Zip'f

Jordy Joaquín Cuan Robledo [1], Eduardo Cantorán Flores [1],
Rafael Pérez Aguirre [1]
[1] Facultad de Ciencias de la Computación, Benemérita Universidad
Autónoma de Puebla, México

El autor de esta fórmula fue George Kingsley Zipf en 1949, él decía que una pequeña cantidad de palabras normalmente tienen una alta frecuencia,.
Zip's mide la frecuencia de una palabra que aparece en un texto. Para ello cuenta las repeticiones de una palabra en todo un documento y lo divide entre el total de palabras.
La fórmula es la siguiente:

$$Pn \sim 1/n^a$$

Para mejores resultados se debe emplear un documento suficientemente extenso.
En el lenguaje español las palabras con mayor frecuencia son los artículos y preposiciones.
A continuación se muestra el código:

```
awk '{
gsub(/[!$%&=?¡¿´¨""*{},_;:()+[\]\x2E-]/, " ", $0);
    for (i=1; i<=NF; i++){
        frecuencia[tolower($i)]++;
        total++;
    }
    next;
}
END{
    for (x in frecuencia) {
        split(x, a, SUBSEP);
        print a[1], "\t"frecuencia[x], "\t"frecuencia[x]/total;
  }
    print "----------------------------"
    print "Total de palabras = " total;
}
' $*
```

Antes de calcular la frecuencias de las palabras contenidas en el documento es necesario limpiar el texto, para ello se utiliza la función gsub la cual quitará los caracteres especiales que delimitamos en cada palabra del texto.
Después recorremos cada registro insertando cada palabra en una tabla hash aumentando su valor de repetición. Luego para cada palabra de la tabla hash se divide su número de repeticiones entre en total de palabras del texto.
Por último para ejecutarlo lo guardamos en un archivo:
./zipf.awk textoPrueba.txt > modeloZipf.txt