Project: Wikidata, Movies, and Success

There is a Wikipedia cousin project WikiData *(https://www.wikidata.org/)* that contains many of the same facts as Wikipedia, but in a structured kind of way. For example, where Wikipedia describes SFU as a university founded in 1965, the WikiData SFU page *(https://www.wikidata.org/wiki/Q201603)* asserts that it is an "instance of" the concept "university" and has "inception" in "1965".

This project idea centres around movies (instances of film *(https://www.wikidata.org/wiki/Q11424)* in WikiData speak).

What makes a movie successful? Is it the genre of the movie? The subject? Specific actors? The country where it was created?

And what is success for a movie? Making money? Getting good reviews from critics or the general audience?

What advice do you have for the film-makers of the future?

Data

Data extracted from WikiData and two other relevant data sets are provided. They are described on a separate page of movie data descriptions.

Possible Questions

There are a lot of questions that are interesting and can be addressed by this data. Among them:

- Do the various criteria for success (critic reviews, audience reviews, profit/loss) correlate with each other? Is there something you can say about better or worse kinds of "success"?
- Can you predict review scores from other data we have about the movie? Maybe genre determines a lot about the success or a movie? Or maybe the actors?
- Does the plot summary predict success in any useful way? (See "Natural Language Processing" below.)
- What specific factors are related to a movie's success? Which are the most related? Is paying Vin Diesel worth it?
- Have any of these things changed over time (depending on the movie's release date)? Maybe people in the past liked documentaries more than we do now.

Getting Results

There are a thousand ways this data could be beaten into producing results. You are welcome to explore and see what interesting facts you can uncover.

While doing this, give a little though to the validity of your results. Are you p-hacking (or doing whatever the ML equivalent is)? If you think it's relevant, address this in your report.

Natural Language Processing

There are a few thousand records with plot summaries from OMDb *(http://www.omdbapi.com/)* . These are definite candidates for some kind of natural language processing *(https://en.wikipedia.org/wiki/Natural-language_processing)* .

Using the tf-idf *(https://en.wikipedia.org/wiki/Tf%E2%80%93idf)* as calculated by TfidfVectorizer *(http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)* is probably a useful tool. The tf-idf results are used to pick out "important" words that appear often in a particular plot, but infrequently in the data set overall. The scikit-learn docs page Working With Text Data *(http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html)* would be a good first stop.

Word2Vec *(https://radimrehurek.com/gensim/models/word2vec.html)* may be another useful tool to look at.

You might also explore predicting the genre(s) a movie is in from the plot summary. Or you could look at the words that are more strongly associated with particular genres (“scream” with horror movies, or “hilarious” with comedies).

Updated Mon June 11 2018, 14:41 by ggbaker.