Cleaning a borked Jupyter notebook

It's a small simple thing, but this might save me (or you) an hour some day. One of my Jupyter notebooks got corrupted - I have some less-than-friendly tweet data in it that not only stopped my notebook from loading, but also crashed my Jupyter instance when I did. I would normally just fix this in the terminal window, but thought it might be nice to share. If you're already familiar with the backend of Jupyter and json file formats, feel free to skip to the next post. And if you've borked a Jupyter file but can still open it, then "clear all" might be a better solution. Otherwise...

Jupyter notebooks are pretty things, but behind the scenes, they're just a Json data object. Here's how to look at that, using Python (in another Jupyter notebook: yay, recursion!):

```
import json  raw = open('examine_cutdown_tweets.ipynb', 'r').read()  raw[
:200]
```

That just read the jupyter notebook file in as text. See those curly brackets: it's formatted as json. If we want to clean the notebook up, we need to read this data in as json. Here's how (and a quick look at its metadata):

```
jin = json.loads(raw)  print(jin.keys())  (jin['metadata'], jin['nbformat
'], jin['nbformat_minor'])
```

Each section of the notebook is in one of the "cells". Let's have a look at what's in the first one:

```
print(len(jin['cells']))  jin['cells'][0].keys()
```

The output from the cell is in "outputs". This is where the pesky killing-my-notebook output is hiding. The contents of the first of these looks like:

```
print(type(jin['cells'][0]['outputs']))  jin['cells'][0]['outputs']
```

At which point, if I delete all the outputs (or the one that I *know* is causing problems), I should be able to read in my notebook and look at the code and comments in it again.

```
for i in range(len(jin['cells'])):     jin['cells'][i]['outputs'] = []  j
in['cells']
```

And write out the cleaned-up notebook:

```
jout = json.dumps(jin)  open('tmp_notebook.ipynb', 'w').write(jout)
```

Finally, the code that does the cleaning (and just that code) is:

```
import json  raw = open('mynotebook.ipynb', 'r').read()  jin = json.loads
(raw)  for i in range(len(jin['cells'])):  jin['cells'][i]['outputs'] = [
]  open('tmp_mynotebook.ipynb', 'w').write(json.dumps(jin))
```

Good luck!