SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries

Amy Pavel[1], **Dan B Goldman**[2], **Björn Hartmann**[1], **Maneesh Agrawala**[3]
[1]University of California, Berkeley {amypavel,bjoern}@cs.berkeley.edu
[2]Adobe Research {dgoldman}@adobe.com
[3]Stanford University maneesh@cs.stanford.edu

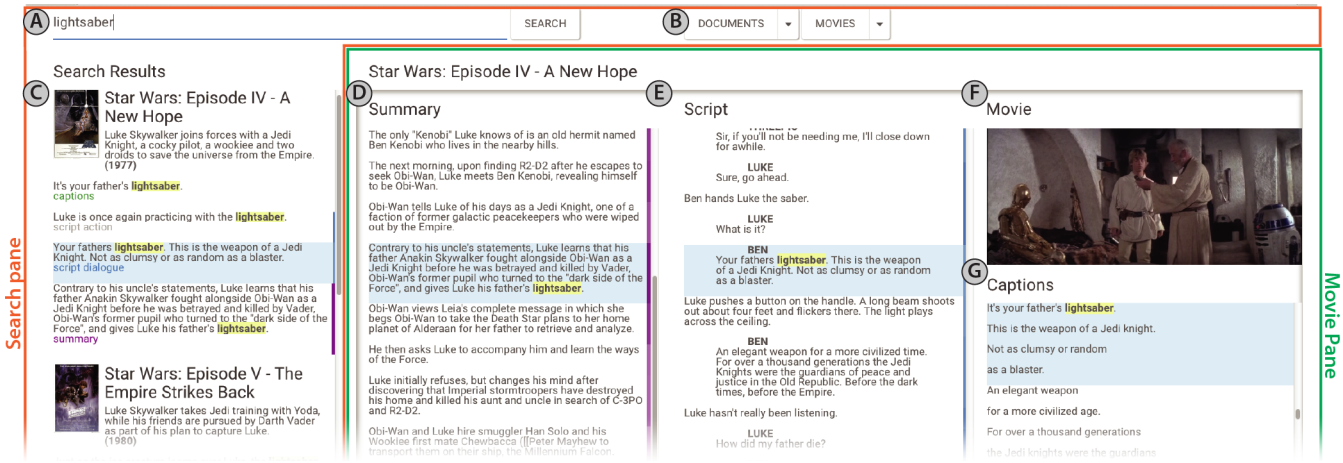

Figure 1. The SceneSkim interface consists of a *search pane* for finding clips matching a query and a *movie pane* for browsing within movies using synchronized documents. The search pane features a keyword search bar (A), search filters (B) and a search results view (C). The movie pane includes the synchronized summary (D), script (E), captions (G), and movie (F).

ABSTRACT

Searching for scenes in movies is a time-consuming but crucial task for film studies scholars, film professionals, and new media artists. In pilot interviews we have found that such users search for a wide variety of clips—e.g., actions, props, dialogue phrases, character performances, locations—and they return to particular scenes they have seen in the past. Today, these users find relevant clips by watching the entire movie, scrubbing the video timeline, or navigating via DVD chapter menus. Increasingly, users can also index films through transcripts—however, dialogue often lacks visual context, character names, and high level event descriptions. We introduce SceneSkim, a tool for searching and browsing movies using synchronized captions, scripts and plot summaries. Our interface integrates information from such sources to allow expressive search at several levels of granularity: Captions provide access to accurate dialogue, scripts describe shot-by-shot actions and settings, and plot summaries contain high-level event descriptions. We propose new algorithms for finding word-level caption to script alignments, parsing text scripts, and aligning plot summaries to scripts. Film studies graduate students evaluating SceneSkim expressed enthusiasm about the usability of the proposed system for their research and teaching.

UIST '15, November 08-11, 2015, Charlotte, NC, USA
ACM 978-1-4503-3779-3/15/11.
http://dx.doi.org/10.1145/2807442.2807502

Author Keywords

video; search user interfaces; script; summary

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Searching for clips in film is a crucial task for film studies researchers, film professionals and media editors who analyze, share and remix video clips. For example, film studies researchers may search for clips containing a particular action, prop, character, or line of dialogue, in order to discover patterns in films. Film professionals find existing examples of settings, props, character performances, and action sequences in order to inspire new projects and communicate desired visual attributes (e.g. animation style, lighting, sets). Movie fans remix existing Hollywood movies and TV shows into "supercuts": montages of repeated elements from existing films such as words, phrases, or clichés.

However, searching and browsing within movies is a time-consuming task. Consider a user who wishes to analyze the context and appearance of lightsabers, in the *Star Wars* movie series. If she knows the films already, she might try to navigate to remembered scenes using DVD chapter menus, or

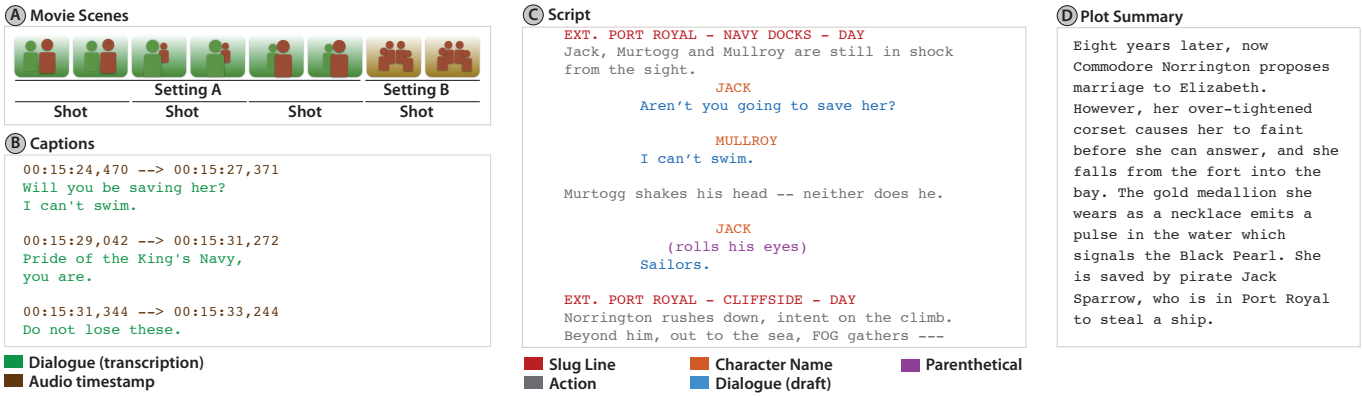

Figure 2. Different documents describe different aspects of a movie: The script (C) contains locations, character names, actions, parentheticals and draft dialogue of the movie (A), in which scenes comprise multiple shots. The captions (B) contain timestamped dialogue as spoken, while the summary (D) contains a high-level description of the plot.

by scrubbing through video timelines. Otherwise, she might have to simply watch all of the movies, taking notes whenever a lightsaber appears.

In practice, the practitioners we interviewed primarily search for clips using those same approaches. Some video viewing interfaces allow users to search and browse videos by text transcripts [3, 6, 15, 25] or caption files [10, 24]. Yet, these documents do not always contain contextual information (e.g. locations, props, actions, or names of characters speaking the dialogue) that may be relevant to these searches. In addition, it can be difficult to browse for a specific scene using transcripts because they lack visual context.

However, it is often easy to find related documents containing these other textual features: Movie scripts identify speakers and contain a written description of the dialogue, actions and settings for each scene. Online movie databases include user generated plot summaries that describe important moments, high level events (e.g. escapes), and character traits (e.g. a hostile character). We call these related datasets *parallel documents*. However, although they contain a wealth of contextual information about the same underlying film, the data in parallel documents are generally not explicitly linked to each other. Nor are they directly linked to the corresponding clips in movies, even though the visual aspects of a scene can be critical to both creation and analysis of films.

The goal of our work is to help users search and browse movies more efficiently by jointly exploiting all of these data sources. To this end, we have developed SceneSkim, a system that automatically aligns the captions, script and plot summary to the audio-visual content of a film. SceneSkim also aligns each parallel document to one another, enabling synchronized browsing of the captions, script and summary for easy access to details or context at any point in the film. SceneSkim leverages the structure of the related documents, allowing users to search over specific document features, such as lines by a particular character, or scene locations. Our implementation introduces new techniques for aligning the summary to the script and the words in the captions to the audio-visual content. In particular, we use unique terms to inform an ordered alignment between the summary and script, and we use caption timestamps to aid forced alignment between caption words and movie audio.

Returning to our lightsaber example, Figure 1 shows how SceneSkim lets our user easily discover the complete context of lightsabers in the *Star Wars* films. First, the user types "lightsaber" into the search bar (A). The search results (C) reveal caption, script and summary sentences describing the appearances of lightsabers in chronological order within each film. The first entries reveal the main character's earliest encounter with a lightsaber, which they views in the video panel (F) by clicking on one such sentence. From there, the user can either browse through that scene by scrolling and clicking in the script (E), or navigate to dialogue about lightsabers with the captions (G), or continue to scroll in the search results (C) to other appearances of lightsabers.

We used the SceneSkim interface to answer queries (e.g. for character appearances, locations, dialogue, and actions) described in existing film studies research [2, 9, 13, 41]. For instance, we were able to manually estimate total screen time of lightsabers across three *Star Wars* movies in about 20 minutes. In an informal evaluation with three film studies researchers, the researchers were able to formulate and answer novel questions using our system. The users reported they would use such a tool for day-to-day tasks in research and teaching, and as an exploration tool. Users positively commented on our system's usability and suggested new features.

DEFINITIONS

To allow users to search and browse movie using document text, we align movies to three parallel documents—scripts, captions, and summaries. Each parallel document provides a unique type of information, enriching searching and browsing capabilities. In this section we define background terminology related to these documents.

A movie, once shot and edited, is comprised of a sequence of continuous videos (i.e. *shots*) that span several locations (i.e. *settings*), and a synchronized audio track (Figure 2A). The audio track contains dialogue, music and sound effects.

Captions transcribe the dialogue, sound effects, and relevant musical cues in the movie (Figure 2B). Editors break the

dialogue transcript into short caption phrases displayed on screen with the corresponding dialogue. A caption file contains *caption phrases* and *timestamps* that link each caption phrase to the time the phrase is spoken in the movie.

A *script*, or screenplay, is a written description of the film that includes the dialogue, actions, and settings for each scene (Figure 2C). At the beginning of each *scene*, screenwriters typically provide a *slug line* which specifies the setting, whether the scene will be filmed outside (EXT.) or inside (INT.), and the time of day (e.g. day, night, morning). *Action* lines describe character actions, camera movement, appearance, and other details. Often, writers use all caps to introduce an object or character for the first time, whereas the rest of the action line is described using regular capitalization. The *dialogue* proposes what each character will say. However, the planned dialogue in the script and the transcribed dialogue in the captions do not match exactly: Editors may remove lines or scenes from the final film, and actors may improvise dialogue. Thus, we refer to script dialogue lines as *draft dialogue.* The *character name* specifies which character will speak the draft dialogue. Script writers also include *parentheticals* to clarify how the character will deliver the dialogue. In general we use the term *script line* to describe a contiguous block of text in the script that all shares the same label, bounded by text with different labels.

Plot summaries give a high level overview of main events, character traits and locations in the movie (Figure 2D). Such plot summaries are typically written by viewers or critics. Plot summaries contain higher-level event descriptions than contained in the script – which describes scene by scene actions – or captions – which directly transcribe the movie dialogue. Unlike the captions and script, summaries leave out events that are not central to main plot lines.

RELATED WORK

We build on two main areas of prior work: video search and navigation tools, and algorithms for aligning videos with their related text documents.

Video Browsing, Search and Navigation

Several systems aim to facilitate video navigation and search by constructing visualizations of video frames, allowing users to click in that visualization to jump to the relevant point in a video [17, 5, 20] or enabling users to see more frames during scrubbing [28, 27]. These techniques all rely on the viewer to recognize visual features of the desired location in a film. In contrast, our method enables users to browse using text, which also surfaces information conveyed in the audio, scene descriptions, character names, and plot information.

A number of commercial video players (e.g., TED) feature a synchronized transcript. Berthouzoz *et al.* [6] and Informedia [10] align a text transcript (or concatenated captions) to the video so that clicking on a word in the transcript navigates to the corresponding point in the video, while scrubbing the video highlights the corresponding part of the transcript. Unlike our work, these methods do not align the original script or a plot summary to the film and therefore do not enable users to search over attributes such as character dialogue, action descriptions, setting names, important events or high level event descriptions.

Many prior systems use domain-specific metadata to browse video – e.g., user interface events for software tutorials [26, 18, 11, 36], sports statistics for athletic events [34, 29], or computer vision and crowdsourcing approaches for informational lecture videos [19, 31, 19, 35, 21], how-to videos [22] and movies [39, 38, 30]. DIVA [25] and Chronoviz [15] align video with time-coded metadata (e.g. user annotations, notes, subtitles) to support exploratory data analysis of raw multimedia streams (e.g., for user studies). We develop new methods for browsing movie clips by aligning existing parallel text documents.

MovieBrowser [30], VideoGrep [24], and the system by Ronfard *et al.* [39, 38] address the problem of browsing for specific clips in movies. MovieBrowser uses computer vision and audio processing to predict action sequences, montages, and dialogue and displays clip classifications on the movie timeline. Although this work allows users to browse clips by clip type, it does not allow browsing by other aspects of clip content (e.g. dialogue, plot). VideoGrep is a command line tool allows users to search videos using timestamped captions in order to create supercuts. But captions contain about 10 words on average, whereas our system facilitates individual word-level indexing, and searching over plot and visual descriptions.

The system by Ronfard *et al.* [39, 38] is most similar to our work. It offers browsing and search capabilities for *The Wizard of Oz* using an exact shot-by-shot script synchronized to corresponding shots in the DVD. Shot-by-shot scripts describe each shot and transition in detail, but such shot-by-shot scripts are rare (we were only able to find one other shot-by-shot script besides *The Wizard of Oz*). In addition, most films deviate from their source screenplays. We create a system that processes approximate scripts in a common format, and surfaces possible misalignments using confidence markers. Our system also allows users to search across multiple movies simultaneously, and uses the captions and summary along with the script to aid searching and browsing tasks.

Algorithms for aligning movie metadata

SceneSkim facilitates searching and browsing over aligned captions, scripts, and summaries. Informedia [10] finds a word-level alignment between captions and broadcast news video by concatenating the captions and aligning the resulting transcript captions to a speech-to-text transcript of the film (similar to Berthouzoz *et al.* [6]). However, speech-to-text methods expect clean speech audio as input, and thus may fail when the audio track includes non-speech sounds characteristic of movies, such as background music and sound effects. Instead we use the alignment method of Rubin *et al.* [40] to align caption phonemes to audio features directly, leveraging caption timestamps to remove extraneous non-speech noise.

Prior work also aligns scripts to movies and TV shows in order to identify characters [14, 37] and actions [8, 23, 12] in the video. We build on this work, using the edit distance

technique first described by Everingham *et al.* [14] for aligning script dialogue to caption dialogue. Tapaswi *et al.* [44] and Zhu *et al.* [47] align movies to their original books, but we focus on aligning movie scripts to plot summaries. Tapaswi *et al.* [43] align plot synopses to the caption dialogue and identify character faces in TV episodes. We align plot synopses to scripts which contain information about locations, objects, and actions not mentioned in the caption dialogue. The prior work in this area develops algorithms primarily for machine learning purposes, and does not build user interfaces for search and browsing that take advantage of such alignments.

CURRENT PRACTICE

To learn about current practices for searching and browsing in movies, we interviewed two individual film studies scholars and a film studies research group, as well as two visual effects professionals. We also analyzed a corpus of 101 supercuts [4].

Film studies researchers typically search to analyze sets of clips and find patterns. In particular, they search for specific actions, props, locations and characters in order to study audio and visual attributes of the corresponding clips. These researchers often return to clips they have seen before for further review. They identify text results of interest through Web search and then both watch and scrub through films to locate clips of interest. One researcher noted that it is easy to miss short events while scrubbing. The research group mentioned they would like to search dialogue by keyword (e.g. to analyze scenes where particular slang terms occurred), or by performing characters (e.g. to study prominence and time given to a character).

Because film studies researchers frequently study visual attributes of scenes, text-based search over scripts alone is insufficient, but accessing film scenes through text search could be very helpful. For instance, one researcher wanted to study the body language of actors performing a common action. Another researcher wanted to study if communication technologies (e.g., cell phones) appeared as focal or peripheral in different films.

Film professionals search for locations and props in order to design new sets or create concept art. Such professionals also return to particular parts of films they have seen before. In addition, they search for character movements and dialogue to cast actors, create new characters, and inform the design of animations. Our interviewees reported that they search for clips on the Web, hoping that someone had uploaded a scene to YouTube or other video sharing sites. As a fallback, they search through their personal collections using DVD chapter menus. After locating clips of interest, they save the clip to review or send the clip to team members.

We also categorized **supercuts** published on supercut.org to infer the practices of media editors and remixers. We found that out of the first 101 supercuts (in alphabetical order by movie title), 60 were based on occurrences of some word, phrase or vocal noise. 28 were based on occurrences of some action, 5 were based on dialogue by a particular speaking character, and 4 were based on appearances of a given character.

To summarize, film studies researchers, visual effects professionals, and media editors all frequently search within films for clips matching a particular query, or browse to return to a particular scene in a movie. Specifically, users search for clips matching the following types of queries:

- Performances by a specific character or sets of characters (e.g. to closely study performance by character, or to watch main events containing the character)
- Locations (e.g. city skyline, living room)
- Actions (e.g. playing video games, car chase)
- Objects (e.g. cell phones, laptops)
- Words or phrases in dialogue (e.g. slang terms)

SCENESKIM INTERFACE

Motivated by these tasks, we developed the SceneSkim interface to support searching for clips that match a query and browsing for specific clips within a movie. We acquired complete caption, script and plot summary sets for 816 movies from existing corpora [45, 33] and by scraping the web (for more detail on this set, see Appendix A). We also purchased and loaded seven movies (*Star Wars* episodes 4-6, *Chinatown*, *Taxi Driver*, *The Descendants*, and *Pirates of the Caribean: Curse of the Black Pearl*) to demonstrate interactions. The interface supports searching and browsing this data set through two main components (Figure 1): the *search pane* and the *movie pane*.

Search pane

The search pane enables faceted search of movie clips through the keyword search bar, search filters, and search result view.

Keyword search: The keyword search bar allows users to specify keywords and concatenate them with the boolean expressions AND and OR. (Figure 1A).

Search filters: The search filters in the "Documents" dropdown menu (Figure 1B) allow users to restrict keyword searches to specific entity types. The user may select one or more entity types to search over (Figure 3).

Search results: Within the search result pane (Figure 1C), we first sort results by movie (Figure 4). Each movie heading contains a movie poster, title, short plot summary, and release

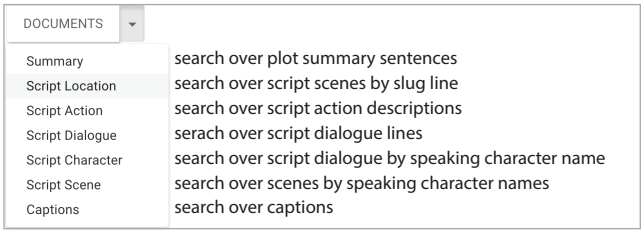

Figure 3. Users can select one or more entity types to search over using the "Documents" dropdown menu, from Figure 1b.

Figure 4. Search results for searching "falcon" across all entity types. Caption results show caption phrases, indexing into the movie at the word level, script dialogue results contain full dialogue lines, script location results show scene headings, and summary results show high level events. The last result is a script action result. As the Falcon is not a character name, we do not see script character results, or script scene results. The search results are also shown in (Figure 1C)

year. We then sort results by entity type (e.g. caption, script dialogue, script location etc.) and then in chronological order.

Each result shows a text snippet with the search terms highlighted, the result type label, and a confidence bar. The color of the confidence bar shows the likelihood that clicking on a script or summary result will jump to the correct part of the film. Darker colors represent better alignments. When the user clicks a search result, the movie pane scrolls all documents to the corresponding sections and plays the movie from that point onwards.

Movie pane

The movie pane allows users to browse within a movie using synchronized documents (Figure 1). From left to right, the movie pane displays the summary, script, and movie with captions below. We display the summary broken into summary sentences, the script broken into lines, and the captions broken into timestamped phrases. We align each document to all others, and by clicking on any document the user can view the corresponding sections in other documents. For example, clicking on a summary sentence will cause the captions and script to scroll and highlight the corresponding script lines and caption phrases. The movie will also play starting at the estimated start time for the summary sentence. Similarly, the documents scroll in synchrony and the movie plays when a user clicks on a script line or caption word.

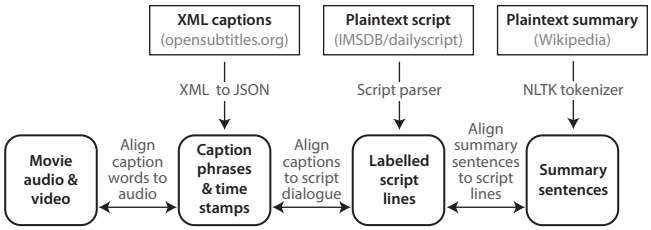

Figure 5. To faciliate searching and browsing using synchronized documents, SceneSkim loads the movie, captions, scripts, and summaries from outside data sources then finds alignments between pairs of documents and the caption and the movie.

As in the search results, each script line and summary sentence also includes a color bar at the right margin of the column, which is used to visualize the confidence of the match between document regions and the movie. For example, sections of the script omitted from the final movie will display a white bar indicating no match. Sections with high confidence matches will appear dark blue, and lighter blue bars indicate low match confidence.

ALGORITHMS

Our interface relies on several algorithms to let users search and browse movie clips using parallel documents (Figure 5). To allow users to access movie clips via caption words, we find a word-level alignment between the caption dialogue and the film. To support browsing the movie based on the script, we parse the script to identify the dialogue and then align the script dialogue to the caption dialogue. We facilitate browsing the movie and script through high level events by aligning the script to the plot summary.

Caption to film word-level alignment

Captions consist of a sequence of transcribed phrases of dialogue, but are not always segmented or labelled by speakers. For instance, in Figure 2B the timestamp of the first caption spans lines for two characters: Jack asks "Will you be saving her?" and Mullroy answers "I can't swim." (Character names are shown in corresponding script lines in Figure 2C.) Thus, to enable users to search and view dialog lines uttered by individual characters, we require a finer-grained synchronization than timestamped phrases. Our system therefore computes a word-level alignment between the captions and the movie.

Our system employs Rubin's implementation [40] of the Penn Phonetics Forced Alignment (P2FA) algorithm [46] to find a word-level alignment between speech audio and the caption text. P2FA first generates expected phonemes from the caption words, and perceptual linear prediction features from the audio. Then, P2FA computes an alignment between the phonemes and audio features using a Hidden Markov Model. However, this method alone does not work well for movies, because movies contain many non-speech sounds—like sound effects and music—that falsely align to phonemes. To overcome this limitation, we constrain alignment to audio containing speech using the caption timestamps. Although timestamps in caption files can be inaccurate, we use them only as approximate markers: Our system groups caption phrases with timestamps less than .25s apart (the approximate

Item	Properties
Slugline	All caps Interior/exterior designation Left margin: 1.5 in.
Action	Left margin: 1.5 in.
Character	All caps Left margin: 4-4.2 in.
Dialogue	After character or parenthetical Left margin: 2.5 in.
Parenthetical	Enclosed in parenthesis After character or dialogue Left margin: 3 in.

Table 1. This table describes the standard format for different types of script elements. The format distinguishes elements using spacing, capitalization, and specific labels such as the interior (INT) or exterior (EXT) labels used in slug lines.

length of a spoken word), and then applies P2FA to align the words in the grouped caption text to the corresponding audio.

Caption-to-film alignment accuracy: We randomly sampled 8 one minute sections from one of the movies in our dataset and created ground truth alignment for the 702 words included in the sampled sections. The start of each word deviates from ground truth by an average of 0.067s ($\sigma = 0.28$) compared to 1.1s ($\sigma = 0.77$) using caption time stamps alone.

Parsing Scripts

Authors format screenplays according to industry standards (Table 1) as detailed in the AMPAS screenplay formatting guide [1]. We exploit this standard to assign action, character, slugline, dialogue, and parenthetical labels to each script line (Figure 2A). However, the formats of the scripts in our dataset deviate from industry standards, because some scripts have been converted from the original PDFs into text documents by third parties, losing some of the original formatting in the process. Our parser relaxes standard requirements by using relative spacing between script elements to determine labels instead of exact margin size. In particular, we assume that the margins of character names are not equal to slugline margins, so that the parser can distinguish these attributes. We additionally assume that the dialogue has a different left margin than the action descriptions. We discard scripts that do not meet these relative spacing requirements.

The output of our parser is a label (e.g. character name, slugline, dialogue, parenthetical, or action) for each script line. The supplementary material contains a detailed description of our script parser.

Parser accuracy: We evaluated our parser by creating ground truth for a randomly selected set of 20 non-empty lines in the original text document from each of 20 movies. We find our parser assigns labels to the text lines with 95.7% accuracy. Most errors are due to our parser's reliance on spacing to classify line type: In this evaluation, the parser mistakenly classified some dialogue lines as action lines in a script containing widely varying spacing for the dialogue lines and character names.

Script dialogue to caption dialogue alignment

Our system matches script dialogue lines to their corresponding caption dialogue phrases in order to allow users to browse the captions and the movie using the script. These two documents do not match exactly—captions contain improvised lines not present in the script; and scenes described in scripts may have been cut from the final movie.

Following the method of Everingham *et al.* [14], our system finds an alignment between the script dialogue and captions documents using the Needleman-Wunsch (NW) algorithm [32]. NW allows for insertions and deletions, which account for most differences between scripts and captions. Our system uses NW to find the highest scoring alignment between the sequence of all script dialogue words and the sequence of all caption dialogue words, where a matching pair counts for 1 point, and any mismatching pair, insertion, or deletion counts for -1 point.

Our interface then uses this alignment to play back dialogue at a line level. When users click on a script dialogue line, the system plays caption lines where one or more words aligned with words in the script line. For instance if a user clicks "Aren't you going to save her?" in the script (Figure 2C), the system would play the caption line "Will you be saving her"(Figure 2B) because the words "you" and "her" match.

Dialogue-to-caption alignment accuracy: To evaluate how well the script dialogue and caption dialogue align in our dataset, we counted the number of caption phrases which had at least one matching word with a script line for 13 randomly selected films, and found an average of 81% ($\sigma = 9$) of caption lines had a matching script line.

To help users navigate around parts of the script that do not match to the final film, our interface includes a confidence bar displaying measures of correspondence. We set the correspondence score by calculating α for each script line and caption phrase match. Then, we set the opacity of the dark blue confidence bar between 0 (for $\alpha = 0$) and 1 (for $\alpha >= 0.5$).

Summary sentence to script alignment

To align the summary to the script our system first splits the summary into sentences using the NLTK sentence tokenizer [7]. Then, the system produces overlapping windows of script lines of length K, which we set to 14 based on alignment accuracy with a small training set. Next, the system finds an ordered alignment between script windows and summary sentences maximizing the TF-IDF [42] similarity of matched pairs.

TF-IDF is a similarity measure used in data mining to compare two documents in the same corpus. In that context, each document is represented as a vector in which each element is the frequency of a term in that document, divided by the frequency of the same term in the corpus as a whole. The similarity between two documents is then computed as the cosine distance between the corresponding vectors. In our context, the TF-IDF for a script window is a vector in which each element is the frequency of a term in that script window, divided by the frequency of the same term in the entire summary and script. The TF-IDF vector for a summary sentence is a vector of elements in which, for each term, the term frequency in the sentence is divided by the term frequency in the entire summary and script.

Thus, for all N script windows $(w_1, w_2, .., w_N)$, and M summary sentences $(s_1, s_2, .., s_M)$ our system finds the TF-IDF vector for each script window and summary sentence. The system constructs matrix T where $T_{i,j}$ is the cosine similarity between the TF-IDF vector for s_i and the TF-IDF vector for w_j. To find an ordered alignment between summary sentences and script windows, our system again uses use Needleman Wunsch, setting the cost, $C_{i,j}$, of matching any pair to $T_{i,j} - mean(T)$ and the cost of skipping a summary sentence or script window to -0.1. We chose -0.1 to balance between allowing insertions/deletions and allowing matches with low TF-IDF scores. In particular, we prefer an insertion or deletion over a match if the match score is less than the mean TF-IDF score minus 0.1. Our TF-IDF-based summary alignment algorithm works better for movies that contain many distinct objects, actions, characters, and location terms for different parts of the movie (i.e. action-driven movies), than for movies that focus on character development (i.e. character-driven movies).

Summary-to-script alignment accuracy: To evaluate the summary sentence algorithm we created ground truth summary-to-script alignments for four movies randomly selected for the most popular genres in our dataset (e.g. drama, action, comedy, romance). For the movies *King Kong*, *Code of Silence*, *American Graffiti* and *Punch Drunk Love* we find that the algorithm's predicted alignment matches at least one ground truth line 82%, 83%, 75% and 57% of the time respectively. In practice, this summary alignment aids users in locating specific clips in the movie by navigating to a nearby region within a long movie. After using the summary alignment to get close to the location of interest, users can then use finer-grained script or caption navigation to find the desired clip.

SEARCHING AND BROWSING WITH SCENESKIM

We used SceneSkim to conduct each type of query mentioned in our interviews with film studies researchers and film professionals. In order to identify realistic queries that might arise in the course of their work, we retrieved specific examples from existing film studies literature about the original *Star Wars* trilogy: *Star Wars Episode IV: A New Hope*, *Star Wars Episode V: The Empire Strikes Back*, and *Star Wars Episode VI: Return of the Jedi*. We instrumented our system to record interaction statistics and completion time while performing each task (Table 2).

Searching for characters: The essay *The Empire Strikes Back: Deeper and Darker* [9] makes the claim that the character Darth Vader becomes more menacing in *Episode V* than he was in *Episode IV*. To explore this hypothesis, we searched for "Vader" in the **summary** across both films. We clicked through the **summary** search results in *Episode IV* to view corresponding important events where Vader occurs. One summary result describes Vader dueling with another character. Clicking on this sentence navigates into the vicinity of the duel. We refine navigation by clicking on an explanation of the duel in the script. Another summary result describes Vader spiraling away in a spaceship. When we click this result, we see Vader spiraling away in a spaceship while yelling "What??" Comparing the results from both movies, *Episode V* depicts Vader in darker lighting with a louder wheezing noise, suggesting that Vader is portrayed in a more menacing fashion in *Episode V*.

Searching for objects: In *Your Father's Lightsaber* [41], Wetman notes that the screen time devoted to the "lightsaber," a futuristic weapon, increases throughout the first trilogy. To investigate this hypothesis, we searched for "lightsaber" in **script actions**, **summary**, **captions** and **script dialogue** across all three movies. Using search results from all three documents, we located all instances of lightsabers in the movie, watched the corresponding clips, and timed the amount of screentime during which lightsabers appeared on screen. We found that scenes with lightsabers did increase through the three movies with 157s of screen time in *Episode IV*, 217s of screen time in *Episode V* and 258s in *Episode VI*.

Searching for dialogue: *Stoicism in the Stars* [13] uses 22 quotes from the first trilogy throughout the essay. We were able to locate and watch all 22 quotes using our system. We found the quotes by searching for quote terms in **captions** and browsing within the captions and scripts when multiple quotes were close to one another.

In the *Star Wars Episodes 4-6 Guide: Themes, Motifs, and Symbols* [2], the author notes that Luke's costumes change from white in *Episode IV*, to grey in *Episode V*, to black in *Episode VI*. Searching for Luke's appearances in **summary** quickly reveals this transition and several other variations of Luke's costumes (e.g., a pilot suit, a brown cloak, a Storm Troopers suit and a tan jacket). Brode and Deyneka [9] describe Hoth, Dagobah, and Cloud City in detail arguing that these locations represent the phases of Luke's journey in *Episode V*. We quickly locate and confirm these descriptions by searching **script locations** for "Hoth", "Dagobah" and "Cloud City". Based on Stephen's essay on stoicism [13] which suggests Jedi's encourage patience while the Dark Side encourages anger, we compare uses of the terms "patience" and "anger" in **captions** across the first trilogy to find that characters associated with the "dark side" explicitly encourage anger 3 times while Jedis discourage it 3 times (both encourage patience). Kaufman's essay [9] suggests that the robot character R2D2 shows affection towards other characters. To examine how R2D2 conveys affection without speech, we search for "R2D2 AND beeps" in **script actions** to find that the robot character shows emotions by beeping in a wide variety of pitches, tones, and durations. Finally, according to Gordon's [9] observation that there are "red lights flashing in the background whenever [Han] Solo and Leia confront each other" we search "Han AND Leia" in **scenes** finding flashing red lights do not occur during at least 4 conversations between Han and Leia.

INFORMAL EVALUATION

We conducted an informal evaluation with three film studies researchers in order to get feedback about how SceneSkim might support their work. We recruited the film studies researchers by e-mailing all film studies graduate students at Berkeley. After an initial 12-minute tutorial, we posed four concrete search tasks. In a subsequent exploratory section,

		Search result clicks			Document clicks				
Label	Task	Summary	Script	Captions	Summary	Script	Captions	Video watched	Completion time
A	Vader	10	0	0	0	3	0	3:48	5:37
B	lightsabers	1	22	1	0	37	0	20:00	21:44
C	Luke's costumes	17	0	0	0	16	0	4:23	5:20
D	22 quotes	0	2	19	0	0	5	4:31	8:36
E	main locations	3	24	0	0	2	0	2:57	4:23
F	anger/patience	0	0	12	0	0	0	00:21	1:37
G	R2D2 beeps	0	13	0	0	3	0	1:23	1:56
H	Han and Leia	0	16	0	0	5	0	2:28	3:10

Table 2. We instrumented our interface to record interactions while answering queries. "Search result clicks" refers to the origin document for search results we clicked in the search pane, while "document clicks" refers to clicks within the movie pane. While answering queries, we typically first clicked a search result from the summary, script, or captions then used the script or captions for fine grained navigation around that location. On average, we watched 5 minutes and 9 seconds of video for each query and spent a total of 6 minutes and 17 seconds completing the task.

users formulated their own questions and answered them using our system. The concluding interview gathered qualitative feedback about the system.

Our search tasks focused on questions about characters, costumes, relationships, and settings in the original *Star Wars* trilogy. For example, we asked: *Given that Chewbacca doesn't speak English, how does the film use growls to convey relationships between Chewbacca and other characters?* Supplementary materials list the complete set of questions. Users successfully answered all the questions in the tasks section using our system in 1-3 searches per question.

In the exploratory section, U1, U2, and U3 used our system to answer new questions. We invited users to use all movies in our system, including the majority of the corpus that included only parallel documents, without the source videos. All three users chose movies with source video. U1 asked in which circumstances the sentence "Are you talking to me?" occurs in *Taxi Driver*. She searched for the words "you talking to me" over **script dialogue** and **captions** only. She found that only the main character uses this phrase, and that it only occurs in the captions and never in the script, suggesting that the actor either improvised the phrase, or it was added in a later script draft. Because the line did not occur in the script, U1 watched the clip for each caption line to definitively identify the speaker each time.

U2 wanted to know if Jack Sparrow's eccentric body movements in *Pirates of the Caribbean* were explicitly written into the screenplay, and if so, how they were described. She searched for "Jack" in **scenes** to retrieve all scenes in which the character occurs, watched the scenes to find where Jack makes these movements, then read the script corresponding to those scenes. She did not see a consistent description of these movements in the script, which suggests that they are most likely the actor's invention.

U3 wanted to know how Tarkin dies in the first *Star Wars* trilogy. He searched for "Tarkin" in **script characters** to find all of Tarkin's lines and navigated to Tarkin's last speaking line. Then, he navigated beyond this using the script to the next time Tarkin appeared in an action description. He clicked on the action description to view Tarkin in a space station shortly before the space station explodes.

U1, U2, and U3 expressed strong interest in using the tool in the future. U1 explained *"[the system] is just really useful. [...] I feel like I would use this a lot, for research, for teaching and for looking at scenes."* U3 mentioned *"if I had the tool open in front of a class room, it would be an incredible teaching tool. [...] You would be able to respond to student's questions, suggestions and requests on the fly."* All three users inquired about plans for releasing the tool and commented positively on its usability.

Users found that each of the different search and browsing features successfully supported particular tasks. Summaries were useful for navigating familiar movies, e.g. for revisiting clips repeatedly. Our users also mentioned uses we had not considered: U1 explained that having captions and script aligned to the movie facilitates pulling quotes for research essays from both documents. U2 explained that having all of the parallel documents aligned to each other encouraged new ideas: *"I like that the four types of information are available simultaneously [...] So there are moments of discovery that I don't think I would come across if I was just picking any one of these to look at at a time."*

Participants also suggested several other areas for improvement: U2 noted that captions are mostly useful for refining sections found in the script, but less useful in isolation. All three users requested the ability to bookmark scenes to play back later. U1 and U2 also suggested showing some visualizations of aggregate statistics to inspire new questions and searches. For instance, U1 recommended showing a preview of all locations or characters in a movie could be helpful, whereas U2 wanted a list of frequently used verbs. U1 and U2 found the confidence bars helpful when deciding which results to click on, and U2 suggested filtering out results with low certainty.

LIMITATIONS AND FUTURE WORK

Our current implementation of SceneSkim has both technical and practical limitations.

Availability of scripts and movies: The availability of scripts currently restricts which movies we can search over in SceneSkim, though we found informally that scripts are increasingly available for new movies. Captions and summaries are more readily available for both older and newer movies. In the future, movie library search may become available through sites like Netflix, which already have a paid usage model, or through micropayment schemes. In fact, video players like Amazon's TV X-Ray[1] already show meta-

[1] http://www.wired.com/2015/04/amazon-xray-fire-tv/

data such as actor names on demand. In the present, individuals will have to load their own collections into our system.

Summary to script alignment: Our informal evaluation suggests that the alignment of summary sentences to script sections is useful high-level navigation, but the technique could be improved substantially. First, our approach does not yet support re-orderings of scenes between script and final movie, which can occur during editing. Summaries can also present events in a slightly different order than the source film, especially when the film contains scenes in which two or more narrative strands are interleaved. In addition, our method does not support a summary sentence that should match multiple non-contiguous script sections. Finally, our manually-constructed ground truth alignments are time-consuming to create (about 1 hour for every 20 summary sentences). Given a larger corpus of ground truth data, we could consider using a machine learning approach for summary alignment instead.

Adding new types of meta data: Some film studies researchers study shot length, lighting, set design, and shot types in order to analyze patterns in videos. Future work could employ computer vision techniques to detect visual attributes to allow users to search over more types of metadata. For example, users could study how a director used a particular shot type to convey a certain concepts by searching for that shot type in all movies by that director.

Adding more visualization capabilities: Currently, we visualize search results by displaying a text snippet and the result type. This visualization supports the tasks outlined in early interviews. However, our interface may reveal more patterns if we added visualizations for frequency of search results over different facets, such as time within the movie, genre, release date, writer, or director.

Adding bookmarks and correcting mistakes: In the informal evaluation, several users pointed out that they would like support for bookmarking, categorizing, and taking notes on video clips in order to keep track of many clips for research or teaching. Future work could add capabilities to support these common user tasks. In addition, since the algorithms do not achieve 100% accuracy, we hope to add tools within the system to correct algorithmic mistakes.

CONCLUSION

In this paper we present SceneSkim, a new interface for searching and browsing videos through aligned captions, scripts, and summaries. Our system allows us to quickly search and browse clips in order to answer queries drawn from from existing film studies analysis. In the informal evaluation, three film studies researchers answered new questions using our tool, and enthusiastically stated they would use this tool for their work.

ACKNOWLEDGMENTS

We thank Andrew Head for creating our video. We also thank Professor Gail De Kosnik's new media research group for their valuable feedback in early stages of this project. Our research is supported by the National Science Foundation IIS grant 1210836, the CITRIS Connected Communities Initiative, and an NDSEG fellowship.

REFERENCES

1. *For a Few Days More* screenplay formatting guide. **http://www.oscars.org/sites/default/files/scriptsample.pdf**. Accessed 2015-03-03.
2. *Star Wars Episodes IV – VI*: Themes, motifs and symbols. **http://www.sparknotes.com/film/starwars/themes.html**. Accessed 2015-04-12.
3. Ted.com. **http://www.ted.com/**. Accessed 2015-04-11.
4. Baio, A., and Bell-Smith, M. supercut.org. **http://supercut.org/**. Accessed 2015-07-17.
5. Barnes, C., Goldman, D. B., Shechtman, E., and Finkelstein, A. Video tapestries with continuous temporal zoom. *ACM Transactions on Graphics (TOG) 29*, 4 (2010), 89.
6. Berthouzoz, F., Li, W., and Agrawala, M. Tools for placing cuts and transitions in interview video. *ACM Transactions on Graphics (TOG) 31*, 4 (2012), 67.
7. Bird, S. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics (2006), 69–72.
8. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. Finding actors and actions in movies. In *Proc. IEEE International Conference on Computer Vision* (2013).
9. Brode, D., and Deyneka, L. *Sex, Politics, and Religion in Star Wars: An Anthology*. Scarecrow Press, 2012.
10. Casares, J., Long, A. C., Myers, B. A., Bhatnagar, R., Stevens, S. M., Dabbish, L., Yocum, D., and Corbett, A. Simplifying video editing using metadata. In *Proceedings of DIS*, ACM (2002), 157–166.
11. Chi, P.-Y., Ahn, S., Ren, A., Dontcheva, M., Li, W., and Hartmann, B. MixT: Automatic generation of step-by-step mixed media tutorials. In *Proceedings of UIST*, ACM (2012), 93–102.
12. Cour, T., Jordan, C., Miltsakaki, E., and Taskar, B. Movie/script: Alignment and parsing of video and text transcription. In *Computer Vision–ECCV 2008*. Springer, 2008, 158–171.
13. Decker, K., and Eberl, J. *Star Wars and Philosophy: More Powerful Than You Can Possibly Imagine*. Popular culture and philosophy. Open Court, 2005.
14. Everingham, M., Sivic, J., and Zisserman, A. Hello! My name is... Buffy – automatic naming of characters in TV video. In *British Machine Vision Conference* (2006).
15. Fouse, A., Weibel, N., Hutchins, E., and Hollan, J. D. Chronoviz: A system for supporting navigation of time-coded data. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, ACM (2011), 299–304.
16. Fritz, B. OMDB API. **http://www.omdbapi.com/**. Accessed 2015-03-03.
17. Goldman, D. B., Curless, B., Seitz, S. M., and Salesin, D. Schematic storyboarding for video visualization and editing. *ACM Transactions on Graphics (Proc. SIGGRAPH) 25*, 3 (2006).
18. Grossman, T., Matejka, J., and Fitzmaurice, G. Chronicle: Capture, exploration, and playback of document workflow histories. In *Proceedings of UIST*, ACM (2010), 143–152.
19. Haubold, A., and Kender, J. R. VAST MM: Multimedia browser for presentation video. In *Proceedings of CIVR*, ACM (2007), 41–48.
20. Jackson, D., Nicholson, J., Stoeckigt, G., Wrobel, R., Thieme, A., and Olivier, P. Panopticon: A parallel video overview system. In *Proceedings of UIST*, ACM (2013), 123–130.
21. Kim, J., Guo, P. J., Cai, C. J., Li, S.-W. D., Gajos, K. Z., and Miller, R. C. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of UIST*, ACM (2014), 563–572.
22. Kim, J., Nguyen, P. T., Weir, S., Guo, P. J., Miller, R. C., and Gajos, K. Z. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of CHI*, ACM (2014), 4017–4026.

23. Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, IEEE (2008), 1–8.

24. Lavigne, S. VideoGrep. **http://lav.io/2014/06/videogrep-automatic-supercuts-with-python/**. Accessed 2015-04-11.

25. Mackay, W. E., and Beaudouin-Lafon, M. Diva: Exploratory data analysis with multimedia streams. In *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co. (1998), 416–423.

26. Matejka, J., Grossman, T., and Fitzmaurice, G. Ambient help. In *Proceedings of CHI*, ACM (2011), 2751–2760.

27. Matejka, J., Grossman, T., and Fitzmaurice, G. Swift: Reducing the effects of latency in online video scrubbing. In *Proceedings of CHI*, ACM (2012), 637–646.

28. Matejka, J., Grossman, T., and Fitzmaurice, G. Swifter: Improved online video scrubbing. In *Proceedings of CHI*, ACM (2013), 1159–1168.

29. Matejka, J., Grossman, T., and Fitzmaurice, G. Video lens: Rapid playback and exploration of large video collections and associated metadata. In *Proceedings of UIST*, ACM (2014), 541–550.

30. Mohamad Ali, N., Smeaton, A. F., and Lee, H. Designing an interface for a digital movie browsing system in the film studies domain. *International Journal of Digital Content Technology and Its Applications 5*, 9 (2011), 361–370.

31. Monserrat, T.-J. K. P., Zhao, S., McGee, K., and Pandey, A. V. NoteVideo: Facilitating navigation of blackboard-style lecture videos. In *Proceedings of CHI*, ACM (2013), 1139–1148.

32. Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology 48*, 3 (1970), 443–453.

33. Noah, D. B. B. O., and Smith, A. Learning latent personas of film characters. ACL (2013).

34. Olsen, D. R., Partridge, B., and Lynn, S. Time warp sports for internet television. *ACM Transactions on Computer-Human Interaction (TOCHI) 17*, 4 (2010), 16.

35. Pavel, A., Reed, C., Hartmann, B., and Agrawala, M. Video digests: A browsable, skimmable format for informational lecture videos. In *Proceedings of UIST*, ACM (2014), 573–582.

36. Pongnumkul, S., Dontcheva, M., Li, W., Wang, J., Bourdev, L., Avidan, S., and Cohen, M. F. Pause-and-play: Automatically linking screencast video tutorials with applications. In *Proceedings of UIST*, ACM (2011), 135–144.

37. Ramanathan, V., Joulin, A., Liang, P., and Fei-Fei, L. Linking people in videos with "their" names using coreference resolution. In *Computer Vision–ECCV 2014*. Springer, 2014, 95–110.

38. Ronfard, R. Reading movies: An integrated DVD player for browsing movies and their scripts. In *Proceedings of the 12th annual ACM international conference on Multimedia*, ACM (2004), 740–741.

39. Ronfard, R., and Thuong, T. T. A framework for aligning and indexing movies with their script. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, IEEE (2003), I–21.

40. Rubin, S., Berthouzoz, F., Mysore, G. J., Li, W., and Agrawala, M. Content-based tools for editing audio stories. In *Proceedings of UIST*, ACM (2013), 113–122.

41. Silvio, C., Vinci, T., Palumbo, D., and Sullivan, C. *Culture, Identities and Technology in the Star Wars Films: Essays on the Two Trilogies*. Critical Explorations in Science Fiction and Fantasy. McFarland & Company, 2007.

42. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation 28*, 1 (1972), 11–21.

43. Tapaswi, M., Bäuml, M., and Stiefelhagen, R. Aligning plot synopses to videos for story-based retrieval. *International Journal of Multimedia Information Retrieval 4*, 1 (2015), 3–16.

44. Tapaswi, M., Bäuml, M., and Stiefelhagen, R. Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 1827–1835.

45. Tiedemann, J. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA) (2012).

46. Yuan, J., and Liberman, M. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America 123*, 5 (2008), 3878.

47. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724* (2015).

APPENDIX A: DATASET

We obtained complete caption, script and plot summary sets for 816 movies by joining three document databases: the caption corpus collected from opensubtitle.org by Tiedemann *et al.* [45], the Wikipedia plot summary corpus collected by by Bamman *et al.* [33], and a script corpus we generated by scraping dailyscript.com and imsdb.com for all scripts in plain text format.

Specifically, we join the script corpus and summary corpus using provided movie titles, then find the IMDB identification number for each movie in our dataset using the OMDB API [16]. For each movie, we use the movie's IMDB identification number to find the corresponding caption file in Tiedemann's caption corpus [45]. We remove all movies that lack a caption file, script file, or summary file, leaving us with 1002 unique movies with all three documents in our dataset. We then remove all movies that do not have a script in the format accepted by our parser, which yields the 816 unique movies in our dataset.

This dataset is biased towards recent movies in wide public release, as captions and summaries are more readily available for such movies. Although our dataset contains movies released between 1928 and 2013, half of the movies in the dataset were released in 1999 or later, and 90% of the movies in our dataset were released in 1979 or later.

Because movies themselves are not freely available, we purchased seven movies (Starwars 4-6, Chinatown, Taxi Driver, The Descendants, Pirates of the Carribean: Curse of the Black Pearl) to demonstrate interactions.