	**0,1948**	–
BM25	0,3916	0,1725	0,1906	-1,57 %
BM25+	0,3903	0,1711	0,1894	-2,07 %
Referencia DUC	0,39	0,1689	0,186	-2,84 %

El resultado de la métrica de similitud coseno también fue satisfactoria, presentando una mejora de 2,54% por sobre el método original.
A su vez, los métodos de secuencias de máxima longitud también mostraron una mejora considerable: cerca del 1,40 % por sobre TextRank.

Implementación

El desarrollo fue hecho completamente en Python, en el marco de un proyecto de código libre. Se publicó como un módulo que se puede descargar desde los repositorios gratuitos y públicos de PyPI.

Las funcionalidades fueron integradas al proyecto Gensim de procesamiento de lenguaje natural y modelaje de tópicos. Nuestra contribución forma parte del entorno desde su versión 0.12.0.

Conclusión

En este trabajo se analizaron variantes al algoritmo de TextRank. A partir del mismo se propusieron e implementaron optimizaciones cuyos resultados fueron significativos: se obtuvo una mejoría del 2,92% por sobre el método original.
Este número es notable si se tiene en cuenta que TextRank por sí solo tiene un desempeño de 2,84% por sobre el estándar de comparación (baseline).
En base a estos resultados proponemos la utilización de BM25 o BM25+ junto con TextRank como herramienta para la generación de resúmenes automáticos.
Los resultados obtenidos demuestran mejoras a la versión original del algoritmo con ganancias extra de tiempo de procesamiento.

Referencias

- Gusfield, D.: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, New York, NY, USA (1997)
- Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain (2004)
- Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004. Association for Computational Linguistics, Barcelona, Spain (July 2004)
- Ouyang, Y., Li, W., Wei, F., Lu, Q.: Learning similarity functions in graph-based document summarization. In: Li, W., Aliod, D.M. (eds.) ICCPOL. Lecture Notes in Computer Science, vol. 5459
- Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference. Brisbane, Australia (1998)
- Rehurek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta (May 2010)
- Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA
- Singhal, A.: Modern information retrieval: A brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24(4), 35–43 (2001)