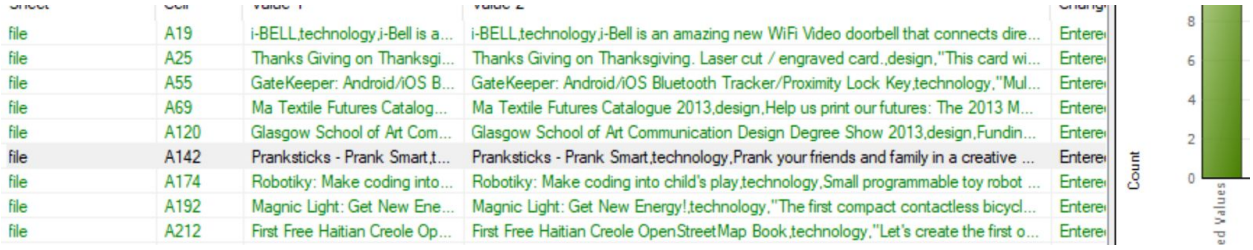
(Anything that was edited was done through a script or function in python or excel to be sure that EVERYTHING is scalable to massive proportions! In other words, nothing is being done by hand and nothing is hard-coded.)

I don't think there is reason to merge these two files specifically because one appears to be just a newer  (more recent) crawl of the same pages. The only differences are things like the headers on comments saying "almost 5 years ago" -> "about 5 years ago" and minor formatting differences. If there's something I misunderstood or a different file they can be merged.

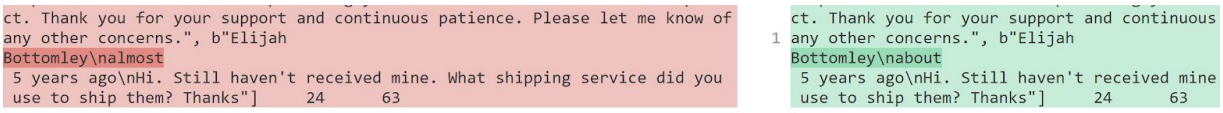
"Kicstarter_master20181115-1.csv" appears to be the newer one, I will clean up and use this one.



| | | | | |
|---|---|---|---|---|
| file | A19 | i-BELL,technology,i-Bell is a... | i-BELL,technology,i-Bell is an amazing new WiFi Video doorbell that connects dire... | Entere |
| file | A25 | Thanks Giving on Thanksgi... | Thanks Giving on Thanksgiving. Laser cut / engraved card.,design,"This card wi... | Entere |
| file | A55 | GateKeeper: Android/iOS B... | GateKeeper: Android/iOS Bluetooth Tracker/Proximity Lock Key,technology,"Mul... | Entere |
| file | A69 | Ma Textile Futures Catalog... | Ma Textile Futures Catalogue 2013,design,Help us print our futures: The 2013 M... | Entere |
| file | A120 | Glasgow School of Art Com... | Glasgow School of Art Communication Design Degree Show 2013,design,Fundin... | Entere |
| file | A142 | Pranksticks - Prank Smart,t... | Pranksticks - Prank Smart,technology,Prank your friends and family in a creative ... | Entere |
| file | A174 | Robotiky: Make coding into... | Robotiky: Make coding into child's play,technology,Small programmable toy robot ... | Entere |
| file | A192 | Magnic Light: Get New Ene... | Magnic Light: Get New Energy!,technology,"The first compact contactless bicycl... | Entere |
| file | A212 | First Free Haitian Creole Op... | First Free Haitian Creole OpenStreetMap Book,technology,"Let's create the first o... | Entere |



https://gyazo.com/7c82065305fc060db7d46b7f26574cae



```
ct. Thank you for your support and continuous patience. Please let me know of
1 any other concerns.", b"Elijah
Bottomley\nalmost
 5 years ago\nHi. Still haven't received mine. What shipping service did you
 use to ship them? Thanks"]     24      63
```

```
ct. Thank you for your support and continuous
1 any other concerns.", b"Elijah
Bottomley\nabout
 5 years ago\nHi. Still haven't received mine
 use to ship them? Thanks"]     24      63
```

https://gyazo.com/02110301424d934b880cc10d9547c80b

Vader seems to be a good choice to try first because it is easy to integrate with python and automate the workflow from pulling data to tokenizing, analyzing, and inserting the results. To get a good number for overall "intensity" I came up with this formula to start with for each sentence:

$$\frac{\text{Positive Intensity}_{normalized\ 0-1} + \text{NegativeIntensity}_{normalized0-1}}{2} - \text{Neutral Intensity}_{normalized\ 0-1}$$

\frac{\mathrm{Positive\: Intensity_{\mathit{normalized\, 0-1}}}+\mathrm{Negative Intensity_{\mathit{normalized 0-1}}}}{2}\; -\; \mathrm{Neutral\: Intensity_{\mathit{normalized\, 0-1}}}

Each sentence would have a value according to this formula then an overall score would be created for the each description by averaging each constituent sentences score. (Sentences are weighted equally.. Perhaps we can weigh by sentence length or use a different formula?)

My rationale for using this formula is that using the “compound” score that VADER outputs might be inappropriate for us because take this for an example:

If a sentence has extreme negative emotion such as “This really sucks” and extreme positive emotion such as “but also this part is really cool” the extremes would “cancel” and result in a compound score closer to 0. To avoid this, we take the raw positive score and raw negative score and add both and average them so that extreme negative emotions and extreme positive emotions BOTH actually contribute to our version of the “extremity” score.

Removed instances of many formatting characters like [ “ \b spaces etc. throughout the document, many were in the descriptions.

Excel data more cleaned up.

| Heading | Category | Brief descr | Start date | End date | Endorseme | Location ci | Location cc | Creator | Creator_id | Description |
|---|---|---|---|---|---|---|---|---|---|---|
| PETite Prin | design | Celebrate t | ######## | ######## | NA | ['Project W | US | erin l. meh | 8.32E+08 | UPDATE 7/9/13', bl can not say Thank You |
| Firm Founc | design | A field guid | ######## | ######## | NA | ['Project W | US | Solo Kota K | 1.57E+09 | In many places around the world, design tc |
| Clover Coo | design | The ultima | ######## | ######## | NA | ['San Franc | US | Alite Desigr | 2.06E+08 | We peeked into your kitchen drawers, and |
| Read Faste | technology | SuperRead | ######## | ######## | NA | ['Ipswich, L | GB | Iain Macas | 1.23E+09 | When we first learn to read we're taught tc |
| Vampire Pr | technology | Get OFF th | ######## | ######## | NA | ['Austin, TX | US | Vampire La | 1.07E+09 | Check out the Great Press our Project has F |
| BE High Scl | technology | Brookfield | ######## | ######## | NA | ['Brookfiel | US | Tim Vrakas | 1.95E+09 | Well, I lied. WE MADE IT!', Brookfield East |
| i-BELL | technology | i-Bell is an | ######## | ######## | NA | ['London, L | GB | I-BELL LTD | 7.29E+08 | I-Bell is the first stand-alone doorbell manu |
| Getting Sta | technology | A real-worl | ######## | ######## | NA | ['Oregon C | US | Kenneth Lc | 6.57E+08 | We hit our initial $5,000 goal (thank you ve |
| Freestyle F | technology | We want tc | ######## | ######## | NA | ['Project W | US | Pavan Bahl | 1.19E+09 | We are opening the doors to the Freestyle |
| Thanks Giv | design | This card w | ######## | ######## | NA | ['Lower Eas | US | Artistic Eng | 4.94E+08 | Hi Friends, Yesterday I received an email fr |
| Laravel Ele | technology | Support th | ######## | ######## | NA | ['Dayton, C | US | Brian Rette | 1.56E+09 | Show your Laravel Support with the first ba |
| the Pebble | technology | Train and n | ######## | ######## | NA | ['Amsterda | NL | Mojo Creat | 7.65E+08 | Mojo Creations develops tactile designer fi |

https://gyazo.com/2ab8d11e2f2e51755132b8fe1aecb053

Used punktSentenceTokenizer to do the main splitting into sentences.

Example of some of the other code:

```
for sentence in tokenized_description:
    result = analyzer.polarity_scores(sentence)
    print(result) #just debugging, can delete
    intensity_score = ((result['neg'] + result['pos'])/2) - result['neu']
    print("Sentence ", i, "intensity score: ", intensity_score)
    total_intensity_score = total_intensity_score + intensity_score
    i = i+1

total_intensity_score = total_intensity_score/i
print("total sentences: ", i)
```

https://gyazo.com/662af70a0e7affd8417a7484405448c0

Example of some of the command line info:

```
Sentence  1 intensity score:  -0.7314999999999999
{'neg': 0.0, 'neu': 0.953, 'pos': 0.047, 'compound': 0.2732}
Sentence  2 intensity score:  -0.9295
{'neg': 0.0, 'neu': 0.591, 'pos': 0.409, 'compound': 0.7371}
Sentence  3 intensity score:  -0.38649999999999995
{'neg': 0.0, 'neu': 0.909, 'pos': 0.091, 'compound': 0.0258}
Sentence  4 intensity score:  -0.8635
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Sentence  5 intensity score:  -1.0
total sentences:  6
total intensity score:  -0.76525
['The biggest risk I foresee is if the supplier for the waterjet parts gets too busy before I pl
 found 2 local shops that have this capability, and will be meeting with them in the next few da
k up plan in place before the end of this project.']
{'neg': 0.104, 'neu': 0.896, 'pos': 0.0, 'compound': -0.2732}
Sentence  0 intensity score:  -0.844
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Sentence  1 intensity score:  -1.0
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Sentence  2 intensity score:  -1.0
total sentences:  3
total intensity score:  -0.948
length of final list:  101
```

https://gyazo.com/76b76a2827697a99e43bb82abdc91e5b

With additional preprocessing the descriptions could be cleaned up a little more when being tokenized and the formula could be adjusted to something that provides less skewed results, but it's a starting place for now.

TODO/Ideas:

- Change the actual formula used to calculate intensity and compare to others
- If we do use the current formula or a similar one, perhaps weighing sentences by length (# of characters) would be a good idea to extract a better picture of the emotion in a given paragraph. The neutral part weighs too much, especially with slightly poorly tokenized sentences.
-