Databases, Scrapping, and APIs

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

After this lesson, you should be able to:

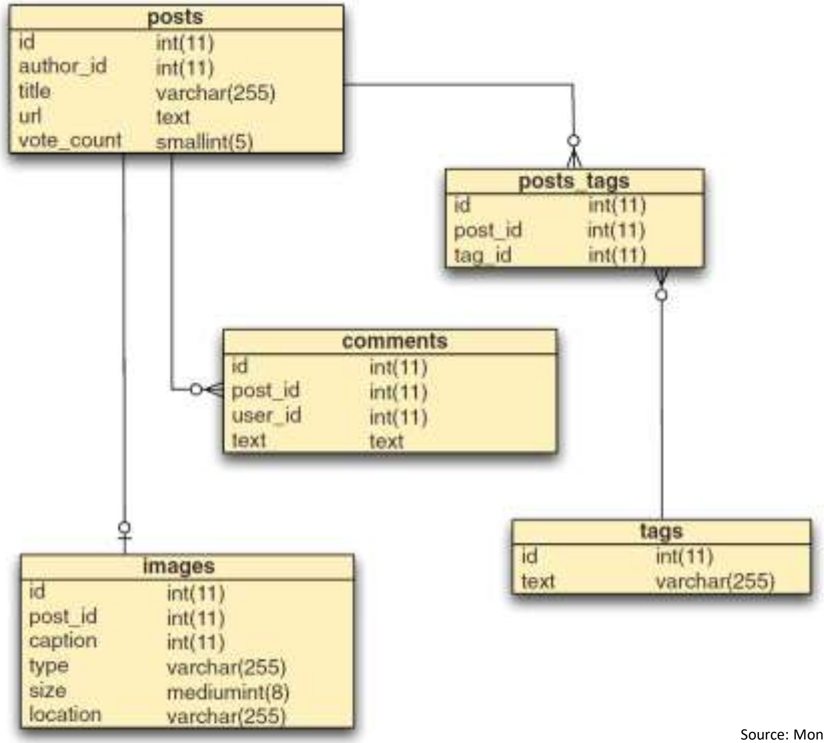
- Understand the uses and differences of the major types of databases, including RDBMS databases (i.e., SQL) and NoSQL databases
- Access databases from *pandas*
- Describe how web scraping works, conceptually and explain practically how to web scraping works using Python
- Define how to approach scraping project data
- Describe APIs and how to make calls and consume API data

Here’s what’s happening today:

- RDBMS/SQL Databases
  - The `SELECT` statement
- NoSQL (Not-Only SQL) Databases
  - NoSQL Classification
  - CAP Theorem
  - ACID vs. BASE
  - Popular NoSQL Databases
  - CRUD and REST
  - Map Reduce
- Scrapping
- Application Programming Interfaces (APIs)

Relational Database Management Systems (RDBMS) and Structured Query Language (SQL)

A relational database links data entities and concepts. E.g., a relational data model for entries on a social news site



posts

id	int(11)
author_id	int(11)
title	varchar(255)
url	text
vote_count	smallint(5)

posts_tags

id	int(11)
post_id	int(11)
tag_id	int(11)

comments

id	int(11)
post_id	int(11)
user_id	int(11)
text	text

tags

id	int(11)
text	varchar(255)

images

id	int(11)
post_id	int(11)
caption	int(11)
type	varchar(255)
size	mediumint(8)
location	varchar(255)

Source: MongoDB in Action

Relational databases are organized into *tables.* Each table has a specific *schema*, a set of rules for what goes in each table

- Each table corresponding to one entity or concept
- A table is made up of rows and columns, similar to a *pandas* dataframe

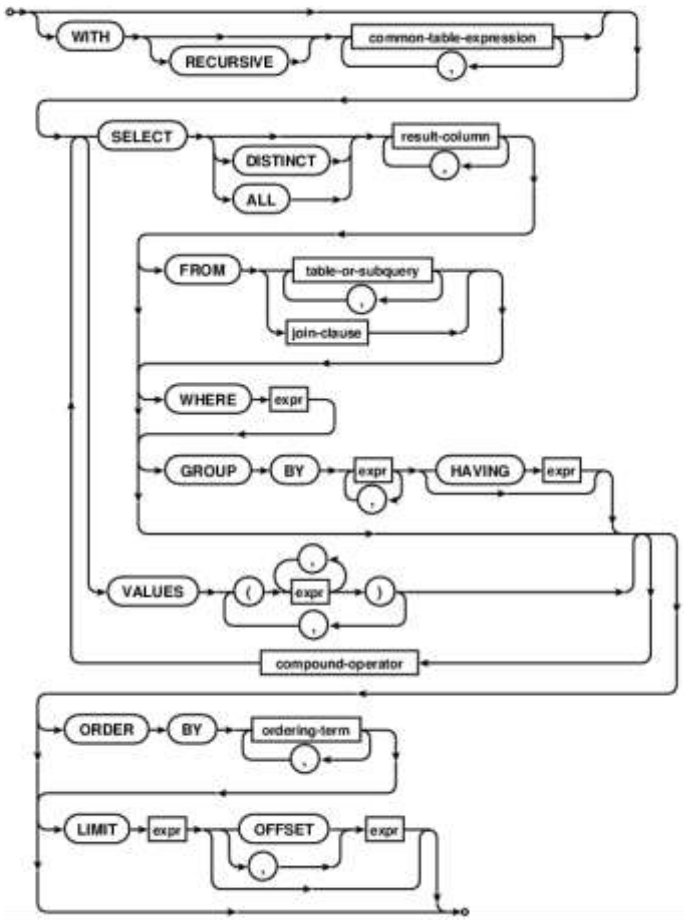
- Schemas specify which columns are contained in the table and what type of data is in each column (e.g., text, integer, or date). This means you can't add text data to an integer column
- For this reason and many others, databases allow for stronger consistency of the data and are often a better solution for data storage

CRUD and SQL

CRUD(*) **(*) the four basic functions of persistent storage**	**SQL Statement**	
Create	CREATE	Create or add new entries
Read	**SELECT**	Read, retrieve, search, or view existing entries
Update	UPDATE	Update or edit existing entries
Delete	DELETE	Delete/deactivate/remove existing entries

SQLite’s **SELECT** Statement

(https://www.sqlite.org/lang_select.html)



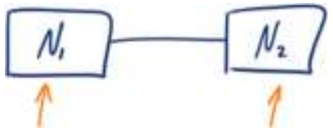
DS

NoSQL (Not-Only SQL) Databases

NoSQL databases fall into four primary classifications

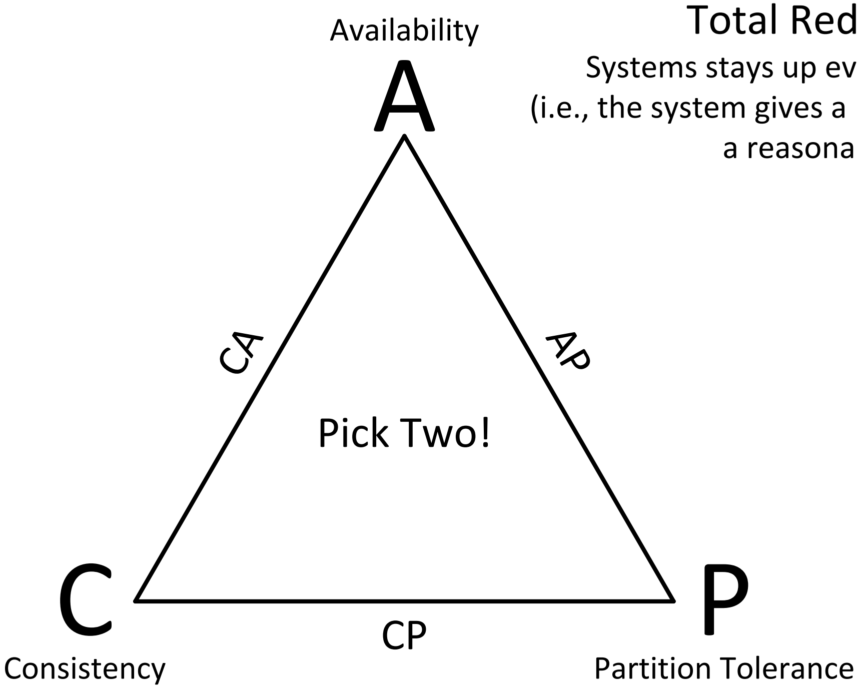
- **Key-value stores** – use a simple data model that pairs a unique key and its associated value in storing data elements. Common uses include storing clickstream data and application logs
- **(Wide)-column stores** (a.k.a., table-style databases) – store data across tables that can have very large numbers of columns. Common uses include Internet search and other large-scale Web applications
- **Document databases** – store data elements in document-like structures that encode information in formats such as JSON. Common uses include content management and monitoring Web and mobile applications
- **Graph databases** – emphasize connections between data elements, storing related "nodes" in graphs to accelerate querying. common uses include recommendation engines and geospatial applications

The CAP theorem states that in the presence of a network partition (P), one has to choose between consistency (C) and availability (A)



Total Redundancy

Systems stays up even when nodes fail
(i.e., the system gives a reasonable response in
a reasonable time)



ACID Transactions

Everyone always have the same view of the
data (regardless of any pending updates or
deletes)

Infinite Scaleout

System stays up even when network between
nodes fail (i.e., only a total network failure can
cause the system to respond incorrectly)

ACID vs. BASE: The pH of Database Transaction Processing

ACID

- **Atomicity** – all operations are performed or none of them are. If one part of the transaction fails, then all fail
- **Consistency** – a transaction must meet all rules defined by the system at all times; there are never any half-completed transactions
- **Isolation** – transactions are independent from each other
- **Durability** – once complete, a transaction cannot be undone

BASE

- **Basically Available** – the system will give and accept queries and give responses even in regards to node failures
- **Soft State** – the data is in a constant state of flux and might be stale
- **Eventual Consistency** – the data will eventually be consistent through all nodes and in all databases, but not every transaction at every moment

AC/AP/CP vs. ACID/BASE

AC

- Small datasets can be both consistent and available but a non-option in distributed systems (networks aren’t completely reliable so you must tolerate partitions)

AP
BASE w/ eventual consistency

- System returns the most recent version of the data it has (which could be stale). The system will also accept writes that can be processed later when the partition is resolved
- Choose AP (over C) when you are flexible on when the data in the system synchronizes

CP
ACID w/ eventual availability

- System waits for a response from the partitioned node which could result in a timeout error
- Choose CP (over A) when you require atomic reads and writes

Popular NoSQL databases(*)

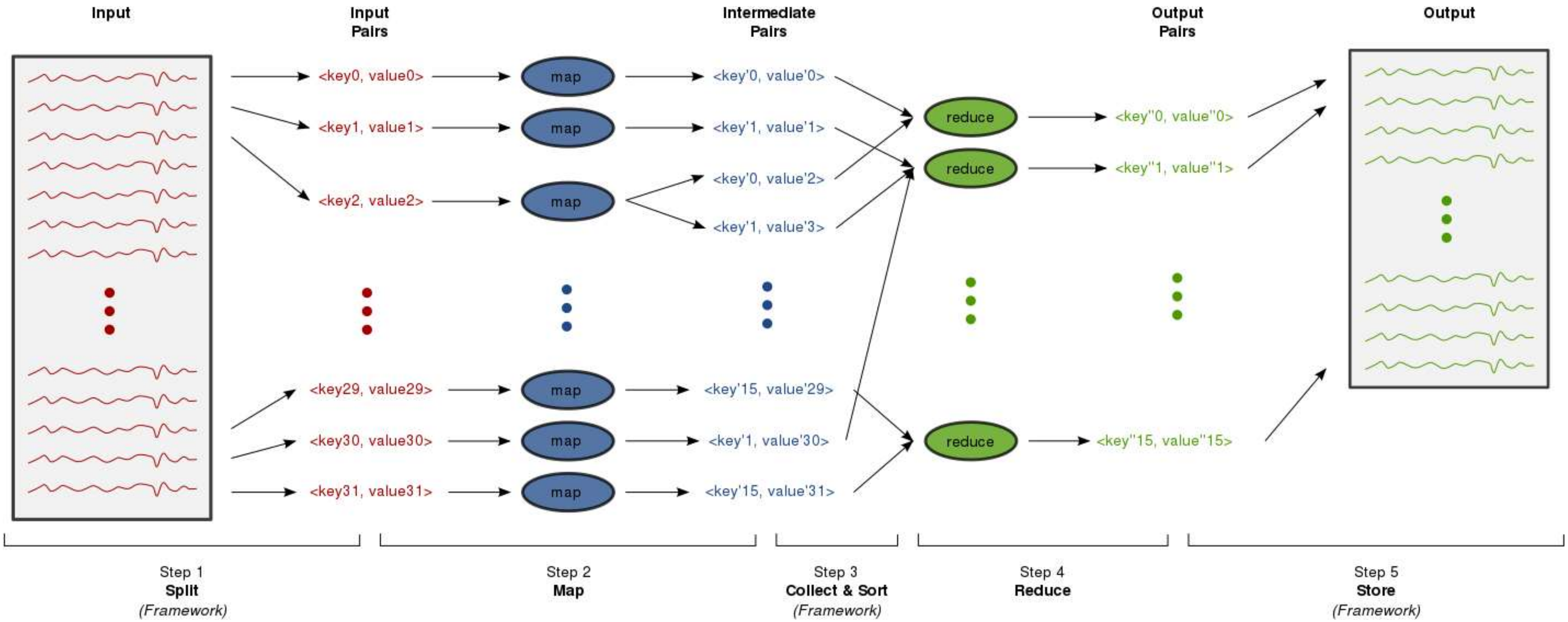
(*) Many systems allows you to tune both the write and the read quorums can be either CP or AP, depending on your needs.

Accumulo	Wide-column store	CP/ACID
Cassandra	Wide-column store	AP/BASE
CouchDB	Document database	AP/BASE
DynamoDB	Key-value store	AP/BASE
HBase	Wide-column store	CP/ACID
MongoDB	Document database	CP/ACID
Neo4j	Graph database	CP/ACID
Redis	Key-value store	CP/ACID
Riak	Key-value store	AP/BASE
SimpleDB	Wide-column store	AP/BASE

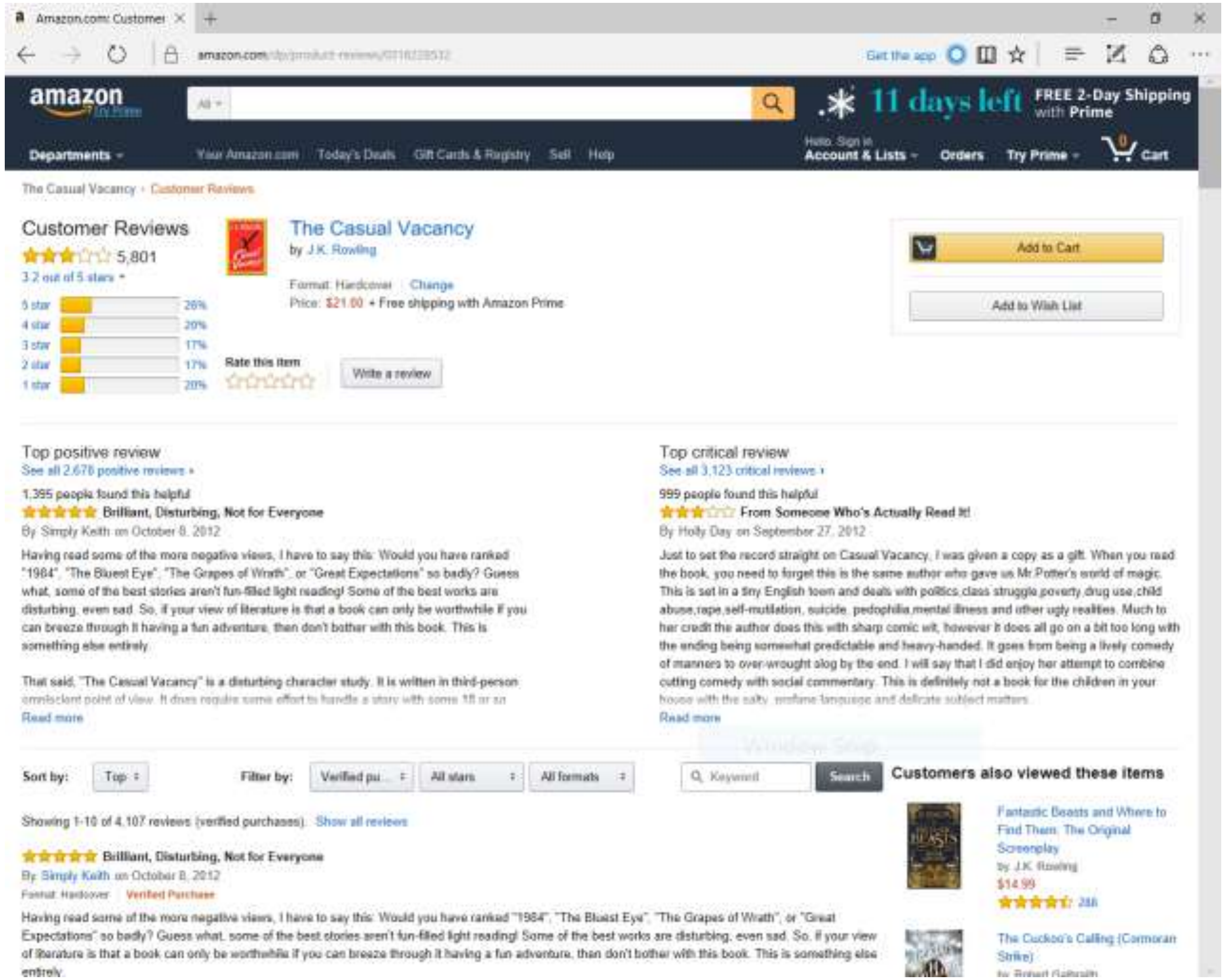
CRUD and REST

CRUD[*] [*] **the four basic functions of persistent storage**	**HTTP**[**] [**] **Methods for RESTful services**	
Create	POST	Create or add new entries
Read	**GET**	Read, retrieve, search, or view existing entries
Update	**POST**	Update or edit existing entries
Delete	DELETE	Delete/deactivate/remove existing entries

Map Reduce


Input
Input Pairs
Intermediate Pairs
Output Pairs
Output
<key0, value0>
<key1, value1>
<key2, value2>
<key29, value29>
<key30, value30>
<key31, value31>
map
map
map
map
map
map
<key'0, value'0>
<key'1, value'1>
<key'0, value'2>
<key'1, value'3>
<key'15, value'29>
<key'1, value'30>
<key'15, value'31>
reduce
reduce
reduce
<key''0, value''0>
<key''1, value''1>
<key''15, value''15>
Step 1
Split
(Framework)
Step 2
Map
Step 3
Collect & Sort
(Framework)
Step 4
Reduce
Step 5
Store
(Framework)

Scraping



The Casual Vacancy › Customer Reviews

Customer Reviews

★★★☆☆ 5,801

3.2 out of 5 stars

5 star	26%
4 star	20%
3 star	17%
2 star	17%
1 star	20%

The Casual Vacancy

by J.K. Rowling

Format: Hardcover | Change

Price: $21.00 + Free shipping with Amazon Prime

Top positive review

See all 2,678 positive reviews ›

1,395 people found this helpful

★★★★★ **Brilliant, Disturbing, Not for Everyone**

By Simply Keith on October 8, 2012

Having read some of the more negative views, I have to say this: Would you have ranked "1984", "The Bluest Eye", "The Grapes of Wrath", or "Great Expectations" so badly? Guess what, some of the best stories aren't fun-filled light reading! Some of the best works are disturbing, even sad. So, if your view of literature is that a book can only be worthwhile if you can breeze through it having a fun adventure, then don't bother with this book. This is something else entirely.

That said, "The Casual Vacancy" is a disturbing character study. It is written in third-person omniscient point of view. It does require some effort to handle a story with some 18 or so

Read more

Top critical review

See all 3,123 critical reviews ›

999 people found this helpful

★★★☆☆ **From Someone Who's Actually Read It!**

By Holly Day on September 27, 2012

Just to set the record straight on Casual Vacancy, I was given a copy as a gift. When you read the book, you need to forget this is the same author who gave us Mr.Potter's world of magic. This is set in a tiny English town and deals with politics,class struggle,poverty,drug use,child abuse,rape,self-mutilation, suicide, pedophilia,mental illness and other ugly realities. Much to her credit the author does this with sharp comic wit, however it does all go on a bit too long with the ending being somewhat predictable and heavy-handed. It goes from being a lively comedy of manners to over-wrought slog by the end. I will say that I did enjoy her attempt to combine cutting comedy with social commentary. This is definitely not a book for the children in your house with the salty, profane language and delicate subject matters.

Read more

Sort by: Top

Filter by: Verified pu... | All stars | All formats

Keyword | Search

Showing 1-10 of 4,107 reviews (verified purchases). Show all reviews

★★★★★ **Brilliant, Disturbing, Not for Everyone**

By Simply Keith on October 8, 2012

Format: Hardcover | Verified Purchase

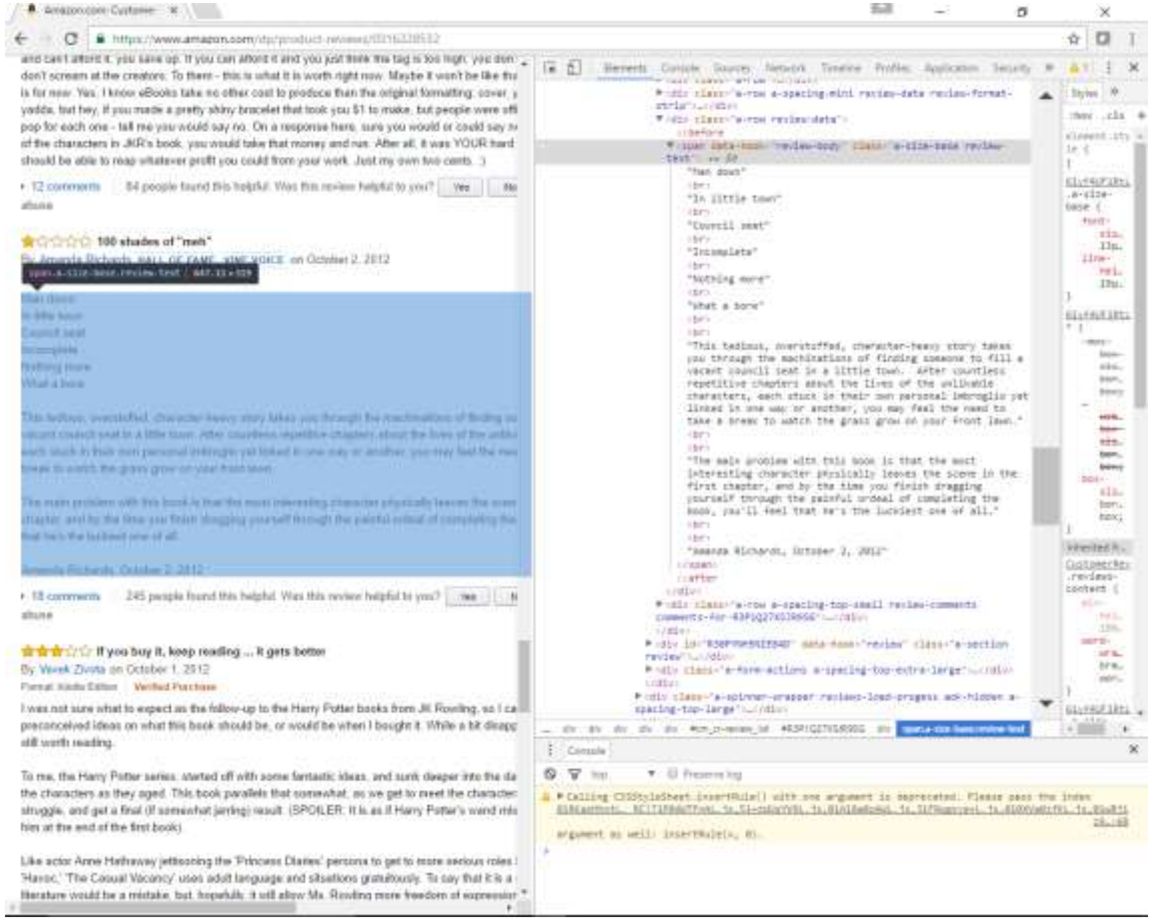
Having read some of the more negative views, I have to say this: Would you have ranked "1984", "The Bluest Eye", "The Grapes of Wrath", or "Great Expectations" so badly? Guess what, some of the best stories aren't fun-filled light reading! Some of the best works are disturbing, even sad. So, if your view of literature is that a book can only be worthwhile if you can breeze through it having a fun adventure, then don't bother with this book. This is something else entirely.

Customers also viewed these items

Fantastic Beasts and Where to Find Them: The Original Screenplay
by J.K. Rowling
$14.99
★★★★½ 288

The Cuckoo's Calling (Cormoran Strike)
by Robert Galbraith

Google Chrome’s inspect element feature is a great tool to scrape web sites



and can't afford it, you save up. If you can afford it and you just think the tag is too high, you don
don't scream at the creators. To them - this is what it is worth right now. Maybe it won't be like thi
is for now. Yes, I know eBooks take no other cost to produce than the original formatting, cover, y
yadda, but hey, if you made a pretty shiny bracelet that took you $1 to make, but people were off
pop for each one - tell me you would say no. On a response here, sure you would or could say no
of the characters in JKR's book, you would take that money and run. After all, it was YOUR hard
should be able to reap whatever profit you could from your work. Just my own two cents. :)

12 comments | 84 people found this helpful. Was this review helpful to you? Yes No
abuse

100 shades of "meh"

By Amanda Richards HALL OF FAME VINE VOICE on October 2, 2012

span.a-size-base.review-text | 447.33 x 329

Man down
In little town
Council seat
Incomplete
Nothing more
What a bore

This tedious, overstuffed, character-heavy story takes you through the machinations of finding so
vacant council seat in a little town. After countless repetitive chapters about the lives of the unlike
each stuck in their own personal imbroglio yet linked in one way or another, you may feel the nee
break to watch the grass grow on your front lawn.

The main problem with this book is that the most interesting character physically leaves the scen
chapter, and by the time you finish dragging yourself through the painful ordeal of completing the
that he's the luckiest one of all.

Amanda Richards, October 2, 2012

18 comments | 245 people found this helpful. Was this review helpful to you? Yes N
abuse

If you buy it, keep reading ... it gets better

By Vivek Zivota on October 1, 2012

Format: Kindle Edition | Verified Purchase

I was not sure what to expect as the follow-up to the Harry Potter books from JK Rowling, so I ca
preconceived ideas on what this book should be, or would be when I bought it. While a bit disapp
still worth reading.

To me, the Harry Potter series started off with some fantastic ideas, and sunk deeper into the da
the characters as they aged. This book parallels that somewhat, as we get to meet the character
struggle, and get a final (if somewhat jarring) result. (SPOILER: It is as if Harry Potter's wand mis
him at the end of the first book)

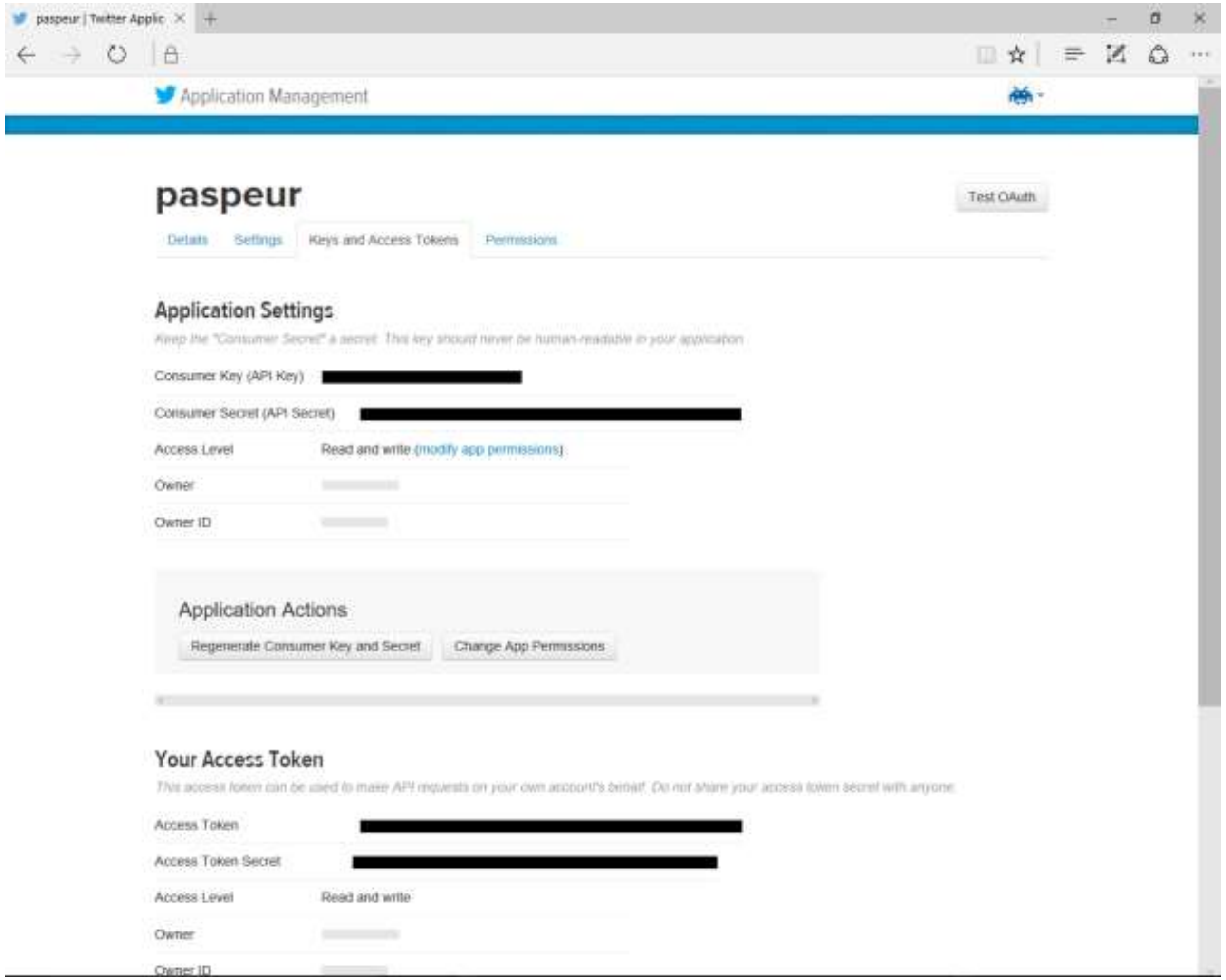
Like actor Anne Hathaway jettisoning the 'Princess Diaries' persona to get to more serious roles
'Havoc,' 'The Casual Vacancy' uses adult language and situations gratuitously. To say that it is a
literature would be a mistake, but, hopefully, it will allow Ms. Rowling more freedom of expressior

```
        ▶<div class="a-row a-spacing-mini review-data review-format-strip">…</div>
        ▼<div class="a-row review-data">
            ::before
          ▼<span data-hook="review-body" class="a-size-base review-text"> == $0
              "Man down"
              <br>
              "In little town"
              <br>
              "Council seat"
              <br>
              "Incomplete"
              <br>
              "Nothing more"
              <br>
              "What a bore"
              <br>
              <br>
              "This tedious, overstuffed, character-heavy story takes
              you through the machinations of finding someone to fill a
              vacant council seat in a little town.  After countless
              repetitive chapters about the lives of the unlikable
              characters, each stuck in their own personal imbroglio yet
              linked in one way or another, you may feel the need to
              take a break to watch the grass grow on your front lawn."
              <br>
              <br>
              "The main problem with this book is that the most
              interesting character physically leaves the scene in the
              first chapter, and by the time you finish dragging
              yourself through the painful ordeal of completing the
              book, you'll feel that he's the luckiest one of all."
              <br>
              <br>
              "Amanda Richards, October 2, 2012"
            </span>
            ::after
          </div>
        ▶<div class="a-row a-spacing-top-small review-comments comments-for-R3P1Q27XOJR9SG">…</div>
        </div>
      ▶<div id="R30PPXR5NIE84D" data-hook="review" class="a-section review">…</div>
      ▶<div class="a-form-actions a-spacing-top-extra-large">…</div>
      </div>
    ▶<div class="a-spinner-wrapper reviews-load-progess aok-hidden a-spacing-top-large">…</div>
```

```
… div div div div div #cm_cr-review_list #R3P1Q27XOJR9SG div span.a-size-base.review-text
```

```
▲ ▶Calling CSSStyleSheet.insertRule() with one argument is deprecated. Please pass the index
  argument as well: insertRule(x, 0).
```

Application Programming Interfaces (APIs)



paspeur

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) [redacted]

Consumer Secret (API Secret) [redacted]

Access Level Read and write (modify app permissions)

Owner

Owner ID

Application Actions

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token [redacted]

Access Token Secret [redacted]

Access Level Read and write

Owner

Owner ID