Data Refiner (Distill)

Goal

```
1.    2012-01-04 00:01:23,180 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block
      blk_-2281137920769708011_1116 src: /127.0.0.1:32981 dest: /127.0.0.1:50010
2.    2012-01-04 00:01:23,184 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /
      127.0.0.1:32981, dest: /127.0.0.1:50010, bytes: 3758, op: HDFS_WRITE,
      cliID: DFSClient_-603743753, offset: 0, srvID:
      DS-292194659-127.0.1.1-50010-1324763300176, blockid:
      blk_-2281137920769708011_1116, duration: 2016056
3.    2012-01-04 00:01:23,185 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for
      block blk_-2281137920769708011_1116 terminating
4.    2012-01-04 00:01:23,291 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block
      blk_3766031435252346505_1117 src: /127.0.0.1:32982 dest: /127.0.0.1:50010
5.    2012-01-04 00:01:23,293 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /
      127.0.0.1:32982, dest: /127.0.0.1:50010, bytes: 265, op: HDFS_WRITE,
      cliID: DFSClient_-603743753, offset: 0, srvID:
      DS-292194659-127.0.1.1-50010-1324763300176, blockid:
      blk_3766031435252346505_1117, duration: 552828
6.    2012-01-04 00:01:23,293 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for
      block blk_3766031435252346505_1117 terminating
7.    2012-01-04 00:01:23,324 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block
      blk_-8044922265890142318_1118 src: /127.0.0.1:32983 dest: /127.0.0.1:50010
8.    2012-01-04 00:01:23,326 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /
      127.0.0.1:32983, dest: /127.0.0.1:50010, bytes: 43, op: HDFS_WRITE, cliID:
      DFSClient_-603743753, offset: 0, srvID:
      DS-292194659-127.0.1.1-50010-1324763300176, blockid:
      blk_-8044922265890142318_1118, duration: 607104
9.    2012-01-04 00:01:23,327 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for
      block blk_-8044922265890142318_1118 terminating
10.   2012-01-04 00:01:23,409 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block
      blk_-965793757262168743_1119 src: /127.0.0.1:32984 dest: /127.0.0.1:50010
```

```
1.  2012-01-04 00:01:23 127.0.0.1:32981 127.0.0.1:50010
2.  2012-01-04 00:01:23 127.0.0.1:32981 127.0.0.1:50010
3.  2012-01-04 00:01:23 NULL NULL
4.  2012-01-04 00:01:23 127.0.0.1:32982 127.0.0.1:50010
5.  2012-01-04 00:01:23 127.0.0.1:32982 127.0.0.1:50010
6.  2012-01-04 00:01:23 NULL NULL
7.  2012-01-04 00:01:23 127.0.0.1:32983 127.0.0.1:50010
8.  2012-01-04 00:01:23 127.0.0.1:32983 127.0.0.1:50010
9.  2012-01-04 00:01:23 NULL NULL
10. 2012-01-04 00:01:23 127.0.0.1:32984 127.0.0.1:50010
```

Goal

```
Reported crime in 'Alaska',
,
2004,+3370.9
2005,+3615
2006,+3582
2007,+3373.9
2008,+2928.3
Reported crime in 'Arizona',
,
2004,+5073.3
2005,+4827
2006,+4741.6
2007,+4502.6
2008,+4087.3

Reported crime in 'Arkansas',
,
2004,+4033.1
2005,+4068
2006,+4021.6
2007,+3945.5
2008,+3843.7

Reported crime in 'California',
,
2004,+3423.9
2005,+3321
2006,+3175.2
2007,+3032.6
2008,+2940.3
```

```
'Alaska',+3370.9,+3615,+3582,+3373.9,+2928.3
'Arizona',+5073.3,+4827,+4741.6,+4502.6,+4087.3
'Arkansas',+4033.1,+4068,+4021.6,+3945.5,+3843.7
'California',+3423.9,+3321,+3175.2,+3032.6,+2940.3
```

Approach: Cleaning by Example

Sample Input

```
1.    2012-01-04 00:01:23,180 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block
      blk_-2281137920769708011_1116 src: /127.0.0.1:32981 dest: /
      127.0.0.1:50010
2.    2012-01-04 00:01:23,185 INFO
      org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for
      block blk_-2281137920769708011_1116 terminating
```

Sample Output

```
1.  2012-01-04 00:01:23 127.0.0.1:32981 127.0.0.1:50010
2.  2012-01-04 00:01:23 NULL NULL
```

Text to be cleaned

```
2012-01-04 00:01:23,180 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-2281137920769708011_1116 src: /
127.0.0.1:32981 dest: /127.0.0.1:50010
2012-01-04 00:01:23,184 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32981, dest: /127.0.0.1:50010,
bytes: 3758, op: HDFS_WRITE, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid:
blk_-2281137920769708011_1116, duration: 2016056
2012-01-04 00:01:23,185 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-2281137920769708011_1116
terminating
2012-01-04 00:01:23,291 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_3766031435252346505_1117 src: /
127.0.0.1:32982 dest: /127.0.0.1:50010
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32982, dest: /127.0.0.1:50010,
bytes: 265, op: HDFS_WRITE, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid:
blk_3766031435252346505_1117, duration: 552828
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_3766031435252346505_1117
terminating
2012-01-04 00:01:23,324 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-8044922265890142318_1118 src: /
127.0.0.1:32983 dest: /127.0.0.1:50010
2012-01-04 00:01:23,326 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32983, dest: /127.0.0.1:50010,
bytes: 43, op: HDFS_WRITE, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid:
blk_-8044922265890142318_1118, duration: 607104
2012-01-04 00:01:23,327 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-8044922265890142318_1118
terminating
2012-01-04 00:01:23,409 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving block blk_-965793757262168743_1119 src: /
127.0.0.1:32984 dest: /127.0.0.1:50010
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace: src: /127.0.0.1:32984, dest: /127.0.0.1:50010,
bytes: 29743, op: HDFS_WRITE, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300176, blockid:
blk_-965793757262168743_1119, duration: 751930
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResponder 0 for block blk_-965793757262168743_1119
terminating
```

Approach: Cleaning by Example

Sample Input

```
1.  "Reported crime in 'Alabama',\n,
    \n2004,+4029.3\n2005,+3900\n2006,+3937\n2007,+3974.9\n2008,+4081.9",
2.  "Reported crime in 'Alaska',\n,
    \n2004,+3370.9\n2005,+3615\n2006,+3582\n2007,+3373.9\n2008,+2928.3”
```

Sample Output

```
1.  "Alabama","+4029.3","+3900","+3937","+3974.9","+4081.9"
2.  "Alaska","+3370.9","+3615","+3582","+3373.9","+2928.3"
```

Text to be cleaned

```
Reported crime in 'Alabama',
,
2004,+4029.3
2005,+3900
2006,+3937
2007,+3974.9
2008,+4081.9

Reported crime in 'Alaska',
,
2004,+3370.9
2005,+3615
2006,+3582
2007,+3373.9
2008,+2928.3

Reported crime in 'Arizona',
,
2004,+5073.3
2005,+4827
2006,+4741.6
2007,+4502.6
2008,+4087.3
```

Building an Inference Model

Step 1: encode examples at semantic level

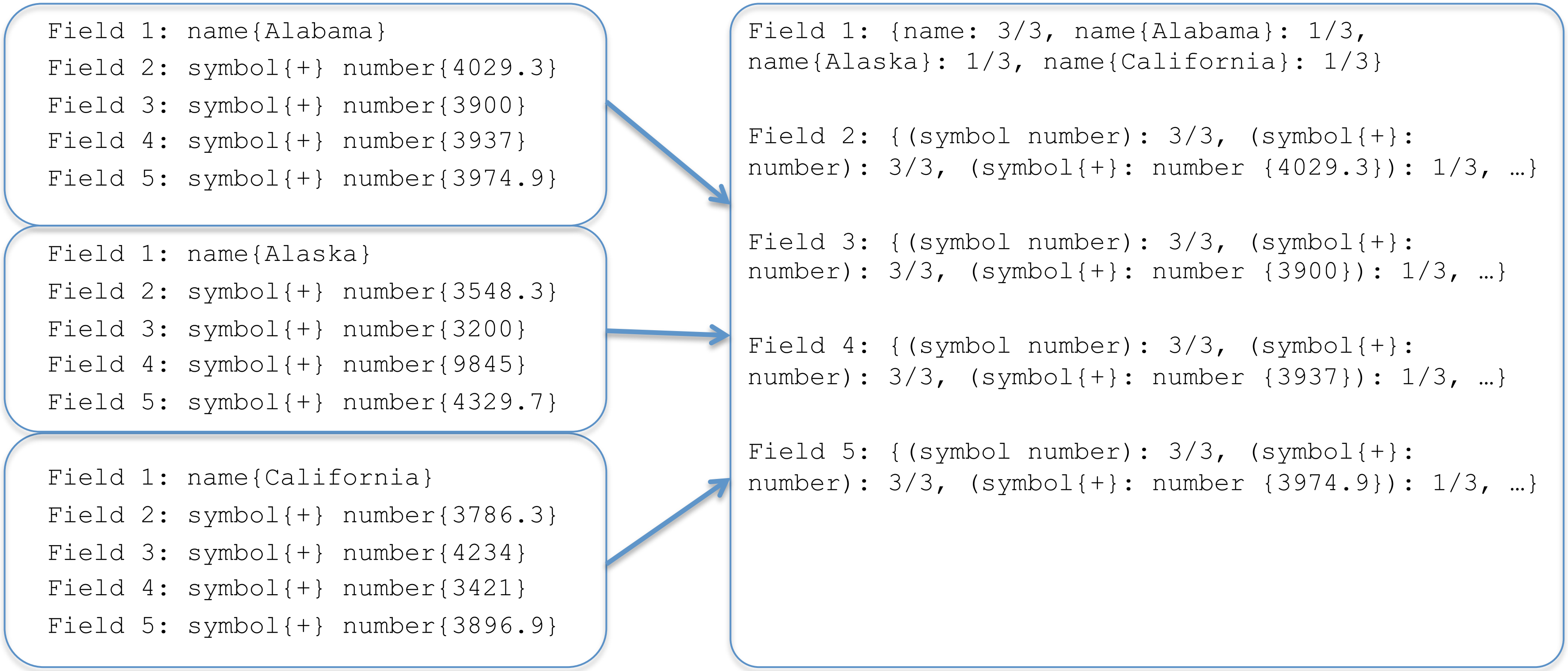
- Example:
  - "recorded crime in 2000" ➔
  - [name {recorded}, Symbol {' '}, name {'crime'}, name {'in'}, Symbol (' '}, number{'2000’}]

```
name{Reported} name{crime} name{in} symbol{’}
name{Alabama} symbol{’} symbol{,} symbol{\n}
symbol{,} symbol{\n}number{2004} symbol{,}
symbol{+}number{4029.3} symbol{\n} number{2005}
symbol{,} symbol{+} number{3900} symbol{\n}
number{2006} symbol{,} symbol{+} number{3937}
symbol{\n} number{2007} symbol{,} symbol{+}
number{3974.9} symbol{\n} number{2008} symbol{,}
symbol{+} number{4081.9}
```

```
Field 1: name{Alabama}
Field 2: symbol{+} number{4029.3}
Field 3: symbol{+} number{3900}
Field 4: symbol{+} number{3937}
Field 5: symbol{+} number{3974.9}
```

Building an Inference Model

Step 2: Build a Histogram on each output field



```
Field 1: name{Alabama}
Field 2: symbol{+} number{4029.3}
Field 3: symbol{+} number{3900}
Field 4: symbol{+} number{3937}
Field 5: symbol{+} number{3974.9}
```

```
Field 1: name{Alaska}
Field 2: symbol{+} number{3548.3}
Field 3: symbol{+} number{3200}
Field 4: symbol{+} number{9845}
Field 5: symbol{+} number{4329.7}
```

```
Field 1: name{California}
Field 2: symbol{+} number{3786.3}
Field 3: symbol{+} number{4234}
Field 4: symbol{+} number{3421}
Field 5: symbol{+} number{3896.9}
```

```
Field 1: {name: 3/3, name{Alabama}: 1/3,
name{Alaska}: 1/3, name{California}: 1/3}

Field 2: {(symbol number): 3/3, (symbol{+}:
number): 3/3, (symbol{+}: number {4029.3}): 1/3, …}

Field 3: {(symbol number): 3/3, (symbol{+}:
number): 3/3, (symbol{+}: number {3900}): 1/3, …}

Field 4: {(symbol number): 3/3, (symbol{+}:
number): 3/3, (symbol{+}: number {3937}): 1/3, …}

Field 5: {(symbol number): 3/3, (symbol{+}:
number): 3/3, (symbol{+}: number {3974.9}): 1/3, …}
```

Building an Inference Model

Step 3: find envelop

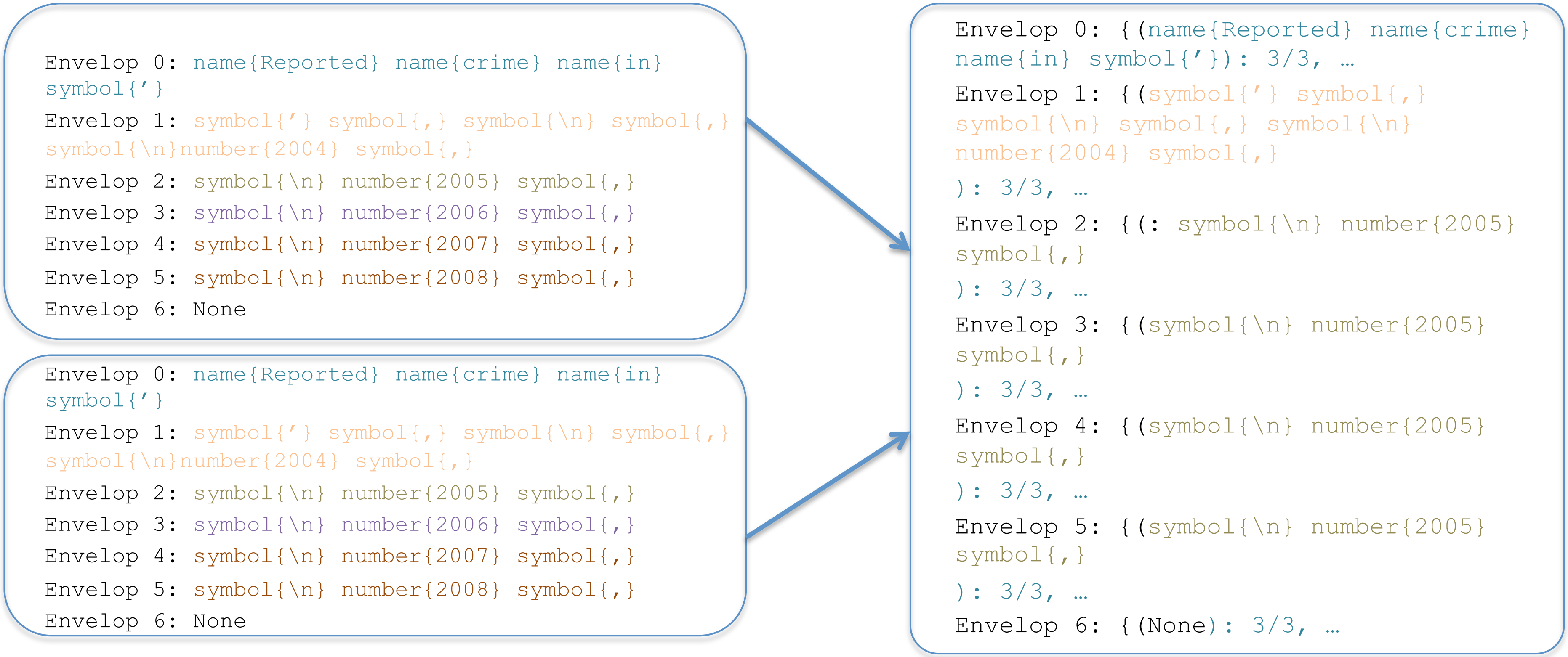
```
name{Reported} name{crime} name{in} symbol{’}
name{Alabama} symbol{’} symbol{,} symbol{\n}
symbol{,} symbol{\n}number{2004} symbol{,}
symbol{+}number{4029.3} symbol{\n} number{2005}
symbol{,} symbol{+} number{3900} symbol{\n}
number{2006} symbol{,} symbol{+} number{3937}
symbol{\n} number{2007} symbol{,}
symbol{+} number{3974.9} symbol{\n}
number{2008} symbol{,} symbol{+} number{4081.9}
```

```
Field 1: name{Alabama}
Field 2: symbol{+} number{4029.3}
Field 3: symbol{+} number{3900}
Field 4: symbol{+} number{3937}
Field 5: symbol{+} number{3974.9}
Field 6: symbol{+} number{4081.9}
```

```
Envelop 0: name{Reported} name{crime} name{in} symbol{’}
Envelop 1: symbol{’} symbol{,} symbol{\n} symbol{,} symbol{\n}number{2004} symbol{,}
Envelop 2: symbol{\n} number{2005} symbol{,}
Envelop 3: symbol{\n} number{2006} symbol{,}
Envelop 4: symbol{\n} number{2007} symbol{,}
Envelop 5: symbol{\n} number{2008} symbol{,}
Envelop 6: None
```

Building an Inference Model

Step 4: Build a Histogram on the envelops



```
Envelop 0: name{Reported} name{crime} name{in}
symbol{’}
Envelop 1: symbol{’} symbol{,} symbol{\n} symbol{,}
symbol{\n}number{2004} symbol{,}
Envelop 2: symbol{\n} number{2005} symbol{,}
Envelop 3: symbol{\n} number{2006} symbol{,}
Envelop 4: symbol{\n} number{2007} symbol{,}
Envelop 5: symbol{\n} number{2008} symbol{,}
Envelop 6: None
```

```
Envelop 0: name{Reported} name{crime} name{in}
symbol{’}
Envelop 1: symbol{’} symbol{,} symbol{\n} symbol{,}
symbol{\n}number{2004} symbol{,}
Envelop 2: symbol{\n} number{2005} symbol{,}
Envelop 3: symbol{\n} number{2006} symbol{,}
Envelop 4: symbol{\n} number{2007} symbol{,}
Envelop 5: symbol{\n} number{2008} symbol{,}
Envelop 6: None
```

```
Envelop 0: {(name{Reported} name{crime}
name{in} symbol{’}): 3/3, …
Envelop 1: {(symbol{’} symbol{,}
symbol{\n} symbol{,} symbol{\n}
number{2004} symbol{,}
): 3/3, …
Envelop 2: {(: symbol{\n} number{2005}
symbol{,}
): 3/3, …
Envelop 3: {(symbol{\n} number{2005}
symbol{,}
): 3/3, …
Envelop 4: {(symbol{\n} number{2005}
symbol{,}
): 3/3, …
Envelop 5: {(symbol{\n} number{2005}
symbol{,}
): 3/3, …
Envelop 6: {(None): 3/3, …
```

Building an Inference Model

Step 5: Generalize

```
Envelop 0: name{Reported} name{crime} name{in} symbol{’}
Envelop 0: name{Reported} name{murder} name{in} symbol{’}
Envelop 0: name{Reported} name{murder} name{around} symbol{’}
Envelop 0: name{Reported} name{burglary} name{in} symbol{’}

➔   name{Reported(1)} name{crime(0.25) | murder(0.5), burglary(0.25)} name{in(0.67) around(0.33)} symbol{’}

Envelop 0: name name name number name
Envelop 0: name name
Envelop 0: name
Envelop 0: number number name
➔  name[*]
```

Extraction

- Given a string to extract:
  - Find the lexical tokens of the string
  - name{Reported} name{crime} name{in} symbol{’} name{Alabama} symbol{’} symbol{,} symbol{\n} symbol{,} symbol{\n}number{2004} symbol{,}symbol{+}number{4029.3} symbol{\n} number{2005} symbol{,} symbol{+} number{3900} symbol{\n} number{2006} symbol{,} symbol{+} number{3937} symbol{\n} number{2007} symbol{,} symbol{+} number{3974.9} symbol{\n} number{2008} symbol{,} symbol{+} number{4081.9}
  - project the tokens onto the output_fields_signature and envelop_signature such that scores (probabilities) are maximum
  - **maximize** Pr(Envelop 0) + Pr(Field 1) + Pr (Envelop 1) + Pr (Field 2) + …

Questions