Introducción a Learning Analytics con ejemplos prácticos

UD 04. Caso práctico 01 - Análisis de resumen (Individual)

Autor: Sergi García Barea

Actualizado Octubre 2020

Licencia

Nomenclatura

A lo largo de este tema se utilizarán distintos símbolos para distinguir elementos importantes dentro del contenido. Estos símbolos son:

📖 **Importante**

❕ **Atención**

💬 **Interesante**

ÍNDICE DE CONTENIDO

UD04. CASO PRÁCTICO 01 - ANÁLISIS DE RESUMEN (INDIVIDUAL)

1. DESCRIPCIÓN DEL CASO PRÁCTICO

En este caso práctico para ilustrar cómo puede aplicarse el análisis de resúmenes, vamos a tomar como ejemplo una de las obras más conocidas y destacadas: “El Quijote”, de Miguel de Cervantes.

En este ejemplo, vamos a plantear dos fuentes para que el profesor obtenga información de que puede esperar de un posible resumen de “El Quijote”:

- El Quijote: obra libre de derechos, obtenida en formato TXT de “Proyecto Gutenberg” http://www.gutenberg.org/ebooks/2000
- Fragmentos de la Wikipedia: https://es.wikipedia.org/wiki/Don_Quijote_de_la_Mancha

Tras ello analizaremos un resumen propuesto a un alumno, cuyo umbral de palabras esperado es de entre **150 y 200 palabras**.

2. HERRAMIENTAS UTILIZADAS DURANTE EL CASO PRÁCTICO

A continuación, indicamos las herramientas utilizadas para realizar este caso práctico:

- Procesamiento de texto online: https://countwordsfree.com/
- Análisis de legibilidad: https://legible.es/
-

> 📖 **Importante:** las herramientas aquí propuestas son una sugerencia. Al realizar procesos de análisis siempre debéis utilizar las herramientas que os hagan sentir más cómodos.

3. EXTRAYENDO INFORMACIÓN DE LAS FUENTES PRINCIPALES

Como se ha realizado esta extracción de información puede verse en el video https://www.youtube.com/watch?v=EVjcsV-cC_g&feature=youtu.be

En dicho video, hemos observado cómo extraer información de diversos textos, usando https://countwordsfree.com/ (cargando fichero TXT del Quijote y cargando TXT del resumen de Wikipedia) y con https://legible.es/ . Los ficheros de los cuales se ha extraído la información se encuentran en **“CasoEstudioUD04-01.zip”**.

3.1 El Quijote (Proyecto Gutenberg)

El texto tiene una longitud de aproximadamente 769 páginas en tamaño A4 con un total de 384352 palabras. Su tiempo de lectura medio estimado es de 32 horas, 1 minuto y 45 segundos. Hemos configurado la aplicación para que elimine en el análisis de palabras usadas las llamadas “stop words” (Palabras vacías, sin significado https://es.wikipedia.org/wiki/Palabra_vac%C3%ADa). También hemos indicado que solo tenga en cuenta palabras con al menos 4 caracteres.

El análisis de palabras más utilizadas en las 20 primeras posiciones es el siguiente:

Puesto	Palabra	Ocurrencias	Caracteres
1	**sancho**	2036	6
2	**quijote**	2003	7
3	**respondió**	1061	9
4	**señor**	1047	5
5	**merced**	881	6
6	**vuestra**	850	7
7	**caballero**	643	9
8	**dios**	516	4
9	**señora**	498	6
10	**cosa**	433	4
11	**allí**	417	4
12	**aquella**	331	7
13	**mundo**	313	5
14	**casa**	312	4
15	**panza**	309	5
16	**digo**	306	4
17	**cura**	298	4
18	**puesto**	293	6
19	**vida**	292	4
20	**mano**	285	4

Observando la frecuencia del uso de palabras, ya podemos detectar algunas candidatas a ser palabras clave. De estas 20 palabras, combinados con el conocimiento experto, podríamos seleccionar “Sancho”, “Quijote”, “caballero”, “dios”, “cura”.

Si seguimos observando la lista hasta las 50 palabras con más apariciones, podemos detectar otras interesantes como “Dulcinea” (puesto 22), “barbero” (puesto 35), “escudero” (puesto 37), “Rocinante” (99).

3.2 El Quijote (Resumen Wikipedia, apartado 1)

Llevando a cabo un análisis similar sobre el resumen del Quijote de Wikipedia (eliminando "stop words" y contando solo palabras con 4 o más letras), obtenemos la siguiente frecuencia de textos.

Puesto	Palabra	Ocurrencias	Caracteres
1	**quijote**	104	2.5 %
2	**sancho**	53	1.1 %
3	**caballero**	28	0.9 %
4	**cervantes**	17	0.5 %
5	**edición**	13	0.3 %
6	**escudero**	12	0.3 %
7	**novela**	12	0.2 %
8	**libros**	11	0.2 %
9	**obra**	11	0.1 %
10	**personaje**	11	0.3 %
11	**aventuras**	10	0.3 %
12	**dulcinea**	10	0.3 %
13	**gigantes**	9	0.2 %
14	**aventura**	8	0.2 %
15	**castillo**	8	0.2 %
16	**ingenioso**	8	0.2 %
17	**mancha**	8	0.2 %
18	**quijote.**	8	0.2 %
19	**continuación**	7	0.3 %
20	**panza**	7	0.1 %

En el análisis vemos en primer lugar que alguna palabra ha sido mal contada al estar unida a un signo de puntuación.

Podríamos haber utilizado alguna página web online tipo https://www.browserling.com/tools/remove-punctuation para eliminar estos simbolos e incluso para un mayor nivel de normalización, haber usado alguna web tipo https://txtformat.com/ o incluso haber aplicado al texto algoritmos de "Stemming"en español (obtener raiz de la palabra).

En cualquier caso, en este ejemplo, al ser residual los casos y estar apoyados por nuestro conocimiento experto, estos pasos no son importantes (sí lo serían en trabajos de "Procesamiento del lenguaje natural").

Sobre lo observado en el texto, en las 20 primeras ocurrencias, solo aparecen 4 de las palabras que hemos visto en el primer caso ("Quijote", "Panza", "Dulcinea", "escudero").

Sin embargo aparecen otras interesantes, como "gigantes" (del episodio donde confunde molinos con gigantes) o "castillo" (referencia a cuando confunde una venta con un castillo).

Si seguimos observando en los siguientes puestos, ya aparecen como relevantes, algunas como "barbero" (puesto 73), "molinos" (puesto 73) , "Rocinante" (puesto 83).

3.3 Palabras clave seleccionadas

En base a este análisis de frecuencias y nuestro conocimiento experto, hemos seleccionado como candidatas a palabras clave:

"Sancho", "Quijote", "caballero", "dios", "cura", "Dulcinea", "barbero", "escudero", "Rocinante", "castillo", "molinos", "gigantes".

Obviamente, esta selección de palabras clave del Quijote, podría mejorarse con el conocimiento experto de la obra, pero de cara a nuestro ejercicio de análisis, vamos a utilizar esas palabras clave.

4. OBJETIVOS PLANTEADOS

En este proceso de análisis del resumen nos planteamos los siguientes objetivos:

- Analizar el rendimiento del resumen individual
  - ¿Se ajusta al umbral esperado?
  - ¿El resumen usa o no determinadas palabras clave?
  - ¿Qué nivel de legibilidad tiene el resumen?

5. MÉTRICAS

Las métricas a extraer de un resumen individual de un alumno, son las siguientes:

- Número de palabras total del texto.
- Número de veces que se ha utilizado cada palabra clave.
- Índices de legibilidad variados.

6. PROCESAMIENTO DE DATOS

Los datos de este resumen se encuentran dentro de **"CasoEstudioUD04-01.zip"** en concreto en el fichero **"ResumenAlumnoQuijote.txt"**.

El texto tiene un total de **164 palabras**. Sobre este texto extraemos los siguientes datos:

Palabra clave	**Ocurrencias**
Sancho	3
Quijote	2
caballero	2
dios	0
cura	0
Dulcinea	1
barbero	0
escudero	0
Rocinante	1
castillo	0
molinos	1
gigantes	1

Además, procesamos el texto del resumen en www.legible.es obtenemos los siguientes indices de legibilidad, así como un tiempo de lectura estimado de 0.8 minutos.

Índice	**Valor**	**Dificultad**
Fernández Huerta	70.74	algo fácil
Gutiérrez	43.59	normal
Szigriszt-Pazos	67.16	bastante fácil
INFLESZ	67.16	bastante fácil
legibilidad µ	63.89	adecuado

7. ANÁLISIS

7.1 Análisis adecuación al umbral propuesto

Este análisis es sencillo. Al ser un texto de 164 palabras, se encuentra dentro del umbral propuesto en el enunciado (entre 150 y 200 palabras).

Análisis: el alumno se ha adecuado a la longitud indicada.

7.2 Análisis sobre palabras clave utilizadas en el resumen

Destacamos que entre las palabras clave propuestas, se han utilizado las palabras:
"Sancho", "Quijote", "caballero", "Dulcinea", "Rocinante", "molinos", "gigantes".

> ❗ **Atención:** detectar unas palabras clave, pueden ser indicio de que un alumno ha tratado determinados temas, pero debe validarse por el docente.

Análisis: se han utilizado 7 de las 12 palabras clave propuestas. Ello es un indicio de que es posible que el alumno haya podido identificar algunas ideas claves: quienes son los protagonistas (Quijote y Sancho), la novela está relacionada con la temática de la caballería, que se existen otros personajes destacados (Dulcinea, Rocinante) o que la presencia de "molinos" y "gigantes" indica que se hace referencia a dicha parte de la novela.

7.3 Análisis sobre palabras clave no utilizadas en el resumen

Destacamos que entre las palabras clave propuestas, no se han utilizado las palabras:
"dios", "cura", "escudero", "barbero", "castillo".

Análisis: no se han utilizado 5 de las 12 palabras clave propuestas. De ello se puede desprender que en el resumen hay algunas ideas que no están reflejadas: la aparición de personajes secundarios como el cura y el barbero, las referencias a dios o la condición de Sancho Panza como escudero de Quijo. También, al no aparecer la palabra "castillo", podemos intuir que no hay ninguna referencia al episodio del castillo y la venta u otros episodios donde aparecieran castillos.

7.4 Análisis de legibilidad y velocidad de lectura

Los distintos índices de legibilidad, indican que es de fácil lectura.

Análisis: el ser de fácil lectura, es indicio que se han utilizado en general palabras y frases cortas. La lectura de este valor como positiva o negativa puede depender del tipo de lenguaje esperado por el alumno. Generalmente, si el resumen tiene el fin de ser utilizado como repaso, la facilidad de lectura es positiva. También dada la longitud propuesta, es posible espere que el resumen no sea muy denso.

8. ACTUACIONES

En este punto, ya con los análisis realizados, vamos a realizar propuestas de actuaciones que debe podría llevar a cabo el profesor. Las propuestas que aquí planteamos son genéricas. Tras realizar cualquier actuación debe realizarse un seguimiento en el tiempo

- Felicitar al alumno por haberse mantenido en el umbral propuesto.
- Felicitar al alumno por haber tenido en cuenta algunos aspectos clave en el resumen.
- Reflexionar como profesor, si se pretendía en la actividad que determinados temas sean tratados por el alumno. Si así era, reflexionar sobre si se ha orientado correctamente para que esos temas fueran reflejados en el resumen:
    - Dios en el Quijote.
    - Condición de Sancho como escudero.
    - Personajes “el cura” y “el barbero”.
    - Episodio del castillo.
- Recabar información hablando con el alumno de porque no ha tratado los temas anteriores.
- Felicitar al alumno si la complejidad del lenguaje era la adecuada.
- Si la complejidad del lenguaje no fuera la adecuada:
    - Revisar si se ha orientado correctamente al alumno en esa dirección.
    - Consultar con el alumno porque no ha usado el lenguaje esperado.

📖 **Importante:** recordad que tras cada actuación finalmente realizada, debe realizarse un seguimiento a través del tiempo para facilitar la evaluación de la misma.

9. BIBLIOGRAFÍA

[1] Speech and language processing (Dan Jurafsky) https://web.stanford.edu/~jurafsky/slp3/

[2] Handbook of Learning Analytics (Charles Lang, George Siemens, Alyssa Wise, Dragan Gašević) https://www.researchgate.net/publication/324687610_Handbook_of_Learning_Analytics