

## 2\_text\_processing

July 12, 2021

### 1 Screencast Code

The follow code is the same used in the "Text Processing" screencast. Run each code cell to see how

```
In [1]: from pyspark.sql import SparkSession
        from pyspark.ml.feature import RegexTokenizer, CountVectorizer, \
            IDF, StringIndexer
        from pyspark.sql.functions import udf
        from pyspark.sql.types import IntegerType

        import re

In [2]: # create a SparkSession: note this step was left out of the screencast
        spark = SparkSession.builder \
            .master("local") \
            .appName("Word Count") \
            .getOrCreate()
```

### 2 Read in the Data Set

```
In [3]: stack_overflow_data = 'Train_onetag_small.json'

In [4]: df = spark.read.json(stack_overflow_data)

In [5]: df.head()

Out[5]: Row(Body="<p>I'd like to check if an uploaded file is an image file (e.g png, jpg, jpeg,
```

### 3 Tokenization

Tokenization splits strings into separate words. Spark has a [Tokenizer](#) class as well as `RegexTokenizer`, which allows for more control over the tokenization process.

```
In [6]: # split the body text into separate words
        regexTokenizer = RegexTokenizer(inputCol="Body", outputCol="words", pattern="\\W")
        df = regexTokenizer.transform(df)
        df.head()

Out[6]: Row(Body="<p>I'd like to check if an uploaded file is an image file (e.g png, jpg, jpeg,
```

## 4 CountVectorizer

```
In [7]: # find the term frequencies of the words
cv = CountVectorizer(inputCol="words", outputCol="TF", vocabSize=1000)
cvmodel = cv.fit(df)
df = cvmodel.transform(df)
df.take(1)
```

```
Out[7]: [Row(Body="<p>I'd like to check if an uploaded file is an image file (e.g png, jpg, jpeg
```

```
In [8]: # show the vocabulary in order of
cvmodel.vocabulary
```

```
Out[8]: ['p',
        'the',
        'i',
        'to',
        'code',
        'a',
        'gt',
        'lt',
        'is',
        'and',
        'pre',
        'in',
        'this',
        'of',
        'it',
        'that',
        'for',
        '0',
        '1',
        'have',
        'my',
        'if',
        'on',
        'but',
        'with',
        'can',
        'not',
        'be',
        'as',
        't',
        'li',
        'from',
        '2',
        's',
        'http',
        'an',
```

'm',  
'strong',  
'new',  
'how',  
'do',  
'com',  
'so',  
'or',  
'at',  
'using',  
'when',  
'am',  
'like',  
'class',  
'id',  
'there',  
'get',  
'are',  
'name',  
'what',  
'any',  
'file',  
'string',  
'data',  
'all',  
'which',  
'want',  
'would',  
'amp',  
'use',  
'java',  
'function',  
'public',  
'some',  
'3',  
'text',  
'error',  
'android',  
'value',  
'c',  
'x',  
'href',  
'you',  
'one',  
'by',  
'user',  
'me',  
'server',

'type',  
'here',  
'way',  
'return',  
'int',  
'will',  
'div',  
'need',  
'then',  
'set',  
'e',  
'system',  
'has',  
'problem',  
'out',  
'php',  
'no',  
'just',  
'4',  
'org',  
'know',  
'html',  
'only',  
'where',  
'page',  
'application',  
'5',  
'thanks',  
'var',  
'br',  
'we',  
'd',  
'should',  
'does',  
'add',  
'n',  
'true',  
've',  
'void',  
'em',  
'was',  
'rel',  
'work',  
'time',  
'other',  
'10',  
'app',  
'null',

'method',  
'b',  
'table',  
'list',  
'now',  
'into',  
'help',  
'end',  
'trying',  
'following',  
'object',  
'view',  
'nofollow',  
'up',  
'example',  
'image',  
'same',  
'create',  
'also',  
'each',  
'something',  
'www',  
'web',  
'first',  
'array',  
'line',  
'script',  
'find',  
'don',  
'run',  
'could',  
'select',  
'about',  
'test',  
'make',  
'form',  
'r',  
'files',  
'tried',  
'ul',  
'net',  
'url',  
'td',  
'self',  
'input',  
'windows',  
'button',  
'see',

'blockquote',  
'database',  
'question',  
'content',  
'else',  
'more',  
'works',  
'xml',  
'6',  
'00',  
'two',  
'8',  
'after',  
'they',  
'possible',  
'false',  
'right',  
'them',  
'y',  
'working',  
'7',  
'width',  
'main',  
'src',  
'try',  
'private',  
'however',  
'version',  
'number',  
'f',  
'result',  
'these',  
'because',  
'project',  
'key',  
'message',  
'why',  
'doesn',  
'used',  
'please',  
'query',  
'import',  
'size',  
'item',  
'call',  
'show',  
'while',  
'title',

'found',  
'been',  
'anyone',  
'change',  
'post',  
'document',  
'users',  
'different',  
'its',  
'start',  
'able',  
'log',  
'access',  
'another',  
'event',  
'case',  
'request',  
'values',  
'update',  
'client',  
'edit',  
'index',  
'9',  
'service',  
'read',  
'without',  
'source',  
'javascript',  
'left',  
'style',  
'open',  
'jquery',  
'img',  
'running',  
'row',  
'h',  
'display',  
'fine',  
'write',  
'site',  
'google',  
'seems',  
'height',  
'click',  
'date',  
'12',  
'static',  
'etc',

'option',  
'path',  
'output',  
'property',  
'20',  
'doing',  
'model',  
'g',  
'default',  
'link',  
'than',  
'through',  
'echo',  
'below',  
'include',  
'even',  
'solution',  
'lib',  
'11',  
'sql',  
'questions',  
'still',  
'program',  
'such',  
'library',  
'getting',  
'exception',  
'created',  
'simple',  
'context',  
'png',  
'your',  
'very',  
'before',  
'apache',  
'both',  
'ol',  
'sure',  
'100',  
'order',  
'asp',  
'command',  
'field',  
'color',  
'window',  
'images',  
'column',  
'load',



'having',  
'thread',  
'background',  
'think',  
'js',  
'wrong',  
'go',  
'point',  
'element',  
'process',  
'length',  
'really',  
'tr',  
'span',  
'being',  
'every',  
'back',  
'current',  
'called',  
'css',  
'label',  
'action',  
'issue',  
'many',  
'info',  
'stack',  
'check',  
'got',  
'top',  
'since',  
'connection',  
'looking',  
'put',  
'second',  
'search',  
'db',  
'local',  
'over',  
'email',  
'above',  
'password',  
'done',  
'api',  
'between',  
'response',  
'build',  
'cannot',  
'alt',

'print',  
'j',  
'well',  
'body',  
'directory',  
'count',  
'description',  
'location',  
'information',  
'next',  
'address',  
'root',  
'good',  
'01',  
'let',  
'control',  
'o',  
'microsoft',  
'part',  
'map',  
'advance',  
'instead',  
'our',  
'v',  
'much',  
'their',  
'best',  
'position',  
'2012',  
'idea',  
'custom',  
'format',  
'mysql',  
'already',  
'say',  
'long',  
'instance',  
'variable',  
'send',  
'13',  
'16',  
'may',  
'currently',  
'results',  
'inside',  
'15',  
'k',  
'header',

'enter',  
'items',  
'correct',  
'home',  
'controller',  
'domain',  
'node',  
'z',  
'override',  
'based',  
'30',  
'seem',  
'group',  
'options',  
'last',  
'ui',  
'added',  
'screen',  
'someone',  
'folder',  
'save',  
'session',  
'website',  
'close',  
'stackoverflow',  
'better',  
'username',  
'python',  
'bit',  
'box',  
'json',  
'menu',  
'login',  
'console',  
'via',  
'l',  
'usr',  
'looks',  
'anything',  
'def',  
'none',  
'https',  
'char',  
'lang',  
'within',  
'multiple',  
'activity',  
'everything',

'appreciated',  
'tag',  
'insert',  
'ajax',  
'had',  
'config',  
'again',  
'understand',  
'parent',  
'final',  
'install',  
'catch',  
'browser',  
'objects',  
'store',  
'os',  
'ideas',  
'double',  
'thing',  
'contains',  
'look',  
'given',  
'jpg',  
'template',  
'font',  
'reference',  
'those',  
'answer',  
'going',  
'connect',  
'debug',  
'frame',  
'imgur',  
'rows',  
'down',  
'ruby',  
'foo',  
'around',  
'always',  
'too',  
'state',  
'remove',  
'block',  
'status',  
'intent',  
'yes',  
'must',  
'errors',

'thank',  
'08',  
'layout',  
'en',  
'did',  
'far',  
'u',  
'core',  
'memory',  
'most',  
'installed',  
'single',  
'w',  
'framework',  
'bar',  
'linux',  
'machine',  
'specific',  
'loop',  
'give',  
'14',  
'delete',  
'begin',  
'take',  
'ip',  
'lot',  
'returns',  
'02',  
'facebook',  
'device',  
're',  
'things',  
'creating',  
'base',  
'ok',  
'nothing',  
'missing',  
'module',  
'fields',  
'host',  
'own',  
'uses',  
'support',  
'androidruntime',  
'might',  
'similar',  
'integer',  
'float',

'actually',  
'figure',  
'50',  
'configuration',  
'2010',  
'alert',  
'tell',  
'us',  
'execute',  
'std',  
'toString',  
'few',  
'empty',  
'note',  
'changes',  
'xmlns',  
'pass',  
'failed',  
'off',  
'settings',  
'network',  
'either',  
'package',  
'interface',  
'2011',  
'eclipse',  
'mode',  
'plugin',  
'25',  
'elements',  
'methods',  
'classes',  
'rb',  
'head',  
'shows',  
'properties',  
'video',  
'18',  
'const',  
'port',  
'pages',  
'break',  
'columns',  
'made',  
'bin',  
'cell',  
'functions',  
'23',

'allow',  
'21',  
'nbsp',  
'generated',  
'q',  
'submit',  
'once',  
'setup',  
'product',  
'grid',  
'22',  
'reason',  
'copy',  
'h2',  
'times',  
'aspx',  
'sub',  
'auto',  
'target',  
'dev',  
'available',  
'04',  
'jar',  
'keep',  
'filter',  
'required',  
'side',  
'convert',  
'tables',  
'handle',  
'center',  
'println',  
'task',  
'selected',  
'join',  
'setting',  
'security',  
'gets',  
'category',  
'17',  
'05',  
'lines',  
'under',  
'though',  
'boolean',  
'gems',  
'localhost',  
'entity',

'isn',  
'space',  
'admin',  
'sun',  
'fix',  
'dll',  
'implement',  
'internal',  
'wondering',  
'nil',  
'init',  
'val',  
'09',  
'parameters',  
'19',  
'binding',  
'rails',  
'txt',  
'maybe',  
'level',  
'super',  
'defined',  
'several',  
'child',  
'non',  
'03',  
'parameter',  
'place',  
'24',  
'sort',  
'needs',  
'language',  
'adding',  
'py',  
'dialog',  
'cache',  
'statement',  
'section',  
'kind',  
'problems',  
'total',  
'27',  
'margin',  
'utf',  
'filename',  
'nsstring',  
'writing',  
'border',



'account',  
'changed',  
'computer',  
'correctly',  
'match',  
'255',  
'tab',  
'bundle',  
'define',  
'didn',  
'basically',  
'warning',  
'solve',  
'release',  
'protected',  
'hr',  
'structure',  
'replace',  
'frac',  
'software',  
'foreach',  
'extends',  
'thought',  
'29',  
'full',  
'byte',  
'word',  
'stored',  
'ie',  
'\_',  
'day',  
'sender',  
'record',  
'clear',  
'never',  
'services',  
'were',  
'pdf',  
'32',  
'stop',  
'07',  
'resources',  
'models',  
'springframework',  
'move',  
'alloc',  
'people',  
'started',

'checked',  
'virtual',  
'attribute',  
'gives',  
'suggestions',  
'container',  
'ubuntu',  
'numbers',  
'old',  
'mail',  
'sample',  
'unknown',  
'bool',  
'small',  
'align',  
'collection',  
'success',  
'visual',  
'person',  
'compile',  
'mean',  
'encoding',  
'standard',  
'iphone',  
'param',  
'append',  
'events',  
'projects',  
'bytes',  
'str',  
'06',  
'rather',  
'written',  
'resource',  
'reading',  
'wrap\_content',  
'exists',  
'details',  
'certain',  
'widget',  
'handler',  
'download',  
'parse',  
'stuff',  
'40',  
'until',  
'expected',  
'documentation',

'variables',  
'socket',  
'git',  
'little',  
'servlet',  
'hello',  
'th',  
'modules',  
'environment',  
'generate',  
'says',  
'development',  
'easy',  
'simply',  
'large',  
'three',  
'31',  
'exe',  
'game',  
'approach',  
'calls',  
'args',  
'normal',  
'bottom',  
'global',  
'views',  
'types',  
'javax',  
'upload',  
'invoke',  
'tags',  
'hibernate',  
'switch',  
'stream',  
'onclick',  
'making',  
'remote',  
'layout\_width',  
'entry',  
'layout\_height',  
'im',  
'namespace',  
'll',  
'whole',  
'he',  
'great',  
'original',  
'textview',

'io',  
'hidden',  
'component',  
'internet',  
'free',  
'comment',  
'obj',  
'links',  
'phone',  
'happens',  
'loaded',  
'developer',  
'messages',  
'testing',  
'names',  
'automatically',  
'whether',  
'mvc',  
'studio',  
'2008',  
'come',  
'play',  
'course',  
'points',  
'200',  
'chrome',  
'xsl',  
'appears',  
'related',  
'pattern',  
'loading',  
'pretty',  
'who',  
'provide',  
'basic',  
'django',  
'require',  
'range',  
'per',  
'player',  
'means',  
'particular',  
'padding',  
'quite',  
'real',  
'tools',  
'forms',  
'util',

'datetime',  
'character',  
'buffer',  
'random',  
'\_post',  
'vector',  
'itself',  
'ms',  
'icon',  
'valid',  
'records',  
'cursor',  
'implementation',  
'achieve',  
'active',  
'drive',  
'yet',  
'goes',  
'calling',  
'runat',  
'max',  
'maps',  
'unable',  
'bind',  
'params',  
'except',  
'takes',  
'displayed',  
'hard',  
'report',  
'share',  
'common',  
'makes',  
'issues',  
'firefox',  
'temp',  
'runs',  
'follows',  
'seen',  
'limit',  
'properly',  
'64',  
'company',  
'characters',  
'doc',  
'applications',  
'step',  
'native',

'tool',  
'h1',  
'attr',  
'sent',  
'exist',  
'external',  
'duplicate',  
'controls',  
'complete',  
'sum',  
'exactly',  
'directly',  
'schema',  
'panel',  
'posts',  
'min',  
'performance',  
'job',  
'dim',  
'difference',  
'matrix',  
'seconds',  
'separate',  
'expression',  
'docs',  
'syntax',  
'examples',  
'textbox',  
'article',  
'book',  
'render',  
'thinking',  
'flash',  
'mac',  
'checkbox',  
'printf',  
'runtime',  
'canvas',  
'existing',  
'28',  
'shared',  
'servers',  
'customer',  
'desktop',  
'buttons',  
'previous',  
'math',  
'master',

```
'000',  
'blog',  
'comes',  
'wordpress']
```

```
In [9]: # show the last 10 terms in the vocabulary  
cvmodel.vocabulary[-10:]
```

```
Out[9]: ['customer',  
         'desktop',  
         'buttons',  
         'previous',  
         'math',  
         'master',  
         '000',  
         'blog',  
         'comes',  
         'wordpress']
```

## 5 Inter-document Frequency

```
In [10]: idf = IDF(inputCol="TF", outputCol="TFIDF")  
idfModel = idf.fit(df)  
df = idfModel.transform(df)  
df.head()
```

```
Out[10]: Row(Body="<p>I'd like to check if an uploaded file is an image file (e.g png, jpg, jpeg
```

## 6 StringIndexer

```
In [11]: indexer = StringIndexer(inputCol="oneTag", outputCol="label")  
df = indexer.fit(df).transform(df)
```

```
In [12]: df.head()
```

```
Out[12]: Row(Body="<p>I'd like to check if an uploaded file is an image file (e.g png, jpg, jpeg
```

```
In [ ]:
```