WHAT TO DO WITH DIRTY DATA

Introduction

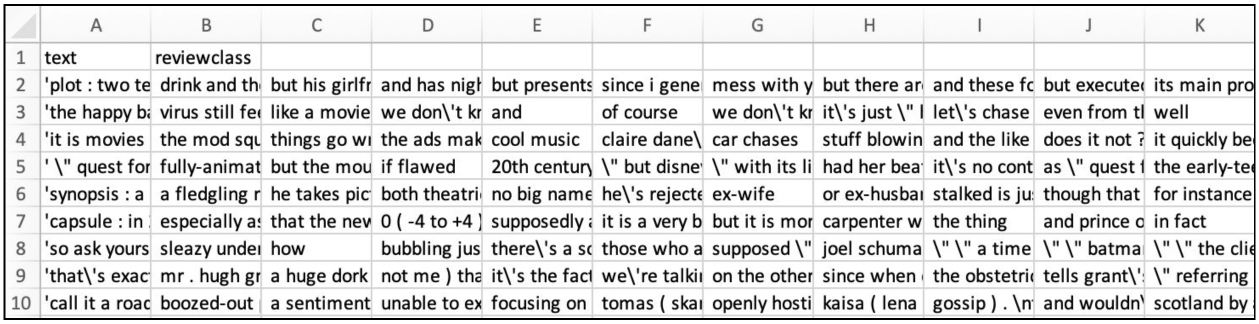
What is data? A better question would be "what **isn't** data?" Data is everything. Quite literally anything can be data. However, in order for it to be data we can use in a paper like this one, we have to take what is often noisy, non-numeric information and carefully clean it (often transforming things like words or photos into tokens and pixels) into something that ends up looking very different from the "data" we started with.

We usually think of "data" as a nice clean spreadsheet. This, however, is a lie. A fantasy. Something you must shoo from your mind immediately or else you will enter into a void of pain and existentialism eventually asking yourself “what even is a number and why do I care?” But, before we leave this dream world entirely, let’s consult the **Dream Data.** This dream data is clearly labeled, filled with mostly numeric, normalized data. There are no missing fields, no strange characters, no instances of nesting and definitely no emojis. This dream data is definitely not missing large swaths of information. It’s not an entire endless matrix of 0s with the occasional 1s. And it is most definitely not a hexadecimal representation of a photo of handwriting from a serial killer. However, to reiterate, all of that is data. So, how can we, as scientists, turn all of this flotsam and jetsam into something our computers can begin to “understand?”

Analysis & Models

ABOUT THE DATA

MoviesRAW.csv is a “dirty” data file. If the “ideal data file” is something like a two column spreadsheet, this is a non-ideal data file. It has many columns for each review and a number of non-alpha characters cluttering each review string. Finally, the label for each review is tacked on to the end, wherever that end may be.



	A	B	C	D	E	F	G	H	I	J	K
1	text	reviewclass									
2	'plot : two te	drink and th	but his girlfr	and has nigh	but presents	since i gene	mess with y	but there ar	and these fc	but execute	its main pro
3	'the happy b	virus still fe	like a movie	we don\'t kr	and	of course	we don\'t kr	it\'s just \" I	let\'s chase	even from tl	well
4	'it is movies	the mod squ	things go wi	the ads mak	cool music	claire dane\	car chases	stuff blowin	and the like	does it not ?	it quickly be
5	' \" quest for	fully-animat	but the mou	if flawed	20th century	\" but disne	\" with its li	had her bea	it\'s no cont	as \" quest f	the early-tee
6	'synopsis : a	a fledgling r	he takes pic	both theatri	no big name	he\'s rejecte	ex-wife	or ex-husba	stalked is ju	though that	for instance
7	'capsule : in	especially as	that the new	0 (-4 to +4)	supposedly	it is a very b	but it is mor	carpenter w	the thing	and prince o	in fact
8	'so ask yours	sleazy unde	how	bubbling jus	there\'s a sc	those who a	supposed \"	joel schuma	\" \" a time	\" \" batma	\" \" the clie
9	'that\'s exac	mr . hugh gr	a huge dork	not me) tha	it\'s the fact	we\'re talki	on the other	since when	the obstetri	tells grant\'	\" referring
10	'call it a roac	boozed-out	a sentiment	unable to ex	focusing on	tomas (ska	openly hosti	kaisa (lena	gossip) . \n	and wouldn\'	scotland by

FIGURE 1: The original csv file

CLEANING THE DATA

There are many different ways up this mountain but for this exercise, we will demonstrate two different approaches. The first way, Ami’s way, takes the data and keeps it in text form while cleaning, ultimately exporting a new, cleaner text file for a fresh import. The second way, Kendra’s way, takes the data and immediately turns it into a pandas dataframe. There are pros and cons to each way.

AMI’S WAY

Overview:

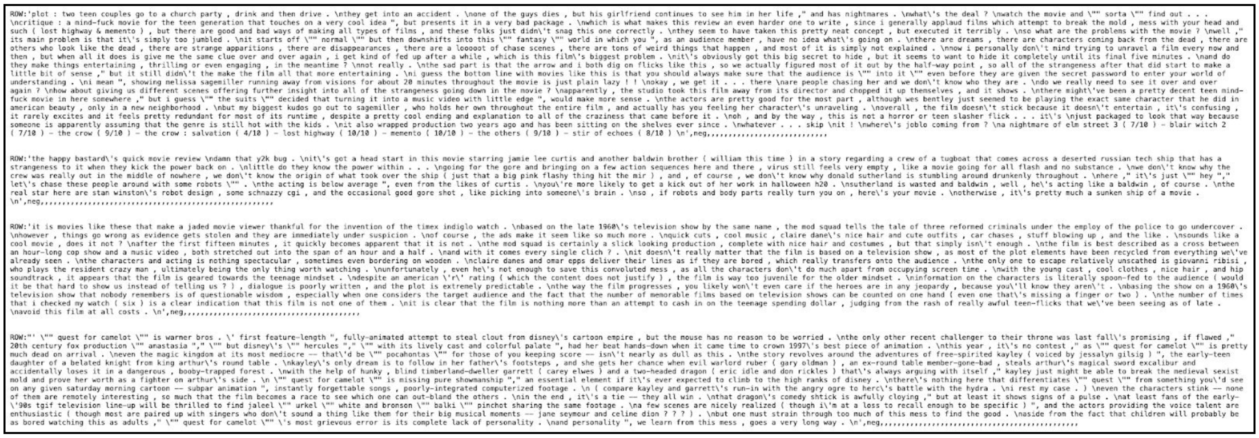
1. Read in the dirty file
2. Prep new clean files
3. Clean the data
    a. For each row in the data, clean the row
    b. For each word in the row, clean the word
4. Export clean data to new clean files
5. Re-import the cleaned data
6. Turn cleaned data into a pandas df

Code to clean each row

```
def display_rows(file_data):
    for row in file_data:
        row = row.lstrip()
        row = row.rstrip()
        row = row.strip()
        raw_row = "\n\nROW:" + row + "\n"
        outfile.write(raw_row)
        row_list = row.split(" ")
        new_list = []
        for word in row_list:
            to_put_in_outfile = "The next word BEFORE is: "+ word +"\n"
            outfile.write(to_put_in_outfile)
            word = clean_word(word)
            if word:
                new_list.append(word)
        label = ''.join(char for char in new_list[-1] if char.isalpha())
        new_list.pop()
        just_text = ' '.join(new_list)
        to_write = label + ',' + just_text + '\n'
        cleanfile.write(to_write)
```

Code to clean each word

```
def clean_word(word):
    word=word.lower()
    word=word.lstrip()
    word=word.lstrip("\\n")
    word=word.strip("\n")
    word=word.replace(",","")
    word=word.replace(" ","")
    word=word.replace("_","")
    word=re.sub('\+', ' ',word)
    word=re.sub('.*\+\n', '',word)
    word=re.sub('zz+', ' ',word)
    word=word.replace("\t","")
    word=word.replace(".","")
    word=word.strip()
    word = word.replace("\\'","")
    if word not in ["", "\\", '"', "'", "*", ":", ";"]:
        if len(word) >= 3:
            if not re.search(r'\d', word): ##remove digits
                return word
```



```
ROW:'plot : two teen couples go to a church party , drink and then drive . \nthey get into an accident . \none of the guys dies , but his girlfriend continues to see him in her life ," and has nightmares . \nwhat\'s the deal ? \nwatch the movie and \"" sorta \"" find out . . .
\ncritique : a mind-fuck movie for the teen generation that touches on a very cool idea ", but presents it in a very bad package . \nwhich is what makes this review an even harder one to write , since i generally applaud films which attempt to break the mold , mess with your head and
such ( lost highway & memento ) , but there are good and bad ways of making all types of films , and these folks just didn\'t snag this one correctly . \nthey seem to have taken this pretty neat concept , but executed it terribly . \nso what are the problems with the movie ? \nwell ,"
its main problem is that it\'s simply too jumbled . \nit starts off \"" normal \"" but then downshifts into this \"" fantasy \"" world in which you ", as an audience member , have no idea what\'s going on . \nthere are dreams , there are characters coming back from the dead , there are
others who look like the dead , there are strange apparitions , there are disappearances , there are a looooot of chase scenes , there are tons of weird things that happen , and most of it is simply not explained . \nnow i personally don\'t mind trying to unravel a film every now and
then , but when all it does is give me the same clue over and over again , i get kind of fed up after a while , which is this film\'s biggest problem . \nit\'s obviously got this big secret to hide , but it seems to want to hide it completely until its final five minutes . \nand do
they make things entertaining , thrilling or even engaging , in the meantime ? \nnot really . \nthe sad part is that the arrow and i both dig on flicks like this , so we actually figured most of it out by the half-way point , so all of the strangeness after that did start to make a
little bit of sense ," but it still didn\'t the make the film all that more entertaining . \ni guess the bottom line with movies like this is that you should always make sure that the audience is \"" into it \"" even before they are given the secret password to enter your world of
understanding . \ni mean ", showing melissa sagemiller running away from visions for about 20 minutes throughout the movie is just plain lazy ! ! \nokay , we get it . . . there \nare people chasing her and we don\'t know who they are . \ndo we really need to see it over and over
again ? \nhow about giving us different scenes offering further insight into all of the strangeness going down in the movie ? \napparently , the studio took this film away from its director and chopped it up themselves , and it shows . \nthere might\'ve been a pretty decent teen mind-
fuck movie in here somewhere ," but i guess \"" the suits \"" decided that turning it into a music video with little edge ", would make more sense . \nthe actors are pretty good for the most part , although wes bentley just seemed to be playing the exact same character that he did in
american beauty , only in a new neighborhood . \nbut my biggest kudos go out to sagemiller , who holds her own throughout the entire film , and actually has you feeling her character\'s unraveling . \noverall , the film doesn\'t stick because it doesn\'t entertain , it\'s confusing ,
it rarely excites and it feels pretty redundant for most of its runtime , despite a pretty cool ending and explanation to all of the craziness that came before it . \noh , and by the way , this is not a horror or teen slasher flick . . . it\'s \njust packaged to look that way because
someone is apparently assuming that the genre is still hot with the kids . \nit also wrapped production two years ago and has been sitting on the shelves ever since . \nwhatever . . . skip \nit ! \nwhere\'s joblo coming from ? \na nightmare of elm street 3 ( 7/10 ) - blair witch 2
( 7/10 ) - the crow ( 9/10 ) - the crow : salvation ( 4/10 ) - lost highway ( 10/10 ) - memento ( 10/10 ) - the others ( 9/10 ) - stir of echoes ( 8/10 ) \n',neg,,,,,,,,,,,,,,,,,,,,,,,,,,,


ROW:'the happy bastard\'s quick movie review \ndamn that y2k bug . \nit\'s got a head start in this movie starring jamie lee curtis and another baldwin brother ( william this time ) in a story regarding a crew of a tugboat that comes across a deserted russian tech ship that has a
strangeness to it when they kick the power back on . \nlittle do they know the power within . . . \ngoing for the gore and bringing on a few action sequences here and there , virus still feels very empty , like a movie going for all flash and no substance . \nwe don\'t know why the
crew was really out in the middle of nowhere , we don\'t know the origin of what took over the ship ( just that a big pink flashy thing hit the mir ) , and , of course , we don\'t know why donald sutherland is stumbling around drunkenly throughout . \nhere ," it\'s just \"" hey ","
let\'s chase these people around with some robots \"" . \nthe acting is below average ", even from the likes of curtis . \nyou\'re more likely to get a kick out of her work in halloween h20 . \nsutherland is wasted and baldwin , well , he\'s acting like a baldwin , of course . \nthe
real star here are stan winston\'s robot design , some schnazzy cgi , and the occasional good gore shot , like picking into someone\'s brain . \nso , if robots and body parts really turn you on , here\'s your movie . \notherwise , it\'s pretty much a sunken ship of a movie .
\n',neg,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,


ROW:'it is movies like these that make a jaded movie viewer thankful for the invention of the timex indiglo watch . \nbased on the late 1960\'s television show by the same name , the mod squad tells the tale of three reformed criminals under the employ of the police to go undercover .
\nhowever , things go wrong as evidence gets stolen and they are immediately under suspicion . \nof course , the ads make it seem like so much more . \nquick cuts , cool music , claire dane\'s nice hair and cute outfits , car chases , stuff blowing up , and the like . \nsounds like a
cool movie , does it not ? \nafter the first fifteen minutes , it quickly becomes apparent that it is not . \nthe mod squad is certainly a slick looking production , complete with nice hair and costumes , but that simply isn\'t enough . \nthe film is best described as a cross between
an hour-long cop show and a music video , both stretched out into the span of an hour and a half . \nand with it comes every single clich ? . \nit doesn\'t really matter that the film is based on a television show , as most of the plot elements have been recycled from everything we\'ve
already seen . \nthe characters and acting is nothing spectacular , sometimes even bordering on wooden . \nclaire danes and omar epps deliver their lines as if they are bored , which really transfers onto the audience . \nthe only one to escape relatively unscathed is giovanni ribisi ,
who plays the resident crazy man , ultimately being the only thing worth watching . \nunfortunately , even he\'s not enough to save this convoluted mess , as all the characters don\'t do much apart from occupying screen time . \nwith the young cast , cool clothes , nice hair , and hip
soundtrack , it appears that the film is geared towards the teenage mindset . \ndespite an american \'r\' rating ( which the content does not justify ) , the film is way too juvenile for the older mindset . \ninformation on the characters is literally spoon-fed to the audience ( would
it be that hard to show us instead of telling us ? ) , dialogue is poorly written , and the plot is extremely predictable . \nthe way the film progresses , you likely won\'t even care if the heroes are in any jeopardy , because you\'ll know they aren\'t . \nbasing the show on a 1960\'s
television show that nobody remembers is of questionable wisdom , especially when one considers the target audience and the fact that the number of memorable films based on television shows can be counted on one hand ( even one that\'s missing a finger or two ) . \nthe number of times
that i checked my watch ( six ) is a clear indication that this film is not one of them . \nit is clear that the film is nothing more than an attempt to cash in on the teenage spending dollar , judging from the rash of really awful teen-flicks that we\'ve been seeing as of late .
\navoid this film at all costs . \n',neg,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,


ROW:"' \"" quest for camelot \"" is warner bros . \' first feature-length ", fully-animated attempt to steal clout from disney\'s cartoon empire , but the mouse has no reason to be worried . \nthe only other recent challenger to their throne was last fall\'s promising , if flawed ,"
20th century fox production \"" anastasia "," \"" but disney\'s \"" hercules "," \"" with its lively cast and colorful palate ", had her beat hands-down when it came time to crown 1997\'s best piece of animation . \nthis year , it\'s no contest ," as \"" quest for camelot \"" is pretty
much dead on arrival . \neven the magic kingdom at its most mediocre -- that\'d be \"" pocahontas \"" for those of you keeping score -- isn\'t nearly as dull as this . \nthe story revolves around the adventures of free-spirited kayley ( voiced by jessalyn gilsig ) ", the early-teen
daughter of a belated knight from king arthur\'s round table . \nkayley\'s only dream is to follow in her father\'s footsteps , and she gets her chance when evil warlord ruber ( gary oldman ) , an ex-round table member-gone-bad , steals arthur\'s magical sword excalibur and
accidentally loses it in a dangerous , booby-trapped forest . \nwith the help of hunky , blind timberland-dweller garrett ( carey elwes ) and a two-headed dragon ( eric idle and don rickles ) that\'s always arguing with itself ," kayley just might be able to break the medieval sexist
mold and prove her worth as a fighter on arthur\'s side . \n \"" quest for camelot \"" is missing pure showmanship "," an essential element if it\'s ever expected to climb to the high ranks of disney . \nthere\'s nothing here that differentiates \"" quest \"" from something you\'d see
on any given saturday morning cartoon -- subpar animation ", instantly forgettable songs , poorly-integrated computerized footage . \n ( compare kayley and garrett\'s run-in with the angry ogre to herc\'s battle with the hydra . \ni rest my case . ) \neven the characters stink -- none
of them are remotely interesting , so much that the film becomes a race to see which one can out-bland the others . \nin the end , it\'s a tie -- they all win . \nthat dragon\'s comedy shtick is awfully cloying ," but at least it shows signs of a pulse . \nat least fans of the early-
\'90s tgif television line-up will be thrilled to find jaleel \"" urkel \"" white and bronson \"" balki \"" pinchot sharing the same footage . \na few scenes are nicely realized ( though i\'m at a loss to recall enough to be specific ) ", and the actors providing the voice talent are
enthusiastic ( though most are paired up with singers who don\'t sound a thing like them for their big musical moments -- jane seymour and celine dion ? ? ? ) . \nbut one must strain through too much of this mess to find the good . \naside from the fact that children will probably be
as bored watching this as adults ," \"" quest for camelot \"" \'s most grievous error is its complete lack of personality . \nand personality ", we learn from this mess , goes a very long way . \n',neg,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

FIGURE 2: An example of the outfile generated by Ami’s way

KENDRA’S WAY

Overview:

1. Read in the dirty file
2. Turn it into a pandas data frame
3. Merge all the rows to get the review text
4. Remove the last characters to get the “labels”
5. Clean the review text

Code to merge all the rows together

```
import pandas as pd
dirtyFile = pd.read_csv('moviereviewRAW.csv')
df = pd.DataFrame()
```

```
df['all'] = dirtyFile[dirtyFile.columns[0:]].apply(
    lambda x: ','.join(x.dropna().astype(str)),
    axis=1)
```

Code to get the label

```
df['label'] = df.apply(lambda x: x['all'][-3], axis=1)
```

In the same way there are many ways up the “how to process files” mountain, there are many ways to clean text. As we discovered in HW1, there are some things we want to keep, some things we want to discard. We rarely want to keep strange characters and unnecessary white spaces.

Code to clean excess characters

```
def clean_rogue_characters(string):
    exclude = ['\\',"\'"]
    string = '.'.join(string.split('\\n'))
    string = ''.join(ch for ch in string if ch not in exclude)
    return string

df['all'] = df['all'].apply( lambda x: clean_rogue_characters(x) )
```

Results

To get ‘results’ for this quick-and-dirty assignment, the researchers compared the ‘dirty data’ to past data sets in their ‘sentiment analysis’ pipeline to answer the question -- does this newly cleaned data behave very similarly, slightly similarly or not at all similarly to a cleaner dataset from the wild?

Text Blob

	Kendra’s Data	**Ami’s Data**	**Cornell Data**	**Dirty Data**	**Joker Data**
CORRECT NEG	5	1	229	227	64
CORRECT POS	0	4	971	972	114

VADER

	Kendra’s Data	Ami’s Data	Cornell Data	Dirty Data	Joker Data
CORRECT NEG	2	3	445	454	64
CORRECT POS	5	3	828	824	114

NLTK

	Kendra’s Data	Ami’s Data	Cornell Data	Dirty Data	Joker Data
CORRECT NEG	--	--	89%	86%	81%
CORRECT POS	--	--	74%	70%	35%
ACCURACY	--	--	81%	77%	58%

Looking at the “Dirty Data” compared to the Cornell data, it’s clear to see that once cleaned, the dirty data performed almost eerily similarly to the clean data. This is without additional NLP cleaning, simply a baseline analysis. For the purposes of this exploration, the dirty data was cleaned enough to perform as well as its cleaned counterpart.

Conclusion

Dirty data is everywhere. Having a pipeline, (or multiple pipelines!) with which to quickly clean, format and export data is essential to being an effective data scientist. The researchers suggest having multiple different (and mutable!) pipelines for different tasks -- is the dirty data coming from the web? Is it littered with HTML? Is the dirty data coming from a poorly formatted csv? There will never be a one-size-fits-all cleaner, however, there will be ways to quickly format the data for easier more in-depth “post cleaning.”