Clasificación Automática de Textos Periodísticos usando Random Forest.

Juan Salinas y Javier Izetta

Facultad de Ingeniería,
Universidad Nacional de Jujuy, Argentina
{juangsalinas90,javierizetta}@gmail.com
http://www.fi.unju.edu.ar

Resumen En los últimos años el periodismo regional, al igual que en todas partes del mundo, pasó de su formato clásico de publicación al electrónico. Así las *webs* de noticias regionales también se ven obligadas a evolucionar y mejorar sus prestaciones a través de una mejor organización y categorización previa de toda la información disponible para el lector. En este trabajo se propone abordar la clasificación automática de textos periodísticos digitales a través del *Aprendizaje Automatizado*. Se presentan dos clasificadores automáticos de textos periodísticos extraídos de páginas webs de noticias del NOA basados en *Random Forest* y se proponen dos técnicas para la reducción de dimensionalidad del espacio de características. Estos clasificadores fueron evaluados con distintas colecciones de noticias demostrando un buen desempeño.

Keywords: categorización automática de textos, random forest, stemming.

1. Introduction

La *Categorización Automática de Texto* (CAT) tiene por objetivo hallar una función óptima de clasificación de documentos a partir de atributos constituidos por palabras que describen cada categoría específica [1]. En la actualidad la CAT cuenta con un amplio campo de aplicación, entre los más destacables podemos encontrar: filtrado y organización del contenido de páginas web [2], detección de correos electrónicos *Spam* [3], filtrado de noticias [4], detección de plagios e identificación de autores [5] y análisis de opinión [6].

En este trabajo se aplica el enfoque de "bolsas de palabras"que es el más utilizado en CAT. En este enfoque se consideran las palabras individuales y su información léxica. Otro aspecto a tener en cuenta es que un texto podría pertenecer a varias, una o todas las categorías especificadas [7]. El presente trabajo se enfoca en la clasificación de única etiqueta, es decir, cada documento de la colección está asociado a solo una clase o categoría.

Actualmente la CAT puede tratarse como un problema de Çlasificación Supervisada". Los métodos computacionales desarrollados para tal fin forman parte de lo que se conoce como *Aprendizaje Automatizado* (AA). Diversos algoritmos

de AA, fueron utilizados para dar solución a la problemática de CAT. Particularmente varios trabajos proponen utilizar árboles de desicíon para la clasificación de textos [8] [9] [10]. En este trabajo se propone utilizar el método *Random Forest* que utiliza un conjunto (o bosque) de árboles de decisión para predecir un valor de salida [11].

2. Clasificación de textos como un problema de AA

Desde el punto de vista del AA, la CAT se puede definir formalmente a partir de la siguiente situación:

Se tiene un conjunto de documentos $D = \{\mathbf{d}_1, ..., \mathbf{d}_i, ..., \mathbf{d}_n\}$, donde cada documento $\mathbf{d}_i$ se expresa como $\mathbf{d}_i = \{t_1, ..., t_j, ..., t_m\} \in \Re^m$. $\mathbf{d}_i$ es un vector en un espacio m-dimensional. $\Re^m$ se denomina *feature space*, espacio de variables o espacio de características, indistintamente. Cada documento tiene asignada una clase o etiqueta real $l(\mathbf{d}_i)$ conocida de antemano. Para un problema con c clases, $l(\mathbf{d}_i)$ puede tomar c valores discretos distintos.

Entonces se tiene por objetivo hallar una función clasificadora $\hat{f}$ tal que para cada documento $\mathbf{d}_i$ se tiene $\hat{f}(\mathbf{d}_i) = l(\mathbf{d}_i)$, donde $\hat{f}(\mathbf{d}_i)$ es la etiqueta o clase asignada por $\hat{f}$ al documento $\mathbf{d}_i$. Además se pretende que $\hat{f}$ pueda clasificar correctamente nuevos documentos no contemplados anteriormente.

3. Random Forest

Se trata de un método desarrollado por Leo Breiman y Adele Cutler en 2001 [11]. Este método utiliza un conjunto (o bosque) de árboles de decisión para predecir un valor de salida. Para la clasificación de un nuevo ejemplo, cada árbol en el ensamble produce un voto, es decir, toma el ejemplo como entrada y produce una salida indicando a que clase pertenece. La decisión final del ensamble se realiza a partir de la clase con mayor cantidad de votos.

4. Construcción de los clasificadores propuestos

La construcción de los clasificadores se abordó a través de dos etapas claramente delimitadas. La primera etapa, a la que suele llamarse etapa de *entrenamiento*, se inicia con la recopilación manual de una serie de textos periodísticos de diarios digitales del noroeste argentino extraídos de la *web* (documentos de entrenamiento). Esta colección se procesa para lograr una representación adecuada para el entrenamiento de los clasificadores. Luego se realiza una reducción del conjunto de características generado por la colección (reducción de dimensionalidad) con el fin de mejorar el rendimiento durante el aprendizaje de los clasificadores. Una vez que éstos fueron entrenados, tiene lugar la segunda etapa, llamada etapa de *prueba*, que consiste en la evaluación del desempeño de los clasificadores con nuevos documentos no considerados durante la etapa anterior. En la Figura 1 es posible observar con más detalle los pasos para la construcción de los clasificadores propuestos.

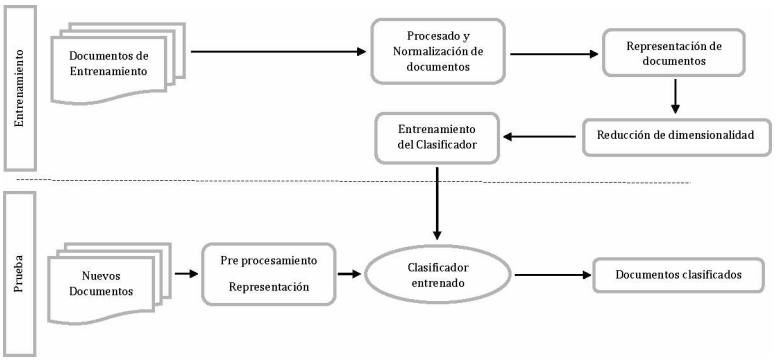


Figura 1: Esquema básico de construcción de los clasificadores.

4.1. Preprocesado y Normalización de documentos

En este paso se busca definir *tokens* o términos, para ello, en los clasificadores propuestos un *token* queda conformado por aquella cadena de caracteres delimitados por espacios en blanco. Además, en esta instancia, se descartan aquellos caracteres tales como símbolos y números ya que no aportan información alguna para la clasificación. Y también aquellos *tokens* o términos identificados como "palabras de parada"(*stopwords*), este conjunto de palabras está constituido por preposiciones, artículos, pronombres, conjunciones, contracciones y ciertos verbos y adverbios. En este trabajo se usó para tal fin el conjunto de *stopwords* para el español definido por el proyecto *Snowball* disponible en [12].

4.2. Representación de documentos

Este paso consiste en la transformación de los documentos en una representación adecuada para que el algoritmo de aprendizaje sea capaz de procesarlos. En este trabajo se propone utilizar para la representación de los textos periodísticos el modelo vectorial propuesto por Salton [13] y el esquema de pesado TF-IDF (*Term Frequency - Inverse Document Frequency*)[14]. En CAT este modelo de representación es uno de los más utilizados debido a sus altas prestaciones cuando se combina a esquemas de pesado y normalización de longitud de documentos [14] [15]. En el modelo vectorial los documentos son formalmente representados a través de vectores cuya dimensión estará dada por la cantidad de términos del vocabulario generado por la colección de documentos.

Cada componente del vector representa la importancia que tiene ese término en el documento y en la colección. Salton propone en [15] calcular los pesos mediante la combinación de la frecuencia relativa de los términos (TF) con la frecuencia inversa en los documentos (IDF), de manera que se tiene:

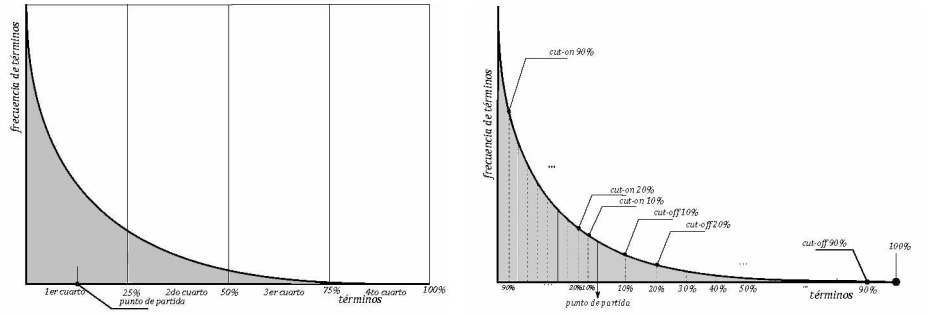
$$TF - IDF\left(t_j, \mathbf{d}_i\right) = f_{ij} \cdot log\left(\frac{N}{df\left(t_j\right)}\right) \quad (1)$$

Donde $df\left(t_j\right)$ es el número de documentos en los que el término t_j aparece, f_{ij} es la frecuencia del término t_j en el documento $\mathbf{d}_i$ y N es el número de documentos en la colección.

4.3. Reducción de dimensionalidad

Es importante aclarar que los dos pasos anteriores (preprocesado y representación documentos) se realizaron de la misma manera para la construcción de los dos clasificadores propuestos. A partir de este paso (reducción de dimensionalidad) en este trabajo se propone dos variantes para la reducción de dimensionalidad. Una basada en la selección de un subconjunto del conjunto de términos originales, alternativa a la que denominaremos CATST. Y otra basada en la transformación del conjunto de términos originales a la que llamaremos CATLT. A continuación se exponen ambas alternativas:

- **Propuesta CATST**
  La ley del mínimo esfuerzo de Zipf [16], comprueba que en una colección de documentos coexisten términos muy pocos frecuentes y específicos para determinados documentos, junto con aquellos términos muy frecuentes que representan la colección de documentos en general. En base a esta ley Luhn [17], afirma que existe un rango de términos que son relevantes para un determinado documento, cuando la tarea es la recuperación de documentos a través de una consulta. Esta misma idea se puede aplicar a CAT, es decir, es posible hallar un rango de términos relevantes para cada categoría. En un problema de clasificación de texto lo que se pretende es encontrar términos que tengan el mayor poder de discriminación entre las categorías. Esto implica centrarnos en términos que sean característicos de cada grupo de documentos pertenecientes a cada categoría, es decir, términos de frecuencia media que no son exclusivamente específicos de uno o muy pocos documentos ni absolutamente generales a toda la colección de documentos. Para encontrar este rango de términos proponemos realizar los siguientes pasos:
  1. Particionar el conjunto de términos originales ordenados de manera decreciente según su frecuencia, en 4 partes iguales. Tomar como "punto de partida"para la determinación del rango, aquel término que se ubica en la parte media del primer cuarto, tal como se puede apreciar en la Figura:2a.
  2. Tomar los términos correspondientes al 10, 20, 30 y sucesivamente hasta un 90 por porciento hacia la izquierda (cut-on) y derecha (cut-off) de este "punto de partida"para formar nueve rangos candidatos, tal como se aprecia en la Figura: 2b.



(a) Punto de partida (b) Determinación de rangos de términos

Figura 2: Propuesta CATST.

3. Entrenar un clasificador utilizando cada rango candidato.
4. Evaluar y seleccionar el rango que mejor desempeño obtenga.

- **Propuesta CATLT**

  Otra alternativa que se propone en este trabajo, es utilizar una técnica de reducción de dimensionalidad basada en la transformación del conjunto de términos originales a través del concepto de *lematización* o por su terminología en inglés, *stemming*. Los algoritmos de lematización de términos son capaces de extraer prefijos y sufijos de palabras que son literalmente diferentes, pero que tienen una raíz en común y que pueden ser consideradas como un mismo término. Cada palabra es "truncada" a su lema o raíz equivalente. Para tal fin en este trabajo se utilizó una adaptación al español del algoritmo de *Porter* [18] [19]. A pesar de que al transformar el espacio de términos en un espacio de raíces este conjunto se reduce notoriamente, se debería considerar solo aquellas raíces que tengan mayor poder de discriminación entre las categorías. A diferencia de los términos originales, cuando se trabaja con raíces, estas últimas tienen mayor poder de discriminación cuando su frecuencia es alta. Para ello proponemos encontrar un rango de raíces de la siguiente manera:

  1. Ordenar las raíces en forma decreciente según su frecuencia de aparición.
  2. Tomar las raíces correspondientes al 10, 20, 30 y sucesivamente hasta un 90 por porciento a partir de aquella raíz cuya frecuencia de aparición sea máxima, para formar nueve rangos candidatos, como se muestra en la Figura:3.
  3. Entrenar un clasificador utilizando cada rango candidato.
  4. Evaluar y seleccionar el rango que mejor desempeño obtenga.

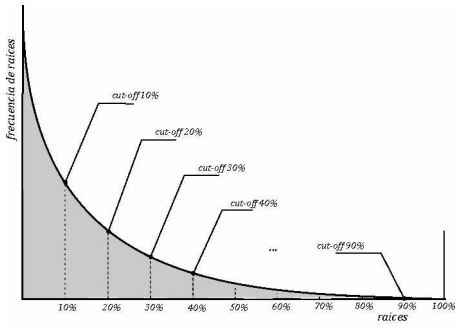


Figura 3: Propuesta CATLT.

4.4. Entrenamiento y prueba de los clasificadores

El entrenamiento y prueba de los clasificadores se llevó a cabo mediante un proceso que realiza dos bucles anidados (Figura:4).

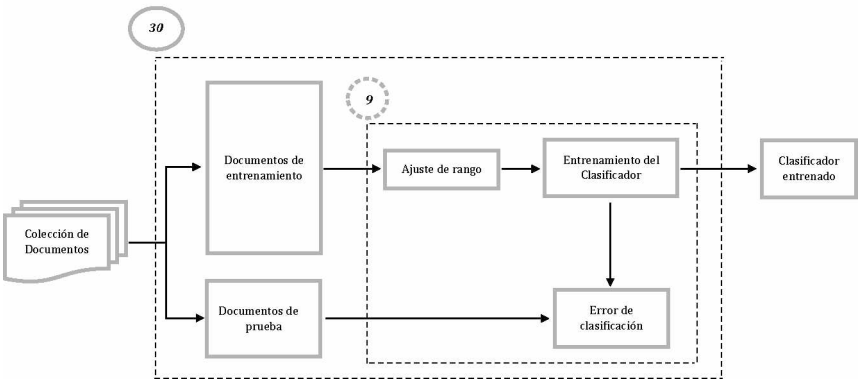


Figura 4: Proceso de entrenamiento y prueba de los clasificadores.

- En el *bucle externo* (recuadro externo en líneas punteadas) se realiza 30 veces la partición de la colección de documentos en un subconjunto de documentos para entrenamiento, seleccionando aleatoriamente un 70 % del total de documentos. El subconjunto de documentos restante (30 % del total de documentos), es utilizado para prueba. De esta manera es posible obtener una mejor estimación del desempeño de los clasificadores propuestos.

- En el *bucle interno* (recuadro interno en líneas punteadas) se entrena un clasificador por cada rango candidato, es decir nueve rangos utilizando ensambles de 100, 500 y 1000 árboles de decisión. Se evalúa el desempeño de cada rango usando el subconjunto de documentos de prueba generado por el *bucle externo*.
- Al finalizar las iteraciones de los dos bucles se selecciona el clasificador entrenado a partir del rango que obtuvo la tasa de error más baja.

5. Experimentación

5.1. Colecciones de noticias

El entrenamiento y evaluación de los clasificadores propuestos se realizó a partir de colecciones de noticias confeccionadas manualmente. El motivo de esta decisión, se debe a la falta de disponibilidad de algún repositorio que contenga alguna con categorías explicitas y en idioma español. Además utilizando los textos originales extraídos de portales de noticias del NOA, se pretende que los clasificadores "aprendan" el estilo de redacción de la zona. Las noticias se obtuvieron a partir de los periódicos digitales más leídos de la región noreste de Argentin [20]. Cada noticia seleccionada aleatoriamente corresponde a un período comprendido entre Octubre de 2015 y Marzo de 2016. Se crearon dos colecciones de documentos, la Tabla 1 muestra los detalles de cada colección.

Tabla 1: Colecciones de noticias: N= cantidad total de noticias, T= cantidad total de términos y C= categorías.

Nombre	*N*	*T*	*C*			
C2PD	200	9084	Policial	Deportes		
C3PES	300	13696	Política	Economía	Salud	

La creación de las colecciones tiene la finalidad de evaluar el desempeño de cada clasificador propuesto en diferentes situaciones. Por un lado, el clasificador puede enfrentarse a un problema de clasificación binaria o multiclase. Por otro lado, las categorías podrían tener muchos términos en común, adicionando complejidad a la colección.

5.2. Resultados y evaluación

- Colección C2PD:
  En la Tabla 2 se exponen los resultados obtenidos a partir de las experimentaciones realizadas sobre la colección C2PD, en particular, se muestra para

Tabla 2: Resultados de los clasificadores con el mejor rango de términos o raíces encontrado sobre la colección C2PD.

Propuesta	*CATST*	*CATLT*
100 árboles		
Cantidad de Características	8733	5202
Error medio de clasificación	0.06825556	0.03532778
Cantidad de caraterísticas en el mejor rango	7860	2601
Tiempo total de ejecución (segundos)	2610.12	561.24
500 árboles		
Cantidad de Características	8733	5202
Error medio de clasificación	0.03679667	0.02134556
Cantidad de caraterísticas en el mejor rango	7859	4681
Tiempo total de ejecución (segundos)	9492.33	2798.16
1000 árboles		
Cantidad de Características	8733	5202
Error medio de clasificación	0.03241778	0.01782056
Cantidad de caraterísticas en el mejor rango	7859	4681
Tiempo total de ejecución (segundos)	16741.23	5631.58

cada ensamble (100, 500 y 1000 árboles de decisión) el error medio de clasificación de 30 corridas de cada clasificador con el mejor rango de términos o raíces encontrado.

En primer lugar se puede observar que la propuesta CATLT al aplicar lematización de términos trabaja con una cantidad de características considerablemente menor que CATST. En segundo lugar al analizar el desempeño de los clasificadores (error medio de clasificación), se puede observar que CATLT (propuesta basada en lematización), obtiene una tasa de error menor a CATST (propuesta basada en la selección de un subconjunto de términos originales). La razón es que este último incluye en el mejor rango encontrado algunos términos con poco poder de discriminación entre las clases. Esto se debe a la dificultad de encontrar un rango que solo contenga términos altamente discriminativos. Para ello, lo que se busca son los términos con una frecuencia de aparición media, ya que éstos son los más informativos para cada clase. Aún así, no todos estos términos van a aportar buena información, llevando en algunos casos a un entrenamiento menos eficaz. Por el contrario en CATLT al trabajar con raíces en vez de términos, el proceso de ajuste del rango es más sencillo ya que solo se debe descartar las raíces con menor frecuencia de aparición. Este proceso lleva a encontrar un rango de raíces con un alto poder de discriminación entre las clases, favoreciendo el entrenamiento del clasificador. Por otro lado es posible observar que con 1000 árboles de decisón se alcanza un mejor desempeño de los clasificadores. En la Figura 5, para el ensamble de 1000 árboles de decisión, se muestran los errores de clasificación obtenidos de 30 corridas en un diagrama de cajas,

que corresponden a los mejores rangos encontrados por cada clasificador. Se puede observar en la gráfica que CATLT produce los errores más bajos y esta diferencia es significativa.

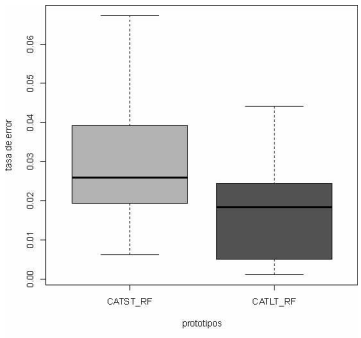


Figura 5: Errores de clasificación para el mejor rango candidato del ensamble de 1000 árboles de decisión.

También se puede observar que el tiempo se ve afectado por la cantidad de características a manipular, llevando menor tiempo de entrenamiento el clasificador CATLT (Tiempo total de entrenamiento, Tabla 2).
Por último las experimentaciones realizadas en la colección de tres categorías (C3PES) ambos clasificadores muestran comportamientos muy similares a las experiencias realizadas en las colecciones de dos categorías (Tabla 3). En términos generales, se pudo comprobar que los clasificadores propuestos son robustos, ya que el desempeño de los mismos se mantiene en las distintas colecciones.

6. Conclusiones y trabajos futuros

En este trabajo se presentó dos clasificadores automáticos de textos periodísticos del noroeste argentino. Los clasificadores desarrollados implementan dos técnicas propuestas para la reducción de dimensionalidad del espacio de características, denominadas CATST (basada en la selección de un sub conjunto de características) y CATLT (basada en lematización).

Después de demostrar el buen desempeño de ambos clasificadores en las diferentes colecciones creadas, nuestros resultados sugieren que:

- La propuesta CATLT produce tasas de errores más bajas que la propuesta CATST en todas las experimentaciones realizadas. Dado que el proceso de ajuste del rango de CATLT, al trabajar con raíces resulta más sencillo que

Tabla 3: Resultados de los clasificadores con el mejor rango de términos y raíces encontrado sobre la colección C3PES.

Propuesta	*CATST*	*CATLT*
100 árboles		
Cantidad de Características	13264	6961
Error medio de clasificación	0.2004667	0.1059481
Cantidad de caraterísticas en el mejor rango	11938	1392
500 árboles		
Cantidad de Características	13264	6961
Error medio de clasificación	0.147003	0.07458296
Cantidad de caraterísticas en el mejor rango	11938	6264
1000 árboles		
Cantidad de Características	13264	6961
Error medio de clasificación	0.1360896	0.06907333
Cantidad de caraterísticas en el mejor rango	11938	6264

en CATST. Esto se debe a que una raíz con frecuencia alta implica que aparece muchas veces en documentos pertenecientes a una determinada clase de la colección, siendo esa raíz representativa para esa clase dentro de la colección. Por el contrario en CATST, encontrar un rango de términos informativos para la clasificación implica centrarse en términos de frecuencia media, que no son exclusivamente específicos de uno o muy pocos documentos, ni absolutamente generales a toda la colección de documentos. Esto hace más complicada la determinación de este rango.

- Al momento de comparar el comportamiento de los ensambles, se puede observar que con 1000 árboles de decisón se alzanza un mejor desempeño de los clasificadores.
- Para concluir, se pudo comprobar la robustez de los prototipos propuestos al mantener un buen desempeño en las distintas colecciones.

Varias vías están abiertas para continuar este trabajo, por supuesto se necesita una evaluación más en profundidad de los clasificadores propuestos incluyendo más colecciones y un análisis comparativo con otras técnicas de reducción de dimensionalidad. Además se podrían emplear otros esquemas de pesado para la representación de relevancia de un término dentro de la colección.

Referencias

1. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47 (2002)
2. Gómez, J. M., de Buenaga Rodríguez, M., and Giráldez, I.: Text categorization for internet content filtering. Inteligencia Artificial: revista iberoamericana de inteligencia artificial, 8(22), 147-160, (2004)

3. Shakeri, S., and Rosso, P.: Spam Detection and Email Classification. Information Assurance and Computer Security. IOS Press, 155, 167, (2006)
4. Amati, G., Aloisi, D. D., Giannini, V., and Ubaldini, F.: A framework for filtering news and managing distributed data. Journal of Universal Computer Science, 3(8), 1007-1021, (1997)
5. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), 538-556, (2009)
6. Lui, A. K. F., Li, S. C., and Choy, S. O.: An evaluation of automatic text categorization in online discussion analysis. In Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on (pp. 205-209), IEEE, (2007)
7. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. 137-142, Springer Berlin Heidelberg (1998)
8. Lewis, D. D., and Ringuette, M.: A comparison of two learning algorithms for text categorization. In Third annual symposium on document analysis and information retrieval, Vol. 33, pp. 81-93, (1994)
9. Apte, C., Damerau, F. J., and Weiss, S. M.: U.S. Patent No. 6,253,169. Washington, DC: U.S. Patent and Trademark Office, (2001)
10. Li, Y. H., and Jain, A. K.: Classification of text documents. The Computer Journal, 41(8), 537-546, (1998)
11. Breiman, L.: Random forests. Machine learning, 45(1), 5-32, (2001)
12. Stopword Spanish Snowball. URL: http://snowball.tartarus.org/algorithms/spanish/stop.txt
13. Salton, G.: The SMART retrieval systemexperiments in automatic document processing. (1971)
14. Salton, G.: Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley. (1989)
15. Salton, G., and Buckley, C.: Term weighting approaches in automatic text retrieval. Information processing and management, 24(5), 513-523, (1988)
16. Zipf, G. K.: Human behavior and the principle of least effort, (1949)
17. Luhn, H. P.: The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165, (1958)
18. Porter, M. F.: An algorithm for suffix stripping. Program, 14(3), 130-137. (1980)
19. Bordignon, F. R. A., and Panessi, W.: Procesamiento de variantes morfolgicas en bsquedas de textos en castellano. Revista Interamericana de Bibliotecologa, 24(1), (2011)
20. Todo Jujuy URL: http://www.todojujuy.com/, Jujuy al momento URL: http://www.jujuyalmomento.com/, Pregón URL: http://www.pregon.com.ar/, El Tribuno de Jujuy URL: http://www.eltribuno.info/jujuy/, El Tribuno de Salta URL: http://www.eltribuno.info/salta/, Informate Salta URL: http://informatesalta.com.ar/, El Intransigente. URL: http://www.elintransigente.com/, La Gaceta: URL: http://www.lagaceta.com.ar/