“The internet is the single most important development in human history since call waiting.”

– Dave Barry

“It would be pretty cool if you could get the wayback machine thing to work.”

– Jaren Haber

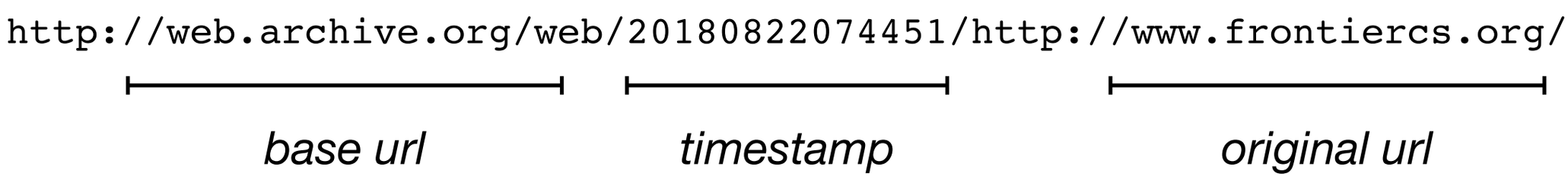
Question 1: How do we avoid scraping bad URLs?

we can define a bad URL as one where we can't get any sensible content, or one where the page doesn't exist. so how do we figure that out? the 'webtext' column gives us insights.

```
# Keep rows where NO WEBTEXT (0 pages) was gathered (i.e., scrape failed)
missing1 = charter_df[charter_df["WEBTEXT"].apply(len) == 0]

# Drop rows where NO WEBTEXT (0 pages) was gathered (i.e., scrape failed)
charter_df = charter_df[charter_df["WEBTEXT"].apply(len) > 0]

# Keep rows where first scraped page is empty
missing2 = charter_df[charter_df["WEBTEXT"].apply(lambda school: len(school[0][3])) == 0]

# Keep rows where all pages are empty
missing3 = charter_df[charter_df["WEBTEXT"].apply(lambda school: sum([len(page[3]) for page in school])) == 0]

missingdf = pd.concat([missing1, missing2, missing3])
```

snippet from scraping_badURLS

Question 2: How do we automatically search the Wayback Machine?




```
['http://web.archive.org/web/20180412015059/http://www.maef.net:80/',
 'http://web.archive.org/web/20170606173900/http://www.kgbsd.org/ketchikancharter',
 'http://web.archive.org/web/20180421031400/http://tongassschool.org:80/',
 'http://web.archive.org/web/20180408075758/http://aquarian.asdk12.org:80/',
 'http://web.archive.org/web/20180806044913/http://winterberrycharterschool.com/',
 'http://web.archive.org/web/20180822074451/http://www.frontiercs.org/',
 'http://web.archive.org/web/20180327115410/http://highlandacademy.asdk12.org:80/']
```

our function processes our existing list of urls to check for their existence, and then constructs the URL

Question 3: How do we integrate this into existing scraping architecture?

Option 1: Create our own new CSV file and run that through existing feed.py method

Didn't work. The feed.py method is built to work only on a certain format with many columns, not just two.

Option 2: Manually create a copy of a working CSV in the same format as a file that worked

Better. It is possible to replace an entire column with a list in pandas - like the generated one.