

Getting Data

[Re-submit Assignment](#)

Due Oct 16 by 11:59pm **Points** 100 **Submitting** a file upload
File Types pdf and html

Create a Jupyter Notebook that solves the following problems, using NumPy, BeautifulSoup, pandas, json, etc.

1) (30 points) Use Python to scrape the article posted to TechCrunch.com and present the articles on the front page in a table with the following pieces of information:

- article title
- publish date
- author
- tags

Note: respect the website and do not repeatedly access the site

Quick Scripts for Practice

2) (5 points) Scrape <https://www.data.gov> (<https://www.data.gov>) to identify the number of data sets available.

3) (10 points) Identify the number of U.S. Senate votes in the 115th Congress that were rejected by a margin of less than 5 votes. Also, provide a list of which votes (#) match this criteria.

Data available at: https://www.senate.gov/legislative/LIS/roll_call_lists/vote_menu_115_1.htm
(https://www.senate.gov/legislative/LIS/roll_call_lists/vote_menu_115_1.htm)

4) (5 points) Give the number of alerts and warnings for international travel given by the US government.

Data available at: <https://travel.state.gov/content/passports/en/alertswarnings.html>
(<https://travel.state.gov/content/passports/en/alertswarnings.html>)

5) (10 points) Report the total number of female babies whose names start with 'K' so far this decade.

Data available at: <https://www.ssa.gov/OACT/babynames/decades/names2010s.html>
(<https://www.ssa.gov/OACT/babynames/decades/names2010s.html>)

6) (40 points) Use Spotify API

You will be looking examining the popularity of songs. The music industry serves a global market and has revenue in the tens of billions. However, the industry is dominated by major production companies and labels. The labels incur costs (production, recording, marketing, etc.) for artists in exchange for a percentage of the profits from album sales, streaming services, tours, merchandise, etc.

A record company will want to turn a profit by working with successful artist's those who ensure a steady revenue stream. However, whether an artist is successful or not is very uncertain. One song of an artist may

be quite successful (high sales, radio and streaming plays) while all other songs by the same artist may be unsuccessful. This phenomenon is often referred to as a one-hit wonder. There are not specific criteria for one-hit wonders but usually this requires a song hitting the top 40 chart just once (e.g., Billboard top 40). Note, this classification is independent of awards, other notoriety or measures of success for an artist.

Examples of Recent and Classic US One-hit Wonders [123](#)


Artist	Single	Year
OMI	"Cheerleader"	2015
Hozier	"Take me to Church"	2014
Bastille	"Pompeii"	2014
The Neighbourhood	"Sweater Weather"	2013
A Great Big World	"Say Something"	2013
Awolnation	"Sail"	2013
Alex Clare	"Too Close"	2012
The Lumineers	"Ho Hey"	2012
Foster The People	"Pumped up Kicks"	2011
Gnarls Barkley	"Crazy"	2006
Chumbawamba	"Tubthumping"	1997
Blind Melon	"No Rain"	1993
a-ha	"Take on Me"	1985
Dexy's Midnight Runners	"Come on Eileen"	1982
Soft Cell	"Tainted Love"	1981
The Knack	My Sharona	1979

Other artists see may have success for a while, but recent releases not as successful as prior albums. For example, the recent albums and songs by No Doubt (Push and Shove, 2012), Smashing Pumpkins (Adore, 1998), Moby (Last Night, 2014), Britney Spears (Britney Jean, 2013), and Christina Aguilera (Bionic, 2014).⁴

You will get data to explore the success of a song

Data

You will be collecting the data to explore and analyze.⁵ You should be looking at the collecting both successful and unsuccessful songs.

As a start to this project, I will give you [data](#)  on songs to include in this analysis. The data was gathered using Billboard's Charts from Jan 1, 2000 - Dec. 31, 2005. The data set includes 350 songs that peaked within the Top 10 during that time. It also has 241 songs that peaked between 30-40 on the charts during that time.

Example Songs:

First Entries in Song List Data

Title	Artist	Peak	Date Entered	Successful
We Belong Together	Mariah Carey	1	4/16/05	1
Yeah!	Usher	1	1/10/04	1

Title	Artist	Peak	Date.Entered	Successful
Smooth	Santana	1	7/31/99	1
Lose Yourself	Eminem	1	10/5/02	1
Independent Women Part I	Destiny's Child	1	9/23/00	1
Dilemma	Nelly	1	7/13/02	1

As shown above this data set has the following pieces of information:

- song title
- artist
- peak (highest spot on the charts)
- date entered (when the song first appeared on the charts)
- successful (criteria described above)

Get Song Attributes

Next, you will need to get some additional information about the songs. You can use the [Spotify](https://www.spotify.com/us/) (<https://www.spotify.com/us/>) / [Echonest](http://the.echonest.com) (<http://the.echonest.com>) APIs to get these features.

A few of the general steps and some sample code to get you started.

1. You will need to sign up for a Spotify account (the free account is fine).
2. Register an application on [Spotify API](https://developer.spotify.com) (<https://developer.spotify.com>) to get credentials. Note, you do not have to complete an application or link it to a website, this is just to allow access to the Web API.

Note Make note of the terms of use for the site. For example, you are not to scrap down their data content. The limited collection of metadata for this project should not be expanded to include additional information. Finally, be sure to obey the data collection limits (e.g., number and size of requests per day).

3. Begin to collect a [song's audio features](https://developer.spotify.com/web-api/get-audio-features/) (<https://developer.spotify.com/web-api/get-audio-features/>). Namely, you will collect the following attributes:

Audio Features from Echo Nest⁶

Key	Value Type	Value Description
acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
analysis_url	string	An HTTP URL to access the full audio analysis of this track. An access token is required to access this data.
danceability	float	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
duration_ms	int	The duration of the track in milliseconds.

Key	Value Type	Value Description
energy	float	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
id	string	The Spotify ID for the track.
instrumentalness	float	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
key	int	The key the track is in. Integers map to pitches using standard Pitch Class notation (https://en.wikipedia.org/wiki/Pitch_class). E.g. 0 = C, 1 = C#/D \flat , 2 = D, and so on.
liveness	float	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
loudness	float	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
mode	int	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
speechiness	float	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
tempo	float	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
time_signature	int	An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
track_href	string	A link to the Web API endpoint providing full details of the track.
type	string	The object type: "audio_features"
uri	string	The Spotify URI for the track.
valence	float	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Missing Data

Are any of the songs not available in Spotify? If so this is an easy example of what to do with missing data, because the entire row will be deleted from the data set.

Deliverables

For the set of songs create a scatterplot for of energy vs. valence.

Note, this data set may be used for future assignments.

Submit your Jupyter notebook as a PDF or HTML file.

References

1. Wikipedia.com [url \(https://en.wikipedia.org/wiki/List_of_2010s_one-hit_wonders_in_the_United_States\)](https://en.wikipedia.org/wiki/List_of_2010s_one-hit_wonders_in_the_United_States) ↵
2. Billboard.com [url \(http://www.billboard.com/articles/news/266487/one-hit-wonders-of-the-2000s-page-1\)](http://www.billboard.com/articles/news/266487/one-hit-wonders-of-the-2000s-page-1) ↵
3. Rollingstone.com [url \(http://www.rollingstone.com/music/pictures/rolling-stone-readers-pick-the-top-10-one-hit-wonders-of-all-time-20110504\)](http://www.rollingstone.com/music/pictures/rolling-stone-readers-pick-the-top-10-one-hit-wonders-of-all-time-20110504) ↵
4. NME.com [url \(http://www.nme.com/photos/30-disastrous-album-flops-from-hitherto-successful-acts/343653#/photo/30\)](http://www.nme.com/photos/30-disastrous-album-flops-from-hitherto-successful-acts/343653#/photo/30) ↵
5. Problem inspired by MIT's EdX Analytics Edge course.↵
6. [developer.spotify.com \(https://developer.spotify.com/web-api/get-audio-features/\)](https://developer.spotify.com/web-api/get-audio-features/) ↵