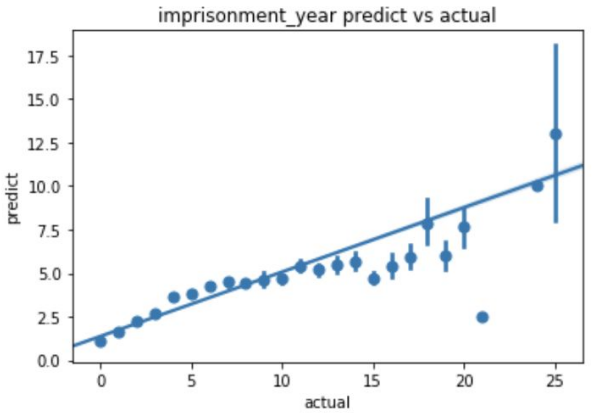
What features are correlated to the length of sentence?

- Studying Chinese court documents from wenshu.gov.cn.

Violet Yao



Structure of Presentation

- Data collection
- Dataset Construction
- Analysis
- Results
- Challenges
- Tools
- Next Steps: How would you take this project forward?
- Advertising

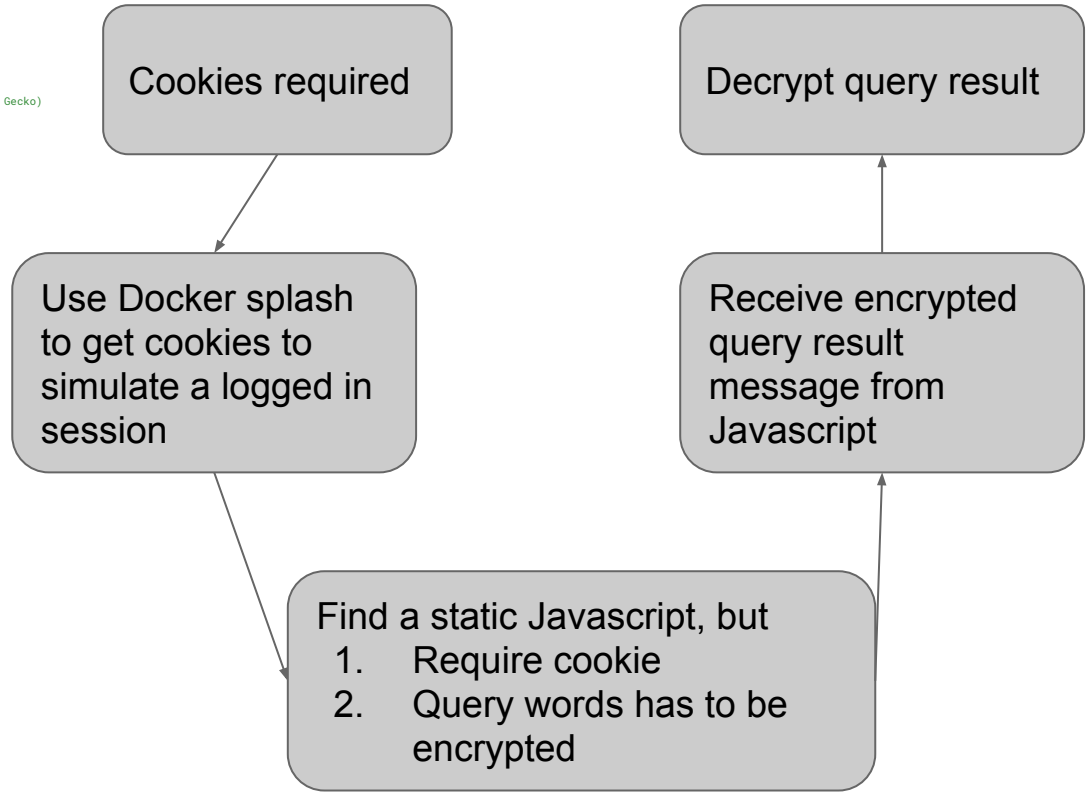
Data Collection

- Data source: http://wenshu.court.gov.cn/

- Data size: collected ~ 1,000,000 cases, ended up using 150,000
  - ```
    In [36]: df_new = df_new.sample(n=150000, random_state=42)
    ```
  - ```
    In [37]: df_new.to_csv("data.csv")
    ```

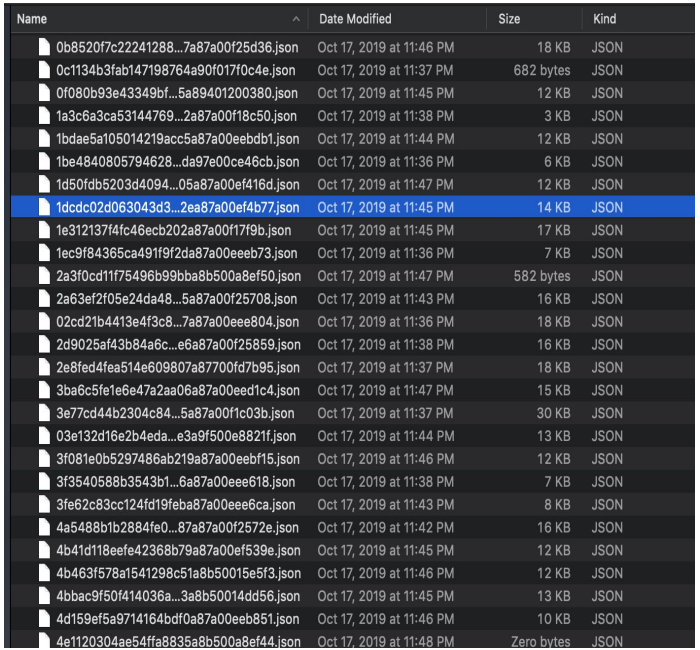
Data Collection - Web Scraping

```
class Wenshu:
#     url : parse.ParseResult = parse.urlparse("http://wenshu.court.gov.cn/website/parse/rest.q4w")
    url = parse.urlparse("http://wenshu.court.gov.cn/website/parse/rest.q4w")
#     url = parse.ParseResult


    def __init__(self):
        self.session = requests.Session()
        self.session.headers.update({
            "User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/76.0.3809.132 Safari/537.36",
        })

    def _request(self, data: dict) -> requests.Response:
        response = self.session.post(self.url.geturl(), data=data)
        if response.status_code != 200:
            raise Exception(response.status_code)

        json_data = response.json()

        plain_text = des3decrypt(cipher_text=json_data["result"],
                                 key=json_data["secretKey"],
                                 iv=datetime.now().strftime("%Y%m%d"))

        result = json.loads(plain_text)
        return result

    def list_page(self, word):
        """input query, return page list"""

        print("Page id: \n", PageID())
        print()

        # the maximum size for one page is 1000
        data = {
            "pageId": PageID(),
            "sortFields": "s50:desc",
            "ciphertext": CipherText(),
            "pageNum": 1,
            "pageSize": 1000,
            "queryCondition": json.dumps([{"key": "s42", "value": "2018"},
                                          {"key": "s33", "value": "上海市"},
                                          {"key": "s8", "value": "02"},
                                          {"key": "s45", "value": word}]),
            "cfg": "com.lawyee.judge.dc.parse.dto.SearchDataDsoDTO@queryDoc",
            "__RequestVerificationToken": RequestVerificationToken(24),
        }

        result = self._request(data)
#         print("list_page: ", result)
        id_lst = list(result['relWemshu'].keys())
#         print(id_lst)
        print("Query word: ", word)
        print("length: ", len(id_lst))
        print()
        return id_lst

    def detail_page(self, doc_id):
        """input """
        print(doc_id)
        data = {
            # ORIGINAL: 4e00b8ae589b4288a725aabe00c0e683
#             "docId": "4e00b8ae589b4288a725aabe00c0e683",
            "docId": doc_id,
            "ciphertext": CipherText(),
            "cfg": "com.lawyee.judge.dc.parse.dto.SearchDataDsoDTO@docInfoSearch",
            "__RequestVerificationToken": RequestVerificationToken(24),
        }

        result = self._request(data)
#         print("detail_page: ", result)
        return result
```



Data Collection

Raw data format: a JSON file per case



Name	Date Modified	Size	Kind
0b8520f7c22241288...7a87a00f25d36.json	Oct 17, 2019 at 11:46 PM	18 KB	JSON
0c1134b3fab147198764a90f017f0c4e.json	Oct 17, 2019 at 11:37 PM	682 bytes	JSON
0f080b93e43349bf...5a89401200380.json	Oct 17, 2019 at 11:45 PM	12 KB	JSON
1a3c6a3ca53144769...2a87a00f18c50.json	Oct 17, 2019 at 11:38 PM	3 KB	JSON
1bdae5a105014219acc5a87a00eebdb1.json	Oct 17, 2019 at 11:44 PM	12 KB	JSON
1be4840805794628...da97e00ce46cb.json	Oct 17, 2019 at 11:36 PM	6 KB	JSON
1d50fdb5203d4094...05a87a00ef416d.json	Oct 17, 2019 at 11:47 PM	12 KB	JSON
1dcdc02d063043d3...2ea87a00ef4b77.json	Oct 17, 2019 at 11:45 PM	14 KB	JSON
1e312137f4fc46ecb202a87a00f17f9b.json	Oct 17, 2019 at 11:45 PM	17 KB	JSON
1ec9f84365ca491f9f2da87a00eeeb73.json	Oct 17, 2019 at 11:36 PM	7 KB	JSON
2a3f0cd11f75496b99bba8b500a8ef50.json	Oct 17, 2019 at 11:47 PM	582 bytes	JSON
2a63ef2f05e24da48...5a87a00f25708.json	Oct 17, 2019 at 11:43 PM	16 KB	JSON
02cd21b4413e4f3c8...7a87a00eee804.json	Oct 17, 2019 at 11:36 PM	18 KB	JSON
2d9025af43b84a6c...e6a87a00f25859.json	Oct 17, 2019 at 11:38 PM	16 KB	JSON
2e8fed4fea514e609807a87700fd7b95.json	Oct 17, 2019 at 11:37 PM	18 KB	JSON
3ba6c5fe1e6e47a2aa06a87a00eed1c4.json	Oct 17, 2019 at 11:47 PM	15 KB	JSON
3e77cd44b2304c84...5a87a00f1c03b.json	Oct 17, 2019 at 11:37 PM	30 KB	JSON
03e132d16e2b4eda...e3a9f500e8821f.json	Oct 17, 2019 at 11:44 PM	13 KB	JSON
3f081e0b5297486ab219a87a00eebf15.json	Oct 17, 2019 at 11:46 PM	12 KB	JSON
3f3540588b3543b1...6a87a00eee618.json	Oct 17, 2019 at 11:38 PM	7 KB	JSON
3fe62c83cc124fd19feba87a00eee6ca.json	Oct 17, 2019 at 11:43 PM	8 KB	JSON
4a5488b1b2884fe0...87a87a00f2572e.json	Oct 17, 2019 at 11:42 PM	16 KB	JSON
4b41d118eefe42368b79a87a00ef539e.json	Oct 17, 2019 at 11:45 PM	12 KB	JSON
4b463f578a1541298c51a8b50015e5f3.json	Oct 17, 2019 at 11:46 PM	12 KB	JSON
4bbac9f50f414036a...3a8b50014dd56.json	Oct 17, 2019 at 11:45 PM	13 KB	JSON
4d159ef5a9714164bdf0a87a00eeb851.json	Oct 17, 2019 at 11:46 PM	10 KB	JSON
4e1120304ae54ffa8835a8b500a8ef44.json	Oct 17, 2019 at 11:48 PM	Zero bytes	JSON

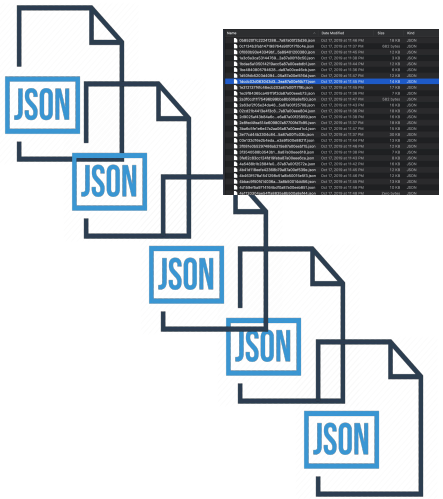
A sample raw data json:

Key-value pairs

{"s1": "□某某危险驾驶一审刑事判决书", "s2": "上海市虹口区人民法院", "s3": "925", "s5": "1d50fdb5203d4094bd05a87a00ef416d", "s6": "01", "s7": "(2017)沪0109刑初1096号", "s8": "刑事案件", "s9": "刑事一审", "s31": "2018-01-09", "s41": "2018-02-01", "s22": "上海市虹口区人民法院\n刑事判决书\n(2017)沪0109刑初1096号", "s23": "上海市虹口区人民检察院以沪虹检诉刑诉〔2017〕XXXX号起诉书指控被告人史某某犯危险驾驶罪，于2017年12月29日向本院提起公诉。本院依法适用简易程序，实行独任审判，公开开庭审理了本案。上海市虹口区人民检察院指派检察员周某出庭支持公诉，被告人史某某到庭参加诉讼。现已审理终结", "s25": "上海市虹口区人民检察院指控：被告人史某某于2017年9月1日凌晨3时20分许，酒后驾驶牌号为鄂A7XXXX的小轿车，行驶至本市虹口区汶水东路、凉城路东约50米处时被民警查获。经现场呼气式酒精测试，酒精含量为135mg／100ml。后被告人史某某被带至医院提取血样。案发后，经上海润家生物医药科技有限公司司法鉴定所鉴定：送检史某某的血样中检出酒精含量为116.3mg／100ml。\n上述事实，被告人史某某在开庭审理过程中无异议，并有证人边某的证言，上海市公安局虹口分局出具的《呼气式酒精测试仪结果打印单》、《当事人血样提取登记表》、《案发经过》、《查获经过》及上海润家生物医药科技有限公司司法鉴定所出具的《司法鉴定意见书》等证据证实，足以认定", "s26": "本院认为，被告人史某某在道路上醉酒驾驶机动车，其行为已构成危险驾驶罪。上海市虹口区人民检察院指控被告人史某某犯危险驾驶罪罪名成立。被告人史某某到案后能如实供述自己的罪行，且能认罪认罚，并在本院审理期间预缴了罚金，确有悔罪表

Dataset Construction

JSON files per case



Pre-processing: only select keys we are interested in; compiled all cases into one JSON

one large JSON file

Pandas dataframe

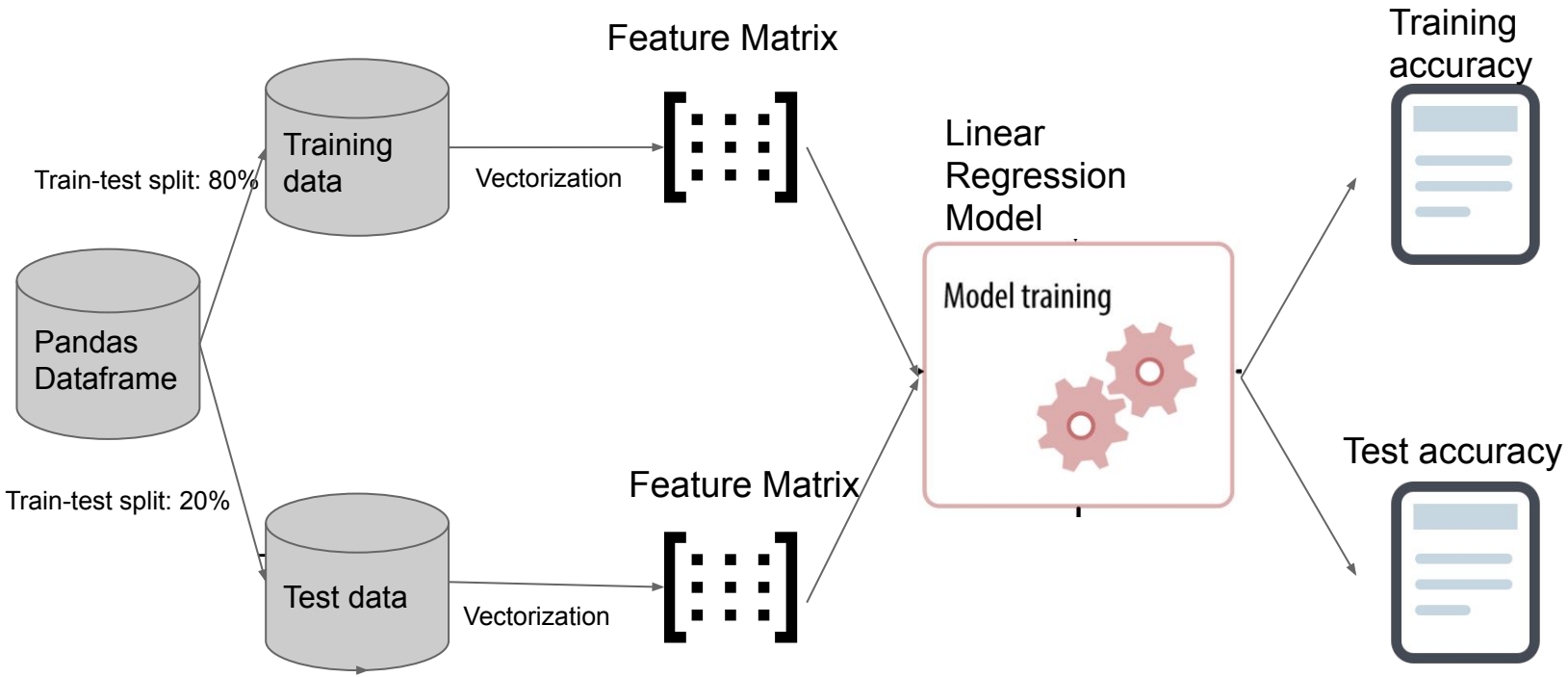
```
In [54]: df.head()
```

Out[54]:

	accusation	defendants	fact	punish_of_money	relevant_articles	term_of_imprisonment	death_penalty	imprisonment	life_imprisonment
0	['受贿']	['杨1某']	安阳县人民检察院指控：2009年至2012年，被告人杨1某利用其在洛阳市中级人民法院工作，先...	0	[383, 388, 383, 386]	{'death_penalty': False, 'imprisonment': 126, ...	False	126	False
1	['组织、强迫、引诱、容留、介绍卖淫', '引诱、容留、介绍卖淫']	['马某某']	大同市新荣区人民检察院指控：从2013年4月起，被告人马某某在新荣区新荣村村委会南自租的宁厉...	0	[359]	{'death_penalty': False, 'imprisonment': 7, '...	False	7	False
2	['非法持有、私藏[枪支、弹药]', '走私、贩卖、运输、制造[毒品]']	['王某']	公诉机关指控，自2010年开始，被告人王某从"傻波"（在逃，另案处理）处购买甲基苯丙胺（冰毒...	3000	[347]	{'death_penalty': False, 'imprisonment': 10, '...	False	10	False
3	['走私、贩卖、运输、制造[毒品]', '容留他人吸毒']	['李1某']	晋中市榆次区人民检察院指控：一、2016年2月11日，被告人李1某在晋中东高速口附近路段经申...	12000	[347, 347, 354]	{'death_penalty': False, 'imprisonment': 102, ...	False	102	False
4	['非法经营']	['彭某']	广西壮族自治区田东县人民检察院指控，2012年10月22日，被告人彭某为获利，在明知运输的物...	30000	[225]	{'death_penalty': False, 'imprisonment': 36, '...	False	36	False

Selected Keys: fact, relevant articles, accusation types, penalty (money, sentence length, if it is death penalty of life imprisonment)

Modeling Overview



Modeling - Feature Matrix construction

Recall research question:

What features correlate to the length of sentence?

- Numeric

```
In [60]: df["punish_of_money"].describe()

Out[60]: count    1.500000e+05
         mean     2.500425e+04
         std      1.479592e+06
         min      0.000000e+00
         25%      0.000000e+00
         50%      0.000000e+00
         75%      6.000000e+03
         max      5.500000e+08
         Name: punish_of_money, dtype: float64
```

- Categorical

```
In [63]: df["accusation"].value_counts()

Out[63]: ['[走私、贩卖、运输、制造]毒品']                                                    4140
         ['抢劫', '盗窃']                                                              2684
         ['盗窃']                                                                    2209
         ['非法占用农用地']                                                             1935
         ['[生产、销售]假药']                                                           1876
         ['[生产、销售]不符合安全标准的食品']                                                1875
         ['非法种植毒品原植物']                                                         1872
         ['重大责任事故']                                                             1866
         ['失火']                                                                    1864
         ['污染环境']                                                                1863
         ['非法行医']                                                                1859
         ['假冒注册商标']                                                             1851
         ['过失致人死亡']                                                             1849
         ['销售假冒注册商标的商品']                                                     1849
         ['拒不执行[判决、裁定]']                                                      1842
         ['[生产、销售][有毒、有害]食品']                                               1842
         ['[组织、强迫、引诱、容留、介绍]卖淫', '[引诱、容留、介绍]卖淫']                      1837
         ['[虚开增值税专用发票、用于骗取出口退税、抵扣税款发票]']                          1834
         ['[走私、贩卖、运输、制造]毒品', '容留他人吸毒']                                1826
         ['非法经营']                                                                1823
```

Melting &
One hot encoding:
pandas.get_dummies()

- Text

Modeling - feature matrix construction - Text

Raw text

Using package Jieba

'安阳县人民检察院指控：2009年至2012年，被告人杨1某利用其在洛阳市中级人民法院工作，先后担任民某副庭长、执行局副局长的便利条件，通过让本院同事在审理案件过程中为案件当事人提供帮助，非法收受他人贿赂现金及购物卡共计36.8万元，个人实得32.8万元。案发后赃款已退。具体如下：1、2009年至2011年期间，被告人杨1某利用其在洛阳市中级人民法院工作，先后担任民某副庭长、执行局副局长的便利条件，让同院法官王某在审理洛阳河科大齿轮制造有限公司诉洛阳科锐机电设备有限公司技术合同纠纷一案和洛阳鸿拓重型齿轮箱有限公司诉洛阳科锐机电设备有限公司技术合同纠纷一案时，为原告提供帮助，先后四次非法收受付某某现金19.5万元和购物卡1万元，杨1某将其中4万元送给王某。2、2011年9月至2012年3月份期间，被告人杨1某利用其先后担任民某副庭长、执行局副局长的便利条件，多次打招呼让民某法官杨某、王某甲在审理河南红旗渠建设集团有限公司诉嵩县中医院工程欠款纠纷一案中给予原告照顾，非法收受朱某某现金15万元及购物卡1.3万元。为认定上述事实，公诉机关提供了被告人供述、证人证言、书证等证据，并据此认为被告人的行为触犯了《中华人民共和国刑法》××××、××××第（一）项、××，应当以××罪追究其刑事责任。提请本院依法判处。'

Tokenized text

"['安阳县', '人民检察院', '指控', '：', '2009', '年', '至', '2012', '年', '，', '被告人', '杨', '1', '某', '利用', '其', '在', '洛阳市', '中级', '人民法院', '工作', '，', '先后', '担任', '民某', '副庭长', '、', '执行局', '副局长', '的', '便利', '条件', '，', '通过', '让', '本院', '同事', '在', '审理案件', '过程', '中为', '案件', '当事人', '提供', '帮助', '，', '非法', '收受', '他人', '贿赂', '现金', '及', '购物', '卡', '共计', '36.8', '万元', '，', '个人', '实得', '32.8', '万元', '。', '案发后', '赃款', '已退', '。', '具体', '如下', '：', '1', '、', '2009', '年', '至', '2011', '年', '期间', '，', '被告人', '杨', '1', '某', '利用', '其', '在', '洛阳市', '中级', '人民法院', '工作', '，', '先后', '担任', '民某', '副庭长', '、', '执行局', '副局长', '的', '便利', '条件', '，', '让', '同院', '法官', '王某', '在', '审理', '洛阳', '河', '科大', '齿轮', '制造', '有限公司', '诉', '洛阳', '科锐', '机电设备', '有限公司', '技术', '合同纠纷', '一案', '和', '洛阳', '鸿拓', '重型', '齿轮箱', '有限公司', '诉', '洛阳', '科锐', '机电设备', '有限公司', '技术', '合同纠纷', '一案', '时', '，', '为', '原告', '提供', '帮助', '，', '先后', '四次', '非法', '收受', '付', '某某', '现金', '19.5', '万元', '和', '购物', '卡', '1', '万元', '，', '杨', '1', '某', '将', '其中', '4', '万元', '送给', '王某', '。', '2', '、', '2011', '年', '9', '月', '至', '2012', '年', '3', '月份', '期间', '，', '被告人', '杨', '1', '某', '利用', '其', '先后', '担任', '民某', '副庭长', '、', '执行局', '副局长', '的', '便利', '条件', '，', '多次', '打招呼', '让民某', '法官', '杨某', '、', '王某', '甲', '在', '审理', '河南', '红旗渠', '建设', '集团', '有限公司', '诉', '嵩县', '中医院', '工程', '欠款', '纠纷', '一案', '中', '给予', '原告', '照顾', '，', '非法', '收受', '朱', '某某', '现金', '15', '万元', '及', '购物', '卡', '1.3', '万元', '。', '为', '认定', '上述事实', '，', '公诉', '机关', '提供', '了', '被告人', '供述', '、', '证人', '证言', '、', '书证', '等', '证据', '，', '并', '据此', '认为', '被告人', '的', '行为', '触犯', '了', '《', '中华人民共和国', '刑法', '》', '×', '×', '×', '×', '、', '×', '×', '×', '×', '第', '（', '一', '）', '项', '、', '×', '×', '，', '应当', '以', '×', '×', '罪', '追究其', '刑事责任', '。', '提请', '本院', '依法', '判处', '。']"

Remove Stop words & punctuations

Tokenized text

"['安阳县', '人民检察院', '指控', ':', '2009', '年', '至', '2012', '年', ',', '被告人', '杨', '1', '某', '利用', '其', '在', '洛阳市', '中级', '人民法院', '工作', ',', '先后', '担任', '民某', '副庭长', '、', '执行局', '副局长', '的', '便利', '条件', ',', '通过', '让', '本院', '同事', '在', '审理案件', '过程', '中为', '案件', '当事人', '提供', '帮助', ',', '非法', '收受', '他人', '贿赂', '现金', '及', '购物', '卡', '共计', '36.8', '万元', ',', '个人', '实得', '32.8', '万元', '。', '案发后', '赃款', '已退', '。', '具体', '如下', ':', '1', '、', '2009', '年', '至', '2011', '年', '期间', ',', '被告人', '杨', '1', '某', '利用', '其', '在', '洛阳市', '中级', '人民法院', '工作', ',

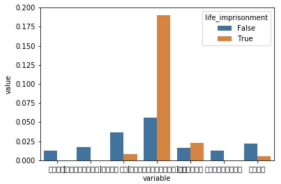
Remove Stopwords & punctuations

A sample code

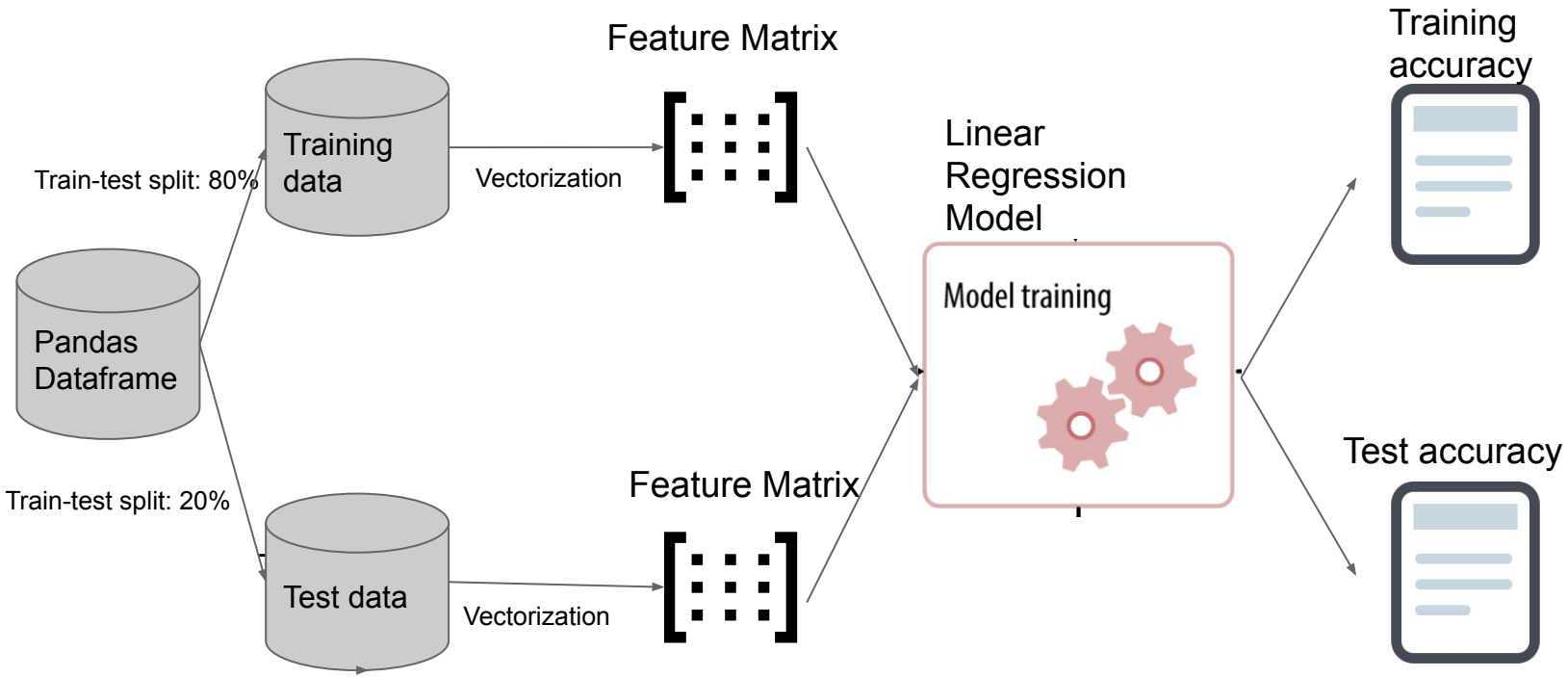
```
add_punc='，。、【】“”：；（）《》‘’{}？！⑦()、%^>℃：.”“^-——=&#@￥'
all_punc=punctuation+add_punc
def sentence_cut(x):
    x = ' '.join(x)
    x=re.sub(r'[A-Za-z0-9]|/d+','',x) #delete numbers and English
letters
    testline=x.split(' ')
    te2=[]
    for i in testline:
        te2.append(i)
        if i in all_punc:
            te2.remove(i)
    return te2
```

	刑事责任	非法	扣押	销售	追究其	犯罪事实	足以认定	毒品	户籍	利用	...	公诉	辩解	经营	信息	银行	工程	骗取	职务
0	1	0	0	0	1	0	0	0	0	0	...	1	1	0	0	0	0	0	0
1	0	0	0	0	0	0	1	0	1	0	...	0	0	0	1	1	0	1	0
2	1	1	1	0	1	1	0	0	0	0	...	1	0	0	0	0	1	0	0
3	1	1	0	1	1	1	0	0	0	0	...	1	0	1	1	1	0	0	0
4	0	0	0	0	0	0	0	1	0	0	...	1	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0
7	1	1	0	0	1	1	0	0	1	0	...	1	0	0	0	1	0	0	0
8	0	0	0	0	0	1	0	0	0	0	...	1	0	1	0	0	0	0	0
9	1	0	0	0	1	0	0	0	0	0	...	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0

Choose ~200 words with high frequency

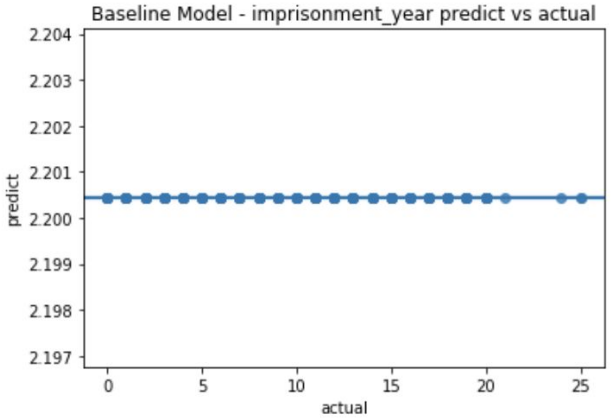


Modeling Overview



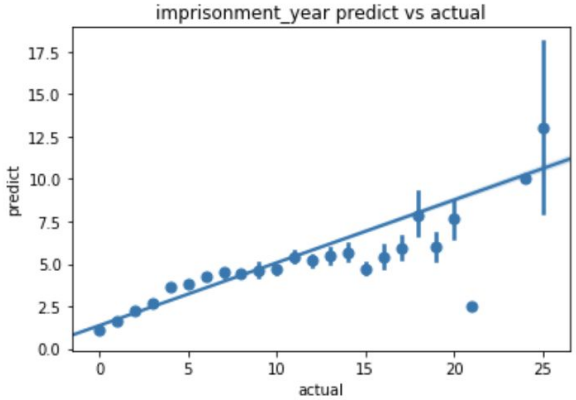
Result

Baseline Model - predicting all to be the mean



VS

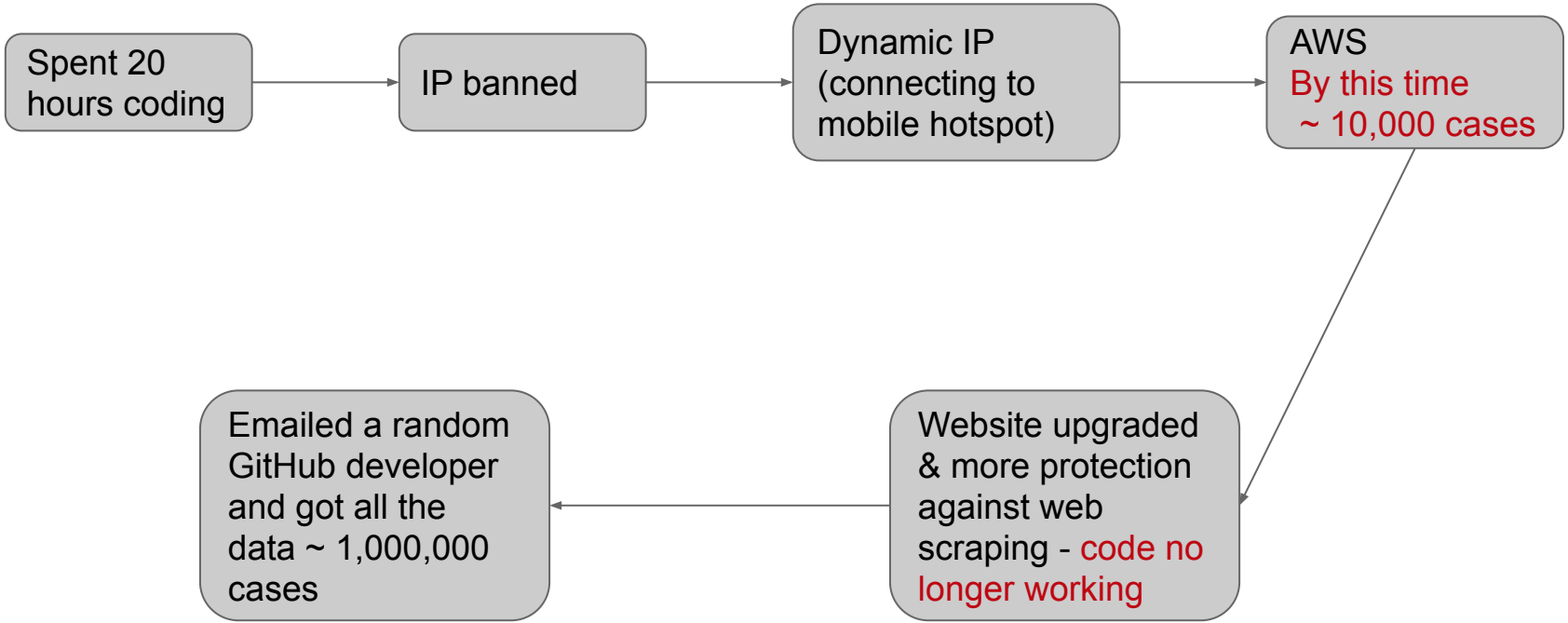
Linear Regression Model - Using 400+ features



Features: amount of fine (numeric); accusation types, relevant articles, words in tokenized fact (one hot encoding)

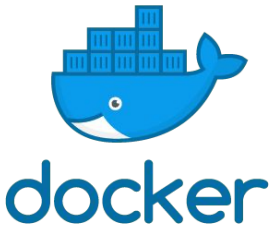
Challenge - Web Scraping

- Wenshu website: no public rules over web scraping & no api for data



Tools Overview

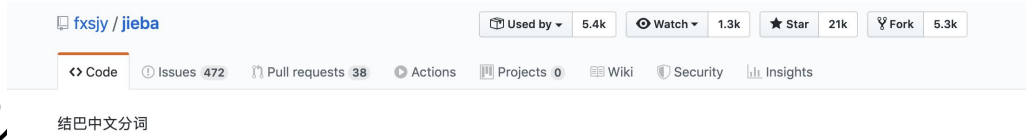
Web Scraping



Data Manipulation

Data Visualization

Machine Learning



Next Steps

- Train a more robust model
  - Include more features
- Implications
  - Get a model, so what?
- Questions

Advertising

- TextXD starts tomorrow
  - https://www.textxd.org/
- Data Discovery Program
  - A perfect way to recruit free undergrad RAs
  - https://data.berkeley.edu/research/discovery
- Looking for RA opportunity
  - Broadly interested in Natural Language Processing, Legal Studies, and AI Safety
  - Resume at https://violetyao.github.io/
  - Email violetyao@berkeley.edu