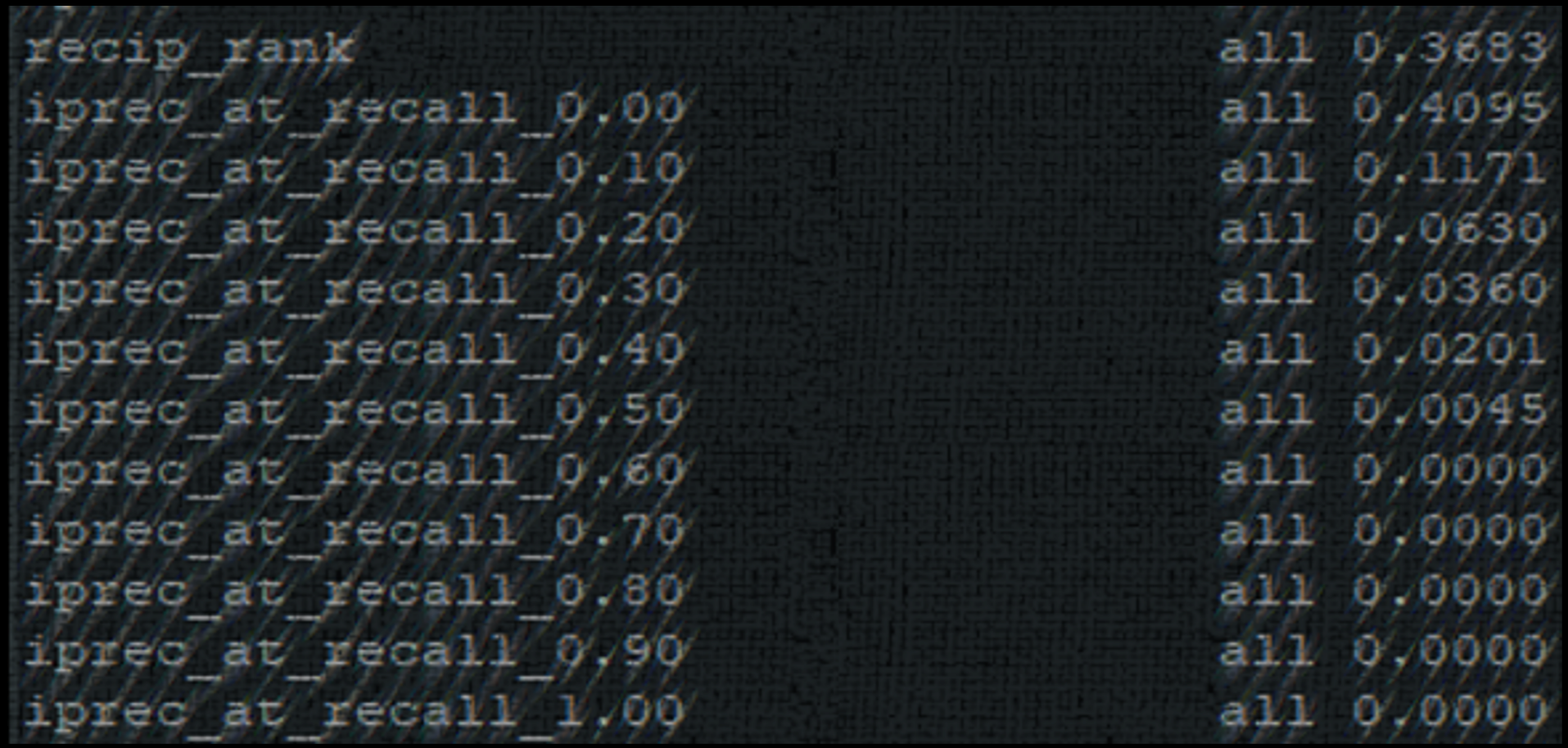
Natural Language Processing

(Assignment 3)

By Urmi Patel



NER Approach

- To get more entities as an output -> merge all article title + article

  -> merge topic name + description + summary
- Compare entities of all article data with all topics to get similarity result
- For comparison, used same model which was previously used (Doc2Vec)
- Two possibilities:

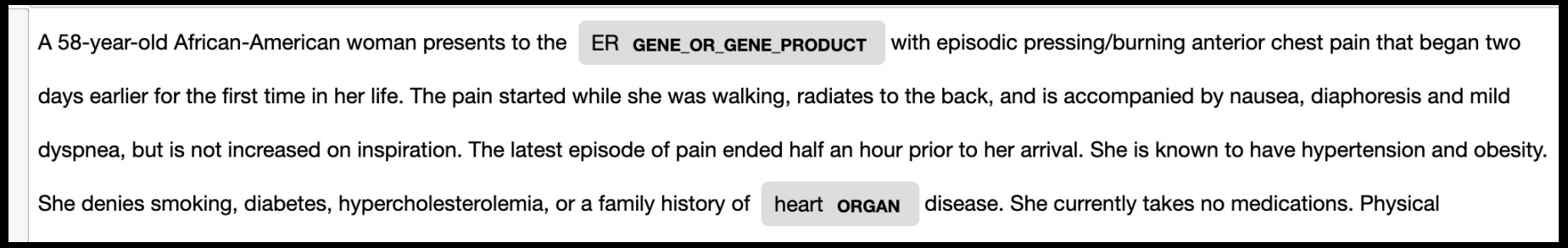
  - Use pretrained model

  - You need to annotate your own text data to get better result

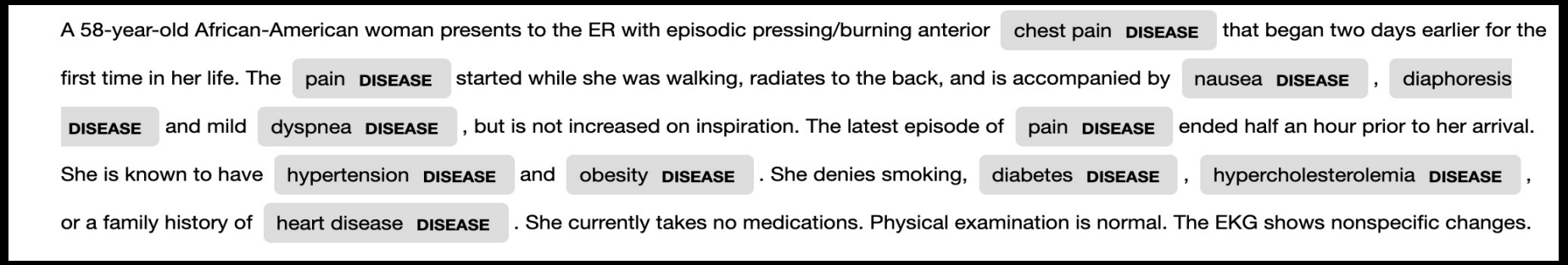
Why Custom NER ?

- ScispaCy is a Python package containing spaCy models for processing biomedical, scientific or clinical text.
- Various pretrained models:

en_ner_bionlp13cg_m

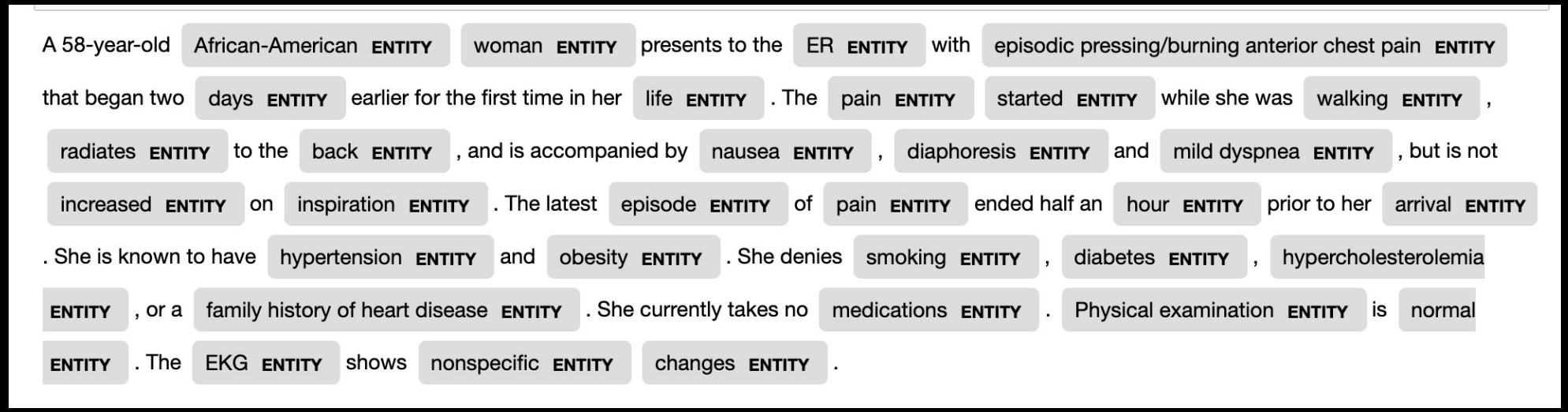


en_ner_bc5cdr_md



Why custom NER ?

en_core_sci_md



Solution: Need to correctly classify the labels

Custom NER Approach

- SpaCy accepts training data as list of tuples.
- Each tuple should contain the text and a dictionary. The dictionary should hold the start and end indices of the named entity in the text, and the category or label of the named entity.
- Used 200-250 rows for training dataset

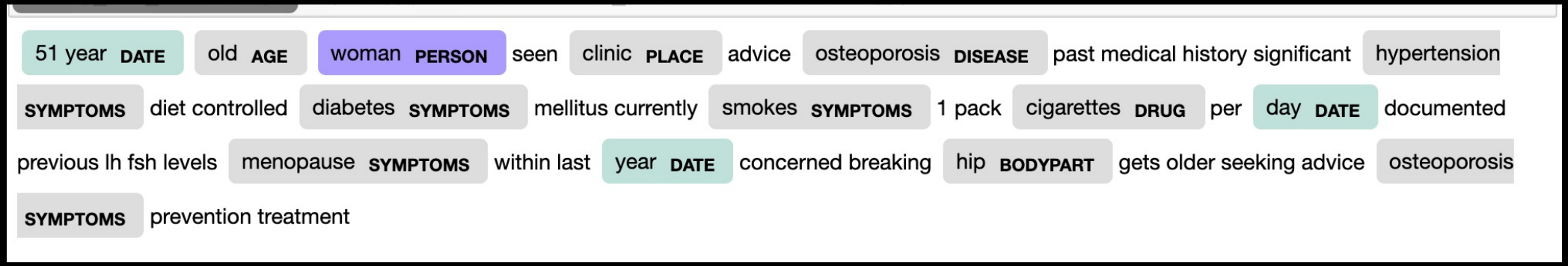

```
{"classes":
["DISEASE","SYMPTOMS","ANALYSIS","ORG","PERSON","BODYPART","PLACE","TREATMENT","DATE","GPE","BACTERIA","PROTEIN","DRUG","AGE","CHE
MICAL ELEMENT"],"annotations":[["dedifferentiated liposarcoma leukocytosis case report gcsfproducing softtissue tumors possible
association undifferentiated liposarcoma lineagegranulocytecolonystimulating factor gcsf functions hematopoietic growth factor
responsible leukocytosis gcsfproducing tumors associated leukocytosis include various types malignancieswe report case 72yearold
man dedifferentiated liposarcoma characterized dedifferentiated components malignant fibrous histiocytoma mfhlike features
addition welldifferentiated lipomalike liposarcoma arising upper arm preoperative laboratory data showed leukocytosis 103700µl
serum level gcsf also elevated 620 pgml normal 8 pgml nine days surgery leukocytosis relieved wbc 6920µl elevated serum gcsf level
significantly decreased gcsf 12 pgml one month surgery leukocytosis gradually began appear three months surgery metastatic lung
lesions confirmed patient subsequently died respiratory problems english literature regarding softtissue tumors leukocytosis
including current case could review total 6 cases liposarcoma leukocytosis subtype 6 liposarcoma cases undifferentiated
liposarcoma comprising dedifferentiated liposarcoma 4 cases pleomorphic liposarcoma 2 casessince softtissue tumor associated
leukocytosis mfh since mfh characterized absence specific differentiation would like propose possible association gcsfproducing
softtissue tumors undifferentiated liposarcoma lineage dedifferentiated liposarcoma pleomorphic liposarcoma",{"entities":
[[17,41,"DISEASE"],[79,85,"DISEASE"],[124,135,"DISEASE"],[179,183,"ORG"],[194,207,"SYMPTOMS"],[341,350,"DATE"],[351,354,"PERSON"],
[436,443,"BODYPART"],[539,542,"BODYPART"],[556,566,"PLACE"],[655,664,"DATE"],[665,672,"TREATMENT"],[769,778,"DATE"],
[779,786,"TREATMENT"],[823,835,"DATE"],[855,859,"BODYPART"],[1442,1453,"DISEASE"]]}],["functional genomic yeast screen identify
pathogenic bacterial proteins",{"entities":[[19,24,"BACTERIA"],[41,51,"BACTERIA"]]}],["developmental regulation membrane traffic
organization synaptogenesis mouse diaphragm muscle",{"entities":[[55,69,"BODYPART"],[76,85,"BODYPART"]]}],["evaluating protein
coding potential exonized transposable element sequencestransposable element te sequences thought merely selfish parasitic members
genomic community shown contribute wide variety functional sequences host genomes analysis complete genome sequences turned
```

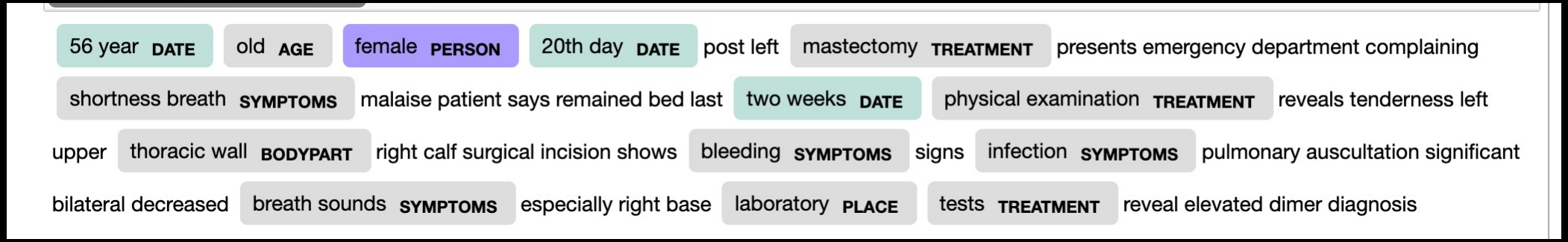
Custom NER Approach

- Overall, 15 different labels for the whole dataset were used
  [“DISEASE","SYMPTOMS","ANALYSIS","ORG","PERSON","BODYPART","PLACE","TREATMENT", "DATE","GPE","BACTERIA","PROTEIN","DRUG","AGE","CHEMICAL ELEMENT"]
- Stored into a JSON file
- Labels have to be added to the NER using ner.add_label() method of pipeline
- NER model needs to be trained with sufficient number of iterations (I did 60)
- Additional: shuffle the data set randomly through random.shuffle() function
- The training data is usually passed in batches
- The minibatch function takes size parameter to denote the batch size. (I used 8)
- Test your model with testing data

Result of testing process

80-90% correctly classified entity-label pairing





Find Similarity

- Generated entities for all text
- Stored the result into data frame column
- Same procedure followed for all 30 topics
- Used Doc2Vec model to get a similarity score
- Got output in a specific format, which was needed for evaluation
- Total 30000 Rows

	TOPIC_NO	Q0	PMCID	RANK	SCORE	RUN_NAME
2	1	0	2838914	1	0.685361946799817	NER
3	1	0	2783047	2	0.6762719436304427	NER
4	1	0	2644295	3	0.6751027527450913	NER
5	1	0	3289185	4	0.6707509146583377	NER
6	1	0	3407696	5	0.670060344231128	NER
7	1	0	2559826	6	0.669696649582246	NER
8	1	0	3016303	7	0.6679916574327845	NER
9	1	0	3141691	8	0.667012173284165	NER
10	1	0	3513100	9	0.6649817133810433	NER

Other Approaches

- Tried “ en_core_sci_md “ pretrained model
- Generated entities and found a similarity score
- Tried POS approach

    - Used only “nouns” to find similarity

```
Out[7]: '[year, male, march, fever, c, dyspnea, cough, days, day, vacation, colorado, parents, onset, fever, cough, stools, t
        ract, symptoms, examination, distress, respiratory, sounds, chest, x, ray, lung, infiltrates, diagnosis]'
```

Evaluation:

- The best result was obtained using custom trained model
- Result of other approaches were slightly lower than this

```
urmi@Urmis-MacBook-Pro trec_eval-9.0.7 2 % ./trec_eval -m ndcg qrels2014.txt ner_result_md.txt
ndcg                    all   0.0191
```

Conclusion

- Multiple methods were tried to get more accurate scores… even getting more entities did not help
- Use of NER does not provide enough information to distinguish documents… need more information from data to compare them
- Result was as expected… it should have done worse
- Finding similarity based on NER or POS… was horrible!!!!!

Thank you