Primera instancia evaluativa de Procesamiento De Lenguaje Natural II

# Clasificador de noticias

# Diaz, Emiliano Andres

# IES

# Resumen

# El clasificador de noticias consiste en, a travez de la cantidad de palabras y el valor asociado a la misma para cada categoría preestablecida, determinar a cual de las mismas se le asocia un texto informativo. Para esto se crearon tres instancias principales, una en la cual se obtiene una base de datos con noticias ya clasificadas por el usuario, otra donde se procesan dichos documentos para obtener las características principales de cada uno y por ultimo se toma el texto a predecir y se lo procesa para poder determinar de manera correcta a la categoría a la que pertenece

Clasificador de noticias

Para poder tener un estándar en las noticias y en los datos, solo se obtuvieron los datos de la pagina web eldiariodecarlospaz.com, aunque se ha probado con predicciones de distintas fuentes teniendo un resultado positivo en la mayoria de los casos por lo que podemos asegurar que es fiable para diversas fuentes.

Otro dato a tener en cuenta es que las únicas librerías utilizadas en todo el proyecto fueron las relacionadas al scraping y base de datos, además de numpy.

# Scraping

En esta sección veremos como se obtuvieron los datos para iniciar la base de datos de la cual poder obtener las características de las distintas categorías.

## Listas de noticias

Lo primero que se hizo fue crear listas de noticias ya clasificadas de una manera rigurosa y buscando que su longitud y estructura sea la otpima. Teniendo asi, noticias con un mínimo de 2 parrafos y con un cuerpo que no conste en citas o sean complementos de noticias anteriores donde la información sea simplemente correciones.

Para esto se ponen los links de las noticias clasificadas en una lista como se muestra a continuación

## Subtítulos

Los subtítulos se escribirán con letra cursive y estarán alineados a la izquierda.

## Subtítulos

Los subtítulos se escribirán con letra cursive y estarán alineados a la izquierda.

## Citas

Las fuentes originales deberán ser documentadas en el cuerpo del artículo, citando los autores y fechas de publicación de las fuentes. La fuente completa aparecerá en la lista de referencias al final del artículo, siguiendo el cuerpo del mismo. Cuando los autores de una fuente que no forman parte de la estructura formal de la oración, tanto los autores y años de publicación aparecerán entre paréntesis, separados por punto y coma, por ejemplo (Smith y Jones, 2001; Anderson, Charles, y Johnson, 2003). Cuando se cita una fuente que tiene tres, cuatro o cinco autores, todos los autores se incluyen la primera vez que la fuente sea citada. Cuando esa fuente se cita de nuevo, se utiliza el apellido del primer autor y "et al.". Vea el ejemplo en el párrafo siguiente.

El uso de este estilo estándar APA "dará lugar a una impresión favorable en su profesor" (Smith, 2001). Esto fue confirmado de nuevo en 2003 por el profesor Anderson (Anderson, Charles & Johnson, 2003).

Cuando se cita una fuente que tiene dos autores, ambos autores se citan cada vez. Si hay seis o más autores a ser citados, usar el apellido del primer autor y "et al.". la primera y cada vez subsiguiente que ha sido citado. Cuando se utiliza una cita directa, siempre incluya el autor, año y número de página como parte de la citación. Una cita de menos de 40 palabras, debe estar encerrado entre comillas dobles y debe ser incorporado en la estructura formal de la sentencia. Una cita más larga de 40 palabras o más, debería aparecer (sin comillas) en formato de bloque con cada línea con sangría de cinco espacios desde el margen izquierdo.1

# Títulos

Utilice títulos y subtítulos para organizar las secciones de su artículo. El título de primer nivel llevara mayúscula en la letra inicial y centrado en la página. No comience una nueva página por cada título.

## Subtítulos

Los subtítulos se escribirán con letra cursive y estarán alineados a la izquierda.

## Citas

Las fuentes originales deberán ser documentadas en el cuerpo del artículo, citando los autores y fechas de publicación de las fuentes. La fuente completa aparecerá en la lista de referencias al final del artículo, siguiendo el cuerpo del mismo. Cuando los autores de una fuente que no forman parte de la estructura formal de la oración, tanto los autores y años de publicación aparecerán entre paréntesis, separados por punto y coma, por ejemplo (Smith y Jones, 2001; Anderson, Charles, y Johnson, 2003). Cuando se cita una fuente que tiene tres, cuatro o cinco autores, todos los autores se incluyen la primera vez que la fuente sea citada. Cuando esa fuente se cita de nuevo, se utiliza el apellido del primer autor y "et al.". Vea el ejemplo en el párrafo siguiente.

El uso de este estilo estándar APA "dará lugar a una impresión favorable en su profesor" (Smith, 2001). Esto fue confirmado de nuevo en 2003 por el profesor Anderson (Anderson, Charles & Johnson, 2003).

Cuando se cita una fuente que tiene dos autores, ambos autores se citan cada vez. Si hay seis o más autores a ser citados, usar el apellido del primer autor y "et al.". la primera y cada vez subsiguiente que ha sido citado. Cuando se utiliza una cita directa, siempre incluya el autor, año y número de página como parte de la citación. Una cita de menos de 40 palabras, debe estar encerrado entre comillas dobles y debe ser incorporado en la estructura formal de la sentencia. Una cita más larga de 40 palabras o más, debería aparecer (sin comillas) en formato de bloque con cada línea con sangría de cinco espacios desde el margen izquierdo.1

# Referencias

Anderson, Charles & Johnson (2003). *The impressive psychology paper.* Chicago: Lucerne Publishing.

Smith, M. (2001). Writing a successful paper. *The Trey Research Monthly*, *53*, 149-150.

Entries are organized alphabetically by surnames of first authors and are formatted with a hanging indent. Most reference entries have three components:

1. Authors: Authors are listed in the same order as specified in the source, using surnames and initials. Commas separate all authors. When there are seven or more authors, list the first six and then use “et al.” for remaining authors. If no author is identified, the title of the document begins the reference.
2. Year of Publication: In parenthesis following authors, with a period following the closing parenthesis. If no publication date is identified, use “n.d.” in parenthesis following the authors.
3. Source Reference: Includes title, journal, volume, pages (for journal article) or title, city of publication, publisher (for book).

# Apéndice

Cada apéndice aparece en una página propia.

# Notas al pie

1La información completa del estilo APA lo puede encontrar en el Manual de Publicación..

Tabla 1

*Escriba el texto de la tabla aquí en cursiva; iniciar una nueva página para cada tabla*

[Inserte la tabla aquí]

# Figuras

*Figura 1.* Título de la figura

[Figuras -Tenga en cuenta que esta página no tiene el título del manuscrito ni número de página]