**ThinkAnalytics Case Study**

**Predicting TV & Movies micro-genres for Recommender Systems**

ThinkAnalytics has been asked by one of their clients to design a recommender system for TV & Movie content. Consider a simple recommender system where a user indicates which micro-genres of TV & Movies they prefer.

Some examples could be “I want to watch…

an emotional comedy drama featuring a strong female lead”

a war movie about the Vietnam war”

a period drama set in the 1940s”

a TV series for children set in a magic world”

a movie set in New York directed by Woody Allen released in 1990”

A recommender system could use this information to recommend some movies and shows from a user’s preferred micro-genres. This obviously means the recommender system needs to know the micro-genre of each content item in its TV & Movie catalogue.

However, TV & Movie catalogue data is often of variable quality and we frequently find that some content items have a synopsis that contains little or no meaningful information to describe the content item – this is typically the case for less well known or specialist content.

**The client has asked ThinkAnalytics to design a system that predicts micro-genres for a piece of TV or Movie content if you have the metadata described in table 2.**

**Terminology & Assumptions**

Micro-genres describe a content item looking at different aspects of it such as the genre, sub-genre, mood or subject.

For this case study, we will focus on the following aspects of a content item:

|  |  |
| --- | --- |
| **Genre** | Fantasy, drama, mystery, etc. |
| **Sub-genre** | Sub-genres are categories of genres. Some examples could be comedy drama or splatter horror. |
| **Mood** | Moods define or imply the emotions that will be evoked in the watcher. For example, scary or thought provoking. |
| **Format** | Movie, TV series etc. |
| **Subject** | Describes the topic of the content item. War, love triangle or apocalypse could be subjects. |
| **Character** | Specifies fictional characters that appear in the content item. For instance, Superman, cowboy or zombie. |
| **Audience** | Refers to the target audience of the content item. The audience could be children, teenagers, adults, etc. |
| **Time period** | Time period is used to describe the time period the content item is set in and it is not related to the release year. |
| **Setting** | Refers to where the action takes place. For example, Paris or imaginary world. |
| **Director** | Director of the content item. |
| **Actors** | Actors that perform in the content item. |
| **Release Year** | E.g. 2016 |

*Table 1. Aspects of the content item to take into consideration*

We will assume the metadata uses a limited collection of genres. For example, IMBD uses 22 main genres as defined here: <http://www.imdb.com/genre/>

A recommender system usually gets information for each content item from a source of TV & Movie metadata. Typically, the metadata contains the following fields for each item:

|  |  |
| --- | --- |
| **Field** | **Description/Example** |
| Title | The main title of the content item |
| Synopsis | A short description about the content item. For TV series, this could be a description about an episode or a general description about the whole series |
| Age Rating | An age or a code indicating the minimum age for the audience |
| Genre | A term that describes the main theme(s) of the content item as described above |
| Credits | Names and roles for actors, presenter, directors etc. |
| Format | Movie, TV series etc. |
| Release Year | E.g. 2016 |

*Table 2. Metadata example.*

There are often many more items in the metadata, such as audio or video format, but for this case study the fields above will be more than sufficient.

Some fields like title and synopsis may contain multiple language variants. However, for this case study let’s assume the titles and synopsis are provided in English only.

You may assume that you have a sample set of TV & Movie metadata which is fully populated with the fields listed above. Hint: you can think of this as a training data set.

**Example: Kill Bill Vol 1**

The following gives three example synopsis for the movie Kill Bill Vol. 1 just to give you a flavour of the quality of metadata we often find. Kill Bill Vol 1 can be classified as a violent action thriller starring Uma Thurman about revenge with a character assassin.

|  |  |  |
| --- | --- | --- |
|  | **Synopsis** | **Comment/Hints** |
| 1 | After awakening from a four-year coma, a former **assassin** wreaks **vengeance** on the team of assassins who betrayed her (IMDB). | From this description we can understand that this movie is about revenge and has a character that is an assassin. |
| 2 | An entire wedding party is **slaughtered** during a dress rehearsal in a rural chapel: the pregnant woman in the **blood-splattered wedding** dress is Black Mamba, better known as The Bride. The **assassin**, Bill, and his circle known as The Vipers left The Bride for dead, but unluckily for them she was merely comatose. Four years later, The Bride suddenly awakens from her coma and realizes what has been done to her. She sets off on a ferociously focused mission, setting out to seek **revenge** on her former master and his **deadly squad of assassins**. One by one, she kills the various members of the assassin group. She saves Bill for last. (Rotten Tomatoes). | This description provides more information, we certainly know that this is a violent movie since they mention blood-splattered wedding. What about the genres? If you had a thesaurus could you map genres to possible synonyms? |
| 3 | The return of renowned filmmaker Quentin Tarantino after an absence of 6 years. **Uma Thurman** stars as The Bride in this movie **full of action**, one-fifth of a team of assassins called DiVAS. When she decides to get married and leave the outfit her boss, Bill (David Carradine), is not a happy man and gets her former colleagues to show up at the wedding; leaving behind a **blood bath** and The Bride shot in the head. Four years later, The Bride sets out on a journey of **revenge** which is ultimately to kill her former boss, Bill. The film was due to be three hours long but was split into two films (hence the name) and Volume 2 concludes the story. | Could you compare the key terms in this synopsis with the key terms on other synopsis in the training set to work out the genre? Are there data mining algorithms and techniques that can help with this? Can other information in the data help us to determine any of the aspects of the micro-genres? |

Your group is asked to produce a short report which is a maximum of 4 pages explaining the design of the recommender system. This report will have two purposes:

1. Explain your design of the recommender system to ThinkAnalytics
2. Be written in such a way that it is possible to cut and paste any of the material directly into a client report being produced by ThinkAnalytics.