# CLASIFICADOR

## DISTANCIAS Y FRECUENCIAS

1. BEST Scoring
   1. Se parte de un texto: **“**The oral **BRAF** inhibitors **vemurafenib** (formerly PLX4032) and **dabrafenib** (formerly GSK2118436) induce a high frequency of tumor regressions in patients with **BRAF** **V600E** mutant metastatic **melanoma** and **vemurafenib** improves overall survival compared with chemotherapy**”.** Utilizando **BEST EE** obtenemos
      1. La mutaciónV600E.
      2. Las posibles relaciones con (V600E, vemurafenib) y (V600E, dabrafenib).
      3. Entidad Biomédicas cercanas: (gen BRAF, enfermedad melanoma)
   2. Buscar en BEST por todas las Entidades Biomédicas de la frase, excluyendo las del mismo tipo de la puesta en la consulta.
   3. BEST es interrogado sobre una consulta lanza el **extractor BEST EE** y obtiene las entidades relacionadas y los artículos de PubMed que se corresponden a la búsqueda.Se guardan los scores de los resultados de las búsquedas en BEST Search:
      1. **BSSM**: Buscar en BEST con el nombre normalizado (V600E)
      2. **BSSA**: Combinar todas con AND (V600E and BRAF and melanoma)
      3. **BSSO**: Combinar todas con OR (V600E or BRAF or melanoma)
      4. **BSSAO**: Combinar la entidad de la consulta con cada una de las entidades relacionadas (V600E and BRAF or V600E and melanoma)
2. FRECUENCIA Scoring
   1. Calculo de las distancias según cuatro características (Silva et al.):
      1. **NTDS** (Nearness to Target Disease Score). Número de co-ocurrencia de una enfermedad y una mutación.
      2. **TDFS** (Target Disease Frequency Score). Frecuencia de la enfermedad objetivo.
      3. **ODFS** (Other Disease Frequency Score). Frecuencia de la enfermedad más frecuente exceptuando la enfermedad objetivo.
      4. **DMCS** (Diseasemutation Co-occurrence Score) Binario que indica cuando una mutación y la enfermedad cercana a la mutación se mencionan en la misma frase.

Con las características de BEST y de FRECUENCIAS se obtienen las distancias para el modelo de clasificación.

## Datasets y Algortimos

1. DATASETS
   1. ENTRENAMIENTO mutación-gen:
      1. **BRONCO.** Corpus de 108 artículos: Mutaciones y sus asociaciones con genes, enfermedades, medicamentos y cell-lines. El corpus se trata con **BEST EE** para obtener una versión estandarizada de los genes y los medicamentos. Adicionalmente se marca si existe o no relación entre los datos de las tuplas.

Desde BRONCO se localiza el nombre normalizado de la Entidad Biomédica “mutación” de la consulta (p.e. de la mutación Val600Glu, BRONCO nos devuelve la mutación normalizada V600E) y se obtienen las entidades relacionadas (BRAF y melanoma).

* + 1. **PUBTATOR**. Corpus de PubMed a partir de los abstract que contienen genes, medicamentos y mutaciones usando PUBTATOR a partir del PMID. Se evitan los 27 millones de documentos. Se buscan frases que contengan pares mutación-gen para los siguientes pasos.
    2. **CLINVAR y COSMIC**. Bases de datos que filtran por PMID con pares mutación-gen. De la lista de PUBTATOR se marcan para el entrenamiento positivo los que aparecen en estas bases de datos.
    3. **UNI-PROT**. Lista de pares mutación-gen considerados erróneos. Se marcan las mutaciones de esta lista para el entrenamiento negativo.
    4. El resto se marcarán como negativos.
  1. ENTRENAMIENTO mutación- medicamento:
     1. **PHARMGKB**. Relación de PMID de PubMed con relaciones mutación-medicamento. Si aparece en esta lista se marca para el entrenamiento positivo y sino para el negativo.
     2. **OncoKB**. 234 relaciones de curaciones de cáncer con esta relación mutación-medicamento. Este DataSet servirá para revisar resultados.

1. WORD2VEC
   1. Se ha entrenado un word2vec de dos maneras:
      1. Con la información de los abstract de PubMed obteniendo 300 dimensiones. La construcción agrupa como palabras simples palabras compuestas del ámbito biomedico: “non-small cell lung cancer” se establece como ““non-small\_cell\_lung\_cancer”.
      2. Con la información de Google News.
      3. Se añaden 10 dimensiones para las entidades biomédicas a comprobar: (Entidades objetivo) mutación, medicamento, gen, (Entidades no objetivo) gen, medicamento, enfermedad,
      4. Se añaden 10 dimensiones para las establecer la posición de una palabra en las 10 dimensiones anteriores.
2. MODELOS DE CLASIFICACIÓN
   1. Clasificador Árbol de decisión.
   2. Clasificador Random Forest.
   3. CNN Clasificador Redes Neuronales Convulacionales: Cada palabra de una frase se empotra en WORD2VECT

El algoritmo con el modelo se construye sobre TensorFlow.

## VARDRUGPUB

1. A partir de la colección de documentos proporcionada por PUBTATOR se extraen todas las frases de los abstracts que incluyan una mutación y un medicamento. A cada frase se le aplica la clasificación del modelo CNN que indicará si existe la relación o no. Adicionalmente se añaden a la tupla el conjunto de entidades biomédicas detectadas en el documento (mutaciones, genes, medicamentos, …)
2. VARDRUGPUB contará con con relaciones entre mutaciones y medicamentos, que contrastadas con OncoKB dan los siguientes resultados:
   1. De las 113 relaciones mutación-medicamento de OncoKB que aparecen en los abstract, 66 aparecen en VARDRUGPUB, 33 no ocurren a nivel de frase, 2 no se han encontrado por problemas del algoritmo NER y 6 directamente no se han encontrado.
   2. Muchas relaciones de OncoKB no se encuentran en los abstract 234 -113 = 121

# Utilidades por generar por el TFM

## PUBMED

([www.ncbi.nlm.nih.gov/pubmed/](https://www.ncbi.nlm.nih.gov/pubmed/), [europepmc.org/](https://europepmc.org/))

Base de datos recopilada a partir de los artículos Médicos. Conviene comprobar si la versión europea es más completa. Cuenta con dos informaciones:

1. Artículos con Abstract del ámbito médico. Se puede obtener con actualizaciones diarias directamente por FTP o por API. Se basa en un XML con toda la información del artículo.
2. Artículos completos gratuitos, conocidos como PMC que permite obtener un fichero comprimido por cada articulo con los PDF, las imágenes. Constituyen un subconjunto de los artículos de PubMed.

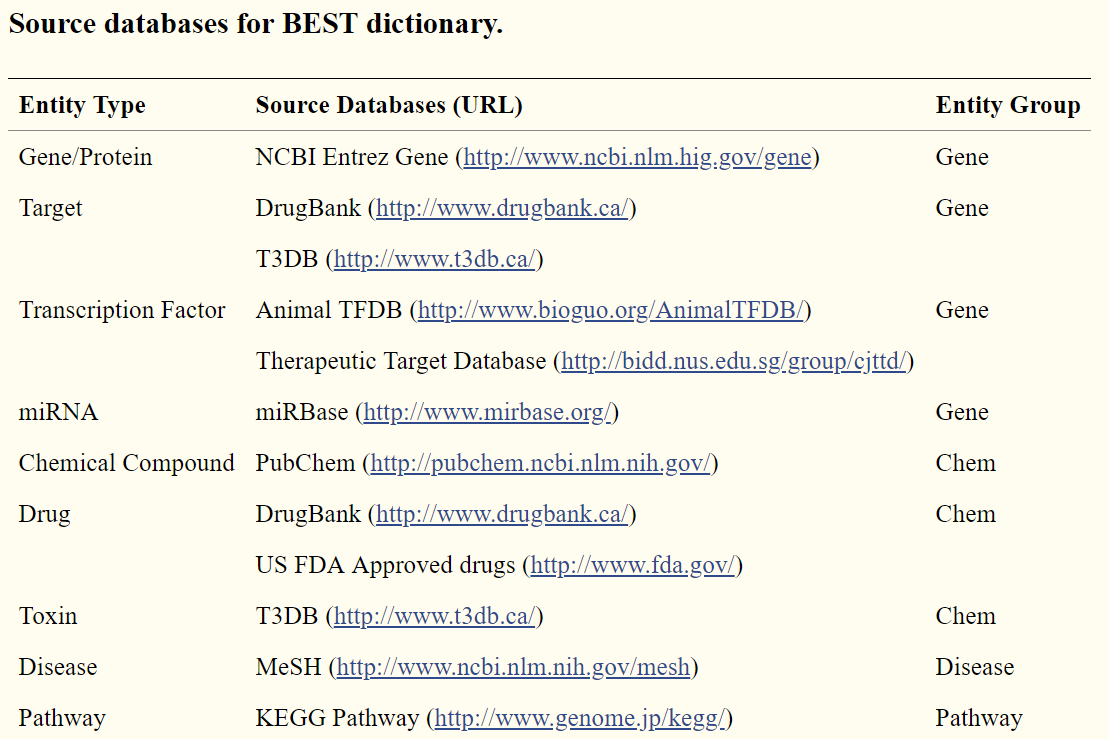
Las tareas necesarias para poder tratar esta información serán:

1. Spider de descarga de la Base inicial
2. Planificador de descarga de las actualizaciones diarias
3. Tratamiento del XML de cada abstract
4. Estructura de la información (BD, dataset, …) para gestionar esta información

## BEST EE

([infos.korea.ac.kr/bioentityextractor/](http://infos.korea.ac.kr/bioentityextractor/))

Se trata de una utilidad de extracción de entidades NER. Se trata de una utilidad que incorpora varios diccionarios para extraer entidades biomédicas. Incorpora la información de las siguientes fuentes:



El objetivo de este componente es contar con un modelo de entidades biomédicas para las diferentes aplicaciones dentro de la arquitectura.

Las tareas necesarias para poder contar con esta utilidad en un entorno NLP basado en Spark serán:

1. La utilidad BEST EE, se utiliza en los cuadros de búsqueda de BEST o de VarDrugPub para facilitar el tratamiento de la búsqueda. Se puede utilizar esta utilidad o realizar una más adecuada ya que en el artículo se han detectado fallos por la no detección de entidades que proceden de esta utilidad.
2. La utilidad BEST EE también se utiliza en el clasificador, por lo que se puede avanzar en dos frentes:
   1. Reutilizar la información distribuida junto a BEST EE dentro de la arquitectura NLP.
   2. Transformar el diccionario de Best EE a un diccionario NER compatible con Spark NLP
   3. Capturar directamente los datos de las fuentes anteriores, junto con otras que se pudieran encontrar y generar el modelo NER compatible con Spark NLP.

En mi opinión se debiera ir hacia un modelo biomédico propio que incluya más relaciones entre las bioentidades y en el que se vaya asentando conocimiento adicional. En este momento solo se cuenta con la lista de cada tipo.

## BEST

([best.korea.ac.kr](http://best.korea.ac.kr), [best.korea.ac.kr/help/BEST\_Guide.pdf](http://best.korea.ac.kr/help/BEST_Guide.pdf))

Se trata de un buscador sobre un índice en Solr. El índice se constituye con la información de PUBMED a la que se anotan campos específicos para las BioEntidades localizadas por BestEE. Se trata por lo tanto de un PipeLine NLP que, a partir de las publicaciones, utilizando un buscador NER rellena un índice que es directamente utilizado usando Solr. La pantalla de búsqueda se ha caracterizado, pero a efectos operativos no se requiere ya que interesa más la consulta por API para consultar la información por los algoritmos.

Las tareas necesarias para poder contar con esta utilidad en un entorno NLP basado en Spark serán:

1. Creación de un pipeline en Spark NLP que anote los artículos de PUBMED. La anotación en el caso de BEST utiliza un simple algoritmo NER (BEST EE) utilizando un diccionario concreto. En el caso del TFM se podía utilizar alguna información procedente de los diccionarios controlados de otras bases de datos (como PharmGKB, HGNC, …).
2. Sobre este pipeline se procesaría toda la base de PubMed a nivel de abstract y/o a nivel de texto completo, obteniendo los elementos requeridos para alimentar un índice Lucene.
3. Vincular el indexador a los resultados encontrados en el proceso anterior y presentarlos para su búsqueda con SOLR.
4. El principal hándicap está en la integración de todos estos diccionarios controlados, que debiera de proporcionar información sobre el nivel de integración que existe en las denominaciones de todos los conceptos relacionados por las bio entidades detectables por nombre.

## VARDRUGPUB

([vardrugpub.korea.ac.kr](http://vardrugpub.korea.ac.kr))

Se trata del entregable final del artículo. Se trata de un proyecto Web creado al efecto que integra utilidades de visualización de información bioinformática ([cytoscape.org/](https://cytoscape.org/)). Se encarga de encontrar la información clasificada utilizando el algoritmo clasificador descrito por el artículo.

Las tareas necesarias para poder contar con esta utilidad en un entorno serán:

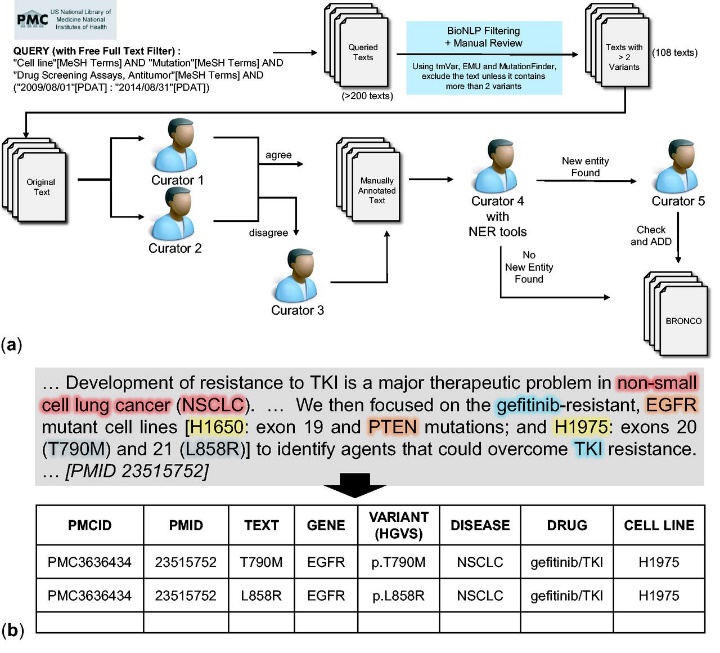
1. Las tareas 1,2 y 3 de BEST, en este caso utilizando el clasificador definido en el marco del TFM.
2. Aquí la complejidad procede de la definición de un clasificador mas o menos sencillo que utilice información de los DataSet que se relatan a continuación. Si se ha logrado una buena integración de los diccionarios controlados, los datasets pueden ofrecer un buen resultado para el clasificador.

El clasificador del artículo está basado en anotaciones sobre las bioentidades, lo que a mi juicio queda un poco limitado para los resultados que se podrían obtener. La existencia de resúmenes de la relación entre bioentidades permitiría hacer un clasificador que calcule distancias en frases/parrafos a partir de resúmenes. Para ello se debiera elaborara un clasificador NLP mas complejo que fuese capaz de transformar frases utilizando las bioentidades y determinados verbos que infieran la relación entre estas.

# Informaciones y DataSets para su tratamiento

## BRONCO

([infos.korea.ac.kr/bronco/](http://infos.korea.ac.kr/bronco/))

Se trata de un Dataset que relaciona entidades Biomédicas con papers de PubMed, que incluye el marcado manual de estas relaciones y su comparación con otros corpus publicados: MutationFinder, Extractor of mutations y tmVar.

Este DataSet con datos desde agosto de 2009 a agosto de 2014, se puede utilizar para relacionar bioentidades de forma correcta para entrenamientos positivos.

Las tareas sobre este artefacto responden a la necesidad de transformar el formato a uno específico para el entrenamiento de Spark NLP y de Spark MLLib. Al incluir relaciones, también puede permitir la generación de un modelo utilizable por Spark GraphX.

## PUBTATOR

([www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/](https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/))

Se trata de una utilidad que ha anotado bioentidades de PubMed, similar a BEST. Permite descargar anotaciones en formato [BioC](http://bioc.sourceforge.net/) XML format. Las anotaciones se pueden descargar directamente, por lo que permite obtener un conjunto de Bioentidades relacionadas con publicaciones. Parece que estos datos se actualizan diariamente pero no de manera incremental. Los menos de 15GB se regeneran con las nuevas publicaciones.

Las tareas sobre este artefacto tendrían que ver con la constitución del modelo NER de detección de bioentidades relacionadas con PubMed. Este modelo se puede hacer sobre datos para entrenamientos o mediante consulta a las Bases de Datos del NIH.

## CLINVAR

([www.ncbi.nlm.nih.gov/projects/clinvar/ClinVarDataDictionary.pdf](https://www.ncbi.nlm.nih.gov/projects/clinvar/ClinVarDataDictionary.pdf))

Base de datos accesible como fichero XML donde se incluyen diariamente con las nuevas anotaciones incorporados por expertos. Cuentan con un vocabulario controlado para enfermedades y sus relaciones con genes.

## COSMIC

([cancer.sanger.ac.uk/cosmic](https://cancer.sanger.ac.uk/cosmic))

Base de datos de mutaciones en genes similar a CLINVAR. La identificación de genes se hace por el proyecto HGNC

## HGNC

([www.genenames.org/tools/search/#!/genes](https://www.genenames.org/tools/search/#!/genes), [www.genenames.org/data/gene-symbol-report/#!/hgnc\_id/HGNC:3467](https://www.genenames.org/data/gene-symbol-report/#!/hgnc_id/HGNC:3467))

Proyecto con el nombrado estándar de los genes. Contiene la estructura de nombres, sinónimos y alias de los genes. Puede ser requerido para un NER propio y completo. Es accesible por REST y se puede descargar una consulta.

## PHARMGKB

([www.pharmgkb.org/downloads](https://www.pharmgkb.org/downloads))

Base de relaciones entre Bioentidades: (mutaciones, enfermedades, genes y medicamentos entre otros), entre otros, con resumen del tipo de interacción encontrada por los autores en un artículo. La información está en fichero plano para ser utilizado. Cuenta con varios diccionarios controlados para genes, mutaciones, …

## OncoKB

([oncokb.org/](http://oncokb.org/))

Se trata de una base de datos con una clasificación de mutaciones de genes, medicamentos y tipos de enfermedad). Las relaciones están clasificadas utilizando niveles que indican desde la evidencia clínica, biológica o aprobada hasta los niveles de resistencia.