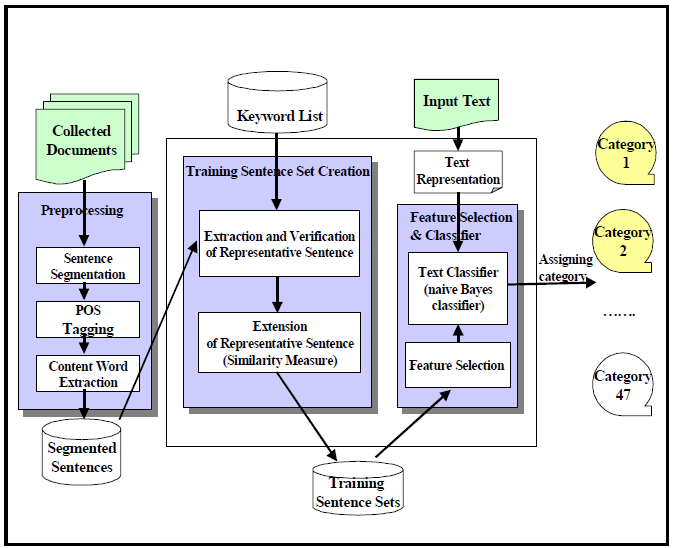
# Método de los coreanos para etiquetar tuits de entrenamiento

## Esquema del sistema.

Figura 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***.

## Python y NPL.

[Pattern](http://www.clips.ua.ac.be/pages/pattern) mining module for the Python programming language.

It has tools for data mining (Google, Twitter and Wikipedia API, a web crawler, a HTML DOM parser), natural language processing (part-of-speech taggers, n-gram search, sentiment analysis, WordNet), machine learning (vector space model, clustering, SVM), network analysis and <canvas> visualization.

Figure Módulo Pattern de Python

NTKL

## Corpora en español.

[Wiki Corpus en español](http://www.cs.upc.edu/~nlp/wikicorpus/)

## Bases de Datos.

<https://www.hatebase.org/about>

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