ENTITY RECOGNITION ON PATENTS

[By Abhilash Gedela]

# Features:

## Feature #1

Provide a RESTFul API to upload an archive of patents and persist the meta-data (Title,Year,Abstract and Entities in Abstract and Text).

## Feature #2

Provide a RESTFul API to delete a template based on the patent id.

# Bonus Features:

Train a custom NER (Named Entity Recognition ) model with DBPedia data provided and compare the entities with the default model provided by Spacy package.

Monodb Document for a Patent:

{"\_id":{"$oid":"5fa84aa113134fd3f6370270"},

"id":325,

"patent\_id":"US06180331B2",

"title":"Photographic element, compound, and process",

"year":2001,

"abstract":"Disclosed is a photographic element comprising a light-sensitive silver halide emulsion layer having associated therewith a cyan \"NB coupler\" having the formula (I):wherein:the term \"NB coupler\" represents a coupler of formula (I) that forms a dye for which the left bandwidth (LBW) using spin-coating is at least 5 nm less than that of the same dye in solution form;Y is H or a coupling-off group;each Z'' and Z\* is an independently selected substituent group where n is 1 to 4 and p is 0 to 2;W2 represents the atoms necessary to complete a carbocyclic ring group; andV is a substituent containing a heterocyclic sulfone or sulfoxide group;provided that the combined sum of the aliphatic carbon atoms in V, all Z'' and all Z\* is at least 8.The element exhibits improved cyan dye hue.",

"default\_entities: ('R1 and R2', 'PRODUCT'), ('0', 'CARDINAL'), ('0', 'CARDINAL'), ('2.Suitable', 'CARDINAL'), ('acyl', 'GPE'), ('acyloxy', 'GPE'), ('alkenyl', 'GPE'), ('alkyl', 'GPE'), ('alkoxy', 'GPE'), ('aryl', 'GPE'), ('carboxy', 'GPE'), ('cyano', 'PERSON'), ('halogen', 'ORG'), ('heterocyclic', 'PERSON'), ('hydroxy', 'PERSON'), ('nitro', 'GPE'), ('oxycarbonyl', 'GPE'), ('oxysulfonyl', 'GPE'), ('sulfonamido', 'PERSON'), ('sulfonyl', 'PERSON'), ('sulfoxide', 'GPE'), ('thio', 'GPE'), ('ureido', 'ORG'), ('alkyl', 'GPE'), ('sulfonyl', 'PERSON'), ('sulfamoyl', 'NORP'), ('nitro', 'PERSON'), ('halogen', 'ORG'), ('R1', 'GPE'), ('R2', 'PRODUCT'), ('5', 'CARDINAL'), ('6', 'CARDINAL'), ('at least one', 'CARDINAL'), ('at least one', 'CARDINAL'), ('benzoxazolyl', 'GPE'), ('chromonyl', 'GPE'), ('furyl', 'GPE'), ('indazolyl', 'ORG'), ('isoquinolyl', 'ORG'), ('morpholinyl', 'GPE'), ('pyranyl', 'GPE'), ('pyrrolidinyl', 'GPE'), ('quinaldinyl', 'GPE'), ('quinazolinyl', 'GPE'), ('quinoxalinyl', 'PERSON'), ('tetrazolyl', 'GPE'), ('tetrahydrofuryl', 'PERSON'), ('thiadiazolyl', 'ORG'), ('thienyl', 'GPE'), ('thiophenyl', 'GPE'), ('W1', 'PERSON'), ('1,2,3-triazole', 'DATE'), ('1,2,4-triazole', 'CARDINAL'), ('one', 'CARDINAL'), ('Chemical Structure', 'ORG'), ('3', 'CARDINAL'), ('1', 'CARDINAL'), ('2;Y', 'CARDINAL'), ('0', 'CARDINAL'), ('4;W1', 'CARDINAL'), ('all Z', 'PRODUCT'), ('III', 'ORG'), ('Chemical Structure', 'ORG'), (\"'4'\", 'DATE'), ('R1 and R2', 'CARDINAL'), ('1', 'CARDINAL'), ('5', 'CARDINAL'), ('atoms;provided', 'DATE'), ('R1', 'GPE'), ('R2', 'PRODUCT'), ('at least 8.Particular', 'CARDINAL'), ('Chemical Structure', 'ORG'), ('R2', 'PRODUCT'), ('8.Typically', 'CARDINAL'), ('R2', 'PRODUCT'), ('Z\"', 'GPE'), ('12 or more with 15', 'CARDINAL'), ('16', 'CARDINAL'), ('Chemical Structure', 'ORG'), ('Chemical Structure', 'ORG'), ('Chemical Structure', 'ORG'), (\"'8'\", 'DATE'), ('Chemical Structure', 'ORG'), (\"'9'\", 'DATE'), ('Chemical Structure', 'ORG'), (\"'10'\", 'DATE'), ('Chemical Structure', 'ORG'), (\"'11'\", 'DATE'), ('Chemical Structure', 'ORG'), (\"'12'\", 'DATE'), ('4-amino-3-methyl', 'QUANTITY'), ('LBW', 'ORG'), ('70 nm', 'QUANTITY'), ('less than 60 nm', 'QUANTITY'), ('less than 650', 'CARDINAL'), ('less than', 'CARDINAL'), ('cyano', 'PERSON'), ('carboxyl', 'PERSON'), ('alkyl', 'GPE'), ('cyclic alkyl', 'GPE'), ('methyl', 'LOC'), ('trifluoromethyl', 'LOC'), ('ethyl', 'LOC'), ('2-butene', 'PERSON'), ('methoxy', 'PERSON'), ('sec', 'ORG'), ('2,4,6-trimethylphenyl', 'DATE'), ('naphthyl', 'GPE'), ('acetamido', 'GPE'), ('2-oxo-5-tetradecylpyrrolin-1-yl', 'CARDINAL'), ('N-methyltetradecanamido', 'PERSON'), ('2,5-dioxo-1-oxazolidinyl', 'GPE'), ('3-dodecyl-2,5-dioxo-1-imidazolyl', 'CARDINAL'), ('N-acetyl-N-dodecylamino', 'ORG'), ('2,4-di', 'CARDINAL'), ('2,5-(di', 'QUANTITY'), ('N-methylureido', 'ORG'), ('N', 'GPE'), ('N', 'GPE'), ('N-dioctadecylureido', 'ORG'), ('N', 'GPE'), ('N', 'ORG'), ('N', 'GPE'), ('N-(2,5-di', 'GPE'), ('methylsulfonamido', 'GPE'), ('benzenesulfonamido', 'GPE'), ('N-methyltetradecylsulfonamido', 'ORG'), ('N', 'GPE'), ('N-methylsulfamoyl', 'ORG'), ('N', 'GPE'), ('N-dipropylsulfamoyl', 'ORG'), ('N-hexadecylsulfamoyl', 'ORG'), ('N', 'GPE'), ('N-methyl', 'ORG'), ('N-dodecylsulfamoyl', 'ORG'), ('N-methylcarbamoyl', 'ORG'), ('N', 'GPE'), ('N-methyl-N-tetradecylcarbamoyl', 'ORG'), ('N', 'ORG'), ('phenoxycarbonyl', 'GPE'), ('butoxycarbonyl', 'GPE'), ('tetradecyloxycarbonyl', 'GPE'), ('ethoxycarbonyl', 'GPE'), ('benzyloxycarbonyl', 'GPE'), ('3-pentadecyloxycarbonyl', 'LOC'), ('sulfonyl', 'PERSON'), ('methoxysulfonyl', 'GPE'), ('octyloxysulfonyl', 'GPE'), ('tetradecyloxysulfonyl', 'GPE'), ('2-ethylhexyloxysulfonyl', 'GPE'), ('methylsulfonyl', 'GPE'), ('octylsulfonyl', 'GPE'), ('2-ethylhexylsulfonyl', 'GPE'), ('dodecylsulfonyl', 'GPE'), ('phenylsulfonyl', 'GPE'), ('4-nonylphenylsulfonyl', 'GPE'), ('methylsulfinyl', 'GPE'), ('octylsulfinyl', 'GPE'), ('2-ethylhexylsulfinyl', 'GPE'), ('dodecylsulfinyl', 'GPE'), ('phenylsulfinyl', 'GPE'), ('4-nonylphenylsulfinyl', 'GPE'), ('thio', 'GPE'), ('ethylthio', 'GPE'), ('octylthio', 'GPE'), ('benzylthio', 'GPE'), ('tetradecylthio', 'GPE'), ('phenylanilino', 'GPE'), ('2-chloroanilino', 'GPE'), ('imino', 'GPE'), ('diethyl', 'GPE'), ('thio group', 'ORG'), ('3', 'CARDINAL'), ('at least one', 'CARDINAL'), ('2-furyl', 'QUANTITY'), ('one', 'CARDINAL'), ('two', 'CARDINAL'), ('up to 48', 'CARDINAL'), ('1', 'CARDINAL'), ('36', 'CARDINAL'), ('less than 24', 'CARDINAL'), ('alkyl', 'GPE'), ('aryl', 'PERSON'), ('8', 'CARDINAL'), ('48', 'CARDINAL'), ('alkyl', 'GPE'), ('aryl', 'GPE'), ('alkylthio', 'GPE'), ('hydroxy', 'PERSON'), ('halogen', 'ORG'), ('alkoxycarbonyl', 'PERSON'), ('aryloxcarbonyl', 'GPE'), ('carboxy', 'GPE'), ('acyl', 'GPE'), ('anilino', 'NORP'), ('alkylsulfonyl', 'GPE'), ('arylsulfonyl', 'GPE'), ('sulfonamido', 'PERSON'), ('1', 'CARDINAL'), ('42', 'CARDINAL'), ('three', 'CARDINAL'), ('three', 'CARDINAL'), ('at least one', 'CARDINAL'), ('at least one', 'CARDINAL'), ('at least one', 'CARDINAL'), ('at least one', 'CARDINAL'), ('at least one', 'CARDINAL'), ('at least one', 'CARDINAL'), ('Research Disclosure', 'GPE'), ('November 1992', 'DATE'), ('Kenneth Mason Publications, Ltd.', 'ORG'), ('Dudley Annex', 'ORG'), ('12a', 'DATE'), ('North Street', 'LOC'), ('Emsworth', 'LOC'), ('ENGLAND', 'GPE'), ('Hatsumi Kyoukai', 'PERSON'), ('94', 'CARDINAL'), ('Mar. 15, 1994', 'DATE'), ('the Japanese Patent Office', 'ORG'), ('Research Disclosure', 'PERSON'), ('June 1994', 'DATE'), ('36230', 'DATE'), ('Research Disclosure', 'PERSON'), ('September 1996', 'DATE'), ('38957', 'DATE'), ('the Research Disclosure', 'ORG'), ('Sections', 'PERSON'), ('V. Various', 'GPE'), ('Sections II', 'WORK\_OF\_ART'), ('VI', 'GPE'), ('Section X(E).Scan', 'LAW'), ('Section', 'ORG'), ('Sections XV', 'PERSON'), ('September 1994', 'DATE'), ('36544', 'ORDINAL'), (('mercaptodiazoles', 'GPE'), ('mercaptooxathiazoles', 'ORG'), ('Chemical Structure', 'ORG'), (\"'13'\", 'DATE'), ('1', 'CARDINAL'), ('about 8', 'CARDINAL'), ('benzyl', 'GPE'), ('phenyl', 'GPE'), ('one', 'CARDINAL'), ('more than one', 'CARDINAL'), ('RII', 'ORG'), ('RIII', 'ORG'), ('1', 'CARDINAL'), ('about 5', 'CARDINAL'), ('1', 'CARDINAL'), ('3', 'CARDINAL'), ('RIV', 'ORG'), ('RV', 'ORG'), ('alkyl', 'GPE'), ('aryl', 'GPE'), ('one', 'CARDINAL'), ('U.S.', 'GPE'), ('4,248,962', 'CARDINAL'), ('4,409,323', 'CARDINAL')",

"custom\_entities":"('than', 'CHEMICAL'), ('sulfate', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('aniline', 'CHEMICAL'), ('than', 'CHEMICAL'), ('acetonitrile', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('polyethylene', 'CHEMICAL'), ('than', 'CHEMICAL'), ('acetonitrile', 'CHEMICAL'), ('phenylsulfone', 'CHEMICAL'), ('sulfur', 'CHEMICAL'), ('sulfur', 'CHEMICAL'), ('purinyl', 'CHEMICAL'), ('quinaldinyl', 'CHEMICAL'), ('benzimidazole', 'CHEMICAL'), ('benzotriazole', 'CHEMICAL'), ('furan', 'CHEMICAL'), ('imidazole', 'CHEMICAL'), ('indazole', 'CHEMICAL'), ('indole', 'CHEMICAL'), ('isoquinoline', 'CHEMICAL'), ('purine', 'CHEMICAL'), ('pyrazole', 'CHEMICAL'), ('pyridine', 'CHEMICAL'), ('pyrimidine', 'CHEMICAL'), ('pyrrole', 'CHEMICAL'), ('quinoline', 'CHEMICAL'), ('thiophene', 'CHEMICAL'), ('pyridine', 'CHEMICAL'), ('pyrimidine', 'CHEMICAL'), ('pyrazole', 'CHEMICAL'), ('aniline', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('sulfur', 'CHEMICAL'), ('ethylene', 'CHEMICAL'), ('2-butene', 'CHEMICAL'), ('pentyl', 'CHEMICAL'), ('acetyl', 'CHEMICAL'), ('acetyl', 'CHEMICAL'), ('benzyloxycarbonyl', 'CHEMICAL'), ('dodecylsulfinyl', 'CHEMICAL'), ('diethylamine', 'CHEMICAL'), ('dodecylamine', 'CHEMICAL'), ('phosphate', 'CHEMICAL'), ('phosphite', 'CHEMICAL'), ('sulfur', 'CHEMICAL'), ('ammonium', 'CHEMICAL'), ('than', 'CHEMICAL'), ('aryloxcarbonyl', 'CHEMICAL'), ('benzothiazole', 'CHEMICAL'), ('those', 'CHEMICAL'), ('catechol', 'CHEMICAL'), ('than', 'CHEMICAL'), ('combine', 'CHEMICAL'), ('formaldehyde', 'CHEMICAL'), ('those', 'CHEMICAL'), ('those', 'CHEMICAL'), ('those', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('bromide', 'CHEMICAL'), ('bromide', 'CHEMICAL'), ('bromide', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('iodide', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('diethylaniline', 'CHEMICAL'), ('diethylaniline', 'CHEMICAL'), ('diethylaniline', 'CHEMICAL'), ('toluidine', 'CHEMICAL'), ('toluene', 'CHEMICAL'), ('than', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('hydroxide', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('dimethylformamide', 'CHEMICAL'), ('carbonate', 'CHEMICAL'), ('toluene', 'CHEMICAL'), ('methanol', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('heptane', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('dimethylformamide', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('pyridine', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('pyridine', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('aniline', 'CHEMICAL'), ('acetate', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('polyethylene', 'CHEMICAL'), ('acetonitrile', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL'), ('acetonitrile', 'CHEMICAL'), ('polyethylene', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('polystyrene', 'CHEMICAL'), ('sulfate', 'CHEMICAL'), ('carbonate', 'CHEMICAL'), ('chloride', 'CHEMICAL'), ('bromide', 'CHEMICAL'), ('ammonium', 'CHEMICAL'), ('Ammonium', 'CHEMICAL'), ('Ammonium', 'CHEMICAL'), ('than', 'CHEMICAL'), ('than', 'CHEMICAL')"}

We can clearly see in the response that the chemical compounds are tagged incorrectly with the built in NER model.

And the ner results with built in models is highly unpredictable.

But the custom model which we trained with DB-Pedia data is tagging correctly.

# Technology and design Decisions:

DJANGO\_REST\_FRAMEWOK :

As it is said to provide a REST API, I have used DRF for API development for the following reasons.

* Web-Browsable API which help in quick development
* In built Authentication
* Supports serialization of both ORM and Non-ORM data sources
* Extensive documentation

REDIS :

Celery uses brokers to pass messages between a Django project and the celery workers.

I have used these asynchronous tasks to handle large number of incoming requests. Where the processing is done on the background.

CELERY:

Celery communicates via messages, usually using a broker to mediate between clients and workers. To initiate a task the client adds a message to the queue, the broker then delivers that message to a worker.

As it was mentioned clearly in the problem statement that the code should handle any number of archives. So I have decided to process the uploaded files asynchronously which can be achieved by celery.

SPACY:

Used spacy to find the entities and trained a custom NER(Named Entity Recognition) model.

It provides the fastest and most accurate syntactic analysis of any NLP library released to date. It also offers access to larger word vectors that are easier to customize. For an app builder mindset that prioritizes getting features done, spacy would be the better choice.

# Scalability and Maintainability

* Application can be scaled with multiple workers where many tasks can be processed simultaneously. And reduce the waiting time of each task in the queue.
* Elastic Search(NO SQL Database) can be integrated where search capabilities of text are enhanced and the text can be processed before saving to the database with analyzers where data preprocessing is faster.
* The custom trained model for NER is stopped early because of computation. If the same model is trained on a GPU ( Spacy supports GPU) we can achieve better results .
* We can use multipart uploading. instead of using local file storage we can use S3 buckets on AWS
* Containerize the services using docker.

# Conclusion:

As asked in the challenge two API’s have been created using DRF (Django Rest Framework)

Entities are read from the text and persisted in the DB

Custom NER model has bee built and deployed in the application

Chemical Entities have been recognized and persisted.

Provided with a JSON with collection of all documents from mongo DB. You can find the output.json file and check the response. JSON format will like:

"\_id": {

"$oid": ""

},

"id": 310,

"patent\_id": "",

"title": “",

"year": ,

“abstract”: “”,

“default\_entities”: “”,

“custom\_entities”:””