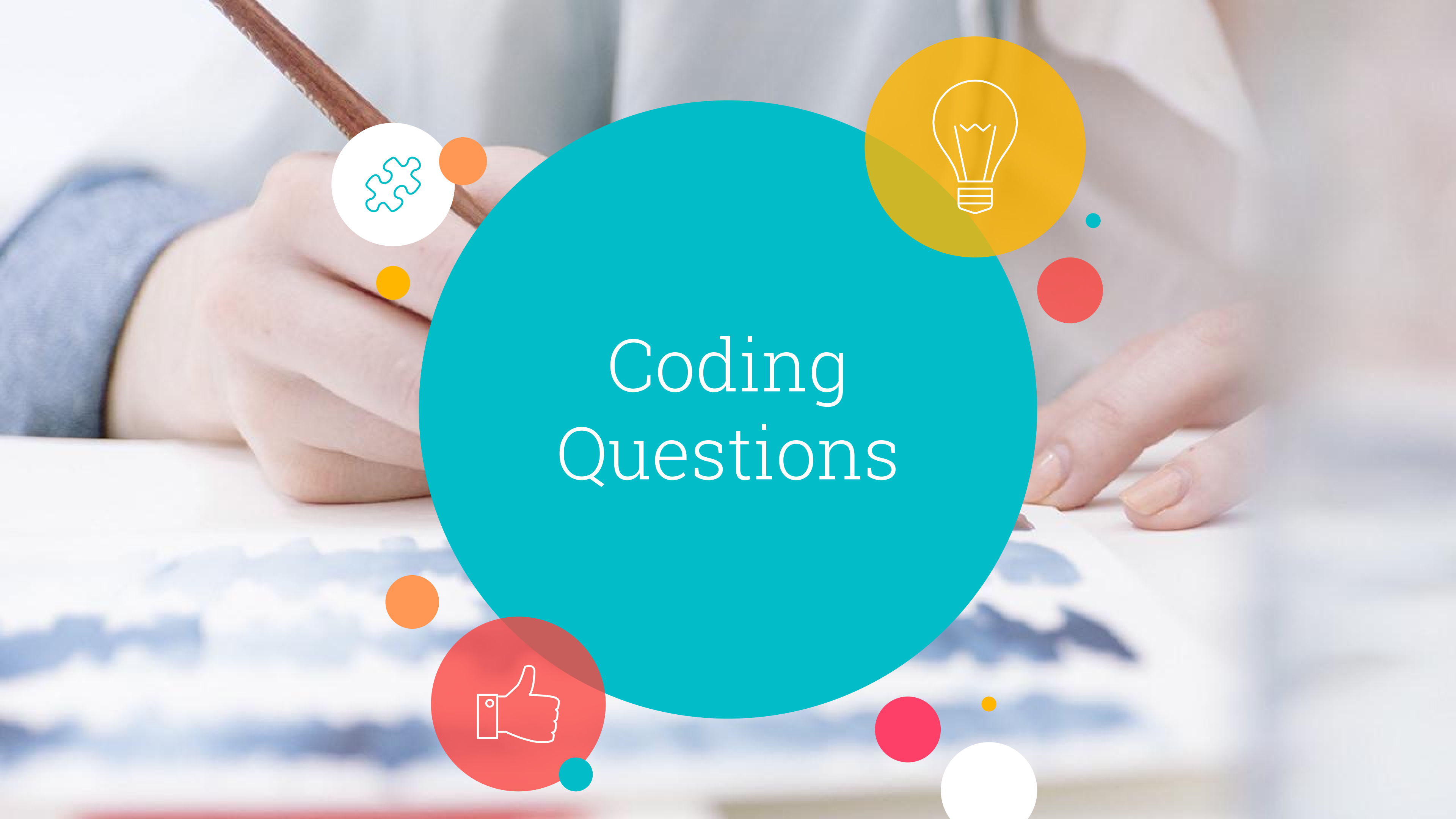


Coding Questions

Crawler

Scrapy

An open source and collaborative python framework for extracting and crawling the data you need from the websites.

Domínios

- Codeforces
- CodeChef
- URI Online Judge
- Sphere Online Judge
- DMOJ
- A^2 Online Judge
- AtCoder
- CS Academy
- Timus Online Judge
- Caribbean Online Judge

Preocupações:

- Respeitar Robots.txt
- Verificar campo Content-Type
- Manter-se nos domínios iniciais
- Evitar sobrecarregar o site

Coding Highlights

Manter-se no domínios iniciais

Início

```
start_urls = [
'http://codeforces.com/',
'https://www.codechef.com/',
'https://urionlinejudge.com,
'http://www.spoj.com/',
'https://dmoj.ca/',
'https://a2oj.com/',
'http://atcoder.jp/',
'https://csacademy.com/',
'http://acm.timus.ru/',
'http://coj.uci.cu',
]
```

Restrição

```
allowed_domains = [
'codeforces.com',
'codechef.com',
'urionlinejudge.com,
'spoj.com',
'dmoj.ca',
'a2oj.com',
'atcoder.jp',
'csacademy.com',
'acm.timus.ru',
'coj.uci.cu',
]
```

Respeitar Robots.txt e Evitar sobrecarregar os sites

```
custom_settings = {
   'USER_AGENT': 'coding-questions-bot
        (github.com/Arthurlpgc/InfoRetrievalProject)',
   'DOWNLOAD_TIMEOUT': '5',
   'DOWNLOAD_MAXSIZE': '1000000',
   'ROBOTSTXT_OBEY': 'True',
   'DOWNLOAD_DELAY': '1',
}
```

Detectar campo Content-Type

```
def parse(self, response):
       if not isinstance(response, HtmlResponse):
           raise scrapy.exceptions.IgnoreRequest()
       else:
         savePage()
         extractContent()
```

Algoritmos de Busca

BFS

Expansão uniforme da fronteira

Harvest Ratio para 200 páginas

Domínio	Harvest Ratio [BFS]
Codeforces	0,00
CodeChef	0,25
URI Online Judge	0,00
Sphere Online Judge	0,28
DMOJ	0,31
A² Online Judge	0,00
AtCoder	0,00
CS Academy	-
Timus Online Judge	0,02
Caribbean Online Judge	-
MEAN:	0,11

Heurística

Expansão controlada da fronteira

Bag of Words counting distance to relevant page

Funções

- Increase Distance
- Reduce Distance
- Maximize Distance
- Minimize Distance

```
 problemset',              self.
('/problemset/page/',      self.in
n('/problemset/tags/',     self.incr
en('/problemset/problem/', self.maxi
ken('mobile',              self.mini
ken('status',              self.mini
en('standings',            self.mini
('submit',                 self.min
'locale=',                 self.m
ets.',                     sel
```

Tentativa 1:

Diferentes bag of words para diferentes domínios

Diferentes Bag of Words para diferentes domínios: Harvest Ratio para 200 páginas

Domínio	Harvest Ratio [BFS]	Harvest Ratio [HEURÍSTICA]
Codeforces	0,00	0,96
CodeChef	0,25	0,98
URI Online Judge	0,00	0,91 (?)
Sphere Online Judge	0,28	0,85
DMOJ	0,31	0,80
A² Online Judge	0,00	0,91
AtCoder	0,00	0,60
CS Academy	-	-
Timus Online Judge	0,02	0,77
Caribbean Online Judge	-	-
MEAN:	0,11	0,85
Increase Percentage:	788,37	

Mesma bag of words: Haverst Ratio para 200 sites de cada domínio

Harvest Ratio : 0.65

Problemas:

- Mesma word presente em dois domínios:
  Em um, increase distance
  Em outro, decrease distance

Solução:

Unir o melhor dos dois mundos

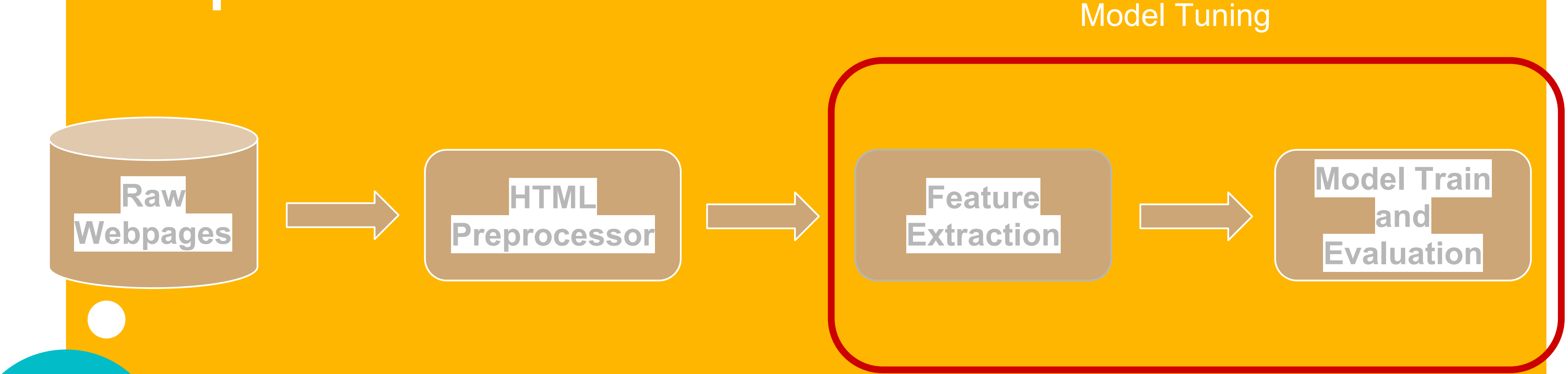
Dúvidas?

Classifier

Ferramentas

- Python 3.6
- Scrapy
- BeautifulSoup
- Scikit-Learn

Pipeline de Texto



Conjunto de Treino

- Codechef
- Codeforces
- Dmoj
- Sphere Online Judge
- URI Online Judge

Avaliam-se os melhores parâmetros realizando busca em grid nesse conjunto.

10 exemplos positivos e 10 negativos de cada site

Conjunto de Teste

- A² Online Judge
- AtCoder
- CS Academy
- Timus Online Judge
- Caribbean Online Judge

Mede-se performance no conjunto de teste após seleção do best fit.

10 exemplos positivos e 10 negativos de cada site

Html preprocessor

- Extrai-se do HTML apenas o texto visível

Tags como “style” e “script” são ignoradas

Model Tuning

- Foram utilizados os classificadores SVM, Random Forest, Regularized Logistic Regression, Multinomial Naive Bayes e Knn.
- É feita uma busca em grade (grid search) para cada modelo, considerando as diferentes features possíveis de serem extraídas e os parâmetros dos classificadores.

Model Evaluation

- 3-fold Stratified Cross Validation
  - Classes estão balanceadas nos folds
  - O score final da combinação de parâmetros é a média entre os scores dos 3 folds
- Métricas Calculadas:
  - Precision Macro
  - Recall Macro
  - F1 Macro
  - Acurácia

Utilizamos o F-score para escolher best fit

Feature Extraction & Selection

- Representação Bag of Words - Word Count Matrix
- Feature Selection - Frequência das Palavras
  - Todas as features (43744), 5000, 10000
- Unigramas ou Bigramas
- Stopwords: 0.5, 0.75, 1.0 (específicas do corpus)
- TF normalizado e TF-IDF
  - Normalização L1 ou L2 para os term vectors

SVM

- C: 0.1, 1, 10, 100
- Gamma: 1/n, 0.001, 0.01, 0.1, 1 n é o número de features
- Kernel: Polynomial, RBF

3841 avaliações

Random Forest

- Max Depth: 3, None - árvore muito cheia pode dar overfitting
- Min Sample Split: 2, 3, 10
- Min Samples Leaf: 1, 3, 10
- Max Features: log2, sqrt, None
- Criterio: Gini - mais rápido que entropia e resultados similares

5185 avaliações

Regularized Logistic Regression

- C: 0.001, 0.01, 0.1, 1, 10, 100, 1000
  - Parâmetro do termo de regularização

1345 avaliações

Multinomial Naive Bayes

- A distribuição multinomial aproxima melhor o modelo Bag of Words escolhido para representar os documentos
- Alpha: 1, 0.1, 0.01, 0.001, 0.0001, 0.00001
    - Smoothing Parameter

575 avaliações

Knn

- n_neighbors: 1, 2, 3, 4
- Weights: com ou sem

-

1345 avaliações

Melhores Resultados (Grid)

	Accuracy	F1-score	Precision	Recall	Fit Time
SVM	0.97	0.97	0.97	0.97	0.29
Random Forest	0.98	0.98	0.98	0.98	0.46
Naive Bayes	0.97	0.97	0.97	0.97	0.38
Logistic Regression	0.98	0.98	0.98	0.98	0.35
Knn	0.9	0.89	0.91	0.9	0.44

Best Fit

- SVM:

'vect__max_features': 5000,
'clf__gamma': 1,
 'clf__kernel': 'poly',
 'vect__ngram_range': (1, 2),
 'clf__C': 10,
'tfidf__use_idf': False,
 'tfidf__norm': 'l2',
 'vect__max_df': 1.0

Best Fit

- Logistic Regression:

'vect__max_df': 0.75,
'vect__max_features': 10000,
'vect__ngram_range': (1, 2),
'tfidf__use_idf': True,
'tfidf__norm': 'l2',
'clf__penalty': 'l1',
'clf__C': 10

Best Fit

- Random forest:

```
'tfidf__norm': 'l2',
'clf__min_samples_leaf': 3,
'clf__min_samples_split': 2,
'clf__criterion': 'gini',
'vect__max_features': 50000,
'clf__max_features': 'sqrt',
'clf__max_depth': None,
 'vect__max_df': 1.0,
'tfidf__use_idf': False,
 'vect__ngram_range': (1, 2)
```

Best Fit

- Naive Bayes

'tfidf__norm': 'l2',
'vect__ngram_range': (1, 2),
'vect__max_features': 50000,
'clf__alpha': 0.0001,
'tfidf__use_idf': True,
'vect__max_df': 0.5

Best Fit

- Knn

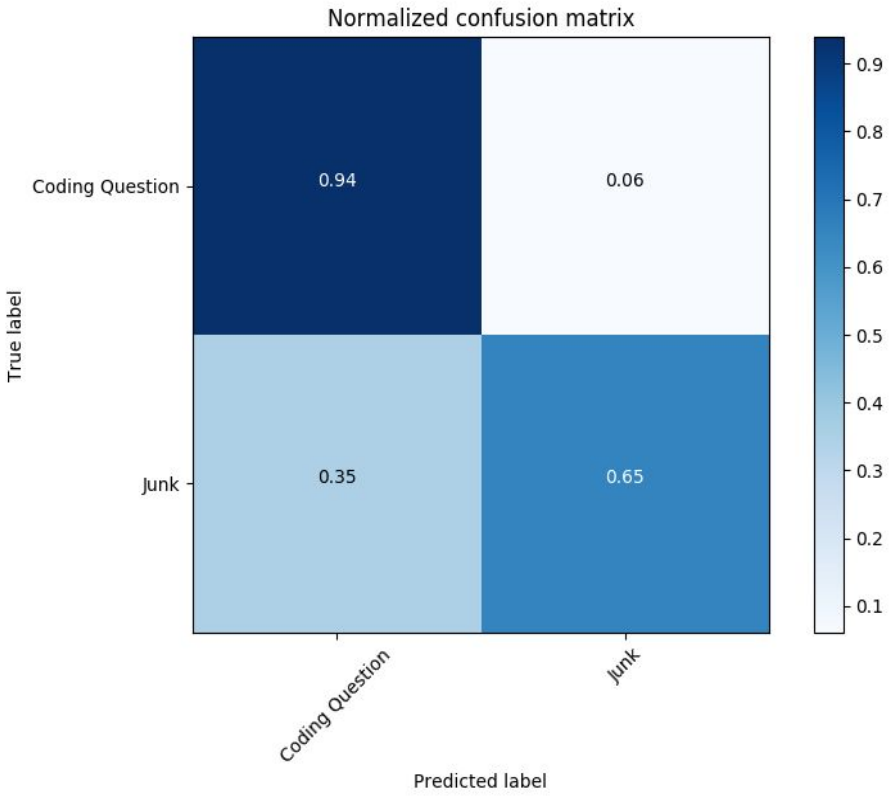
'vect__max_df': 0.75,
'vect__ngram_range': (1, 2),
'clf__n_neighbors': 4,
'tfidf__norm': 'l2',
'tfidf__use_idf': False,
'vect__max_features': 5000,
'clf__weights': 'uniform'

Resultados Teste

- SVM treinada com o best fit e testada com os 5 novos sites não presentes no conjunto de treino

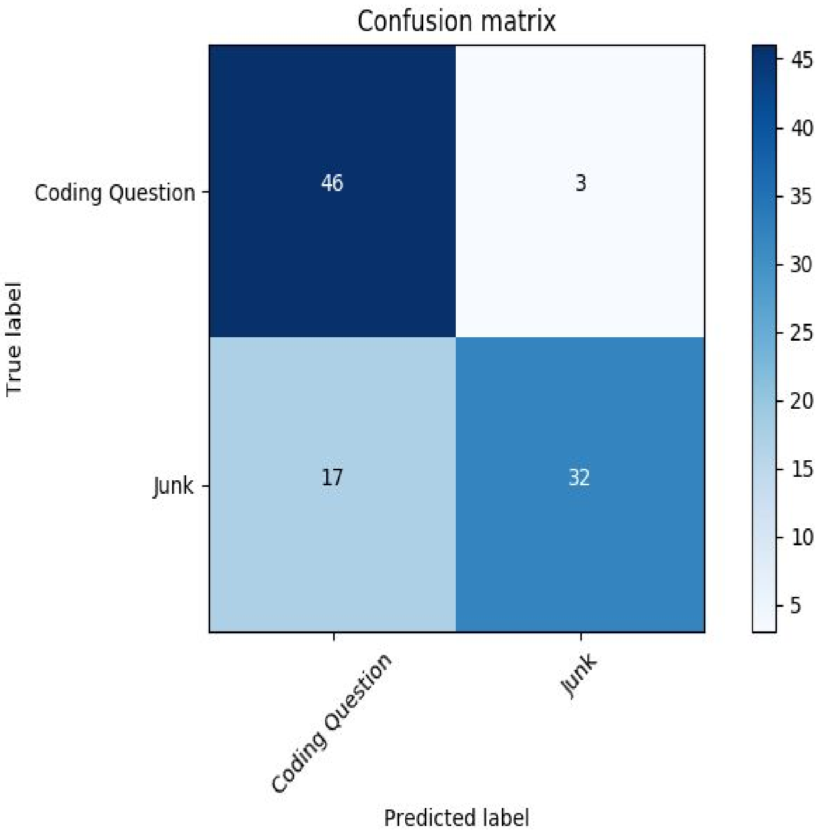
	Precision	Recall	F1-score	Support
Junk	0.73	0.94	0.82	49
Coding Question	0.91	0.65	0.76	49
avg / total	0.82	0.80	0.79	98

Resultados Teste



Normalized confusion matrix

True label \ Predicted label	Coding Question	Junk
Coding Question	0.94	0.06
Junk	0.35	0.65



Confusion matrix

True label \ Predicted label	Coding Question	Junk
Coding Question	46	3
Junk	17	32

SVM foi treinada com os exemplos de todos os sites para ser integrada ao crawler e com os parâmetros do best fit.

Dúvidas?

Extractor

A2OJ

```
{
   "time-limit": 3000, The TL should be between  2-3 S
   "statement": "\n A2OJ © Copyright 2011-2016  Ahmed Aly  All Rights Reserved. A2OJ uses
Sphere Engine ™ © by  Sphere Research Labs .",
   "props": {
              "Sample Input": "\n \n 5",
              "Input Format": "\n \n \nYour program ",
              "Output Format": "\n \n \nFor each ",
              "Added by": "\n \n \n ahmed",
              "Partial score": "\n \n \nNo",
              "Time Limit": "\n \n \n2",
              "Problem Statement": "\n \n \nOmar is ",
              "Sample Output": "\n \n 34",
              "Added at": "\n \n \n2014"
   },
   "memory-limit": -1, Ok, não tinha na pagina
   "title": "Omar" Ok
}
```

a2oj Stats

a2oj	P	R	F1
Title	100%	100%	1
Time	90%	90%	0.9
Memory	-	-	-
Statement	0%	0%	0

AtCoder

```
{
  "time-limit": 5000, Ok
  "statement": null,
  "props": {
          "niku ": " meat",
          "following format": " \n S ",
          "Score ": "  100 ",
          "tako ": " octopus",
          "yaki ": " grilled"
  },
  "memory-limit": 256, Ok
  "title": "Snuke's favorite YAKINIKU" Ok
}
```

AtCoder Stats

AtCoder	P	R	F1
Title	100%	100%	1
Time	0%	0%	0
Memory	100%	100%	1
Statement	100%	100%	1

CodeChef

```
{
  "time-limit": 500,Ok
  "statement": null,
  "props": {
          "Languages": " \n ADA",
          "now is": "  07",
          "Tags": " \n \n                    cakewalk",
          "Author": " \n 6",
          "Tester": " \n 6",
          "such that": "\n\n \n 1",
          "Your IP": "  187.112.30.97 ",
          "Problem Code": "  LIKECS01 ",
          "Time Limit": " \n 0.5 secs ",
          "Source Limit": " \n 50000 Bytes ",
          "Editorial": " \n https",
          "Date Added": " \n 8"
  },
  "memory-limit": 1536, Fixed for whole codechef, not extracted in common way
  "title": "Subsequence Equality" Ok
}
```

CodeChef Stats

Codechef	P	R	F1
Title	100%	100%	1
Time	100%	100%	1
Memory	-	-	-
Statement	100%	100%	1

CodeForces

```
{
   "time-limit": 2000,Ok
   "statement": null,
   "props": {
            "Server time": " Sep",
            "following way": " the origin ",
            "satellites locations": "     Input The ",
            "following types": "          1 x "
   },
   "memory-limit": 256,Ok
   "title": "Satellites"Ok
}
```

Codeforces Stats

Codeforces	P	R	F1
Title	100%	100%	1
Time	100%	100%	1
Memory	100%	100%	1
Statement	100%	100%	1

Caribean OJ

```
{
   "time-limit": 2000,Ok
   "statement": null,
   "props": {errors
            "Sub": "  34626 ",
            "Total Time": "  60000 MS ",
            "one integer": " the sum ",
            "mero inteiro": " a soma ",
            "por espacio": "  A",
            "hour archive": " Problem",
            "Test Time": " \n                                        12000 MS",
            "Memory": "  937 MB ",
            "Output": "  64 MB ",
            "mero entero": " la suma ",
            "Size": " \n                                   9 KB"
   },
   "memory-limit": 62,Ok
   "title": "A+B Problem"Ok
}
```

CsAcademy

```
{
   "time-limit": 15000,Far wrong, should be 1000 ms as on props, no idea what the RE got her
   "statement": null,
   "props": {
            "Time limit": "  1000 ms ",
            "Memory limit": "  128 MB "
   },
   "memory-limit": 128,Ok
   "title": "3-divisible Pairs"Ok
}
```

CsAcademy Stats

CsAcademy	P	R	F1
Title	100%	100%	1
Time	0%	0%	0
Memory	100%	100%	1
Statement	100%	100%	1

Dmoj

```
{
  "time-limit": 500,Wrong should be 5000 ms as in props
  "statement": null,
  "props": {
          "top": " 27px",
          "Re": " Inconsistant Speed ",
          "Points": " 3",
          "Memory limit": " 256M ",
          "Time limit": " 5.0s ",
          "margin": " 80px auto "
  },
  "memory-limit": 256,Ok
  "title": "A Plus B"Ok
}
```

Dmoj Stats

Dmoj	P	R	F1
Title	100%	100%	1
Time	0%	0%	0
Memory	100%	100%	1
Statement//	100%	100%	1

Spoj

```
{
  "time-limit": 5000,Should be 10000
  "statement": null,
  "props": {
            "Resource": " Douglas Adams",
            "Last edit": " 2017",
            "Cluster": " \n \n Cube ",
            "Added by": " mima ",
            "Languages": " All ",
            "Time limit": " 10s",
            "Source limit": " 50000B ",
            "Memory limit": " 1536MB ",
            "Output": " \n1",
            "Date": " 2004",
            "Input": " \n1"
  },
  "memory-limit": 1536,Ok
  "title": "TEST - Life, the Universe, and Everything"Ok
}
```

Spoj Stats

Spoj	P	R	F1
Title	100%	100%	1
Time	37.5%	37.5%	0.375
Memory	100%	100%	1
Statement	100%	100%	1

Timus

```
{
  "time-limit": 1000,Ok
  "statement": null,
  "props": {
          "Tags": "  problem for ",
          "Problem Author": "  Pavel Atnashev ",
          "Difficulty": " 17 ",
          "Memory limit": " 64 MB ",
          "Time limit": " 1.0 second ",
          "position": "absolute"
  },
  "memory-limit": 64,Ok
  "title": "A+B Problem"Ok
}
```

Timus Stats

Timus	P	R	F1
Title	100%	100%	1
Time	100%	100%	1
Memory	100%	100%	1
Statement	0%	0%	0

URI

```
{
    "time-limit": -1,Page had it, html hadnt
    "statement": null,
    "props": {},
    "memory-limit": -1,Ok Not in Page
    "title": "Area of a Circle"Ok
}
```

URI Stats

Uri	P	R	F1
Title	100%	100%	1
Time	-	-	-
Memory	-	-	-
Statement	-	-	-