

# Introducción al tratamiento de textos (unix, python, perl, nltk, ...)

---



Lingüística Computacional  
Ferran Pla

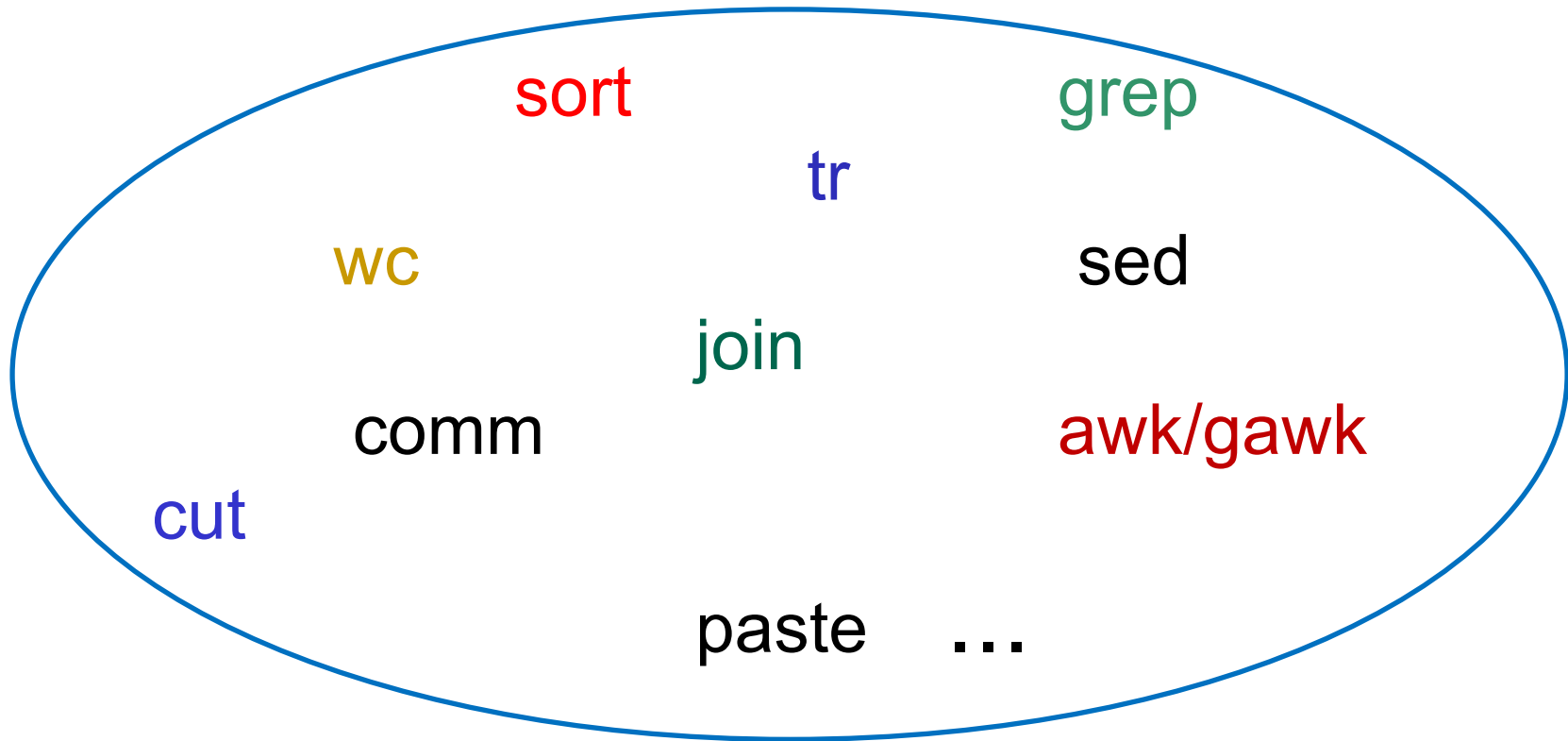
# Introducción

---

- Para el tratamiento de grandes colecciones de textos es conveniente el uso de lenguajes potentes y “fáciles” para la manipulación de cadenas.
- En este tema nos vamos a centrar en presentar, sin que sea esto un curso de programación, el lenguaje Python, incidiendo en las estructuras/clases y métodos útiles para la manipulación de textos.

# "Unix for Poets"

---



**Unix for Poets**  
Kenneth Ward Church  
AT&T Bell Laboratories

<http://www.stanford.edu/class/cs124/kwc-unix-for-poets.pdf>  
<http://ufal.mff.cuni.cz/~hladka/tutorial/UnixforPoets.pdf>

# Ejemplo

Salgo de #VeoTV , que día más largooooo ...

@PauladeLasHeras No te libraras de ayudar me/nos . Besos y gracias@marodriguezb Gracias MAR

Off pensando en el regalito Sinde , la que se va de la SGAE cuando se van sus corruptos .

Intento no sacar conclusiones ( lo intento ) Conozco a alguien q es adicto al drama ! Ja ja ja te suena d algo !

Toca @crackoviadeTV3 . Grabación dl especial Navideño ... Mari crismas !

Rajoy , 3-1 para el Madrid ; Zapatero para " su " Barça y Rubalcaba evita " mojarse " - ABC . es <http://t.co/LxBXidLx> via @abc\_es

Eso es ;-) RT @ccifuentes : @JuananSanzNunez Por supuesto que nos veremos en #Sevilla , un besazo ;-) @mariviromero Eso es ;-) )

Veeeeenga ..... Hagamos porra !!! Quién se lleva el partido ??

```
cat text.tx | gawk '{for (i=1;i<=NF;i++) print $i;}'  
| sort | uniq -c | sort -nr>freq.txt
```

```
cat text.tx | gawk '{for (i=1;i<=NF;i++) print $i;}'  
| sort | uniq -c | sort -nr | grep '.*e$'
```

freq.tx

4 es

4 "

4 .

4 ,

3 se

3 que

3 el

3 de

3 !

2 y

# Python

---

**<http://python.org/>**

## **Python 3.x**

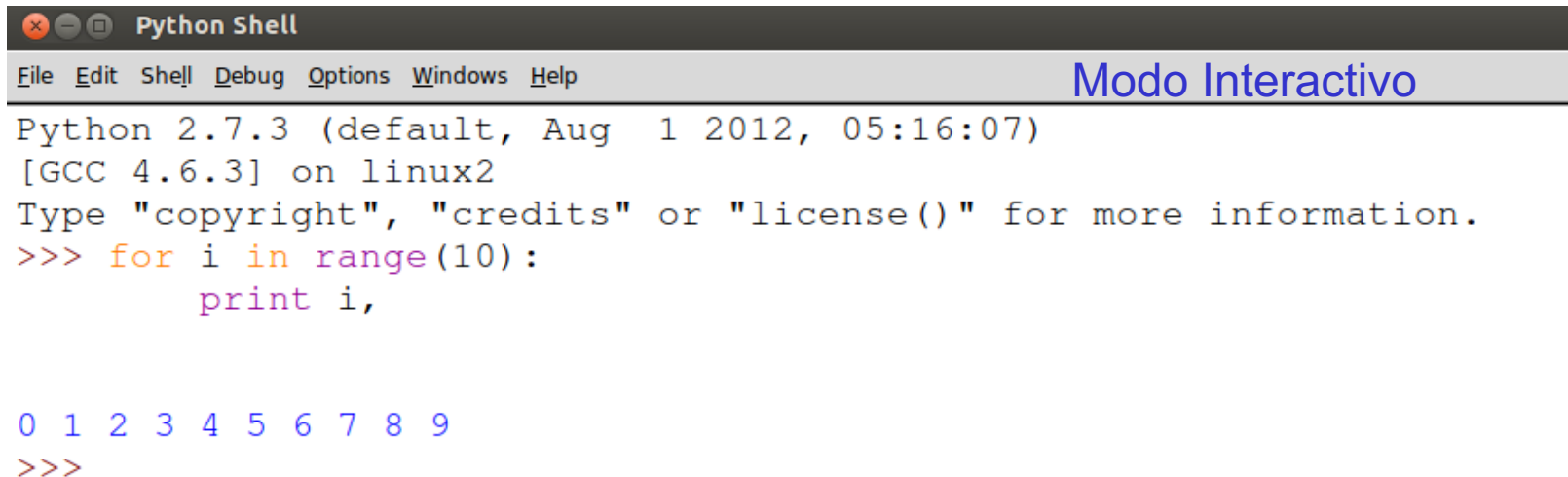
**<https://docs.python.org/3/>**

- Lenguaje creado por **Guido Van Rossum** en Holanda en 1991.
- Lenguaje interpretado multi-plataforma, orientado a objetos.
- Resulta muy adecuado para la manipulación de textos.

## **Tutoriales**

**Buscad por la red**

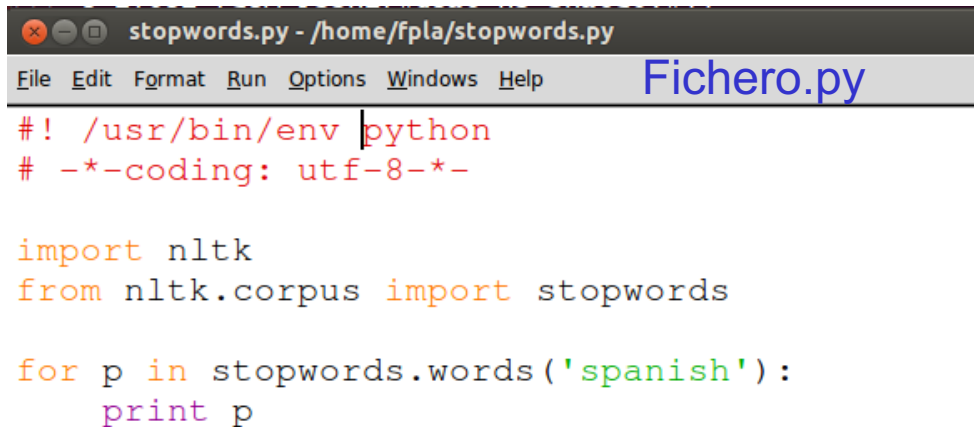
# ¿Cómo editar/ejecutar un programa Python?



The screenshot shows a window titled "Python Shell" with a menu bar (File, Edit, Shell, Debug, Options, Windows, Help) and a status bar indicating "Modo Interactivo". The main text area displays the Python version (2.7.3), GCC version (4.6.3), and the operating system (linux2). It shows a prompt for user input, followed by a loop that prints numbers from 0 to 9.

```
Python 2.7.3 (default, Aug 1 2012, 05:16:07)
[GCC 4.6.3] on linux2
Type "copyright", "credits" or "license()" for more information.
>>> for i in range(10):
    print i,

0 1 2 3 4 5 6 7 8 9
>>>
```



The screenshot shows a text editor window titled "stopwords.py - /home/fpla/stopwords.py" with a menu bar (File, Edit, Format, Run, Options, Windows, Help) and a status bar indicating "Fichero.py". The main text area displays a Python script that imports the nltk corpus and prints the words from the 'spanish' corpus.

```
#!/usr/bin/env python
# -*-coding: utf-8 -*-

import nltk
from nltk.corpus import stopwords

for p in stopwords.words('spanish'):
    print p
```

## ¿Qué editor utilizar?

- Python lleva su propio IDLE
- Elige tu editor preferido, en tu SO.
- Importante: que gestione bien la indentación, la sintaxis y la codificación del texto

# Distribución python: Anaconda

---

- <https://docs.anaconda.com/anaconda/>

## Anaconda Distribution

*Open Data Science Core*

Anaconda® is a package manager, an environment manager, a Python distribution, and a collection of [over 720 open source packages](#). It is free and easy to install, and it offers [free community support](#).

Get the [Anaconda Cheat Sheet](#) and then [download Anaconda](#).

Don't want the huge collection of 720 software packages? Get [Miniconda](#).

# NLTK

---

- <http://www.nltk.org/>

## NLTK 3.2.4 documentation

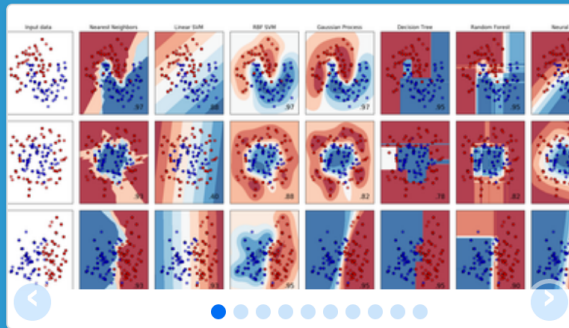
[NEXT](#) | [MODULES](#) | [INDEX](#)

## Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](#) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](#).



<http://scikit-learn.org/stable/>



## scikit-learn

*Machine Learning in Python*

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

### Classification

Identifying to which category an object belongs to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, ... — Examples

### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, ridge regression, Lasso, ... — Examples

### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, ... — Examples

### Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization. — Examples

### Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** grid search, cross validation, metrics. — Examples

### Preprocessing

Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** preprocessing, feature extraction. — Examples