Deep learning model using the Cornell movie data
Phase 0

- **Metadata: -**

  - 220,579 conversational exchanges between 10,292 pairs of movie characters
  - involves 9,035 characters from 617 movies
  - in total 304,713 utterances

  - gender provided for 3,774 characters

  Files description:

  - movie_titles_metadata.txt

  - movie_characters_metadata.txt

  - movie_lines.txt

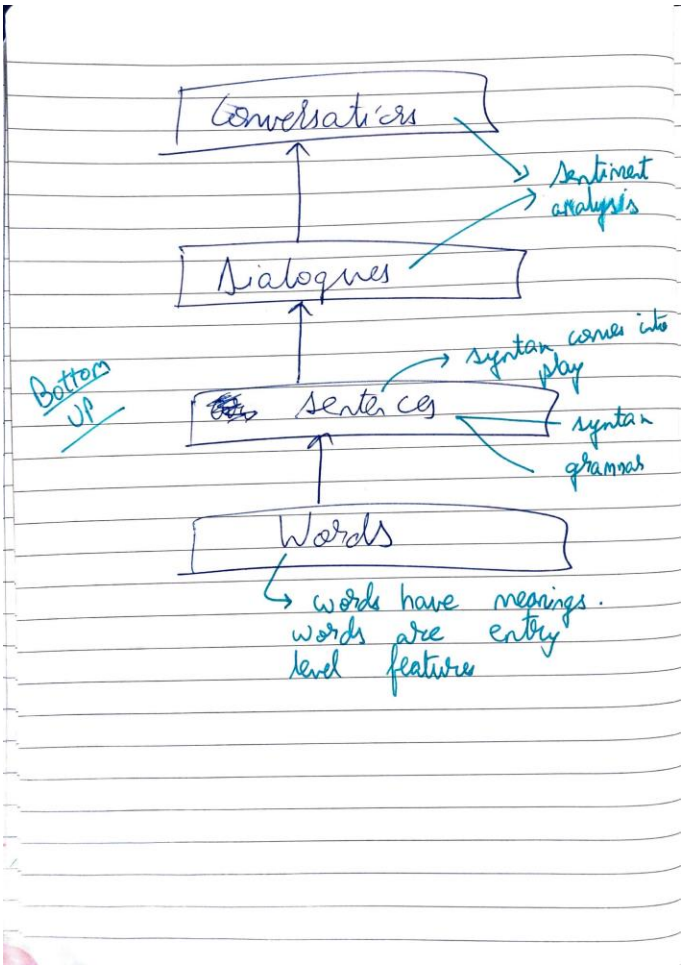
  - movie_conversations.txt

  - raw_script_urls.txt

- **A conversation:**

Conversations can be broken down into dialogues which are formed using sentences which in turn are formed using words. We start by observing the words(the fundamental level) and recognizing their meaning. They are stored as entry-level features/characteristics.

Words form sentences. This is where syntax and grammar(morphology) come into play. Sentences form dialogues and dialogues form conversations. Sentiment analysis is performed on these conversations and dialogues to judge the tone of said conversation.



The approach used in this form of analysis of a conversation is a bottom-up approach.

- Why bottom-up?

  A single word can have different meanings. Words together with syntax and grammar can lead to less ambiguity as to which class that word belongs to. For example – the word ‘bank’ could either mean a financial institution or the side of a river. Such confusion is minimized using syntax and grammar.

  If one were to use a top-down approach, the sentiment of the conversation could be misjudged and in turn, the literal meaning of words.

- **Our dataset:**

The attributes present in each of the **five** text files is straight forward to understand except for the “movie_conversations.txt” file. Each entry in the last attribute consists of a list containing IDs representing the dialogues that go together.

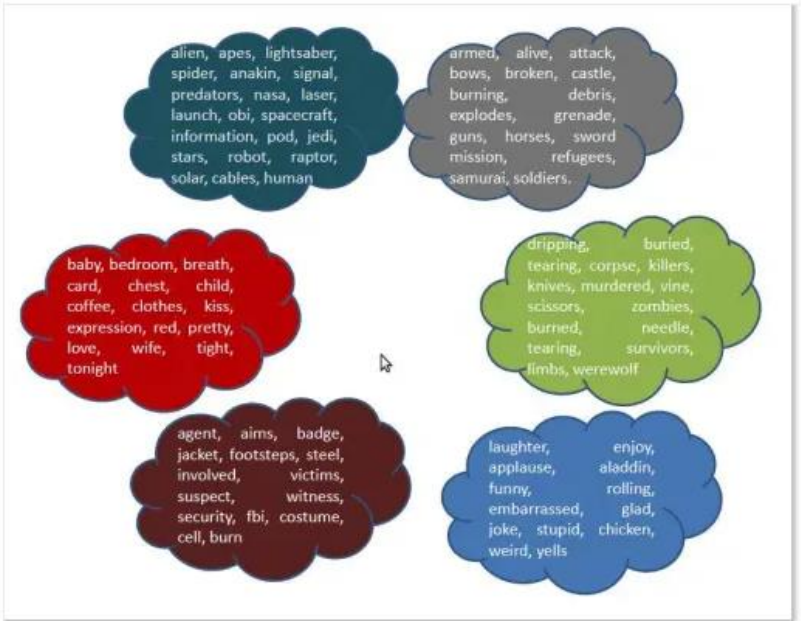
```
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L194', 'L195', 'L196', 'L197']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L198', 'L199']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L200', 'L201', 'L202', 'L203']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L204', 'L205', 'L206']
u0 +++$+++ u2 +++$+++ m0 +++$+++ ['L207', 'L208']
```

In the first entry of the above screenshot, ['L194', 'L195', 'L196', 'L197'] represents the IDs of the dialogues forming a conversation.

- **Intuition for pattern finding:**

  - Patterns found in conversations in a movie based on gender could help in finding its genre. **For example** – A movie with conversations between people of two different genders could indicate that it is a romantic one. On the other hand, a movie with conversations between two males could indicate that it is an action based movie.

  - The occurrence of certain words can signal the sentiment of dialogues which can be used in determining the genre of a movie. For example – A word like ‘poop’ will most likely be used under a goofy/comical context while a word like ‘kill’ will most likely be used under a context filled with adrenaline-fueled context(action).



- The frequency of occurrence of exclamation marks, interjections and disfluencies present in dialogues could help in determining the genre of a movie. For example – More interjections might mean a movie's genre is comedy/action/war.

- Length and continuity of dialogues signal genres too. Monologues lead to connotations of seriousness, wit and descriptions of one's environments(drama, biography, fantasy).

- Number of dialogues present in a conversation along with the average length of a dialogue can indicate the setting of a scene as well. For example – a high number of dialogues under a certain context can signal high energy among the subjects meaning the genre of the movie could, with a high probability, be action or adventure.

- **<u>Problem in finding patterns across movies:</u>**

- Each movie present in our dataset has different conversations. Each conversation has a different context to it(even within the same movie). Finding patterns within these conversations and in turn with a movie becomes the same as finding patterns across movies. The methods/criterias appear to be the same.

- A high overlap of patterns within and across movies with the same genre is expected to be observed but nothing, as of yet, can be said for patterns across movies of different genres.