

# Extracción de Información de Noticias UTFSM

Felipe Barriga Richards  
felipe@felipebarriga.cl  
2473559-1  
<http://blog.felipebarriga.cl>

31 de octubre de 2010

## Resumen

En este documento se describira la metodología utilizada para extraer n-gramas de las noticias publicadas en el portal de Informatica de la UTFSM (<http://www.inf.utfsm.cl>) y su posterior uso en un grafo.

**Keywords:** ngrams extraction news nodes edges graph

## 1. Introducción

Para un mejor entendimiento de los pasos seguidos se hara referencia a los scripts utilizados durante el proceso con una breve descripción de estos. Todos los scripts se encontrarán junto a este documento en archivos separados.

## 2. Scripts Utilizados

Los scripts utilizados fueron los siguientes (en orden secuencial):

- **download\_news.sh:** Script utilizado para descargar las noticias
- **convert\_html\_text.sh:** Script utilizado para la conversión de las páginas html descargadas en texto plano
- **split\_them.sh:** Script encargado de la división de noticias contenidas en cada página (genera multiples archivos conteniendo noticias a partir de una página html)
- **clean.sh:** Script utilizado para limpiar archivos de texto (dejar todo en minúsculas, eliminar saltos de linea, eliminar multiples espacios entre otros)
- **ngrams.sh:** Script que ejecuta **text2ngram** para cada archivo conteniendo noticias y guarda la salida en ngrams.dat
- **count\_co-occurrences.php:** Script para contar las co-ocurrencias de n-gramas en un texto (determina la afinidad de los n-gramas)

## 3. Pasos a Seguir

### 3.1. Preparar Archivos

#### 3.1.1. Descargar Noticias

Descargar archivos html con noticias y almacenarlos en **data/**<sup>1</sup>

#### 3.1.2. Conversión de HTML a texto plano

Ejecutar **convert\_html\_text.sh** lo que generara un archivo txt por cada archivo html

#### 3.1.3. Separar Noticias en diferentes archivos

Ejecutar **split\_them.sh** lo que generará un archivo txt por cada noticia

#### 3.1.4. Limpiar archivos de texto

Ejecutar **clean.sh** para limpiar los archivos de noticia y dejarlos lineales, sin saltos de linea y en minusculas.

#### 3.1.5. Union de Archivos

Unir todos los archivos en uno solo (`cat *.txt > whole.txt`) y limpiarlo (ver filtro en **clean.sh**)

### 3.2. Extracción de n-gramas

#### 3.2.1. Extraer los n-gramas

- `text2ngram -n2 -f15 whole.txt > 2-grams.lst`
- `text2ngram -n3 -f10 whole.txt > 3-grams.lst`
- `text2ngram -n4 -f5 whole.txt > 4-grams.lst`

#### 3.2.2. Ordenar en order de frecuencia los n-gramas

- `sort -n -k 3 -r 2-grams.lst > 2-grams-sorted.lst`
- `sort -n -k 4 -r 3-grams.lst > 3-grams-sorted.lst`
- `sort -n -k 5 -r 4-grams.lst > 4-grams-sorted.lst`

---

<sup>1</sup>El script para bajar noticias ya no funciona por cambio de formato en sitio web

### 3.2.3. Mostrar los n-gramas más frecuentes

- `head -n 500 2-grams-sorted.lst > 2-grams-sorted-500.lst`
- `head -n 500 3-grams-sorted.lst > 3-grams-sorted-500.lst`
- `head -n 500 4-grams-sorted.lst > 4-grams-sorted-500.lst`

## 3.3. Análisis de n-gramas

Los archivos que contienen los diversos n-gramas (*2-grams-sorted-500.lst*, *3-grams-sorted-500.lst*, *4-grams-sorted-500.lst*) fueron analizados con la finalidad de encontrar datos de interés.

Dentro de los n-gramas buscados se encuentran los que hagan alusión a profesores, lugares y roles. Esta información fue analizada manualmente y los n-gramas seleccionados fueron guardados en un archivo de texto plano (*final\_words.lst*). Con éste archivo generamos manualmente el archivo *nodes.csv*.

## 3.4. Creación de Grafo

### 3.4.1. Contar Co-Ocurrencias

Ejecutar `count_co-occurrences.php` para contar las co-ocurrencias y serán guardadas en el archivo *edges.csv*

### 3.4.2. Gephi

Cargar todos los datos en Gephi<sup>2</sup> y trabajarlos.

---

<sup>2</sup><http://gephi.org/>