Miniproyecto N ° 1
Introducción al Análisis de Textos Biomédicos

Profesora: [1]Rosa Figueroa, **Alumnos ayudantes:** [2]Christopher Flores

{[1]rosa.figueroa, [2]christopher.flores}@biomedica.udec.cl

18 de Abril de 2018

En este miniproyecto deberán procesar un texto para generar algunas estadísticas según los contenidos vistos en clases. El texto puede ser descargado del siguiente enlace y deberá ser procesado (considerar sólo la tabla ASCII) para responder las siguientes preguntas:

a) ¿Cuál es el *token* que contiene más vocales?

b) ¿Cuál es el *token* que contiene más consonantes?

c) ¿Cuál es el *token* más largo?

d) ¿Cuáles son los 10 *tokens* más frecuentes?

e) ¿Cuáles son las 10 *tokens* menos frecuentes?

f) ¿Cuántos *tokens* corresponden a sólo a números?

g) ¿Cuántos *tokens* contienen letras y números?

h) ¿Cuáles son las 10 *stopwords* más frecuentes en el *corpus*?

i) Gráfico de la Ley de *Zipf*. Concluir al respecto.

j) ¿Cuál es la variación porcentual de la diversidad léxica del corpus luego de aplicar *stemming*?. Concluir al respecto.

k) ¿Cuál es la variación porcentual de la diversidad léxica del corpus luego de eliminar las *stopwords*?. Concluir al respecto.

Suba todos los archivos necesarios para la revisión de su mini-proyecto al *classroom* del curso antes del **02 de Mayo hasta las 23:59 hrs.** en un archivo `Apellido1_Apellido2.zip`. Si tiene algún inconveniente, envíe los archivos a `rosa.figueroa@biomedica.udec.cl` con copia a `christopher.flores@biomedica.udec.cl`. Además, deberá entregar en secretaría de biomédica un resumen de su trabajo (una hoja) en formato IEEE [1] (resumen, introducción, materiales y métodos, resultados, discusión y conclusión, referencias). Atrasos en la entrega serán penalizados según lo indica el *syllabus* del curso. La copia total o parcial será calificada con la nota mínima según lo indica el reglamento. Para este mini-proyecto se aceptarán trabajos de un máximo de 2 personas.

```
#Leer un archivo de texto
with open('nombre_archivo', 'r') as a:
    texto = a.read()
```

[1] `http://ieeeauthorcenter.ieee.org/create-your-ieee-article/use-authoring-tools-and-ieee-article-templates/ieee-article-templates/templates-for-transactions/`