Silver Screen - Requirements, Vision, and Scope

Overview

Silver Screen will be a web application which allows its users to gather and visualise popular movie related sentiments based on Twitter tweets. Metrics which describe sentiment can be broken down into categories such as positivity, controversy, anger, variation, etc. Visualisations of analyzed tweets could also possibly be filtered by location, date submitted, or other parameters associated with the tweets. Data will be analysed using natural language processing, primarily in Python, and visualized through charts and a review score on the web front-end.

Motivation/Opportunity

We are building this product because Twitter contains a vast, untapped supply of information which can be useful to a variety of industries, one of which being the movie industry. By consolidating this information, we can turn a relatively disorganized social network into a tool for understanding public reactions. This information can be used by the movie industry to fill gaps in their understanding about the success of a movie and by individuals looking to find new movies to watch.

Though there has been a large amount of research into using machine learning to process text based information and determine the sentiment in it, there are currently no existing products that present the results of these analysis techniques in a manner that is easy to access and understand. Websites like IMDB and Rotten Tomatoes provide a way to view movie ratings generated through user contribution or professional reviews; however, our product will provide a larger and more accurate database of reviews as it takes into account the opinions of people who are not inclined to take the time to write reviews on other websites.

Problem Statement

The Problem Of	Inaccurate movie reviews and disorganization of information contained within Twitter
Affects	Persons interested in the performance or popular opinion of movies
The impact of which is	People might be unclear as to how their demographic feels about a movie
A successful solution would be	A software system that concisely and accurately portrays popular opinions on movies

Product Position Statement

For	Movie viewers, industry entities, and vendors (theatre-owners)
Who	Are looking for general public opinion on movies
Our System	Is a web service (all software)
That	Allows users to consolidate, quantify, and visualize popular opinions on movies through tweets found on Twitter
Unlike	Other websites like IMDB and Rotten Tomatoes which focus on reviews by professionals or others who are invested in the industry
Our Product	Will be user friendly, accurate, free from selection bias, and fully available online

User Demographics

Movie-goers: The primary audience of our site is standard movie-goers who wish to use more accurate, crowd-sourced reviews to find movies better suited to their specific demographic. As movies appeal to an extremely wide range of the population, our site must attempt to target users of all ages and technical skill levels. These generally non-technical users expect seamless implementations similar to the ones implemented by other popular review sites (such as IMDB or Rotten Tomatoes), which are presented in a visually appealing, easy to follow format.

Industry Professionals: The movie industry can use Silver Screen in order to gauge public reaction to movie announcements, trailers, and the movies themselves. With an accurate algorithm, this information will augment focus groups, providing them with a much larger sample reaction to the information that they release. These users are generally more technically savvy than standard movie-goers, and are expecting the presentation of robust data which relates to their query.

Theatre Operators: Theatre Operators can use Silver Screen to identify which movies are trending with their target demographic. They can then use this information to decide which movies to play in their future showings. These users are often somewhat-technically inclined, yet they still require a clean interface to ensure that the demographics filters they will use to judge movies are thorough and truly representative of different subsets of the population.

Feature List

- Web-based search interface
- Automatic analysis of the general impression of a movie on Twitter, including the generation of a review score
- A comprehensive front page which presents trending movies and box-office hits
- Visualisation of Twitter data in user-friendly graphs and charts
- Display of sentiment based on detailed demographics breakdowns
- Display of comparisons between how Twitter impressions compare to a movie's gross revenue and other sites' review scores
- Storage and display of historical records regarding to people's responses to a movie

Constraints

- Language: Given the scope of our project, and the limitations of the natural language toolkit we will rely on, it will only be possible to process tweets written in English. Of course, this may affect the accuracy or scope of our results.
- Accuracy of Language Toolkits: While the tools used in sentiment analysis allow for a correct determination of sentiment in a majority of cases, they are far from perfect. As such, the review scores we generate might not be completely accurate and will likely have to be "fuzzed".
- Available data: It is quite likely that people will not have tweeted about certain movies very much. Therefore, for certain search queries, the system may not receive enough information from Twitter to make accurate statements about people's views on that movie.

Scope and Limitations

- Due to the difficulties stemming from translation and parsing, analysis will be restricted to English tweets only
- Silver Screen will not allow for users to create accounts in order to avoid a large amount of login system-based overhead
- When analyzing movies which are part of major franchises, review scores might have to be assigned to the franchise rather than individual movies. One example would be the 'Harry Potter' franchise - tweets regarding a specific movie (such as the 'Chamber of Secrets') would likely get mixed up with tweets regarding different films in the same franchise due to the short nature of tweets

Assumptions and Dependencies

This project will rely heavily on Tweets and thus will make extensive use of the Twitter API. We will also be working under the assumption that there are a statistically significant number of meaningful Tweets about most movies the users of the application would be likely to search for. We also assume that we will be able to request and process enough tweets in order to keep sentiment scores for movies updates in a timely manner.

Given that we will be breaking down and analyzing Tweets using Python's Natural Language Toolkit, we will be relying on its abilities, and our own, to perform meaningful and reasonably accurate analysis of subjective information.

We have chosen to develop our application using Python's Django framework which allows us to focus on algorithm design and data extraction over development specifically for the web.

Use Cases

Use Case 1: User requests general sentiment information about a movie

Stakeholders and Interests	The user is the stakeholder here. He/she is looking for accurate popular perspective on the movie
Preconditions	The user has loaded the webpage and entered his/her query into a search box (it is well-formed), no login required
Postconditions	The user interface will display the sentiment data in a user-friendly manner.
Main Success Scenario	The user is shown basic information about the movie (release date, title etc.), an overall sentiment score, controversy rating, a word cloud of hashtags used in the tweets related to the movie, and tweets made by actors or influential (verified accounts) people related to the movie
Extensions and Alternative Flows	This data could be stored/cached for later use
Open Issues	What do we do if there are few/no Tweets about the movie?

Use Case 2: User requests historical information about a movie

Stakeholders and Interests	The user is the stakeholder. He/she wishes to view the performance and popular opinion of the movie over a specific period of time
Preconditions	The user has entered the site and entered his/her query (it is well-formed), no login required
Postconditions	The user interface will display the sentiment data in a user-friendly manner
Main Success Scenario	The user is presented clear information accumulated from an analysis of a statistically significant amount of data over the given time period. The user is shown basic information about the movie (release date, title etc.), an overall sentiment score, controversy rating, a word cloud of hashtags used in the tweets related to the movie, and tweets made by actors or influential (verified accounts) people related to the movie specific to the requested time period
Extensions and Alternative Flows	This data could be stored/cached for later use
Open Issues	What if there is not enough data about this movie for this time period?

Use Case 3: User requests listing of movies by a metric (such as revenue)

Stakeholders and Interests	The user is the stakeholder. He/she wishes to view the performance and popular opinion of the movie grouped by a certain metric
Preconditions	The user has loaded the webpage and entered his/her query into a search box (it is well-formed), no login required
Postconditions	The user interface will display the sentiment data in a user-friendly manner
Main Success Scenario	The user is presented with the list of the top 10 movies with the highest scores under that metric.
Extensions and Alternative Flows	This data could be stored/cached for later use
Open Issues	What should be done if there is not enough

	available data to group by the given metric for that movie?

Non-Functional Requirements

Performance Requirements

Due to the volatile nature of Twitter, review scores will need to be frequently updated (i.e. every 15 minutes) in order for the score to be an accurate indication of the movie's popularity. While the exact details of how long the analysis of an individual movie will take is unknown to us at this time, it is apparent that we need to take steps to avoid propagating processing delays to the users.

Safety Requirements

One possible safety issue is the use of people's tweets and their misinterpretation by our software. As we will only be using public tweets which are, as specified in Twitter's terms and agreements, public property and as our interpretation of individual tweets will not be visible by the user, we believe this issue should be fairly contained. The user may be able to view a short list of movie-related tweets that our software has taken in; however, their individual contribution to the ranking of a movie will not be visible.

Security Requirements

Our system will guard against users submitting many queries within a small window of time. We will also protect against SQL injection and other common attack methods.

Software Quality Attributes

We are not considering adding an API for our software so it will not be adaptable by others to perform their own analysis using our algorithm. However, if we wish we can easily expand the Twitter analysis to analysis of other sites after project completion and presentation.

Our software is readily available on a website and it can be accessed on a PC or capable mobile device. Information should be presented to users in an attractive and simple, yet informative manner.

We have a scale of correctness which correlates with the amount of feedback we obtain about a movie. The more opinions we get about a movie, the better we can form a correct representation of the relevant public opinion.

We have a few dependencies that will need to remain intact throughout the software lifecycle - mainly the Twitter API, Django and Python versions. In order to keep our software maintainable we will need to make sure that these versions are tracked.