NLP Project Tasks

-meet up@
4:30 / 5pm

~~1. Read in scripts + scrape~~
~~-try PDF Miner~~
~~PyPDF2~~

-make subdata frames

2. Explore each scrpt -pre-load
   - -# of words
   - # of main characters
   - # of words per character
   - # of settings
   - # unsaid cues

} store these somewhere
-Pos → then isupper?

→ unsaid cues?

↓ put unsaid cues in a col

↓ put count of character names in a col

↓ put counts of char speaking

talking is indented

2.A Add genre field to table
→ import table

3. Subsets
   - Text per character
   - -locations

4. Clean text -clean, then Pos → more clean
   —or extract based on regex → then do PoS
   - find ways to remove what we don't want
   - Tokenize → Lemmatize
   - Remove Stopwords

} use NLTK

-Remove "noise"
-Extract key parts? + identify what is considered key?

Sentiment Analysis - look @ HW 4

-use ANEW to score words

→ any customization needed?

→ calculate scores

→ split by genre

→ split by winner loser

→ split by actor / main role?

⑥ Topic Modeling   TF-IDF?

→ put words in BoW + use sklearn

⇒ shar words in each topic

→ which method do we want to use?

↳ NMF?

→ split by genre?

→ split by winner loser?

⑦ Semantic Similarity - Lable

→ winners vs losers.

~~thr~~ → genres?

⇓ ? these similar

---

⑧ K-Means Clustering

→ KNN

Visuals - Screen shots of diff formats of PDFs

- word cloud for - genres
  - winners vs losers

graphs → - sentiment across genres graph
  - sentiment winner vs losers

Other - If time Classification Model

→ use 4 unused scripts → maybe current films that could be nominated for 2020 Oscars?