Lab4

March 11, 2024

1 Lab 4, Group 21

1.0.1 Question 1. a

```python
[20]: from nltk.corpus.reader import ConllCorpusReader
      train = ConllCorpusReader('./CONLL2003', 'train.txt', ['words', 'pos',
       ↪'ignore', 'chunk'])


      training_features = []
      training_gold_labels = []
      for token, pos, ne_label in train.iob_words():
          a_dict = {
              'word': token,
              'pos': pos
          }
          training_features.append(a_dict)
          training_gold_labels.append(ne_label)
```

```python
[21]: training_features, training_gold_labels
```

```python
[21]: ([{'word': 'EU', 'pos': 'NNP'},
        {'word': 'rejects', 'pos': 'VBZ'},
        {'word': 'German', 'pos': 'JJ'},
        {'word': 'call', 'pos': 'NN'},
        {'word': 'to', 'pos': 'TO'},
        {'word': 'boycott', 'pos': 'VB'},
        {'word': 'British', 'pos': 'JJ'},
        {'word': 'lamb', 'pos': 'NN'},
        {'word': '.', 'pos': '.'},
        {'word': 'Peter', 'pos': 'NNP'},
        {'word': 'Blackburn', 'pos': 'NNP'},
        {'word': 'BRUSSELS', 'pos': 'NNP'},
        {'word': '1996-08-22', 'pos': 'CD'},
        {'word': 'The', 'pos': 'DT'},
        {'word': 'European', 'pos': 'NNP'},
        {'word': 'Commission', 'pos': 'NNP'},
        {'word': 'said', 'pos': 'VBD'},
```

```
{'word': 'on', 'pos': 'IN'},
{'word': 'Thursday', 'pos': 'NNP'},
{'word': 'it', 'pos': 'PRP'},
{'word': 'disagreed', 'pos': 'VBD'},
{'word': 'with', 'pos': 'IN'},
{'word': 'German', 'pos': 'JJ'},
{'word': 'advice', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'consumers', 'pos': 'NNS'},
{'word': 'to', 'pos': 'TO'},
{'word': 'shun', 'pos': 'VB'},
{'word': 'British', 'pos': 'JJ'},
{'word': 'lamb', 'pos': 'NN'},
{'word': 'until', 'pos': 'IN'},
{'word': 'scientists', 'pos': 'NNS'},
{'word': 'determine', 'pos': 'VBP'},
{'word': 'whether', 'pos': 'IN'},
{'word': 'mad', 'pos': 'JJ'},
{'word': 'cow', 'pos': 'NN'},
{'word': 'disease', 'pos': 'NN'},
{'word': 'can', 'pos': 'MD'},
{'word': 'be', 'pos': 'VB'},
{'word': 'transmitted', 'pos': 'VBN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'sheep', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'Germany', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'representative', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'the', 'pos': 'DT'},
{'word': 'European', 'pos': 'NNP'},
{'word': 'Union', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'veterinary', 'pos': 'JJ'},
{'word': 'committee', 'pos': 'NN'},
{'word': 'Werner', 'pos': 'NNP'},
{'word': 'Zwingmann', 'pos': 'NNP'},
{'word': 'said', 'pos': 'VBD'},
{'word': 'on', 'pos': 'IN'},
{'word': 'Wednesday', 'pos': 'NNP'},
{'word': 'consumers', 'pos': 'NNS'},
{'word': 'should', 'pos': 'MD'},
{'word': 'buy', 'pos': 'VB'},
{'word': 'sheepmeat', 'pos': 'NN'},
{'word': 'from', 'pos': 'IN'},
{'word': 'countries', 'pos': 'NNS'},
```

```
{'word': 'other', 'pos': 'JJ'},
{'word': 'than', 'pos': 'IN'},
{'word': 'Britain', 'pos': 'NNP'},
{'word': 'until', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'scientific', 'pos': 'JJ'},
{'word': 'advice', 'pos': 'NN'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'clearer', 'pos': 'JJR'},
{'word': '.', 'pos': '.'},
{'word': '"', 'pos': '"'},
{'word': 'We', 'pos': 'PRP'},
{'word': 'do', 'pos': 'VBP'},
{'word': "n't", 'pos': 'RB'},
{'word': 'support', 'pos': 'VB'},
{'word': 'any', 'pos': 'DT'},
{'word': 'such', 'pos': 'JJ'},
{'word': 'recommendation', 'pos': 'NN'},
{'word': 'because', 'pos': 'IN'},
{'word': 'we', 'pos': 'PRP'},
{'word': 'do', 'pos': 'VBP'},
{'word': "n't", 'pos': 'RB'},
{'word': 'see', 'pos': 'VB'},
{'word': 'any', 'pos': 'DT'},
{'word': 'grounds', 'pos': 'NNS'},
{'word': 'for', 'pos': 'IN'},
{'word': 'it', 'pos': 'PRP'},
{'word': ',', 'pos': ','},
{'word': '"', 'pos': '"'},
{'word': 'the', 'pos': 'DT'},
{'word': 'Commission', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'chief', 'pos': 'JJ'},
{'word': 'spokesman', 'pos': 'NN'},
{'word': 'Nikolaus', 'pos': 'NNP'},
{'word': 'van', 'pos': 'NNP'},
{'word': 'der', 'pos': 'FW'},
{'word': 'Pas', 'pos': 'NNP'},
{'word': 'told', 'pos': 'VBD'},
{'word': 'a', 'pos': 'DT'},
{'word': 'news', 'pos': 'NN'},
{'word': 'briefing', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'He', 'pos': 'PRP'},
{'word': 'said', 'pos': 'VBD'},
{'word': 'further', 'pos': 'JJ'},
{'word': 'scientific', 'pos': 'JJ'},
```

{'word': 'study', 'pos': 'NN'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'required', 'pos': 'VBN'},
{'word': 'and', 'pos': 'CC'},
{'word': 'if', 'pos': 'IN'},
{'word': 'it', 'pos': 'PRP'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'found', 'pos': 'VBN'},
{'word': 'that', 'pos': 'IN'},
{'word': 'action', 'pos': 'NN'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'needed', 'pos': 'VBN'},
{'word': 'it', 'pos': 'PRP'},
{'word': 'should', 'pos': 'MD'},
{'word': 'be', 'pos': 'VB'},
{'word': 'taken', 'pos': 'VBN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'European', 'pos': 'NNP'},
{'word': 'Union', 'pos': 'NNP'},
{'word': '.', 'pos': '.'},
{'word': 'He', 'pos': 'PRP'},
{'word': 'said', 'pos': 'VBD'},
{'word': 'a', 'pos': 'DT'},
{'word': 'proposal', 'pos': 'NN'},
{'word': 'last', 'pos': 'JJ'},
{'word': 'month', 'pos': 'NN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'EU', 'pos': 'NNP'},
{'word': 'Farm', 'pos': 'NNP'},
{'word': 'Commissioner', 'pos': 'NNP'},
{'word': 'Franz', 'pos': 'NNP'},
{'word': 'Fischler', 'pos': 'NNP'},
{'word': 'to', 'pos': 'TO'},
{'word': 'ban', 'pos': 'VB'},
{'word': 'sheep', 'pos': 'NN'},
{'word': 'brains', 'pos': 'NNS'},
{'word': ',', 'pos': ','},
{'word': 'spleens', 'pos': 'NNS'},
{'word': 'and', 'pos': 'CC'},
{'word': 'spinal', 'pos': 'JJ'},
{'word': 'cords', 'pos': 'NNS'},
{'word': 'from', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'human', 'pos': 'NN'},
{'word': 'and', 'pos': 'CC'},
{'word': 'animal', 'pos': 'NN'},

```
{'word': 'food', 'pos': 'NN'},
{'word': 'chains', 'pos': 'NNS'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'a', 'pos': 'DT'},
{'word': 'highly', 'pos': 'RB'},
{'word': 'specific', 'pos': 'JJ'},
{'word': 'and', 'pos': 'CC'},
{'word': 'precautionary', 'pos': 'JJ'},
{'word': 'move', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'protect', 'pos': 'VB'},
{'word': 'human', 'pos': 'JJ'},
{'word': 'health', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'Fischler', 'pos': 'JJR'},
{'word': 'proposed', 'pos': 'VBN'},
{'word': 'EU-wide', 'pos': 'NNP'},
{'word': 'measures', 'pos': 'VBZ'},
{'word': 'after', 'pos': 'IN'},
{'word': 'reports', 'pos': 'NNS'},
{'word': 'from', 'pos': 'IN'},
{'word': 'Britain', 'pos': 'NNP'},
{'word': 'and', 'pos': 'CC'},
{'word': 'France', 'pos': 'NNP'},
{'word': 'that', 'pos': 'WDT'},
{'word': 'under', 'pos': 'IN'},
{'word': 'laboratory', 'pos': 'NN'},
{'word': 'conditions', 'pos': 'NNS'},
{'word': 'sheep', 'pos': 'NN'},
{'word': 'could', 'pos': 'MD'},
{'word': 'contract', 'pos': 'VB'},
{'word': 'Bovine', 'pos': 'NNP'},
{'word': 'Spongiform', 'pos': 'NNP'},
{'word': 'Encephalopathy', 'pos': 'NNP'},
{'word': '(', 'pos': '('},
{'word': 'BSE', 'pos': 'NNP'},
{'word': ')', 'pos': ')'},
{'word': '--', 'pos': ':'},
{'word': 'mad', 'pos': 'JJ'},
{'word': 'cow', 'pos': 'NN'},
{'word': 'disease', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'But', 'pos': 'CC'},
{'word': 'Fischler', 'pos': 'NNP'},
{'word': 'agreed', 'pos': 'VBD'},
{'word': 'to', 'pos': 'TO'},
{'word': 'review', 'pos': 'VB'},
```

```
{'word': 'his', 'pos': 'PRP$'},
{'word': 'proposal', 'pos': 'NN'},
{'word': 'after', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'EU', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'standing', 'pos': 'NN'},
{'word': 'veterinary', 'pos': 'JJ'},
{'word': 'committee', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': 'mational', 'pos': 'JJ'},
{'word': 'animal', 'pos': 'NN'},
{'word': 'health', 'pos': 'NN'},
{'word': 'officials', 'pos': 'NNS'},
{'word': ',', 'pos': ','},
{'word': 'questioned', 'pos': 'VBD'},
{'word': 'if', 'pos': 'IN'},
{'word': 'such', 'pos': 'JJ'},
{'word': 'action', 'pos': 'NN'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'justified', 'pos': 'VBN'},
{'word': 'as', 'pos': 'IN'},
{'word': 'there', 'pos': 'RB'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'only', 'pos': 'RB'},
{'word': 'a', 'pos': 'DT'},
{'word': 'slight', 'pos': 'JJ'},
{'word': 'risk', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'human', 'pos': 'JJ'},
{'word': 'health', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'Spanish', 'pos': 'NNP'},
{'word': 'Farm', 'pos': 'NNP'},
{'word': 'Minister', 'pos': 'NNP'},
{'word': 'Loyola', 'pos': 'NNP'},
{'word': 'de', 'pos': 'NNP'},
{'word': 'Palacio', 'pos': 'NNP'},
{'word': 'had', 'pos': 'VBD'},
{'word': 'earlier', 'pos': 'RBR'},
{'word': 'accused', 'pos': 'VBN'},
{'word': 'Fischler', 'pos': 'NNP'},
{'word': 'at', 'pos': 'IN'},
{'word': 'an', 'pos': 'DT'},
{'word': 'EU', 'pos': 'JJ'},
{'word': 'farm', 'pos': 'NN'},
{'word': 'ministers', 'pos': 'NNS'},
```

```
{'word': "'", 'pos': 'POS'},
{'word': 'meeting', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'causing', 'pos': 'VBG'},
{'word': 'unjustified', 'pos': 'JJ'},
{'word': 'alarm', 'pos': 'NN'},
{'word': 'through', 'pos': 'IN'},
{'word': '"', 'pos': '"'},
{'word': 'dangerous', 'pos': 'JJ'},
{'word': 'generalisation', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': '"', 'pos': '"'},
{'word': '.', 'pos': '.'},
{'word': 'Only', 'pos': 'RB'},
{'word': 'France', 'pos': 'NNP'},
{'word': 'and', 'pos': 'CC'},
{'word': 'Britain', 'pos': 'NNP'},
{'word': 'backed', 'pos': 'VBD'},
{'word': 'Fischler', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'proposal', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'The', 'pos': 'DT'},
{'word': 'EU', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'scientific', 'pos': 'JJ'},
{'word': 'veterinary', 'pos': 'JJ'},
{'word': 'and', 'pos': 'CC'},
{'word': 'multidisciplinary', 'pos': 'JJ'},
{'word': 'committees', 'pos': 'NNS'},
{'word': 'are', 'pos': 'VBP'},
{'word': 'due', 'pos': 'JJ'},
{'word': 'to', 'pos': 'TO'},
{'word': 're-examine', 'pos': 'VB'},
{'word': 'the', 'pos': 'DT'},
{'word': 'issue', 'pos': 'NN'},
{'word': 'early', 'pos': 'RB'},
{'word': 'next', 'pos': 'JJ'},
{'word': 'month', 'pos': 'NN'},
{'word': 'and', 'pos': 'CC'},
{'word': 'make', 'pos': 'VB'},
{'word': 'recommendations', 'pos': 'NNS'},
{'word': 'to', 'pos': 'TO'},
{'word': 'the', 'pos': 'DT'},
{'word': 'senior', 'pos': 'JJ'},
{'word': 'veterinary', 'pos': 'JJ'},
{'word': 'officials', 'pos': 'NNS'},
```

```
{'word': '.', 'pos': '.'},
{'word': 'Sheep', 'pos': 'NNP'},
{'word': 'have', 'pos': 'VBP'},
{'word': 'long', 'pos': 'RB'},
{'word': 'been', 'pos': 'VBN'},
{'word': 'known', 'pos': 'VBN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'contract', 'pos': 'NN'},
{'word': 'scrapie', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': 'a', 'pos': 'DT'},
{'word': 'brain-wasting', 'pos': 'JJ'},
{'word': 'disease', 'pos': 'NN'},
{'word': 'similar', 'pos': 'JJ'},
{'word': 'to', 'pos': 'TO'},
{'word': 'BSE', 'pos': 'NNP'},
{'word': 'which', 'pos': 'WDT'},
{'word': 'is', 'pos': 'VBZ'},
{'word': 'believed', 'pos': 'VBN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'have', 'pos': 'VB'},
{'word': 'been', 'pos': 'VBN'},
{'word': 'transferred', 'pos': 'VBN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'cattle', 'pos': 'NNS'},
{'word': 'through', 'pos': 'IN'},
{'word': 'feed', 'pos': 'NN'},
{'word': 'containing', 'pos': 'VBG'},
{'word': 'animal', 'pos': 'NN'},
{'word': 'waste', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'British', 'pos': 'JJ'},
{'word': 'farmers', 'pos': 'NNS'},
{'word': 'denied', 'pos': 'VBN'},
{'word': 'on', 'pos': 'IN'},
{'word': 'Thursday', 'pos': 'NNP'},
{'word': 'there', 'pos': 'EX'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'any', 'pos': 'DT'},
{'word': 'danger', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'human', 'pos': 'JJ'},
{'word': 'health', 'pos': 'NN'},
{'word': 'from', 'pos': 'IN'},
{'word': 'their', 'pos': 'PRP$'},
{'word': 'sheep', 'pos': 'NN'},
{'word': ',', 'pos': ','},
```

```
{'word': 'but', 'pos': 'CC'},
{'word': 'expressed', 'pos': 'VBD'},
{'word': 'concern', 'pos': 'NN'},
{'word': 'that', 'pos': 'IN'},
{'word': 'German', 'pos': 'JJ'},
{'word': 'government', 'pos': 'NN'},
{'word': 'advice', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'consumers', 'pos': 'NNS'},
{'word': 'to', 'pos': 'TO'},
{'word': 'avoid', 'pos': 'VB'},
{'word': 'British', 'pos': 'JJ'},
{'word': 'lamb', 'pos': 'NN'},
{'word': 'might', 'pos': 'MD'},
{'word': 'influence', 'pos': 'VB'},
{'word': 'consumers', 'pos': 'NNS'},
{'word': 'across', 'pos': 'IN'},
{'word': 'Europe', 'pos': 'NNP'},
{'word': '.', 'pos': '.'},
{'word': '"', 'pos': '"'},
{'word': 'What', 'pos': 'WP'},
{'word': 'we', 'pos': 'PRP'},
{'word': 'have', 'pos': 'VBP'},
{'word': 'to', 'pos': 'TO'},
{'word': 'be', 'pos': 'VB'},
{'word': 'extremely', 'pos': 'RB'},
{'word': 'careful', 'pos': 'JJ'},
{'word': 'of', 'pos': 'IN'},
{'word': 'is', 'pos': 'VBZ'},
{'word': 'how', 'pos': 'WRB'},
{'word': 'other', 'pos': 'JJ'},
{'word': 'countries', 'pos': 'NNS'},
{'word': 'are', 'pos': 'VBP'},
{'word': 'going', 'pos': 'VBG'},
{'word': 'to', 'pos': 'TO'},
{'word': 'take', 'pos': 'VB'},
{'word': 'Germany', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'lead', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': '"', 'pos': '"'},
{'word': 'Welsh', 'pos': 'NNP'},
{'word': 'National', 'pos': 'NNP'},
{'word': 'Farmers', 'pos': 'NNP'},
{'word': "'", 'pos': 'POS'},
{'word': 'Union', 'pos': 'NNP'},
{'word': '(', 'pos': '('},
```

```
{'word': 'NFU', 'pos': 'NNP'},
{'word': ')', 'pos': ')'},
{'word': 'chairman', 'pos': 'NN'},
{'word': 'John', 'pos': 'NNP'},
{'word': 'Lloyd', 'pos': 'NNP'},
{'word': 'Jones', 'pos': 'NNP'},
{'word': 'said', 'pos': 'VBD'},
{'word': 'on', 'pos': 'IN'},
{'word': 'BBC', 'pos': 'NNP'},
{'word': 'radio', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'Bonn', 'pos': 'NNP'},
{'word': 'has', 'pos': 'VBZ'},
{'word': 'led', 'pos': 'VBN'},
{'word': 'efforts', 'pos': 'NNS'},
{'word': 'to', 'pos': 'TO'},
{'word': 'protect', 'pos': 'VB'},
{'word': 'public', 'pos': 'JJ'},
{'word': 'health', 'pos': 'NN'},
{'word': 'after', 'pos': 'IN'},
{'word': 'consumer', 'pos': 'NN'},
{'word': 'confidence', 'pos': 'NN'},
{'word': 'collapsed', 'pos': 'VBD'},
{'word': 'in', 'pos': 'IN'},
{'word': 'March', 'pos': 'NNP'},
{'word': 'after', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
{'word': 'British', 'pos': 'JJ'},
{'word': 'report', 'pos': 'NN'},
{'word': 'suggested', 'pos': 'VBD'},
{'word': 'humans', 'pos': 'NNS'},
{'word': 'could', 'pos': 'MD'},
{'word': 'contract', 'pos': 'VB'},
{'word': 'an', 'pos': 'DT'},
{'word': 'illness', 'pos': 'NN'},
{'word': 'similar', 'pos': 'JJ'},
{'word': 'to', 'pos': 'TO'},
{'word': 'mad', 'pos': 'JJ'},
{'word': 'cow', 'pos': 'NN'},
{'word': 'disease', 'pos': 'NN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'eating', 'pos': 'VBG'},
{'word': 'contaminated', 'pos': 'VBN'},
{'word': 'beef', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'Germany', 'pos': 'NNP'},
{'word': 'imported', 'pos': 'VBD'},
```

```
{'word': '47,600', 'pos': 'CD'},
{'word': 'sheep', 'pos': 'NN'},
{'word': 'from', 'pos': 'IN'},
{'word': 'Britain', 'pos': 'NNP'},
{'word': 'last', 'pos': 'JJ'},
{'word': 'year', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': 'nearly', 'pos': 'RB'},
{'word': 'half', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'total', 'pos': 'JJ'},
{'word': 'imports', 'pos': 'NNS'},
{'word': '.', 'pos': '.'},
{'word': 'It', 'pos': 'PRP'},
{'word': 'brought', 'pos': 'VBD'},
{'word': 'in', 'pos': 'IN'},
{'word': '4,275', 'pos': 'CD'},
{'word': 'tonnes', 'pos': 'NNS'},
{'word': 'of', 'pos': 'IN'},
{'word': 'British', 'pos': 'JJ'},
{'word': 'mutton', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': 'some', 'pos': 'DT'},
{'word': '10', 'pos': 'CD'},
{'word': 'percent', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'overall', 'pos': 'JJ'},
{'word': 'imports', 'pos': 'NNS'},
{'word': '.', 'pos': '.'},
{'word': 'Rare', 'pos': 'NNP'},
{'word': 'Hendrix', 'pos': 'NNP'},
{'word': 'song', 'pos': 'NN'},
{'word': 'draft', 'pos': 'NN'},
{'word': 'sells', 'pos': 'VBZ'},
{'word': 'for', 'pos': 'IN'},
{'word': 'almost', 'pos': 'RB'},
{'word': '$', 'pos': '$'},
{'word': '17,000', 'pos': 'CD'},
{'word': '.', 'pos': '.'},
{'word': 'LONDON', 'pos': 'NNP'},
{'word': '1996-08-22', 'pos': 'CD'},
{'word': 'A', 'pos': 'DT'},
{'word': 'rare', 'pos': 'JJ'},
{'word': 'early', 'pos': 'JJ'},
{'word': 'handwritten', 'pos': 'JJ'},
{'word': 'draft', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
```

```
{'word': 'a', 'pos': 'DT'},
{'word': 'song', 'pos': 'NN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'U.S.', 'pos': 'NNP'},
{'word': 'guitar', 'pos': 'NN'},
{'word': 'legend', 'pos': 'NN'},
{'word': 'Jimi', 'pos': 'NNP'},
{'word': 'Hendrix', 'pos': 'NNP'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'sold', 'pos': 'VBN'},
{'word': 'for', 'pos': 'IN'},
{'word': 'almost', 'pos': 'RB'},
{'word': '$', 'pos': '$'},
{'word': '17,000', 'pos': 'CD'},
{'word': 'on', 'pos': 'IN'},
{'word': 'Thursday', 'pos': 'NNP'},
{'word': 'at', 'pos': 'IN'},
{'word': 'an', 'pos': 'DT'},
{'word': 'auction', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'some', 'pos': 'DT'},
{'word': 'of', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'late', 'pos': 'JJ'},
{'word': 'musician', 'pos': 'NN'},
{'word': "'s", 'pos': 'POS'},
{'word': 'favourite', 'pos': 'JJ'},
{'word': 'possessions', 'pos': 'NNS'},
{'word': '.', 'pos': '.'},
{'word': 'A', 'pos': 'DT'},
{'word': 'Florida', 'pos': 'NNP'},
{'word': 'restaurant', 'pos': 'NN'},
{'word': 'paid', 'pos': 'VBD'},
{'word': '10,925', 'pos': 'CD'},
{'word': 'pounds', 'pos': 'NNS'},
{'word': '(', 'pos': '('},
{'word': '$', 'pos': '$'},
{'word': '16,935', 'pos': 'CD'},
{'word': ')', 'pos': ')'},
{'word': 'for', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'draft', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': '"', 'pos': '"'},
{'word': 'Ai', 'pos': 'VBZ'},
{'word': "n't", 'pos': 'RB'},
{'word': 'no', 'pos': 'DT'},
```

```
{'word': 'telling', 'pos': 'VBG'},
{'word': '"', 'pos': '"'},
{'word': ',', 'pos': ','},
{'word': 'which', 'pos': 'WDT'},
{'word': 'Hendrix', 'pos': 'NNP'},
{'word': 'penned', 'pos': 'VBN'},
{'word': 'on', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
{'word': 'piece', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'London', 'pos': 'NNP'},
{'word': 'hotel', 'pos': 'NN'},
{'word': 'stationery', 'pos': 'NN'},
{'word': 'in', 'pos': 'IN'},
{'word': 'late', 'pos': 'JJ'},
{'word': '1966', 'pos': 'CD'},
{'word': '.', 'pos': '.'},
{'word': 'At', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'end', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
{'word': 'January', 'pos': 'NNP'},
{'word': '1967', 'pos': 'CD'},
{'word': 'concert', 'pos': 'NN'},
{'word': 'in', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'English', 'pos': 'JJ'},
{'word': 'city', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'Nottingham', 'pos': 'NNP'},
{'word': 'he', 'pos': 'PRP'},
{'word': 'threw', 'pos': 'VBD'},
{'word': 'the', 'pos': 'DT'},
{'word': 'sheet', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'paper', 'pos': 'NN'},
{'word': 'into', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'audience', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': 'where', 'pos': 'WRB'},
{'word': 'it', 'pos': 'PRP'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'retrieved', 'pos': 'VBN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
```

{'word': 'fan', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'Buyers', 'pos': 'NNS'},
{'word': 'also', 'pos': 'RB'},
{'word': 'snapped', 'pos': 'VBD'},
{'word': 'up', 'pos': 'RB'},
{'word': '16', 'pos': 'CD'},
{'word': 'other', 'pos': 'JJ'},
{'word': 'items', 'pos': 'NNS'},
{'word': 'that', 'pos': 'WDT'},
{'word': 'were', 'pos': 'VBD'},
{'word': 'put', 'pos': 'VBN'},
{'word': 'up', 'pos': 'RP'},
{'word': 'for', 'pos': 'IN'},
{'word': 'auction', 'pos': 'NN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'Hendrix', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'former', 'pos': 'JJ'},
{'word': 'girlfriend', 'pos': 'NN'},
{'word': 'Kathy', 'pos': 'NNP'},
{'word': 'Etchingham', 'pos': 'NNP'},
{'word': ',', 'pos': ','},
{'word': 'who', 'pos': 'WP'},
{'word': 'lived', 'pos': 'VBD'},
{'word': 'with', 'pos': 'IN'},
{'word': 'him', 'pos': 'PRP'},
{'word': 'from', 'pos': 'IN'},
{'word': '1966', 'pos': 'CD'},
{'word': 'to', 'pos': 'TO'},
{'word': '1969', 'pos': 'CD'},
{'word': '.', 'pos': '.'},
{'word': 'They', 'pos': 'PRP'},
{'word': 'included', 'pos': 'VBD'},
{'word': 'a', 'pos': 'DT'},
{'word': 'black', 'pos': 'JJ'},
{'word': 'lacquer', 'pos': 'NN'},
{'word': 'and', 'pos': 'CC'},
{'word': 'mother', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'pearl', 'pos': 'NN'},
{'word': 'inlaid', 'pos': 'VBD'},
{'word': 'box', 'pos': 'NN'},
{'word': 'used', 'pos': 'VBN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'Hendrix', 'pos': 'NNP'},
{'word': 'to', 'pos': 'TO'},

```
{'word': 'store', 'pos': 'VB'},
{'word': 'his', 'pos': 'PRP$'},
{'word': 'drugs', 'pos': 'NNS'},
{'word': ',', 'pos': ','},
{'word': 'which', 'pos': 'WDT'},
{'word': 'an', 'pos': 'DT'},
{'word': 'anonymous', 'pos': 'JJ'},
{'word': 'Australian', 'pos': 'JJ'},
{'word': 'purchaser', 'pos': 'NN'},
{'word': 'bought', 'pos': 'VBD'},
{'word': 'for', 'pos': 'IN'},
{'word': '5,060', 'pos': 'CD'},
{'word': 'pounds', 'pos': 'NNS'},
{'word': '(', 'pos': '('},
{'word': '$', 'pos': '$'},
{'word': '7,845', 'pos': 'CD'},
{'word': ')', 'pos': ')'},
{'word': '.', 'pos': '.'},
{'word': 'The', 'pos': 'DT'},
{'word': 'guitarist', 'pos': 'NN'},
{'word': 'died', 'pos': 'VBD'},
{'word': 'of', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
{'word': 'drugs', 'pos': 'NNS'},
{'word': 'overdose', 'pos': 'NN'},
{'word': 'in', 'pos': 'IN'},
{'word': '1970', 'pos': 'CD'},
{'word': 'aged', 'pos': 'VBN'},
{'word': '27', 'pos': 'CD'},
{'word': '.', 'pos': '.'},
{'word': 'China', 'pos': 'NNP'},
{'word': 'says', 'pos': 'VBZ'},
{'word': 'Taiwan', 'pos': 'NNP'},
{'word': 'spoils', 'pos': 'VBZ'},
{'word': 'atmosphere', 'pos': 'NN'},
{'word': 'for', 'pos': 'IN'},
{'word': 'talks', 'pos': 'NNS'},
{'word': '.', 'pos': '.'},
{'word': 'BEIJING', 'pos': 'VBG'},
{'word': '1996-08-22', 'pos': 'CD'},
{'word': 'China', 'pos': 'NNP'},
{'word': 'on', 'pos': 'IN'},
{'word': 'Thursday', 'pos': 'NNP'},
{'word': 'accused', 'pos': 'VBD'},
{'word': 'Taipei', 'pos': 'NNP'},
{'word': 'of', 'pos': 'IN'},
{'word': 'spoiling', 'pos': 'VBG'},
```

```
{'word': 'the', 'pos': 'DT'},
{'word': 'atmosphere', 'pos': 'NN'},
{'word': 'for', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
{'word': 'resumption', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'talks', 'pos': 'NNS'},
{'word': 'across', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'Taiwan', 'pos': 'NNP'},
{'word': 'Strait', 'pos': 'NNP'},
{'word': 'with', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
{'word': 'visit', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'Ukraine', 'pos': 'NNP'},
{'word': 'by', 'pos': 'IN'},
{'word': 'Taiwanese', 'pos': 'NNP'},
{'word': 'Vice', 'pos': 'NNP'},
{'word': 'President', 'pos': 'NNP'},
{'word': 'Lien', 'pos': 'NNP'},
{'word': 'Chan', 'pos': 'NNP'},
{'word': 'this', 'pos': 'DT'},
{'word': 'week', 'pos': 'NN'},
{'word': 'that', 'pos': 'WDT'},
{'word': 'infuriated', 'pos': 'VBD'},
{'word': 'Beijing', 'pos': 'NNP'},
{'word': '.', 'pos': '.'},
{'word': 'Speaking', 'pos': 'VBG'},
{'word': 'only', 'pos': 'JJ'},
{'word': 'hours', 'pos': 'NNS'},
{'word': 'after', 'pos': 'IN'},
{'word': 'Chinese', 'pos': 'JJ'},
{'word': 'state', 'pos': 'NN'},
{'word': 'media', 'pos': 'NNS'},
{'word': 'said', 'pos': 'VBD'},
{'word': 'the', 'pos': 'DT'},
{'word': 'time', 'pos': 'NN'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'right', 'pos': 'RB'},
{'word': 'to', 'pos': 'TO'},
{'word': 'engage', 'pos': 'VB'},
{'word': 'in', 'pos': 'IN'},
{'word': 'political', 'pos': 'JJ'},
{'word': 'talks', 'pos': 'NNS'},
{'word': 'with', 'pos': 'IN'},
{'word': 'Taiwan', 'pos': 'NNP'},
```

```
{'word': ',', 'pos': ','},
{'word': 'Foreign', 'pos': 'NNP'},
{'word': 'Ministry', 'pos': 'NNP'},
{'word': 'spokesman', 'pos': 'NN'},
{'word': 'Shen', 'pos': 'NNP'},
{'word': 'Guofang', 'pos': 'NNP'},
{'word': 'told', 'pos': 'VBD'},
{'word': 'Reuters', 'pos': 'NNP'},
{'word': ':', 'pos': ':'},
{'word': '"', 'pos': '"'},
{'word': 'The', 'pos': 'DT'},
{'word': 'necessary', 'pos': 'JJ'},
{'word': 'atmosphere', 'pos': 'NN'},
{'word': 'for', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'opening', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'talks', 'pos': 'NNS'},
{'word': 'has', 'pos': 'VBZ'},
{'word': 'been', 'pos': 'VBN'},
{'word': 'disrupted', 'pos': 'VBN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'Taiwan', 'pos': 'NNP'},
{'word': 'authorities', 'pos': 'NNS'},
{'word': '.', 'pos': '.'},
{'word': '"', 'pos': '"'},
{'word': 'State', 'pos': 'NN'},
{'word': 'media', 'pos': 'NNS'},
{'word': 'quoted', 'pos': 'VBD'},
{'word': 'China', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'top', 'pos': 'JJ'},
{'word': 'negotiator', 'pos': 'NN'},
{'word': 'with', 'pos': 'IN'},
{'word': 'Taipei', 'pos': 'NNP'},
{'word': ',', 'pos': ','},
{'word': 'Tang', 'pos': 'NNP'},
{'word': 'Shubei', 'pos': 'NNP'},
{'word': ',', 'pos': ','},
{'word': 'as', 'pos': 'IN'},
{'word': 'telling', 'pos': 'VBG'},
{'word': 'a', 'pos': 'DT'},
{'word': 'visiting', 'pos': 'JJ'},
{'word': 'group', 'pos': 'NN'},
{'word': 'from', 'pos': 'IN'},
```

{'word': 'Taiwan', 'pos': 'NNP'},
{'word': 'on', 'pos': 'IN'},
{'word': 'Wednesday', 'pos': 'NNP'},
{'word': 'that', 'pos': 'IN'},
{'word': 'it', 'pos': 'PRP'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'time', 'pos': 'NN'},
{'word': 'for', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'rivals', 'pos': 'NNS'},
{'word': 'to', 'pos': 'TO'},
{'word': 'hold', 'pos': 'VB'},
{'word': 'political', 'pos': 'JJ'},
{'word': 'talks', 'pos': 'NNS'},
{'word': '.', 'pos': '.'},
{'word': '"', 'pos': '"'},
{'word': 'Now', 'pos': 'RB'},
{'word': 'is', 'pos': 'VBZ'},
{'word': 'the', 'pos': 'DT'},
{'word': 'time', 'pos': 'NN'},
{'word': 'for', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'two', 'pos': 'CD'},
{'word': 'sides', 'pos': 'NNS'},
{'word': 'to', 'pos': 'TO'},
{'word': 'engage', 'pos': 'VB'},
{'word': 'in', 'pos': 'IN'},
{'word': 'political', 'pos': 'JJ'},
{'word': 'talks', 'pos': 'NNS'},
{'word': '…', 'pos': ':'},
{'word': 'that', 'pos': 'WDT'},
{'word': 'is', 'pos': 'VBZ'},
{'word': 'to', 'pos': 'TO'},
{'word': 'end', 'pos': 'VB'},
{'word': 'the', 'pos': 'DT'},
{'word': 'state', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'hostility', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': '"', 'pos': '"'},
{'word': 'Thursday', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'overseas', 'pos': 'JJ'},
{'word': 'edition', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'People', 'pos': 'NNPS'},

```
{'word': "'s", 'pos': 'POS'},
{'word': 'Daily', 'pos': 'JJ'},
{'word': 'quoted', 'pos': 'VBN'},
{'word': 'Tang', 'pos': 'NNP'},
{'word': 'as', 'pos': 'IN'},
{'word': 'saying', 'pos': 'VBG'},
{'word': '.', 'pos': '.'},
{'word': 'The', 'pos': 'DT'},
{'word': 'foreign', 'pos': 'JJ'},
{'word': 'ministry', 'pos': 'NN'},
{'word': "'s", 'pos': 'POS'},
{'word': 'Shen', 'pos': 'NNP'},
{'word': 'told', 'pos': 'VBD'},
{'word': 'Reuters', 'pos': 'NNP'},
{'word': 'Television', 'pos': 'NNP'},
{'word': 'in', 'pos': 'IN'},
{'word': 'an', 'pos': 'DT'},
{'word': 'interview', 'pos': 'NN'},
{'word': 'he', 'pos': 'PRP'},
{'word': 'had', 'pos': 'VBD'},
{'word': 'read', 'pos': 'VBN'},
{'word': 'reports', 'pos': 'NNS'},
{'word': 'of', 'pos': 'IN'},
{'word': 'Tang', 'pos': 'NNP'},
{'word': "'s", 'pos': 'POS'},
{'word': 'comments', 'pos': 'NNS'},
{'word': 'but', 'pos': 'CC'},
{'word': 'gave', 'pos': 'VBD'},
{'word': 'no', 'pos': 'DT'},
{'word': 'details', 'pos': 'NNS'},
{'word': 'of', 'pos': 'IN'},
{'word': 'why', 'pos': 'WRB'},
{'word': 'the', 'pos': 'DT'},
{'word': 'negotiator', 'pos': 'NN'},
{'word': 'had', 'pos': 'VBD'},
{'word': 'considered', 'pos': 'VBN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'time', 'pos': 'NN'},
{'word': 'right', 'pos': 'RB'},
{'word': 'for', 'pos': 'IN'},
{'word': 'talks', 'pos': 'NNS'},
{'word': 'with', 'pos': 'IN'},
{'word': 'Taiwan', 'pos': 'NNP'},
{'word': ',', 'pos': ','},
{'word': 'which', 'pos': 'WDT'},
{'word': 'Beijing', 'pos': 'NNP'},
{'word': 'considers', 'pos': 'VBZ'},
```

```
{'word': 'a', 'pos': 'DT'},
{'word': 'renegade', 'pos': 'NN'},
{'word': 'province', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'China', 'pos': 'NNP'},
{'word': ',', 'pos': ','},
{'word': 'which', 'pos': 'WDT'},
{'word': 'has', 'pos': 'VBZ'},
{'word': 'long', 'pos': 'RB'},
{'word': 'opposed', 'pos': 'VBN'},
{'word': 'all', 'pos': 'DT'},
{'word': 'Taipei', 'pos': 'NNP'},
{'word': 'efforts', 'pos': 'NNS'},
{'word': 'to', 'pos': 'TO'},
{'word': 'gain', 'pos': 'VB'},
{'word': 'greater', 'pos': 'JJR'},
{'word': 'international', 'pos': 'JJ'},
{'word': 'recognition', 'pos': 'NN'},
{'word': ',', 'pos': ','},
{'word': 'was', 'pos': 'VBD'},
{'word': 'infuriated', 'pos': 'VBN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'a', 'pos': 'DT'},
{'word': 'visit', 'pos': 'NN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'Ukraine', 'pos': 'NNP'},
{'word': 'this', 'pos': 'DT'},
{'word': 'week', 'pos': 'NN'},
{'word': 'by', 'pos': 'IN'},
{'word': 'Taiwanese', 'pos': 'NNP'},
{'word': 'Vice', 'pos': 'NNP'},
{'word': 'President', 'pos': 'NNP'},
{'word': 'Lien', 'pos': 'NNP'},
{'word': '.', 'pos': '.'},
{'word': 'China', 'pos': 'NNP'},
{'word': 'says', 'pos': 'VBZ'},
{'word': 'time', 'pos': 'NN'},
{'word': 'right', 'pos': 'RB'},
{'word': 'for', 'pos': 'IN'},
{'word': 'Taiwan', 'pos': 'NNP'},
{'word': 'talks', 'pos': 'NNS'},
{'word': '.', 'pos': '.'},
{'word': 'BEIJING', 'pos': 'VBG'},
{'word': '1996-08-22', 'pos': 'CD'},
{'word': 'China', 'pos': 'NNP'},
{'word': 'has', 'pos': 'VBZ'},
{'word': 'said', 'pos': 'VBD'},
```

```
{'word': 'it', 'pos': 'PRP'},
{'word': 'was', 'pos': 'VBD'},
{'word': 'time', 'pos': 'NN'},
{'word': 'for', 'pos': 'IN'},
{'word': 'political', 'pos': 'JJ'},
{'word': 'talks', 'pos': 'NNS'},
{'word': 'with', 'pos': 'IN'},
{'word': 'Taiwan', 'pos': 'NNP'},
{'word': 'and', 'pos': 'CC'},
{'word': 'that', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'rival', 'pos': 'JJ'},
{'word': 'island', 'pos': 'NN'},
{'word': 'should', 'pos': 'MD'},
{'word': 'take', 'pos': 'VB'},
{'word': 'practical', 'pos': 'JJ'},
{'word': 'steps', 'pos': 'NNS'},
{'word': 'towards', 'pos': 'IN'},
{'word': 'that', 'pos': 'DT'},
{'word': 'goal', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'Consultations', 'pos': 'NNS'},
{'word': 'should', 'pos': 'MD'},
{'word': 'be', 'pos': 'VB'},
{'word': 'held', 'pos': 'VBN'},
{'word': 'to', 'pos': 'TO'},
{'word': 'set', 'pos': 'VB'},
{'word': 'the', 'pos': 'DT'},
{'word': 'time', 'pos': 'NN'},
{'word': 'and', 'pos': 'CC'},
{'word': 'format', 'pos': 'NN'},
{'word': 'of', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'talks', 'pos': 'NNS'},
{'word': ',', 'pos': ','},
{'word': 'the', 'pos': 'DT'},
{'word': 'official', 'pos': 'NN'},
{'word': 'Xinhua', 'pos': 'NNP'},
{'word': 'news', 'pos': 'NN'},
{'word': 'agency', 'pos': 'NN'},
{'word': 'quoted', 'pos': 'VBN'},
{'word': 'Tang', 'pos': 'NNP'},
{'word': 'Shubei', 'pos': 'NNP'},
{'word': ',', 'pos': ','},
{'word': 'executive', 'pos': 'JJ'},
{'word': 'vice', 'pos': 'NN'},
{'word': 'chairman', 'pos': 'NN'},
```

{'word': 'of', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'Association', 'pos': 'NNP'},
{'word': 'for', 'pos': 'IN'},
{'word': 'Relations', 'pos': 'NNP'},
{'word': 'Across', 'pos': 'IN'},
{'word': 'the', 'pos': 'DT'},
{'word': 'Taiwan', 'pos': 'NNP'},
{'word': 'Straits', 'pos': 'NNP'},
{'word': ',', 'pos': ','},
{'word': 'as', 'pos': 'IN'},
{'word': 'saying', 'pos': 'VBG'},
{'word': 'late', 'pos': 'RB'},
{'word': 'on', 'pos': 'RB'},
{'word': 'Wednesday', 'pos': 'NNP'},
{'word': '.', 'pos': '.'},
{'word': 'German', 'pos': 'NNP'},
{'word': 'July', 'pos': 'NNP'},
{'word': 'car', 'pos': 'NN'},
{'word': 'registrations', 'pos': 'NNS'},
{'word': 'up', 'pos': 'RB'},
{'word': '14.2', 'pos': 'CD'},
{'word': 'pct', 'pos': 'JJ'},
{'word': 'yr', 'pos': 'NN'},
{'word': '/', 'pos': 'SYM'},
{'word': 'yr', 'pos': 'NN'},
{'word': '.', 'pos': '.'},
{'word': 'FRANKFURT', 'pos': 'NNP'},
{'word': '1996-08-22', 'pos': 'CD'},
{'word': 'German', 'pos': 'JJ'},
{'word': 'first-time', 'pos': 'JJ'},
{'word': 'registrations', 'pos': 'NNS'},
{'word': 'of', 'pos': 'IN'},
{'word': 'motor', 'pos': 'NN'},
{'word': 'vehicles', 'pos': 'NNS'},
{'word': 'jumped', 'pos': 'VBD'},
{'word': '14.2', 'pos': 'CD'},
{'word': 'percent', 'pos': 'NN'},
{'word': 'in', 'pos': 'IN'},
{'word': 'July', 'pos': 'NNP'},
{'word': 'this', 'pos': 'DT'},
{'word': 'year', 'pos': 'NN'},
{'word': 'from', 'pos': 'IN'},
…],
['B-ORG',
 'O',
 'B-MISC',

```
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'B-PER',
'I-PER',
'B-LOC',
'O',
'O',
'B-ORG',
'I-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'B-ORG',
'I-ORG',
```

```
'O',
'O',
'O',
'B-PER',
'I-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-ORG',
'O',
'O',
```

```
'O',
'B-PER',
'I-PER',
'I-PER',
'I-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-ORG',
'I-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-ORG',
'O',
'O',
'B-PER',
'I-PER',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-PER',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'I-MISC',
```

```
'I-MISC',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'O',
'B-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
```

```
'O',
'O',
'B-PER',
'I-PER',
'I-PER',
'O',
'O',
'O',
'B-PER',
'O',
'O',
'B-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'B-LOC',
'O',
'B-PER',
'O',
'O',
'O',
'O',
'B-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'B-ORG',
'I-ORG',
'I-ORG',
'I-ORG',
'I-ORG',
'O',
'B-ORG',
'O',
'O',
'B-PER',
'I-PER',
'I-PER',
'O',
'O',
'B-ORG',
'I-ORG',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-PER',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'B-PER',
'I-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'I-MISC',
'I-MISC',
'I-MISC',
'O',
'O',
'O',
'B-PER',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'B-LOC',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-PER',
'O',
'O',
'O',
'B-PER',
'I-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-PER',
'O',
'O',
'O',
'B-PER',
'I-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'B-LOC',
```

```
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'B-LOC',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'I-LOC',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'B-MISC',
'O',
'O',
'B-PER',
'I-PER',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'B-ORG',
'I-ORG',
'O',
'B-PER',
'I-PER',
'O',
'B-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
```

```
'O',
'O',
'O',
'B-LOC',
'O',
'B-PER',
'I-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-ORG',
'I-ORG',
'I-ORG',
'O',
'B-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-ORG',
'O',
'B-ORG',
'I-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'B-PER',
```

```
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'B-LOC',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'B-ORG',
'O',
'O',
'O',
'B-PER',
'I-PER',
'O',
'O',
'O',
'O',
'O',
'O',
'B-ORG',
'I-ORG',
'I-ORG',
'I-ORG',
'I-ORG',
'I-ORG',
'I-ORG',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-MISC',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-LOC',
'O',
'B-MISC',
'O',
'O',
'O',
```

```
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    …])
```

1. b

```python
[22]:  from collections import Counter

       test = ConllCorpusReader('./CONLL2003', 'test.txt', ['words', 'pos', 'ignore',␣
        ↪'chunk'])
       test_features = []
       test_gold_labels = []
       for token, pos, ne_label in test.iob_words():
           a_dict = {
               'word': token,
               'pos': pos
           }
           test_features.append(a_dict)
           test_gold_labels.append(ne_label)

       num_train_instances = len(training_features)
       num_test_instances = len(test_features)

       train_labels_freq = Counter(training_gold_labels)
       test_labels_freq = Counter(test_gold_labels)
```

b) i) and ii)

```python
[23]:  num_train_instances, num_test_instances, train_labels_freq, test_labels_freq
```

```
[23]:  (203621,
        46435,
        Counter({'O': 169578,
                 'B-LOC': 7140,
                 'B-PER': 6600,
                 'B-ORG': 6321,
                 'I-PER': 4528,
                 'I-ORG': 3704,
                 'B-MISC': 3438,
                 'I-LOC': 1157,
```

```
                'I-MISC': 1155}),
  Counter({'O': 38323,
                'B-LOC': 1668,
                'B-ORG': 1661,
                'B-PER': 1617,
                'I-PER': 1156,
                'I-ORG': 835,
                'B-MISC': 702,
                'I-LOC': 257,
                'I-MISC': 216}))
```

b) iii) The datasets are not balanced with respect to the distribution of NERC labels. The majority of instances are tagged as 'O', indicating they are not part of a named entity.

The training and test data are consistent in terms of the relative frequencies of different labels However, the exact counts of each label differ due to the different sizes of the datasets.

c)

```
[24]:  from sklearn.feature_extraction import DictVectorizer

       combined_features = training_features + test_features
       vectorizer = DictVectorizer(sparse=True)
       vectorized_features = vectorizer.fit_transform(combined_features)

       train_vectorized = vectorized_features[:num_train_instances]
       test_vectorized = vectorized_features[num_train_instances:]

       train_vectorized.shape, test_vectorized.shape
```

```
[24]: ((203621, 27361), (46435, 27361))
```

d)

- The SVM classifier performs exceptionally well on non-entity ('O') labels due to their pre-dominance in the dataset, showing high precision and recall.

- It struggles with 'I-PER' labels, likely due to over-predicting this category, and shows lower recall for 'B-PER' and 'B-ORG', indicating missed instances of these entities. These variations suggest the model's performance is influenced by the imbalance in label distribution and the distinctiveness of entity patterns in the training data.

```
[25]:  from sklearn import svm
       from sklearn.metrics import classification_report

       lin_clf = svm.LinearSVC()
       lin_clf.fit(train_vectorized, training_gold_labels)
       test_predictions = lin_clf.predict(test_vectorized)
```

45

```
classification_report_test = classification_report(test_gold_labels,␣
  ↪test_predictions)
```

/home/temp/.local/lib/python3.10/site-packages/sklearn/svm/_classes.py:31:
FutureWarning: The default value of `dual` will change from `True` to `'auto'`
in 1.5. Set the value of `dual` explicitly to suppress the warning.
  warnings.warn(

[26]: `print(classification_report_test)`

	precision	recall	f1-score	support
B-LOC	0.81	0.78	0.79	1668
B-MISC	0.78	0.66	0.72	702
B-ORG	0.79	0.52	0.63	1661
B-PER	0.86	0.44	0.58	1617
I-LOC	0.62	0.53	0.57	257
I-MISC	0.57	0.59	0.58	216
I-ORG	0.70	0.47	0.56	835
I-PER	0.33	0.87	0.48	1156
O	0.98	0.98	0.98	38323
accuracy			0.92	46435
macro avg	0.72	0.65	0.65	46435
weighted avg	0.94	0.92	0.92	46435

e)

[14]: `#couldnt finish this part`

1.0.2 2. a)

[20]:
```python
import pandas as pd
from sklearn.feature_extraction import DictVectorizer

path = './ner_dataset.csv'
kaggle_dataset = pd.read_csv(path, encoding='ISO-8859-1', on_bad_lines='skip')

df_train = kaggle_dataset[:100000]
df_test = kaggle_dataset[100000:120000]

train_features = df_train[['Word', 'POS']].to_dict('records')
train_labels = df_train['Tag'].tolist()

test_features = df_test[['Word', 'POS']].to_dict('records')
test_labels = df_test['Tag'].tolist()
```

```
all_features = train_features + test_features

vectorizer = DictVectorizer()
features_encoded = vectorizer.fit_transform(all_features)

train_labels = df_train['Tag'].tolist()
test_labels = df_test['Tag'].tolist()
```

[22]: `train_labels, test_labels`

[22]: (['O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'B-geo',
    'O',
    'O',
    'O',
    'O',
    'O',
    'B-geo',
    'O',
    'O',
    'O',
    'O',
    'O',
    'B-gpe',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',

```

```
'O',
'O',
'O',
'O',
'O',
'O',
'B-per',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'B-geo',
'O',
'O',
'B-geo',
'O',
'O',
'B-org',
'I-org',
'I-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'B-tim',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'O',
'B-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'B-org',
'I-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'B-art',
'I-art',
'O',
'O',
'B-gpe',
'O',
'O',
'B-per',
```

```
'I-per',
'I-per',
'O',
'B-tim',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
```

```
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'B-org',
'I-org',
'O',
'B-per',
'I-per',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'O',
'B-geo',
'I-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'B-gpe',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'B-per',
'I-per',
'I-per',
'I-per',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'B-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'I-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'I-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'I-org',
```

```
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'B-gpe',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'B-tim',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'B-geo',
'I-geo',
'I-geo',
'I-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
```

```
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
'O',
'B-geo',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
'B-per',
'I-per',
'O',
'B-gpe',
'O',
'O',
'B-tim',
'O',
'B-per',
'I-per',
'O',
```

```
      'O',
      'O',
      'B-geo',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'B-org',
      'I-org',
      'I-org',
      'O',
      'O',
      'O',
      'O',
      'B-per',
      'O',
      'O',
      …],
     ['O',
      'O',
      'O',
      'B-geo',
      'I-geo',
      'O',
      'O',
      'O',
      'B-tim',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'O',
      'B-geo',
      'O',
      'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'B-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
```

```
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'B-geo',
'O',
'B-org',
'I-org',
'I-org',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'B-per',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'I-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'B-org',
'I-org',
'I-org',
'B-per',
'I-per',
'I-per',
'I-per',
'I-per',
'I-per',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-per',
'I-per',
'I-per',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-per',
'I-per',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-per',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'B-tim',
'I-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'B-per',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-per',
'I-per',
'I-per',
'O',
'O',
'O',
'O',
'B-per',
'I-per',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'B-tim',
'I-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'I-geo',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'B-org',
'I-org',
'I-org',
'I-org',
'I-org',
'I-org',
'I-org',
'I-org',
'O',
'O',
'B-tim',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
'O',
'B-org',
'O',
'O',
'B-org',
'I-org',
'I-org',
'I-org',
'I-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'I-org',
'I-org',
'I-org',
'I-org',
'O',
'B-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
```

```
'O',
'B-gpe',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'B-org',
'O',
'B-geo',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'B-tim',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'B-geo',
'I-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'B-geo',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'B-tim',
'I-tim',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'B-per',
'I-per',
'I-per',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'B-geo',
'O',
'O',
'O',
'B-gpe',
'B-per',
'I-per',
'I-per',
'O',
'O',
'O',
'O',
'B-per',
'I-per',
'O',
```

```
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-tim',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-geo',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-gpe',
'O',
'O',
'O',
'O',
```

```
'O',
'O',
'O',
'O',
'B-per',
'I-per',
'O',
'O',
'O',
'B-org',
'I-org',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'O',
'B-org',
'O',
'O',
'O',
```

```
    'O',
    'O',
    'B-gpe',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'B-org',
    'I-org',
    'I-org',
    'I-org',
    'I-org',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    'O',
    …])
```

```python
[25]: from sklearn.svm import LinearSVC
      from sklearn.metrics import classification_report

      num_train_instances = len(train_features)
      train_encoded = features_encoded[:num_train_instances]
      test_encoded = features_encoded[num_train_instances:]
      svm_model = LinearSVC(random_state=42, max_iter=10000)
```

```
svm_model.fit(train_encoded, train_labels)
test_predictions = svm_model.predict(test_encoded)
print(classification_report(test_labels, test_predictions))
```

/home/temp/.local/lib/python3.10/site-packages/sklearn/svm/_classes.py:31:
FutureWarning: The default value of `dual` will change from `True` to `'auto'`
in 1.5. Set the value of `dual` explicitly to suppress the warning.
  warnings.warn(

	precision	recall	f1-score	support
B-art	0.00	0.00	0.00	4
B-eve	0.00	0.00	0.00	0
B-geo	0.80	0.76	0.78	741
B-gpe	0.96	0.92	0.94	296
B-nat	1.00	0.50	0.67	8
B-org	0.64	0.51	0.57	397
B-per	0.81	0.53	0.64	333
B-tim	0.91	0.76	0.83	393
I-art	0.00	0.00	0.00	0
I-eve	0.00	0.00	0.00	0
I-geo	0.74	0.50	0.60	156
I-gpe	1.00	0.50	0.67	2
I-nat	0.80	1.00	0.89	4
I-org	0.65	0.44	0.53	321
I-per	0.42	0.90	0.57	319
I-tim	0.41	0.08	0.14	108
O	0.98	0.99	0.99	16918
accuracy			0.94	20000
macro avg	0.60	0.49	0.52	20000
weighted avg	0.95	0.94	0.94	20000

/home/temp/.local/lib/python3.10/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/home/temp/.local/lib/python3.10/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Recall
is ill-defined and being set to 0.0 in labels with no true samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/home/temp/.local/lib/python3.10/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.

```
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/home/temp/.local/lib/python3.10/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Recall
is ill-defined and being set to 0.0 in labels with no true samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/home/temp/.local/lib/python3.10/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning:
Precision is ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/home/temp/.local/lib/python3.10/site-
packages/sklearn/metrics/_classification.py:1509: UndefinedMetricWarning: Recall
is ill-defined and being set to 0.0 in labels with no true samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

The model demonstrates strong performance on common labels such as 'O' for non-entity text and 'B-geo' for geographical entities, indicating it effectively learns from the dataset's prevalent examples. However, it struggles with rare entities like 'B-art' and 'I-eve', and with distinguishing parts of entities, as seen in 'I-per' and 'I-tim', suggesting challenges in capturing the nuanced context and rarity within the dataset.