# Proyecto Final
# Categorización Automática de artículos de la Wikipedia

Tratamiento de Datos
Máster en Ing. de Telecomunicación
Curso 2018/2019

December 1, 2018

## 1 Introducción

En este proyecto, los alumnos harán uso de los conocimientos y técnicas adquiridos durante el curso para resolver una tarea de aprendizaje sobre documentos textuales. Los alumnos trabajarán individualmente o por parejas sobre documentos descargados de la Wikipedia en inglés, y las tareas a realizar incluirán necesariamente:

- Procesado y homogeneización de textos
- Modelado de tópicos con el algoritmo LDA
- Técnicas de clasificación automática

El proyecto tiene una valoración máxima de $2,5$ puntos. Consta de dos partes:

- Proyecto básico: $1,75$ puntos
- Extensión: $0,75$ puntos.

A continuación se indican los requisitos de cada una de las partes:

## 2 Proyecto básico

El proyecto básico consistirá en desarrollar un clasificador binario que permita discriminar entre los documentos de dos categorías diferentes de artículos de la wikipedia en inglés. Para ello, se completarán los pasos siguientes:

1. Selección de dos categorías de trabajo, y descarga de corpus de datos, incluyendo al menos 100 documentos por categoría.

2. Preprocesado de artículos.

3. Modelado (conjunto) de los artículo descargados utilizando el algoritmo de extracción de tópicos *Latent Dirichlet Allocation (LDA)*

4. Clasificación binaria de los documentos tomando sus subcategorías como etiquetas, y la distribución de tópicos del LDA como representación de entrada. Análisis de las prestaciones del clasificador.

## 2.1 Descarga de datos

Para la realización del proyecto básico, se ha asignado a cada alumno una "**categoría base**" diferente de la Wikipedia. La asignación de categorías se encuentra disponible en Aula Global.

Puede encontrar información general sobre las categorías de artículos de la wikipedia aquí. Todas las categorías base asignadas son sub-sub-categorías de la supercategoría principal "*Fundamental Categories*".

Su primera tarea consistirá en seleccionar dos "**categorías de trabajo**". Las categorías de trabajo pueden ser las dos categorías base asignadas al grupo (en el caso de que se trabaje en grupo de dos personas) o bien cualesquiera subcategorías de éstas, con la única condición de que una de las categorías de trabajo no puede ser subcategoría de la otra.

El segundo paso consistirá en descargar un conjunto suficientemente amplio de artículos de la wikipedia relativos a cada categoría de trabajo. Debe descargar al menos 100 documentos de cada categoría de trabajo, aunque es recomendable utilizar una cantidad algo mayor.

Es posible hacer la descarga de datos de forma parcialmente manual utilizando la herramienta de exportación de la Wikipedia. Si el número de artículos directamente asignados a su categoría de trabajo es insuficiente, deberá descargar documentos de sus subcategorías. La herramienta de exportación, así como esta página de exploración de categorías de puede facilitarle esta tarea.

Alternativamente, puede utilizar algunas librerías de python orientadas a la descarga de datos de la wikipedia y otras wikis. En particular:

- Wikipedia-api. Posiblemente la más recomendable. Incluye métodos para descargar todas las entradas asociadas a una categoría.

- wikipedia.

También tiene alguna información sobre la automatización de descargase en la propia Wikipedia (véase, por ejemplo, aquí, o aquí).

### 2.2 Preprocesado de textos

Tras las descarga de datos, deberá extraer cada artículo y aplicar las técnicas de limpieza y homogeneización de los corpus de datos estudiadas en la asignatura.

Por otra parte, observe que al final de cada artículo de la Wikipedia suele aparecer una sección titulada "Categories", con una lista de categorías y subcategorías a las que pertenece. Se pretende determinar si el contenido del artículo proporciona información discriminativa para separar las categorías, sin utilizar la información contenida en esta sección. Por este motivo, debe eliminarla para que no tenga influencia sobre la entrada al clasificador

### 2.3 Modelado de tópicos

La distribución de tópicos de un documento servirá de entrada al clasificador binario. Para ello, necesitará que todos los documentos de las dos categorías utilicen la misma representación. Por este motivo, debe aplicarse un modelo de tópicos conjunto sobre las dos categorías.

Las prestaciones del diseño final dependerán del número de tópicos elegido. Analice la influencia de este parámetro y seleccione el número de tópicos más adecuado siguiendo algún procedimiento de validación.

### 2.4 Clasificación

Para clasificar los artículos de la Wikipedia, puede utilizar cualquier algoritmo de clasificación que considere oportuno. Recuerde que para evaluar las prestaciones del clasificador no puede utilizar datos utilizados durante el entrenamiento. Aplique la metodología de análisis de datos estudiada en la asignatura.

## 3 Extensión

El trabajo de extensión es libre: deberá ampliar el proyecto básico en la dirección que considere más oportuna:

- Automatización de procesos
- Estudio comparativo de diferentes algoritmos de clasificación
- Estudio comparativo con clasificadores no basados en modelos de tópicos.
- Explorar el potencial de técnicas de NLP como el uso de bigramas, part-of-speech tagging, tesauros, etc, (explotando, por ejemplo, la funcionalidad disponible en la librería `NLTK` de Python).
- ...

Tome esta lista como una mera sugerencia, puede elegir cualquier otro tema siempre que encaje dentro del ámbito de la asignatura. En el trabajo de extensión se valorará la creatividad y originalidad en la elección. Si tiene dudas sobre la idoneidad de la extensión elegida, consulte con el profesor. Tenga en cuenta, en todo caso, que el trabajo de extensión constituye un punto sobre la nota final. Evite embarcarse en trabajos de extensión demasiado ambiciosos que puedan comprometer la entrega en plazo del proyecto.

Si el trabajo de extensión lo justifica, puede utilizar otras subcategorías de trabajo (dentro de las categorías base asignadas).

# 4 Entrega

Los alumnos deberán proporcionar los siguientes entregables para la evaluación del proyecto final:

1. Memoria descriptiva del trabajo realizado en formato .pdf y una extensión máxima de 11 páginas (excluyendo únicamente portada y referencias).
2. Script de Python con el código implementado debidamente comentado

Alternativamente a la memoria, los alumnos pueden entregar un notebook de python que resuelva el proyecto y describa la metodología empleada, los experimentos realizados y sus resultados. La extensión máxima del Notebook resuelto y exportado a formato .pdf no podrá exceder de 40 páginas.

La memoria no debe incluir en ningún caso el código implementado, pero sí debe constar de cuatro apartados principales:

- Proyecto básico (max. 8 páginas de memoria, 32 págs de versión pdf del notebook)
- Extensión (max. 2 páginas de memoria, 8 págs de versión pdf del notebook)
- Manual de usuario del código (max. 1 página).
- Reconocimiento de autorías. Inexcusablemente, la memoria debe respetar el principio de reconocimiento de autorías. Si ha utilizado fragmentos de código ajenos o cualquier material procedente de fuentes externas, debe especificarlo claramente en la memoria.

# 5 Evaluación

El proyecto se evaluará de acuerdo con los criterios siguientes:

- Proyecto básico ($1,75$ puntos)
  - Metodología ($0,6$)

  - Calidad de la memoria $(0,75)$
  - Calidad del código $(0,2)$
  - Reproducibilidad de los resultados $(0,2)$

- Extensión $(0,75$ puntos$)$
  - Originalidad $(0,3)$
  - Calidad del trabajo $(0,45)$

La entrega se realizará vía Aula Global. La fecha límite será el **Domingo, 23 de diciembre**, a las 23,55 horas.