Un ejercicio "real" de procesamiento de datos.

Visión general del problema:

Tenemos un sistema de alineamiento automático de voz y texto que asigna a cada palabra de un texto el instante en que comienza a pronunciarse en el correspondiente fichero de audio, medido en centésimas de segundo del siguiente modo:

```
THE 1200000
RECORDS 6800000
FOR 12100000
THE 12100000
STOCK 13500000
...
```

Donde las cifras son enteras por estar escalada por el factor 10.000.000, es decir 1200000 se corresponde con el instante 0,12 segundos

El número de palabras puede ser muy grande (no lo es demasiado en el caso de los ficheros a utilizar que tienen 32.043 palabras)

Para valorar si el sistema ha realizado bien su trabajo se dispone de una referencia con el mismo formato (es decir un fichero semejante en el que las marcas de tiempo son correctas)

Lo que queremos por tanto es caracterizar la "bondad" del alineamiento mediante el análisis de la distribución de diferencias entre lo correcto y lo obtenido automáticamente (también necesitaremos de otros mecanismos según se vea la necesidad, como detectar zonas problemáticas, etc. … pero esto no forma parte del planteamiento de la práctica)

CLASE HISTOGRAMA

Para esto necesitaremos de una clase capaz de calcular histogramas con un par de constructores:

```
Histogram(float[] data, float barWidth, int numBars);
Histogram(float[] data, float barWidth, float maxValue);
```

Para poder cambiar la representación con los mismos datos, permitiremos cambiar el ancho y número de barras

```
void setBars(float barWidth, int numbars);
void setBars(float barWidth, float maxValue);
```

Los valores de cada barra se podrán obtener mediante

```
float[] getValues();
```

Así como los centros de las barras mediante

```
float[] getCenters();
```

Y los puntos de corte mediante

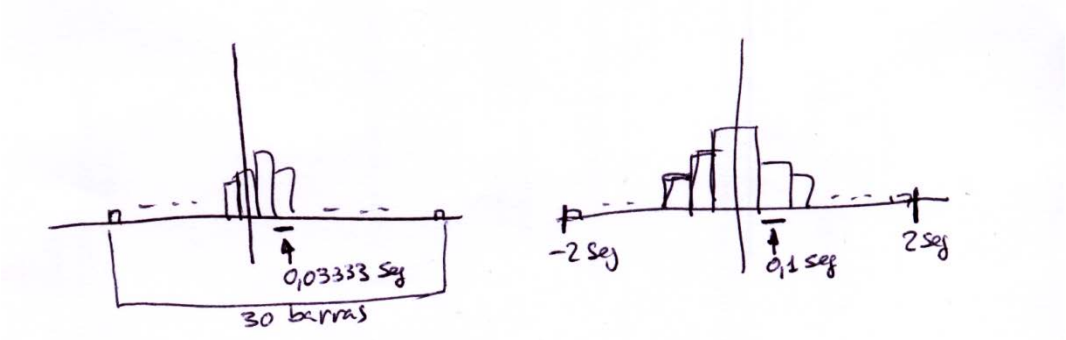
```
float[] getBoundaries();
```

El método toString se reescribirá para que liste pares (centro, valor)

PROGRAMA DE PRUEBA

El programa de prueba recogerá **mediante una rutina** la lista de diferencias de tiempo para cada palabra entre la referencia y el obtenido automáticamente a partir de los dos ficheros con el formato indicado (se trata de quedarse con la segunda columna de cada fichero y obtener la resta –no es preciso comprobar que las palabras coinciden puesto que eso se asegura de antemano-). Los dos ficheros se darán como argumentos de entrada al método main.

A continuación se usará la clase Histograma para calcular el procentaje de errores menores de 0.5 segundos, el de errores menores de 2 segundos, y para obtener 2 histogramas: uno con barras de 0.03333 segundos de ancho y 30 barras, y otro con barras de ancho 0.1 segundo y hasta 2 segundos



"VUELTA DE TUERCA"

Permitamos que el histograma pueda realizarse sobre un intervalo de las muestras (no sobre todo el conjunto de datos) indicando el tiempo de inicio y de final establecido en el fichero de referencia.

Esto servirá para analizar secciones que están listadas en un nuevo fichero de los accesibles en la web para realizar el ejercicio.