Natural Language Understanding, Generation and Machine Translation - Week 9 - Movie Summarisation

Antonio León Villares

April 2023

Contents

Based on:

- *Movie Summarisation via Sparse Graph Construction*

1 Movie Theory

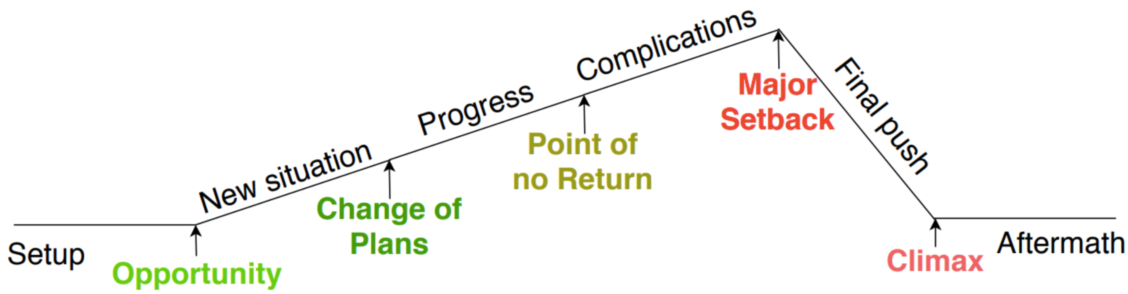
1.1 Movie Summarisation

- **Why is movie summarisation important (or at least useful)?**
  - **movies** and **TV shows** have become **omnipresent** in modern society:
    * $6,000 - 7,000$ movies/series/shows in Netflix
    * $24,000$ movies and $2,100$ shows in Amazon
    * we watch $\approx 3.5h$ of TV **per day**
  - however, the process of movie selection can be difficult:
    * wide availability of streaming services
    * 70% of people struggle to decide on what to watch
    * need ≈ 15 minutes to decide what to watch
  - if we could find ways of **summarising** these movies/shows in an **adaptable** way, this could be quite helpful:
    * generate an alternative trailer
    * generate a 10 minute movie summary
    * show the 3 most action-packed/funny shots
- **Why is movie summarisation difficult?**
  1. **Technical Complexity**: movies incorporate a lot of information (images, speech, sound), in a long format ($> 1.5h$). Moreover, movies aren't **linear**: they represent **interconnected** stories (i.e flashbacks) which need to be understood together.
  2. **Lack of Data**: unlike with **news summarisation**, we can't rely on the first few **scenes** to generate meaningful summaries. In general, no such **simple heuristics** are available. Moreover, whilst there are **gold-standard summaries** available for **text** (i.e Wikipedia synopyses), no such information exists for **video** (trailers don't count as they don't summarise movies, they just try to entice viewership)

1.2 Turning Points

- **What are turning points?**
  - **key narrative moments** in movies
  - these often guide and signal plot changes
  - they help **segment** the narrative into **thematic units**
- **Why are turning points important?**
  - identifying **turning points** can lead to finding the **critical moments** in movies
  - thus, **turning points** can be strong candidates for scenes which should be included in movie summaries
- **Is there a standardised way of measuring or understanding turning points?**
  - generally, the **number** or **meaningfulness** of **turning points** will depend on the movie

– however, over the years, we've been able to identify **common patterns** with regards to turning points
– many movies (here referring to Holywood movies), particularly of certain genres, adhere fairly well to this framework



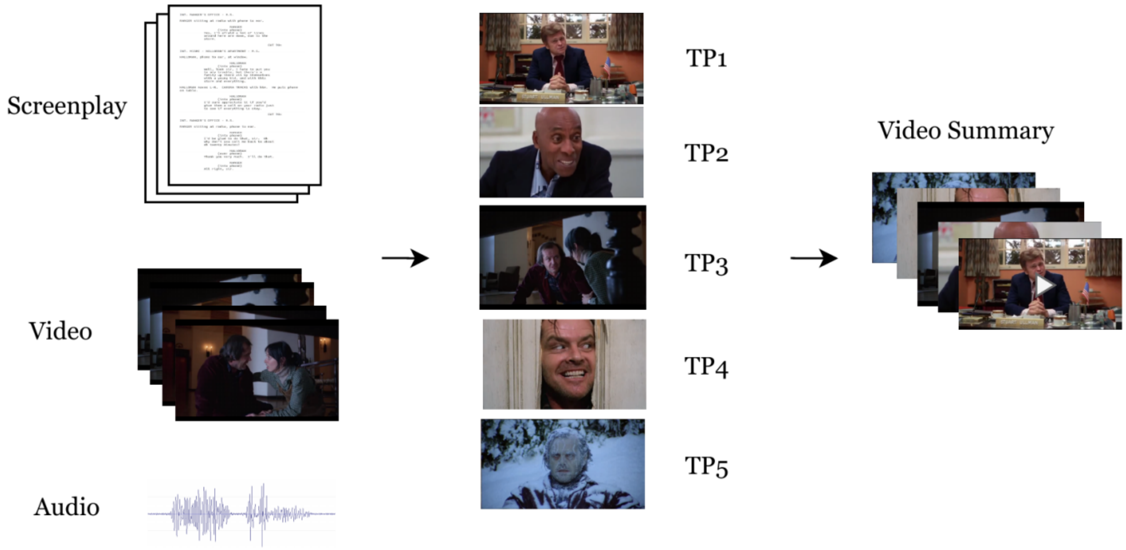
  * **Opportunity**: introductory event. Occurs after setting presentation and background of main characters (i.e the protagonist meets the love interest)
  * **Change of Plans**: main goal of story is defined, and action begins increasing (i.e the love interest gets engaged)
  * **Point of No Return**: event pushes main character(s) to fully commit to their goal (i.e protagonist and love interest decide to continue together)
  * **Major Setback**: everything falls apart (temporarily or permanently) (i.e love interest is pregnant and must get married)
  * **Climax**: final event of main story, moment of resolution, "the" spoiler (i.e everything gets resolved, and protagonist and love interest get to be together)

2 Multimodal Movie Summarisation

We discuss the approach to summarisation presented by Papalampidi, Keller and Lapata in Movie Summarisation via Sparse Graph Construction.

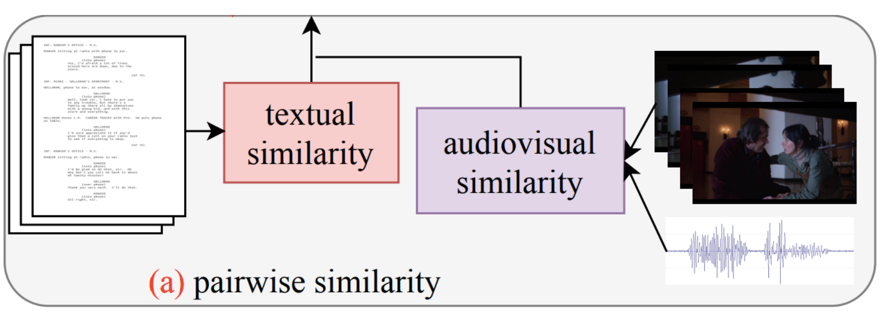
2.1 Problem Formulation

- **What approaches have been tried before?**
  – in Movie Script Summarization as Graph-Based Scene Extraction and Screenplay Summarization Using Latent Narrative Structure they seek to summarise movies solely based on the **screenplay**
  – there have been attempts at **video** summarisation, but only for **isolated video clips**
- **How does this approach differ from previous ones?**
  1. **Multi-Modal**: summaries are based on both **screenplay** and **video** (audio + images)
  2. **Length**: the summaries focus **on the whole movie**, not isolated clips. This is particularly challenging, since movies nowadays are getting longer and more convoluted.
  3. **TP Identification**: to generate the summaries, the task is **reformulated** in terms of identifying **turning points**, since these signal key narrative moments of the story



2.2 Graph-Based Turning Point Identification

- **How are movies represented in this model?**
  - use **sparse graphs** to represent the different **scenes**
  - **edges** showcase the **similarity** between scenes
- **Why are graphs used for this?**
  - **Contextualisation**: as discussed, movie development is **non-linear**: many different scenes might be **interconnected**, despite occurring at different moments (i.e substoreies, flashbacks). **Graphs** flexibly represent these **complex interactions**.
  - **Navigation**: **graphs** are more **interpretable**, in terms of understanding both how scenes are **interrelated**, and to understand how movies are **structured**.
- **How is the graph generated?**
  - different **pre-trained** models are used to generate screenplay, audio and visual **representations** for a given scene
  - a **similarity score** between scenes is computed, by considering both **textual** and **audiovisual** similarity

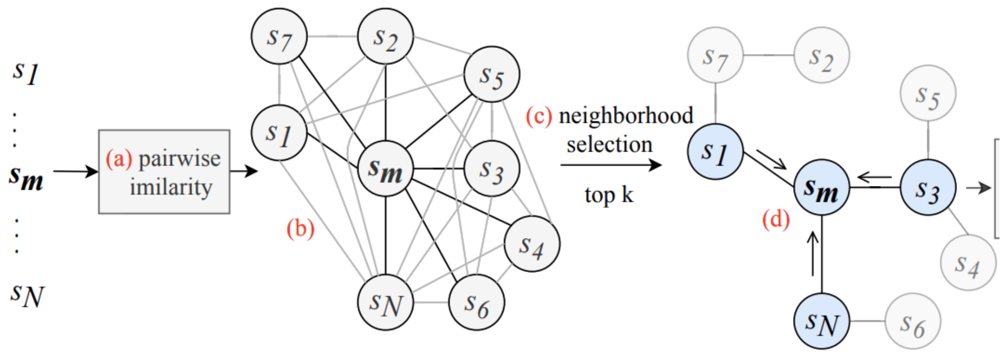


  - explicitly, the similarity between 2 scenes s_i, s_j is:

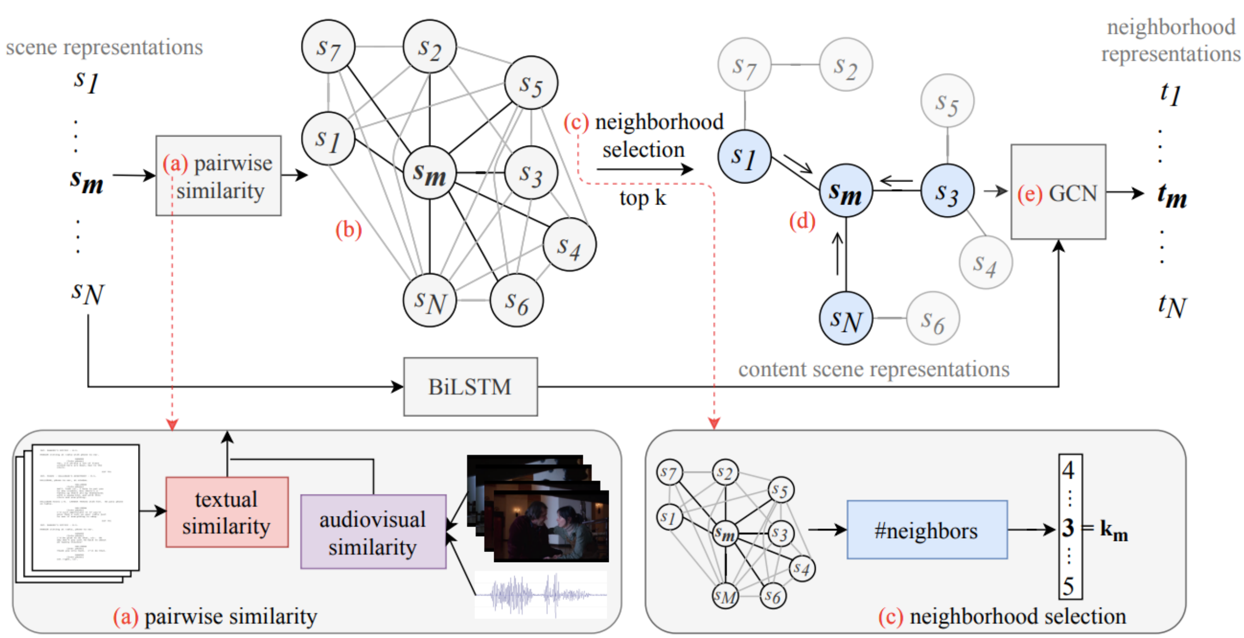
$$e_{ij} = u_{ij} \left(\tanh(W_i \underline{v}_i + \underline{b}_i)^T \tanh(W_j \underline{v}_j + \underline{b}_j) \right) + b_{ij}$$

  where u_{ij} denotes **audiovisual** similarity, and $\underline{v}_i, \underline{v}_j$ are the **textual** representations for the scenes.

- the resulting **graph** is then made **sparse**, by, for each node, only preserving the k most similar edges (for this they used $k = 6$)



- **How are textual and audiovisual scenes aligned?**
  - an **alignment algorithm** is used which seeks to correlate **subtitles** with **screenplay dialogue**
- **How are scenes represented by the model?**
  - scene representation occurs in 2 steps:
    1. **Contextualised Representations**: the **screenplay** representations are passed through a **bidirectional LSTM**, with the **forward** and **backward** hidden representations being **concatenated** to generate a **contextualised scene representation** $\underline{c}_i$
    2. **Neighbourhood Representation**: the **graph** (represented as an ajdacency matrix) alongside the **contextualised representations** are passed through a **Graph Convolutional Network**, which generates a **neighbourhood representation** $\underline{t}_i$ for a scene, which encompasses information about the **immediate** neighbours of s_i, alongside the **contextual representation** (which should contain temporal information not present within the graph representation)
  - then, the final scene representation $\underline{s}_i$ is obtained by concatenating $\underline{c}_i$ and $\underline{t}_i$



- **How can we determine if a scene is a training point?**
  - we train 5 different classifiers (one for each **turning point**)
  - based on the **scene** representations, these all get passed through the classifiers, which determines whether the scene is one of the **turning points**
  - to **train** the **classifiers**, we use **TRIPOD**: a **hand-labelled** dataset containing:
    * 122 movies
    * 17,150 scenes
    * 1,600 **training point** scenes
  - we consider up to 3 scenes per **training point**, which generates summaries of 10-15 minutes

2.3 Results

- **How can the summaries be evaluated, based on turning points?**
  - we can count the number of **training points** found
  - this can be done in 3 ways:
    1. **Total Agreement (TA)**: percentage of correctly identified TP scenes
    2. **Partial Agreement (PA)**: percentage of TP scenes for which **at least** one **gold-standard** scene is identified
    3. **Distance (D)**: minimum distance (in number of scenes) between **predicted** and **gold-standard** set of scenes for a given TP, normalised by **screenplay length**

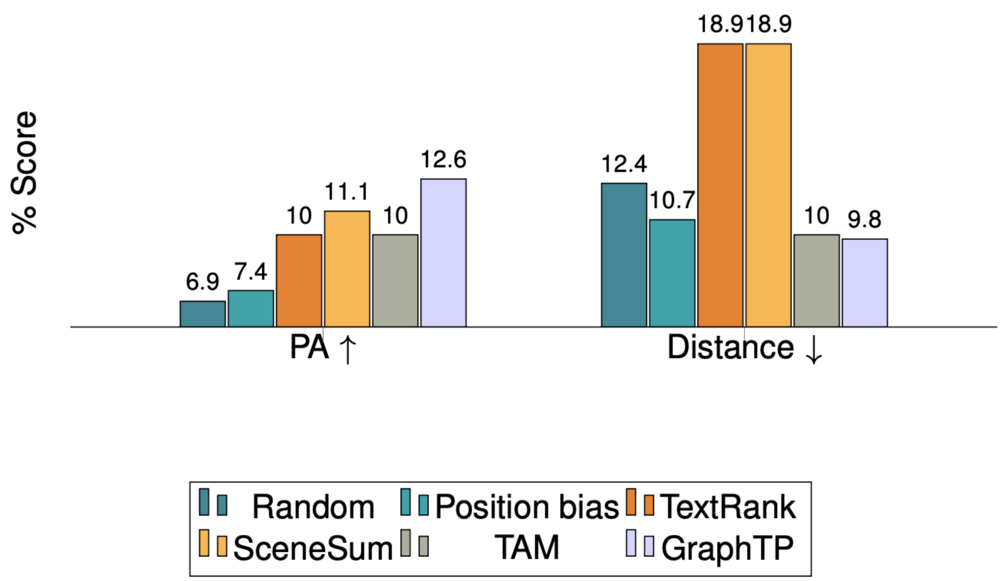


Figure 1: **Random** picks out 5 random scenes from the whole screenplay. **Position bias** selects the first 5 scenes as the summary. **TextRank** is a grpah based summarisation model, whilst **SceneSum** is a previous movie summarisation technique. **TAM** is this approach without the graph, whereas **GraphTP** is this approach with a graph. By far, the best model (in terms of PA and D) is this technique when using a graph.

- **What is one particular difficulty of assessing movie summarisation?**
  - movies are **extremely long**
  - multiple hours would be required to **watch the movie**, and then assess how well a bunch of summaries did at capturing the key points of a movie

– there is also the issue that **different people** will find **different scenes** as most important

- **How can we evaluate the summaries, with regards to capturing the key moments of the movie?**
  – **annotators** can be used to identify **key moments** in the movie
  – these can be compared with the **turning points** found by the model

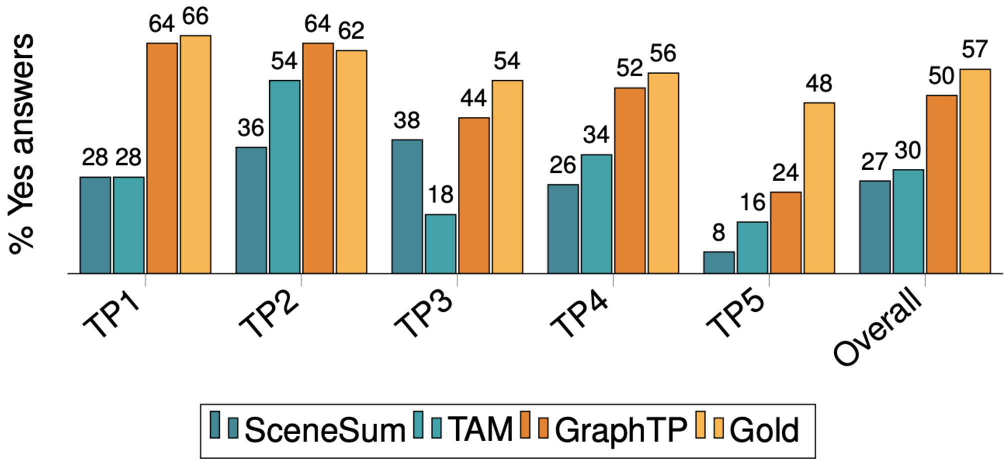


Figure 2: Once again, **GraphTP** seems to perform best for human annotators, approximating or even surpassing the gold labels. Notice how for TP5 (the climax), performance is worst. Nonetheless, this seems to show that the generated summaries generally contain the key information.

- **Can we extract any useful information from the generated graphs?**
  – movies were grouped into **four broad genre catagories**
  – only a subset of the graph was considered, containing the **turning point scenes**, alongside the immediate neighbours
  – **node connectivity** was computed for the resulting graphs, at the **turning point** level (a measure of resilience of the graph to being "split" into subgraphs)

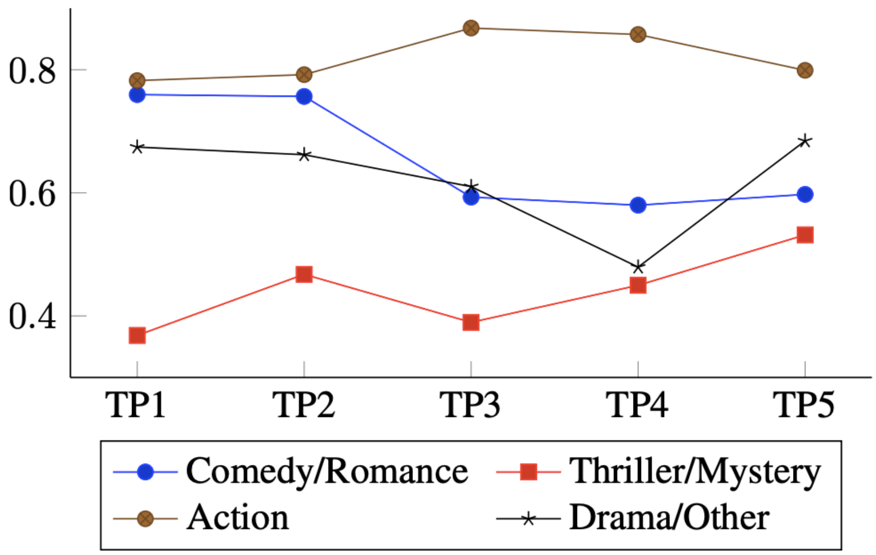


Figure 3: Generally, **actions/comedies** have **high connectivity**, indicating higher interconnectedness between nodes, and thus, more **coherent** stories. On the other hand, **thrillers/dramas** show **low connectivity**, indicating lower interconnectedness between nodes, and thus, more **complex** stories.

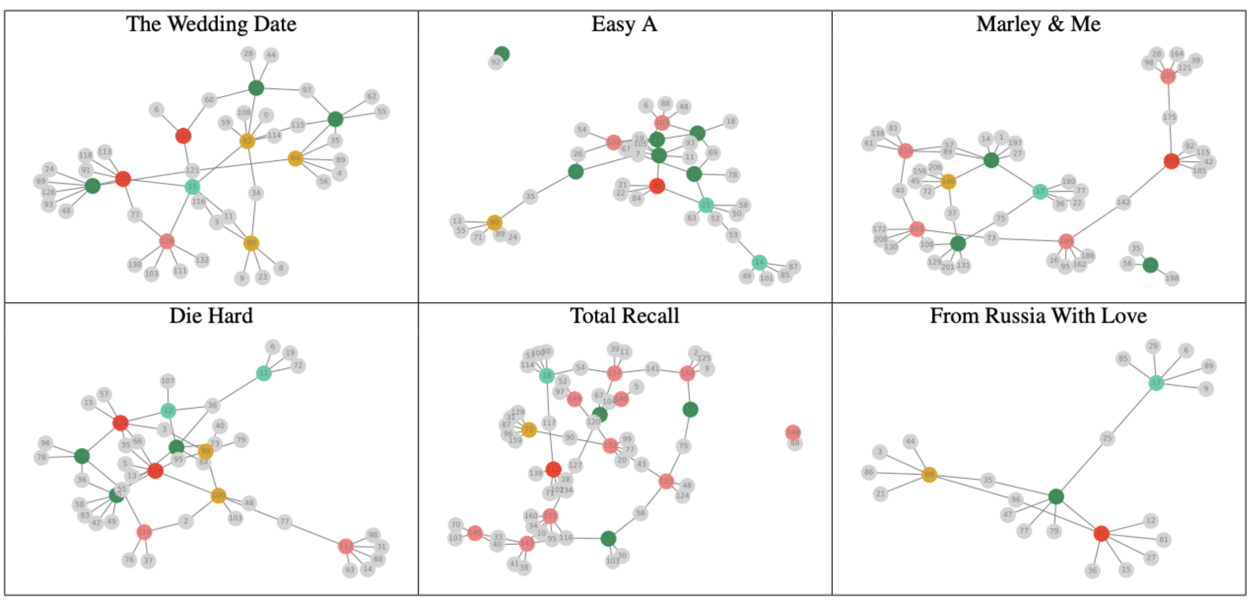


Figure 4: Visualisation of the graph for comedy/romance (first row) and action (second row). There are connections between all turning point scenes (generally), and the graphs seem more cohesive.

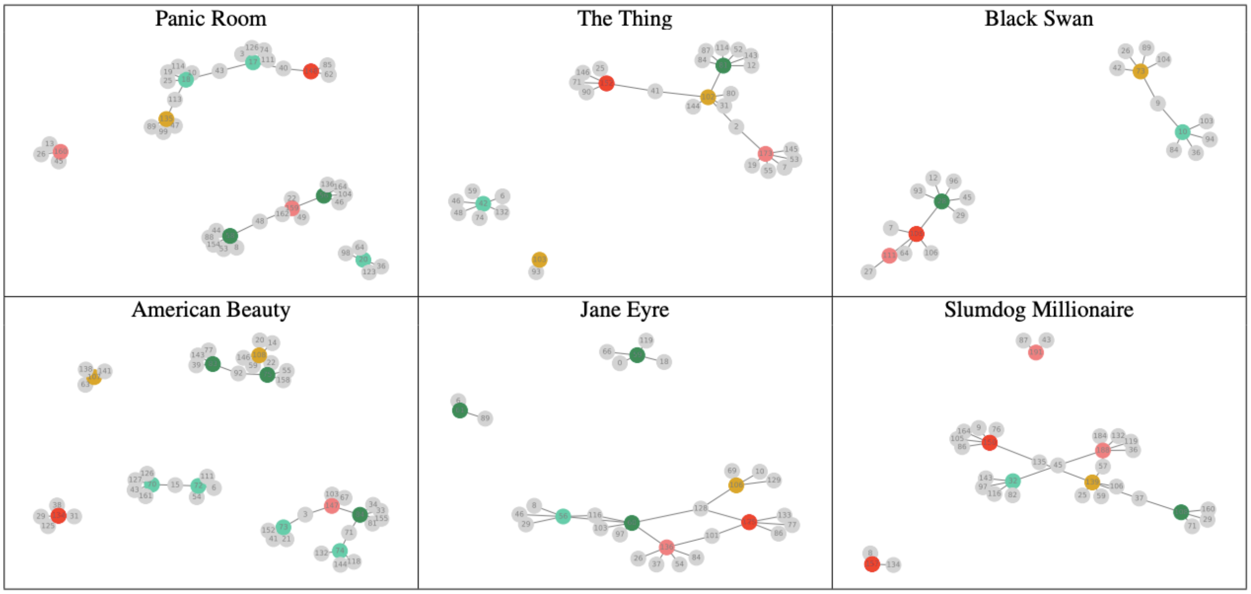


Figure 5: Visualisation of the graph for thrillers/mysteries (first row) and drama/other (second row). The connections are a lot more sparse, with many isolate subgraphs.