Column Projection

Zihua

Feature Engineering: Column Types

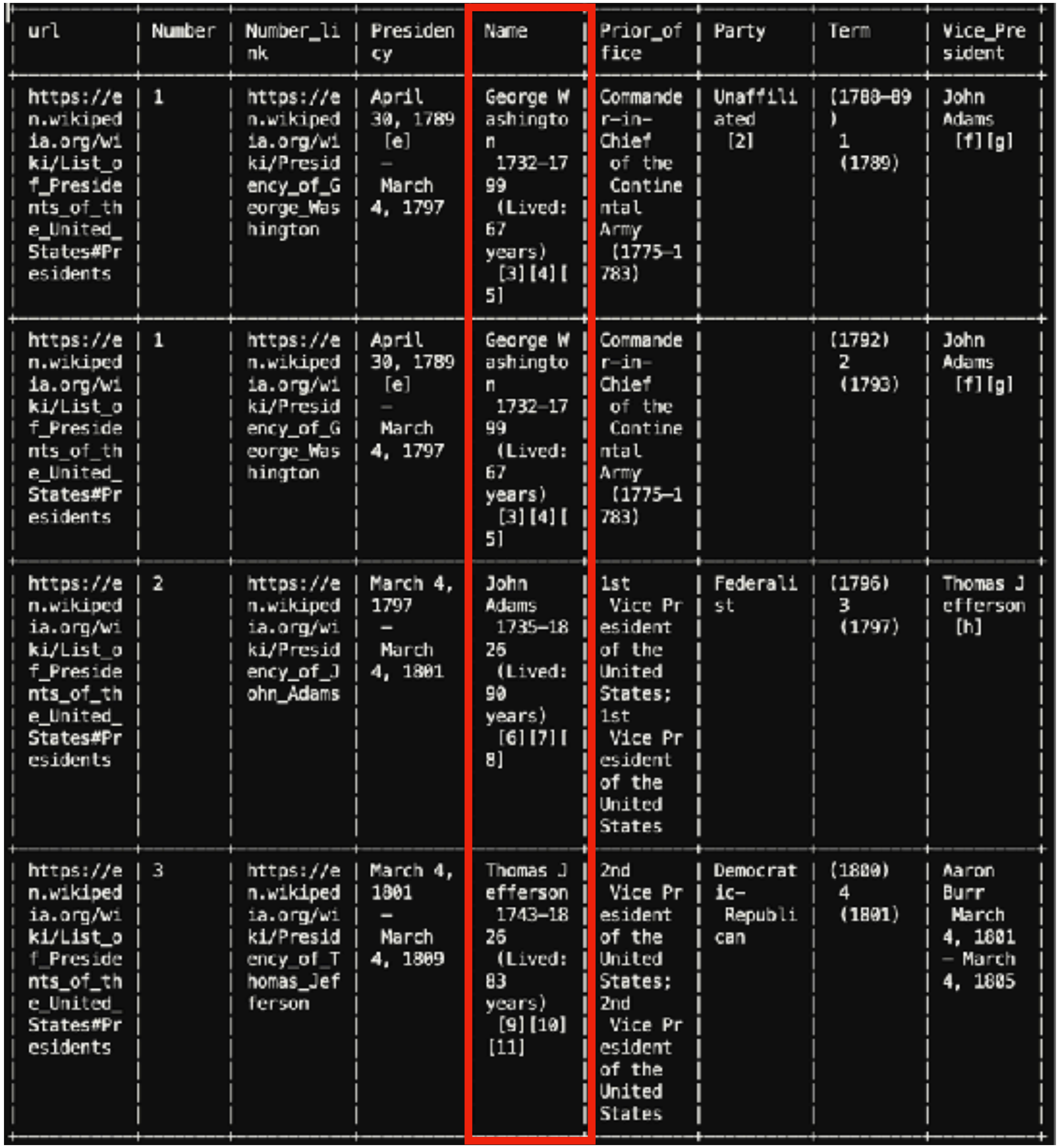
- Text
- Numerical
- Currency
- Percentage
- Datetime
- Boolean
- URL

Feature Engineering: Column Types

- Cons of Rule-Based Approach
    - Low Robustness
    - Overfitting on the training set
    - Feature vector is one-hot, while columns may contain multiple types of information

Feature Engineering: Column Types

If use rule-based algorithm, then "Name" column should be attached with type "Text", so the datetime information in the column would be lost in the one hot vector.

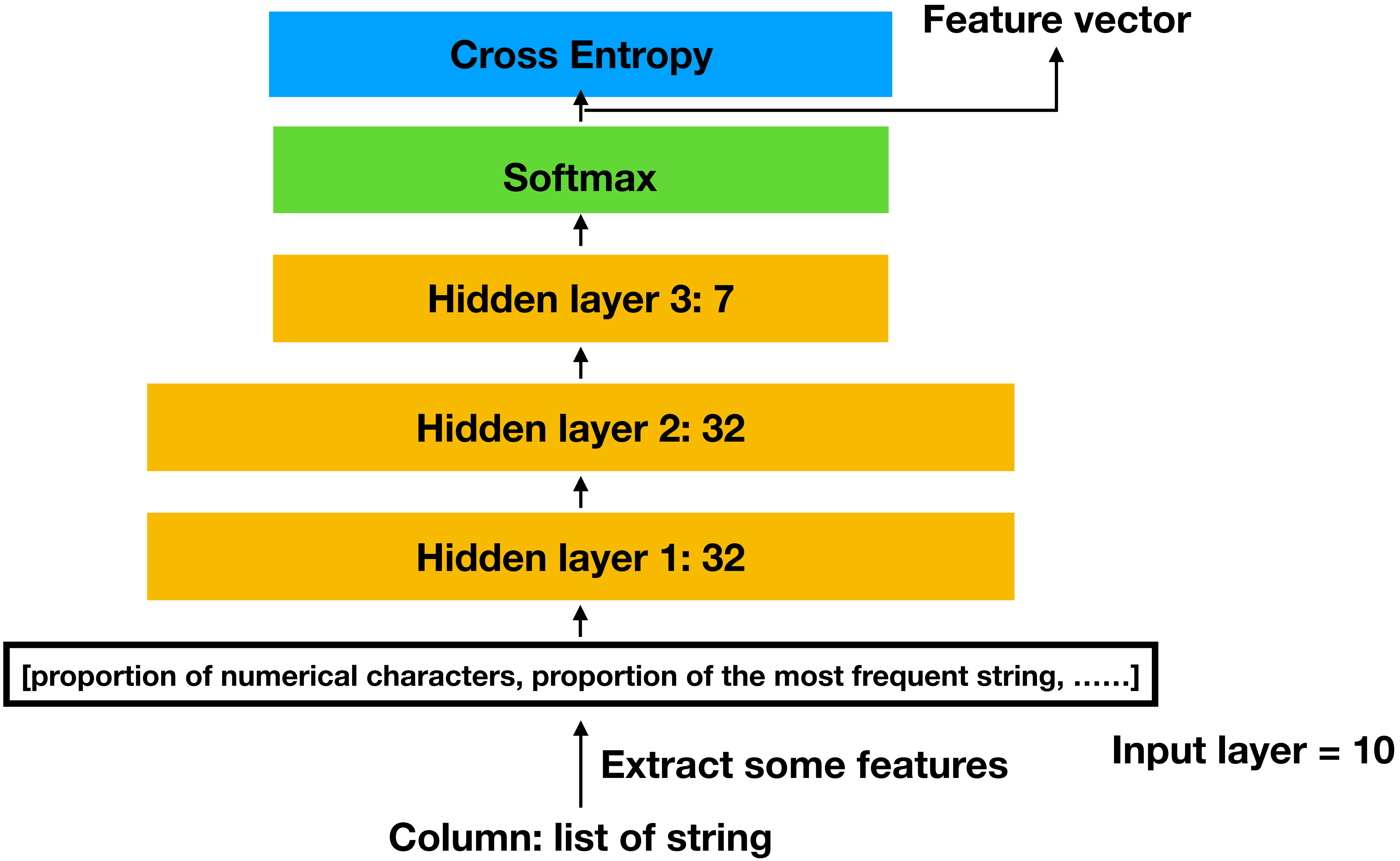


```
+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
| url       | Number    | Number_li | Presiden  | Name      | Prior_of  | Party     | Term      | Vice_Pre  |
|           |           | nk        | cy        |           | fice      |           |           | sident    |
+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
| https://e | 1         | https://e | April     | George W  | Commande  | Unaffili  | (1788–89  | John      |
| n.wikiped |           | n.wikiped | 30, 1789  | ashingto  | r-in-     | ated      | )         | Adams     |
| ia.org/wi |           | ia.org/wi |  [e]      | n         | Chief     |   [2]     |  1        |   [f][g]  |
| ki/List_o |           | ki/Presid |  –        |  1732–17  |  of the   |           |  (1789)   |           |
| f_Preside |           | ency_of_G |  March    | 99        |  Contine  |           |           |           |
| nts_of_th |           | eorge_Was | 4, 1797   |  (Lived:  | ntal      |           |           |           |
| e_United_ |           | hington   |           | 67        | Army      |           |           |           |
| States#Pr |           |           |           | years)    |  (1775–1  |           |           |           |
| esidents  |           |           |           |   [3][4][ | 783)      |           |           |           |
|           |           |           |           | 5]        |           |           |           |           |
+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
| https://e | 1         | https://e | April     | George W  | Commande  |           | (1792)    | John      |
| n.wikiped |           | n.wikiped | 30, 1789  | ashingto  | r-in-     |           |  2        | Adams     |
| ia.org/wi |           | ia.org/wi |  [e]      | n         | Chief     |           |  (1793)   |   [f][g]  |
| ki/List_o |           | ki/Presid |  –        |  1732–17  |  of the   |           |           |           |
| f_Preside |           | ency_of_G |  March    | 99        |  Contine  |           |           |           |
| nts_of_th |           | eorge_Was | 4, 1797   |  (Lived:  | ntal      |           |           |           |
| e_United_ |           | hington   |           | 67        | Army      |           |           |           |
| States#Pr |           |           |           | years)    |  (1775–1  |           |           |           |
| esidents  |           |           |           |   [3][4][ | 783)      |           |           |           |
|           |           |           |           | 5]        |           |           |           |           |
+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
| https://e | 2         | https://e | March 4,  | John      | 1st       | Federali  | (1796)    | Thomas J  |
| n.wikiped |           | n.wikiped | 1797      | Adams     |  Vice Pr  | st        |  3        | efferson  |
| ia.org/wi |           | ia.org/wi |  –        |  1735–18  | esident   |           |  (1797)   |   [h]     |
| ki/List_o |           | ki/Presid |  March    | 26        | of the    |           |           |           |
| f_Preside |           | ency_of_J | 4, 1801   |  (Lived:  | United    |           |           |           |
| nts_of_th |           | ohn_Adams |           | 90        | States;   |           |           |           |
| e_United_ |           |           |           | years)    | 1st       |           |           |           |
| States#Pr |           |           |           |   [6][7][ |  Vice Pr  |           |           |           |
| esidents  |           |           |           | 8]        | esident   |           |           |           |
|           |           |           |           |           | of the    |           |           |           |
|           |           |           |           |           | United    |           |           |           |
|           |           |           |           |           | States    |           |           |           |
+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
| https://e | 3         | https://e | March 4,  | Thomas J  | 2nd       | Democrat  | (1800)    | Aaron     |
| n.wikiped |           | n.wikiped | 1801      | efferson  |  Vice Pr  | ic-       |  4        | Burr      |
| ia.org/wi |           | ia.org/wi |  –        |  1743–18  | esident   |  Republi  |  (1801)   |  March    |
| ki/List_o |           | ki/Presid |  March    | 26        | of the    | can       |           | 4, 1801   |
| f_Preside |           | ency_of_T | 4, 1809   |  (Lived:  | United    |           |           | – March   |
| nts_of_th |           | homas_Jef |           | 83        | States;   |           |           | 4, 1805   |
| e_United_ |           | ferson    |           | years)    | 2nd       |           |           |           |
| States#Pr |           |           |           |  [9][10]  |  Vice Pr  |           |           |           |
| esidents  |           |           |           | [11]      | esident   |           |           |           |
|           |           |           |           |           | of the    |           |           |           |
|           |           |           |           |           | United    |           |           |           |
|           |           |           |           |           | States    |           |           |           |
+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+-----------+
```

Feature Engineering: Column Types

- Machine Learning approach
  - Manually labeled all the columns in the training set
  - Trained a MLP for classification
  - Generate a probability distribution over all types for each column

Feature Engineering: Column Types



Feature vector

Cross Entropy

Softmax

Hidden layer 3: 7

Hidden layer 2: 32

Hidden layer 1: 32

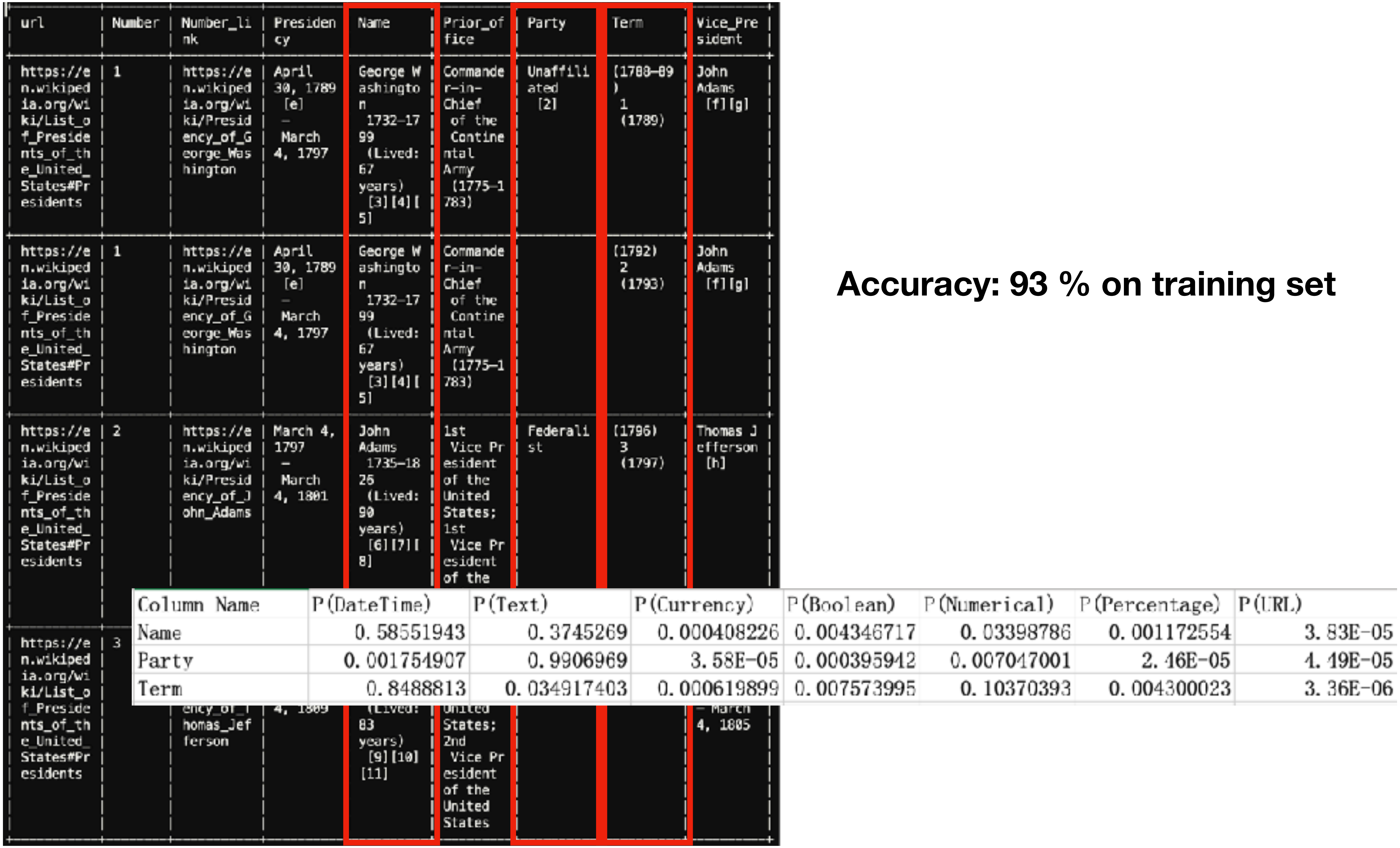
[proportion of numerical characters, proportion of the most frequent string, ……]

Extract some features

Input layer = 10

Column: list of string

Feature Engineering: Column Types



```
+-----------+--------+-----------+-----------+----------+-----------+------------+-----------+-----------+
| url       | Number | Number_li | Presiden  | Name     | Prior_of  | Party      | Term      | Vice_Pre  |
|           |        | nk        | cy        |          | fice      |            |           | sident    |
+-----------+--------+-----------+-----------+----------+-----------+------------+-----------+-----------+
| https://e | 1      | https://e | April     | George W | Commande  | Unaffili   | (1788–89  | John      |
| n.wikiped |        | n.wikiped | 30, 1789  | ashingto | r-in-     | ated       | )         | Adams     |
| ia.org/wi |        | ia.org/wi |  [e]      | n        | Chief     |  [2]       |  1        |  [f][g]   |
| ki/List_o |        | ki/Presid |  –        |  1732–17 |  of the   |            |  (1789)   |           |
| f_Preside |        | ency_of_G |  March    | 99       |  Contine  |            |           |           |
| nts_of_th |        | eorge_Was | 4, 1797   |  (Lived: | ntal      |            |           |           |
| e_United_ |        | hington   |           | 67       | Army      |            |           |           |
| States#Pr |        |           |           | years)   |  (1775–1  |            |           |           |
| esidents  |        |           |           |  [3][4][ | 783)      |            |           |           |
|           |        |           |           | 5]       |           |            |           |           |
+-----------+--------+-----------+-----------+----------+-----------+------------+-----------+-----------+
| https://e | 1      | https://e | April     | George W | Commande  |            | (1792)    | John      |
| n.wikiped |        | n.wikiped | 30, 1789  | ashingto | r-in-     |            | 2         | Adams     |
| ia.org/wi |        | ia.org/wi |  [e]      | n        | Chief     |            | (1793)    |  [f][g]   |
| ki/List_o |        | ki/Presid |  –        |  1732–17 |  of the   |            |           |           |
| f_Preside |        | ency_of_G |  March    | 99       |  Contine  |            |           |           |
| nts_of_th |        | eorge_Was | 4, 1797   |  (Lived: | ntal      |            |           |           |
| e_United_ |        | hington   |           | 67       | Army      |            |           |           |
| States#Pr |        |           |           | years)   |  (1775–1  |            |           |           |
| esidents  |        |           |           |  [3][4][ | 783)      |            |           |           |
|           |        |           |           | 5]       |           |            |           |           |
+-----------+--------+-----------+-----------+----------+-----------+------------+-----------+-----------+
| https://e | 2      | https://e | March 4,  | John     | 1st       | Federali   | (1796)    | Thomas J  |
| n.wikiped |        | n.wikiped | 1797      | Adams    |  Vice Pr  | st         | 3         | efferson  |
| ia.org/wi |        | ia.org/wi |  –        |  1735–18 | esident   |            | (1797)    |  [h]      |
| ki/List_o |        | ki/Presid |  March    | 26       | of the    |            |           |           |
| f_Preside |        | ency_of_J | 4, 1801   |  (Lived: | United    |            |           |           |
| nts_of_th |        | ohn_Adams |           | 90       | States;   |            |           |           |
| e_United_ |        |           |           | years)   | 1st       |            |           |           |
| States#Pr |        |           |           |  [6][7][ |  Vice Pr  |            |           |           |
| esidents  |        |           |           | 8]       | esident   |            |           |           |
|           |        |           |           |          | of the    |            |           |           |
+-----------+--------+-----------+-----------+----------+-----------+------------+-----------+-----------+
| https://e | 3      |
| n.wikiped |        |
| ia.org/wi |        |
| ki/List_o |        |
| f_Preside |        | ency_of_T | 4, 1809   |  (Lived: | United    |            |           | – March   |
| nts_of_th |        | homas_Jef |           | 83       | States;   |            |           | 4, 1805   |
| e_United_ |        | ferson    |           | years)   | 2nd       |            |           |           |
| States#Pr |        |           |           |  [9][10] |  Vice Pr  |            |           |           |
| esidents  |        |           |           | [11]     | esident   |            |           |           |
|           |        |           |           |          | of the    |            |           |           |
|           |        |           |           |          | United    |            |           |           |
|           |        |           |           |          | States    |            |           |           |
+-----------+--------+-----------+-----------+----------+-----------+------------+-----------+-----------+
```

Column Name	P(DateTime)	P(Text)	P(Currency)	P(Boolean)	P(Numerical)	P(Percentage)	P(URL)
Name	0.58551943	0.3745269	0.000408226	0.004346717	0.03398786	0.001172554	3.83E-05
Party	0.001754907	0.9906969	3.58E-05	0.000395942	0.007047001	2.46E-05	4.49E-05
Term	0.8488813	0.034917403	0.000619899	0.007573995	0.10370393	0.004300023	3.36E-06

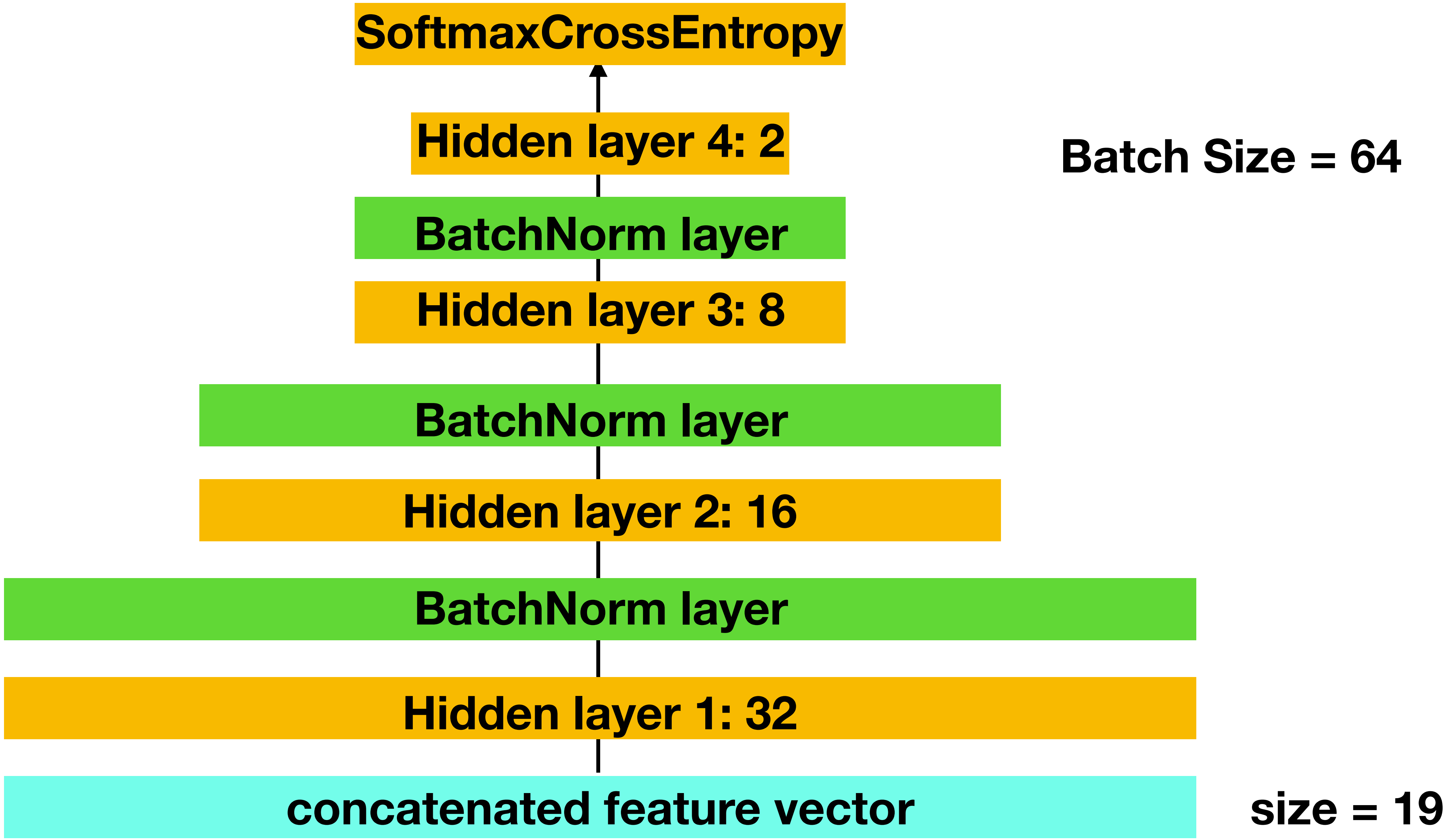
Accuracy: 93 % on training set

Column projection

- Number of columns (Lucile)
- Word2vec proximities (Lucile)
- Column Types (Zihua)
- Question Types (Nithin)—> Transform to one-hot vector

Column projection

For training set, there are total 2046 cases, 273 positive (label = 1), 1773 negative (label = 0), so copy positive cases 6 times


SoftmaxCrossEntropy

Hidden layer 4: 2

BatchNorm layer

Hidden layer 3: 8

BatchNorm layer

Hidden layer 2: 16

BatchNorm layer

Hidden layer 1: 32

concatenated feature vector

Batch Size = 64

size = 19

Column projection

Training set

	Postitive	Negative
True	222	1407
False	51	366
	273	1773

Test set

	Postitive	Negative
True	49	394
False	17	99
	66	493