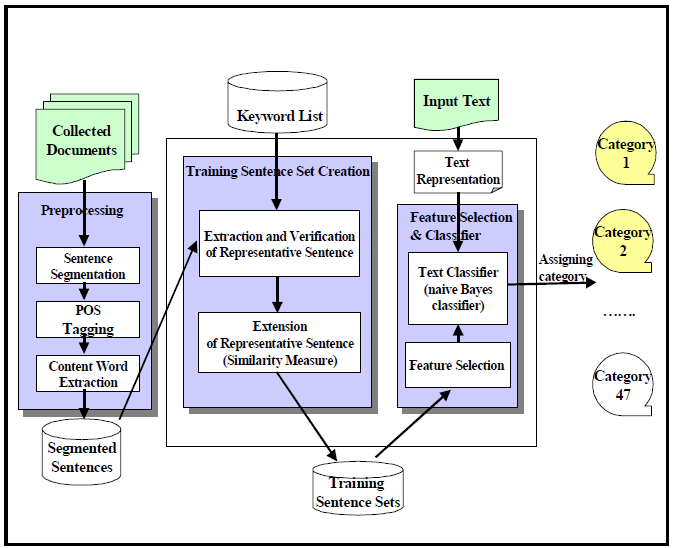
# Método de los coreanos para etiquetar tuits de entrenamiento

## Esquema del sistema.

Figura 1 Arquitectura del sistema coreano

* 1. **Preprocesado.**

Cada tuit se segmenta en *oraciones* (sentencias).

Se extraen los nombres contenidos de cada sentencia usando un *POS Tagger[[1]](#footnote-1)*.

* 1. **Crear conjuntos de palabras claves de entrenamiento.**

Definimos *manualmente* palabras clave para cada categoría (más de una).

Las *sentencias* que contienen *palabras clave* predefinidas se eligen como inicialmente representativas. El resto de sentencias se consideran no clasificadas, esto es, no relevantes para la asignación de clase al texto.

Estas sentencias no clasificadas se tratan de escalar a la categoría representativa mediante la aplicación de *medidas de similaridad*.

* + 1. ***Extracción y verificación de sentencias representativas***.

1. <http://www.nltk.org/book/ch05.html> [↑](#footnote-ref-1)