# Práctica 1

Procesamiento de Lenguaje Natural
Facultad de Ingeniería, UNAM

**Objetivo**: Obtener la curva de Zipf de los tipos de un corpus en escala logarítmica y determinar el parámetro $\alpha$ correspondiente a esa distribución.
Pasos a seguir:

1. Escoger un corpus de cualquier idioma y de un tamaño mayor a 10 000 tokens (se puede tomar este corpus de la paquetería *nltk.corpus*).
2. Limpiar el corpus: eliminar signos de puntuación, de interrogación, admiración y elementos no léxicos.
3. Aplicar un algoritmo de Stemming a los tokens limpios.
4. Obtener las frecuencias de los tipos en el corpus.
5. Ordenar por el rango estadístico de mayor a menor.
6. Graficar el diagrama de dispersión rango-frecuencia en escala logarítmica.
7. Obtener el parámetro de la distribución de Zipf, $\alpha$ (a partir de un procedimiento de regresión).