# Tareas abordables en el TFM

## Proceso de Extracción de PUBMED (literatura procedente de MEDLINE)

El objetivo de esta tarea es contar con la base sobre la que realizar cualquier operación posterior en el marco de la Bioinformática basada en el procesamiento del Lenguaje Natural.

Para cualquier trabajo a posteriori relacionado con el ámbito del NLP sobre los conceptos de genes, mutaciones, fármacos y enfermedades, el punto de partida es contar con toda la literatura en local, incluyendo toda la información posible junto con sus abstracts y los textos completos. La primera tarea abordable sería constituir un DataSet en un entorno HDFS con las informaciones de PUBMED y PMC, con una carga inicial y los mecanismos para realizar los incrementos diarios, semanales, …

Esta tarea se corresponde con un proceso de extracción que se puede realizar con alguna de las herramientas usadas en el Máster (Flume, OOZIE, …)

## Proceso de estandarización de entidades biomédicas

El objetivo de esta tarea es contar con los vocablos estándares y sus acepciones para poder ser utilizados en cualquier procedimiento posterior. Se debiera de partir de los estándares siguientes: (<https://www.genenames.org/useful/links/>)

* HGNC: *Nombrado estándar de genes*
* HGVS y HGMD: *Nombrado de mutaciones genéticas humanas*
* CIE: *Clasificación Internacional de las Enfermedades*
* GSDD: *Gold Standard Drug Database*

Con estos datasets relacionados se puede proceder a la aplicación de herramientas que permitan la vinculación entre artículos y las entidades biomédicas.

Desconozco el nivel de estandarización de las diferentes bases de datos con respecto a estos u otros estándares. Espero obtener más información próximamente.

## Proceso de anotación de documento basados en vocabularios controlados ya existentes para poder hacer comprobaciones posteriores.

Como se ha visto en el resumen se cuenta con múltiples DataSet que vinculan modificaciones genéticas, fármacos, enfermedades, … que se usan para validar procesos de minería como el desarrollo de clasificadores basados en IA, redes neuronales, etc.

El objetivo de este proceso sería anotar el DataSet de publicaciones con la información aportada por estos dataset (BRONCO, MutationFinder, Extractor of mutations, tmVar, pubtator, clinvar, cosmic, oncoKB, pharmgKB, …).

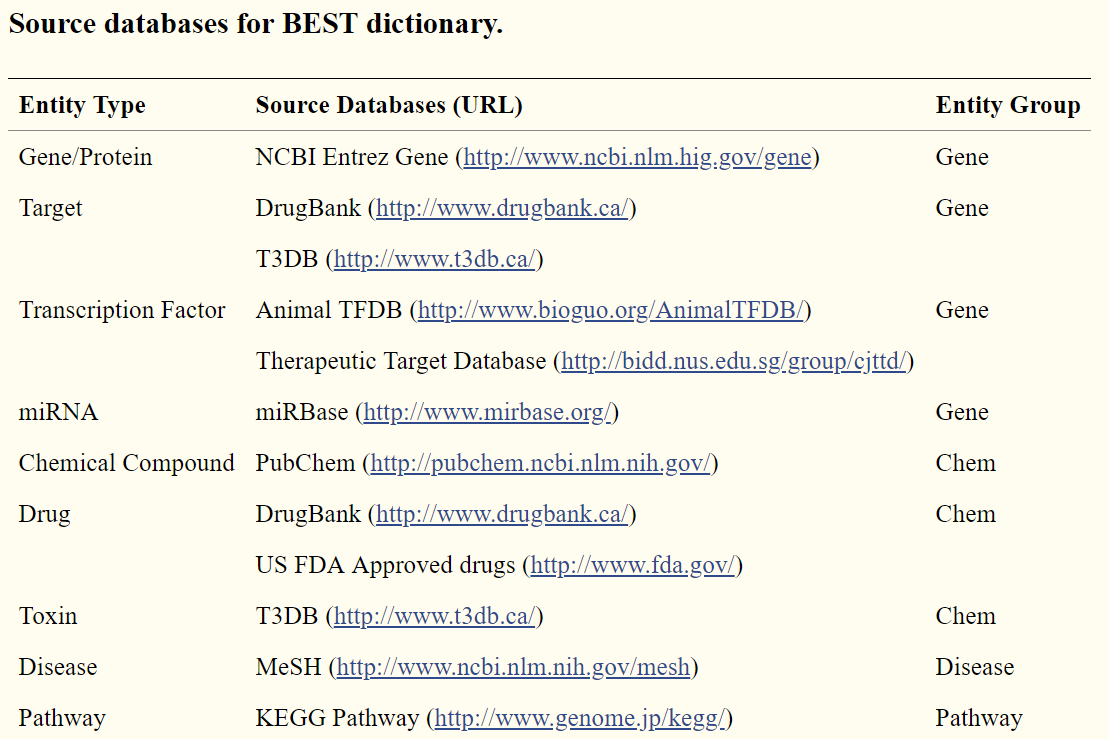
## Proceso de anotación de las entidades biomédicas con nombres no estándar.

El objetivo es incorporar a los diccionarios anteriores, nomenclatura procedente de otras denominaciones, localizados por diversas fuentes de datos, al estilo de lo propuesto en BEST EE. De esta manera se mejorarán los vocabularios controlados para la localización de entidades (Procesos NER en NLP).

El mecanismo propuesto es añadir a cada dataset, las columnas de cada trabajo incluido con denominaciones adicionales y cualquier información. Adicionalmente de cada tupla se deberá anexar una columna resumen que permita a un proceso NER utilizar toda la información para extraer la entidad de un texto. Esta columna puede seguir la sintaxis del diccionario BEST EE, para de esta manera ya contar con el procesador NER.

En todo caso sería aconsejable trasladar el proceso BEST EE a un modelo compatible con Spark NLP para poder incorporarlo a pipelines NLP.

De cara al TFM, se debiera de crear un mecanismo para incluir vocablos controlados para entidades biomédicas, grupos y relaciones entre ellas.



## Proceso de anotación de documento basado en NLP con el clasificador NER descrito.

Con este paso se contaría con un proceso de extracción de entidades (NER) de un texto/documento. Sin mucha mas problemática se podría generar un dataset de PubMed anotado con información genómica directamente extraída del contenido.

El objetivo sería desarrollar un pipeline en Spark NLP que produzca un dataset reducido con textos vinculados de alguna manera a los vocablos estándar anteriores.

## Buscador sobre índices creados sobre los dataset anotados

Este es el paso final. Tanto BEST como VarDrugPub responden al mismo tipo de solución de visualización de datos. El buscador junto con la navegación por las relaciones tanto entre entidades biomédicas de los vocabularios relacionados, como sobre documentación relacionada se deben de realizar con un objetivo claro, que a tenor de lo hablado en este momento no esta disponible. En este sentido la solución obtenible sería una implantación simple de SolR con los índices sobre los datasets anotados anteriores