Capstone Project 2: Project Proposal

What Pulp Fiction, Star Wars, Macbeth, and Siddhartha have in common

1. **What is the problem you want to solve?**

According to Global English Editing, India, Thailand, and China are the countries with best reading habits in 2018, spending on reading between 8:00 and 10:42 hours per week. The U.S., with 5:42 hours, is in 22nd position. 37% of American adults with a high school degree or less and 7% of college graduates have reported not reading a book in any format in the past year.

In the age of Netflix and instantaneous entertainment, seems to be people is still reading, but it is hard to keep the people interested in reading classic, historical and universal literature. How we can improve this reading behavior? Past year, Netflix developed around 50 literary projects (turning novels into series). Cinematographic adaptations bring books to people and could inspire them to read the books later.

This project tries to find relations between movies and books: genres, type of fictitious worlds, the extension of the stories, historical contexts, physical places where the plots are developed, for instance. This kind of relationships could help us to recommend books depending on the people preferences or give us a guide about how to predict the success of the metamorphosis from novels to cinematographic adaptations and discovering what genres would achieve more acceptance by the audience.

Additionally, we'll use historical contexts and physical spaces where books were built for mapping the documents in innovative and educational ways: if you don't show curiosity of books, at least you can connect them with historical events and understands the links with the present. The goal is to achieve a representation of different cultures throughout the literature and looking for innovative visualizations to show the results.

2. **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?**

The inspiration for this project is educational. We are trying to understand the impressions that books generate in population to find innovative ways to teach them and bring books to them. Also, we'll use all the available information on our database to predict trends related to how to decide what books could be interesting purposes to cinematographic adaptations. The educational sector, companies associated with book sell and interested in tunning books into audio-visual projects, could be interested in these results.

3. **What data are you using? How will you acquire the data?**

- Goodreads Datasets
    - Meta-Data: A complete book graph of 2.36M books: titles, ratings, publication date, number of pages, similar books, authors, description, language, etc.
    - Book Review Texts: 1.38M scraped records with detailed review text: users, reviews, sentences, ratings and if the review contains a spoiler or not.
- Movies:
    - Movies Dataset: 45.000 movies, and 26 M ratings from 270.000 users: titles, overviews, home pages, genres, release_date, popularity, runtime, collections, languages, ratings and more could be found there.
    - HuluRaw: top 1000 most popular Hulu shows (includes a description of shows, genres, links to more descriptions, information about seasons and extra notes about episodes)
- Books and Movie Blogs:
    - If descriptions are not enough to extract the information required, we'll apply web scrapping on movie and book blogs, to get more text to analyze.

4. **Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

The diagram below (Figure 1) shows the flow of data and all the processes included in the project.

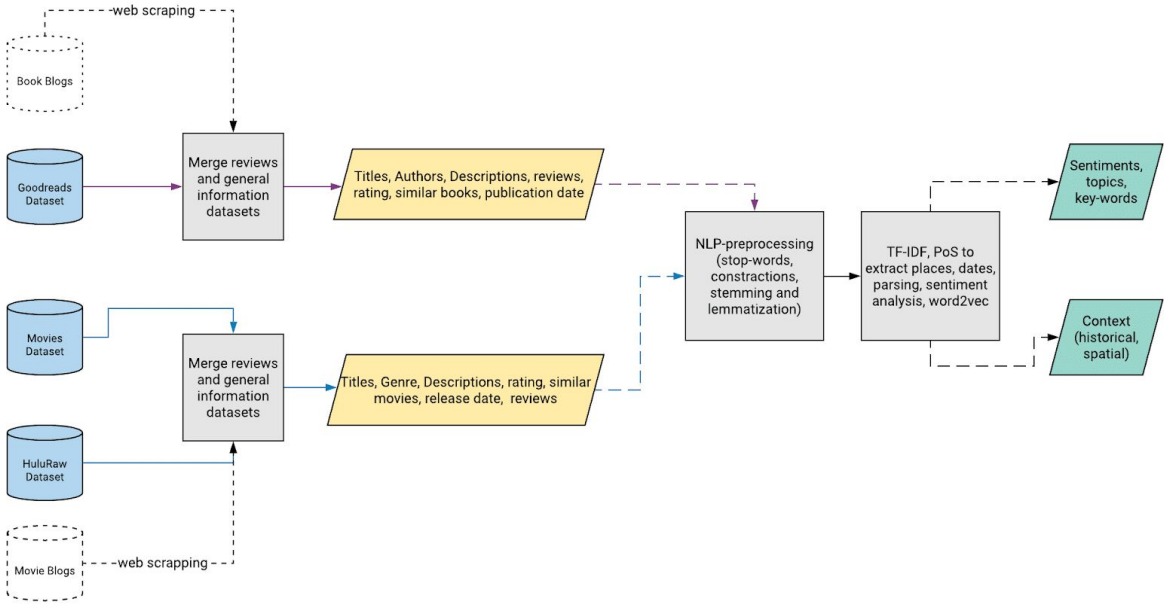

Figure 1: Flow of information and early pipeline of the project

5. **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

All of them. As in Capstone 1 methodology, we'll deliver a Wrangling Report, followed by first findings focused on Statistic techniques and a Data Story. Later, all of them could be found in the Milestone Report. More in-depth findings are included in the ML report and finally, all the main discoveries will be on the Final Report and Presentation. Additionally, we want to do an article for Medium or my personal blog about this Capstone.

6. References

1. Mengting Wan, Julian McAuley, "Item Recommendation on Monotonic Behavior Chains", in *RecSys'18*. [bibtex]
2. Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley, "Fine-Grained Spoiler Detection from Large-Scale Review Corpora", in *ACL'19*.
3. Global English Editing: link