

# Recuperação de informação

Coleta e busca de entidades estruturadas em um domínio



# Equipe & Tarefas

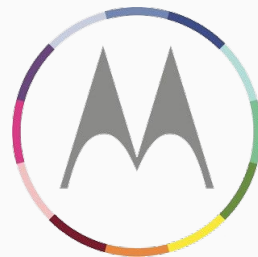
Isaac Douglas    Webcrawler

Wellington Felix    Classificação e extração

# Domínio: Smartphones

## Sites:

1. Banggood
2. Apple
3. Samsung
4. Asus
5. LG
6. Xiaomi
7. Motorola
8. Sony
9. Blu
10. HTC



# Task 1

Localização de páginas relevantes

# Files = 9.195

	Naïve Bayes	Decision Tree	SVM	Logistic Regression	MLP
Relevants	78.61%	10.32%	1.11%	3.39%	2.02%
Not Relevants	21.39%	89.68%	98.89%	96.61%	97.98%

# Task 2

Detecção de páginas com instâncias (classificação)

# Resultados | Sem Feature Selection

# Features = 12495

	Naïve Bayes	Decision Tree	SVM	Logistic Regression	MLP
Accuracy	67.00%	85.50%	86.50%	89.00%	88.50%
Precision	0.70	0.87	0.98	0.95	0.98
Recall	0.79	0.84	0.75	0.82	0.79
Training time	0.0896s	0.14s	0.377s	0.0615s	61.6s

## Resultados | Feature Selection (Frequência)

# Features = 3000

	Naïve Bayes	Decision Tree	SVM	Logistic Regression	MLP
Accuracy	67.00%	83.00%	85.50%	88.00%	87.00%
Precision	0.71	0.88	0.98	0.95	0.98
Recall	0.79	0.76	0.73	0.80	0.76
Training time	0.016s	0.0406s	0.0841s	0.0156s	8.81s



## Resultados | Feature Selection (Information Gain)

# Features = 1000

	Naïve Bayes	Decision Tree	SVM	Logistic Regression	MLP
Accuracy	57.00%	77.00%	65.00%	80.00%	89.00%
Precision	0.54	0.83	0.79	0.77	0.89
Recall	0.93	0.72	0.53	0.85	0.89
Training time	0.00788s	0.0125s	0.0146s	0.0148s	4.03s

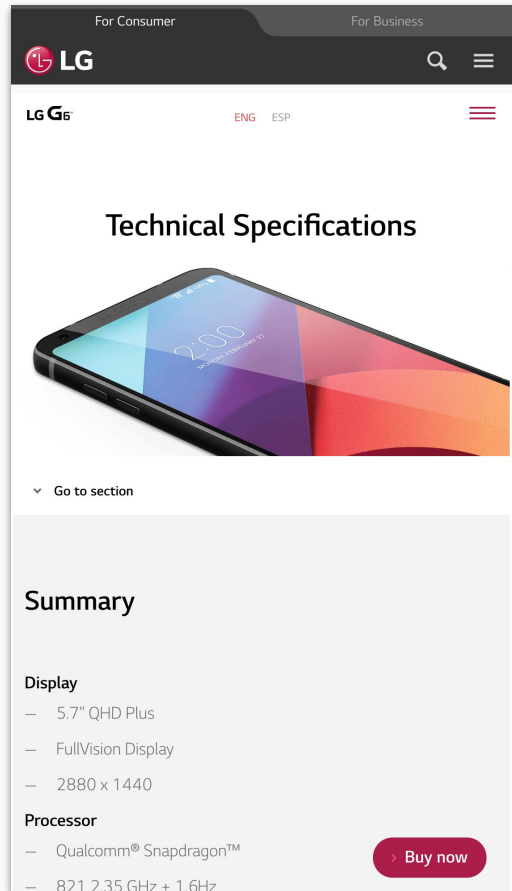
# Feature Selection (Information Gain)

'camera', 'mp', 'gb', 'sim', 'lte', 'ram', 'mobile', 'battery', 'gsm', 'fps', 'memory', 'photos', 'snapdragon', 'microsd', 'mah', 'fingerprint', 'capture', 'octa', 'dual', 'display', 'phones', 'selfie', 'video', 'phone', 'gps', 'rom', 'smartphone', 'fdd', 'qualcomm', 'sensor', 'recording', 'rear', 'resolution', 'core', 'nfc', 'selfies', 'umts', 'main', 'ghz', 'carrier', 'glass', 'com', 'light', 'card', 'support', 'hspa', 'network', 'face', 'slow', 'aperture', 'aac', 'pda', 'buy', 'flash', 'nano', 'unlocked', 'cdma', 'processor', 'glonass', 'gorilla', 'technology', 'bluetooth', 'super', 'motion', 'focus', 'specifications', 'device', 'image', 'dimensions', 'fi', 'proximity', 'flac', 'wav', 'mode', 'usb', 'supported', 'ppi', 'wi', 'storage', 'sharp', 'account', 'autofocus', 'hdr', 'panorama', 'charge', 'available', 'stabilization', 'new', 'weight', 'talk', 'detection', 'internal', 'learn', 'nougat', 'bands', 'wireless', 'cat', 'plus', 'hd', 'android', 'free', 'warranty', 'videos', 'cameras', 'corning', 'mkv', 'tv', 'connectivity', 'images', 'time', 'pictures', 'power', 'high', 'samsung', 'ambient', 'coverage', 'performance', 'inch', 'gprs', 'liquid', 'ois', 'water', 'screen', 'mi', 'accessories', 'lens', 'pixel', 'networks', 'capacity', 'subject', 'reviews', 'program', 'auto', 'shipping', 'wide', 'vary', 'led', 'google', 'gp', 'wet', 'hid', 'hotspot', 'mhz', 'damage', 'yes', 'limited', 'pay', 'accelerometer', 'service', 'htc', 'format', 'hours', 'apple', 'care', 'financing', 'change', 'intel', 'quad', 'shots', 'apps', 'ogg', 'galaxy', 'terms', 'addition', 'micro', 'continue', 'features', 'help', 'photo', 'lg', 'calling', 'days', 'standby', 'cell', 'exposure', 'cover', 'pro', 'great', 'cpu', 'mm', 'type', 'moto', 'charging', 'mac', 'adaptive', 'conditions', 'adreno', 'webm', 'tdd', 'edge', 'latest', 'software', 'product', 'user', 'energy', 'required', 'payments', 'iphone', 'depends', 'unlock', 'angle', 'ac', 'replacement', 'formats', 'apply', 'connect', 'mbps', 'timer', 'pixels', 'credit', 'purchase', 'volte', 'picture', 'zoom', 'convenience', 'offer', 'actual', 'close', 'life', 'note', 'fast', 'provider', 'compass', ...

# Task 3

Extração de instâncias

# Extrator genérico



Specs	
Screen size	5.7"
Camera resolution	13MP
Screen resolution	2880 x 1440
Battery capacity	3,300 mAh
RAM	4GB
Internal Memory	32GB
Processor speed	2.35 GHz
Weight	5.74 oz

# Resultados

	Samsung	LG	Apple	Sony	Xiaomi	Asus	Motorola	Blu	HTC	Banggood
Precision	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	0.50	0.75	1.00	0.62	0.87	0.87	0.87	0.87
F-Measure	1.00	1.00	0.61	0.86	1.00	0.77	0.93	0.93	0.93	0.93

# Recuperação de informação

Coleta e busca de entidades estruturadas em um domínio

