MPST: A Corpus of **M**ovie **P**lot **S**ynopses with **T**ags

Sudipta Kar
Suraj Maharjan
A. Pastor López-Monroy
Thamar Solorio

Department of Computer Science
University of Houston

UNIVERSITY of HOUSTON

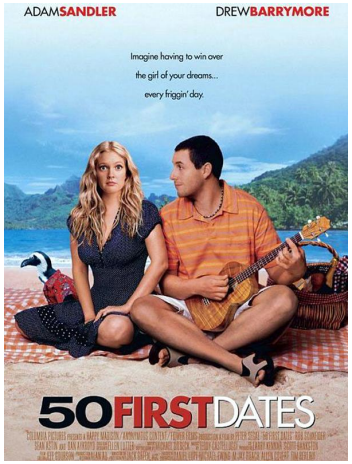
11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Miyazaki (Japan)

RiTUAL

Outline

- Introduction
- The MPST Corpus
- Predicting Tags from Synopses
- Experiments and Results
- Conclusion and Future Work

Folksonomy



comedy, prank, entertaining, romantic, flashback

fantasy, murder, cult, violence, horror, insanity

- Coined by Vander Wal in 2005

The Curious Case of Folksonomy

- Got attention from audience = A lot of tags
- Out of attention/ new/ upcoming = Emptiness
  * Hard to reach audience
- No tags for ~34% movies from ~130K movies in IMDB (Top movies of 22 genres)

CAN A.I. HELP???

Tagging Movies Automatically

- Use the plot to generate tags
- Summary of the storyline + availability
- Is this movie
  - Suspenseful? Boring? Romantic? Inspiring? Brainwashing?

Henry Roth is a veterinarian at Sea Life Park on the island of Oahu, Hawaii. He has a reputation of womanizing female tourists and does not display any interest in committing to a serious relationship. Henry's closest friends are Ula, a marijuana-smoking Islander; his assistant Alexa, whose gender is unclear; Willy, his pet African penguin; and Jocko, a walrus.

One day Henry's boat breaks down while he is sailing around Oahu. He goes to the Hukilau Café to wait for the Coast Guard. There he sees a young woman named Lucy Whitmore, who makes architectural art with her waffles. Henry thinks she is a local, which prevents him from introducing himself, but the next day he comes back. Lucy and he hit it off instantly and she asks him to meet her again tomorrow morning.
When Henry goes back to the café, Lucy does not have any recollection of ever meeting him. The restaurant owner Sue……..

In Quest of a Dataset

- MovieLens 20M Dataset[1]
  - Tagspace contains metadata, too specific tags
- MM-IMDB[2]
  - Plot summaries are too short
- CMU Movie Summary Corpus[3]
  - Lacks IMDB information to assign tags
- ScriptBase[4]
  - Full scripts of ~1200 movies

1.Harper, F Maxwell and Konstan, Joseph A. (2016). *The movielens datasets: History and context*. ACM.
2.Arevalo, J., Solorio, T., y Gómez, M. M., and González, F. A. (2017). Gated multimodal units for information fusion. In *5th International Conference on Learning Rep resentations (ICLR) 2017 -Workshop Track*.
3. Bamman, D., O'Connor, B., and Smith, N. A. (2013). Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria, August. Association for Computational Linguistics.
4. John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado, May–June. Association for Computational Linguistics.

Issues with Tag Spaces

- Difference in perspective of users
- Synonymous tags
- Incompleteness

Titanic (1997)

historical event, self sacrifice, iceberg, told in flashback, love affair, rich woman, poor man, leak, love at first sight, year 1912, orchestra ...

Pulp Fiction (1994)

Nonlinear timeline, black comedy, overdose, neo noir, bondage, corpse, car accident, drug use, drug dealing, shootout, drug dealer, drug overdose ...

*** Tag examples are collected from IMDB**

Outline

- Introduction
- The MPST Corpus
- Predicting Tags from Synopses
- Experiments and Results
- Conclusion and Future Work

Requirements

- Easy to understand
- Storyline related tags
- No redundancy in tagset

  cult, cult classic,
  cult film, cult movie → CULT

- Not too specific for a movie
- Well represented tagset
- Adequate and noise free synopses

✓ paranormal
✓ suspenseful
✓ inspiring
✓ thought-provoking
× two word title
× beating
× crying
× falling from height
× three act structure
× 1940s
× london england
× animal in title
× scantily clad female

Data Collection

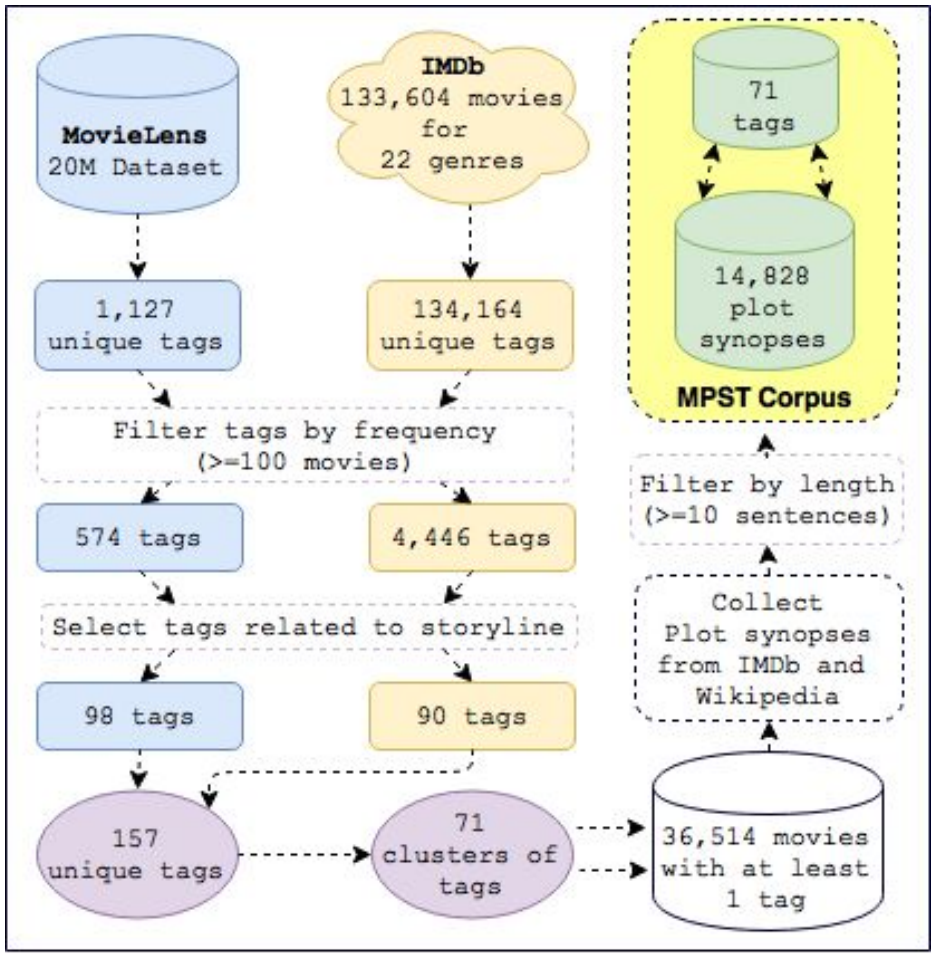
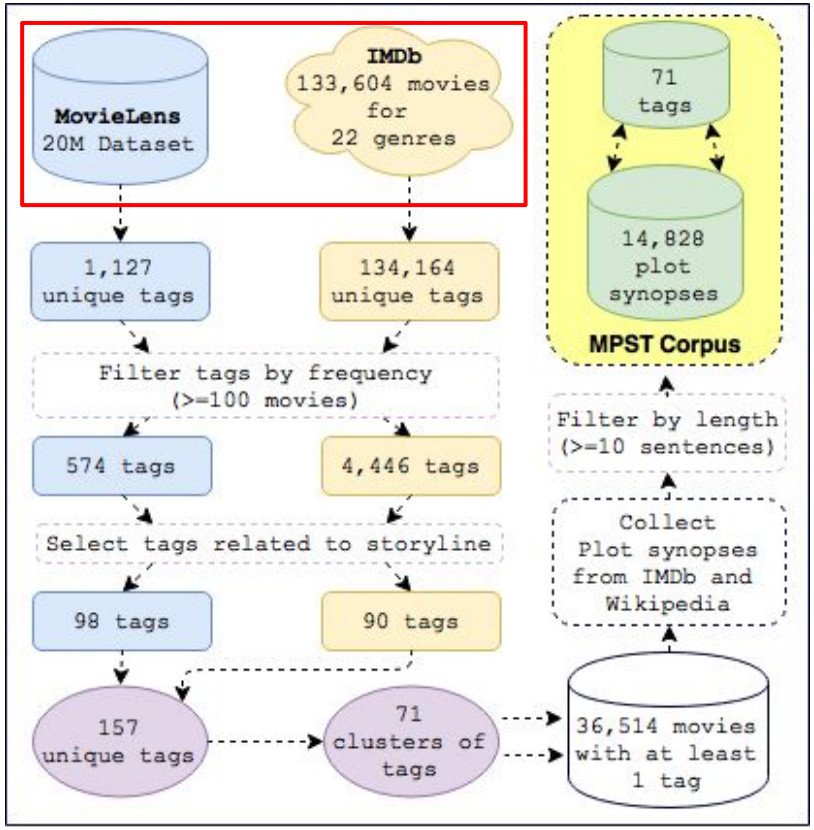
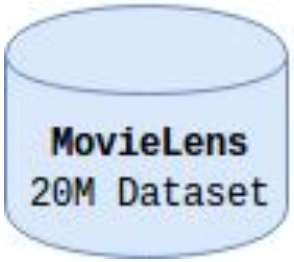


Figure 1: Overview of the data collection process

Data Collection





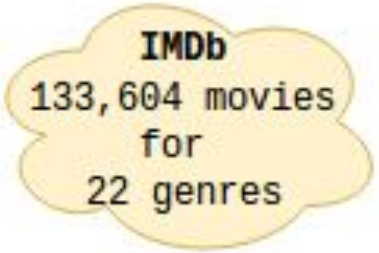
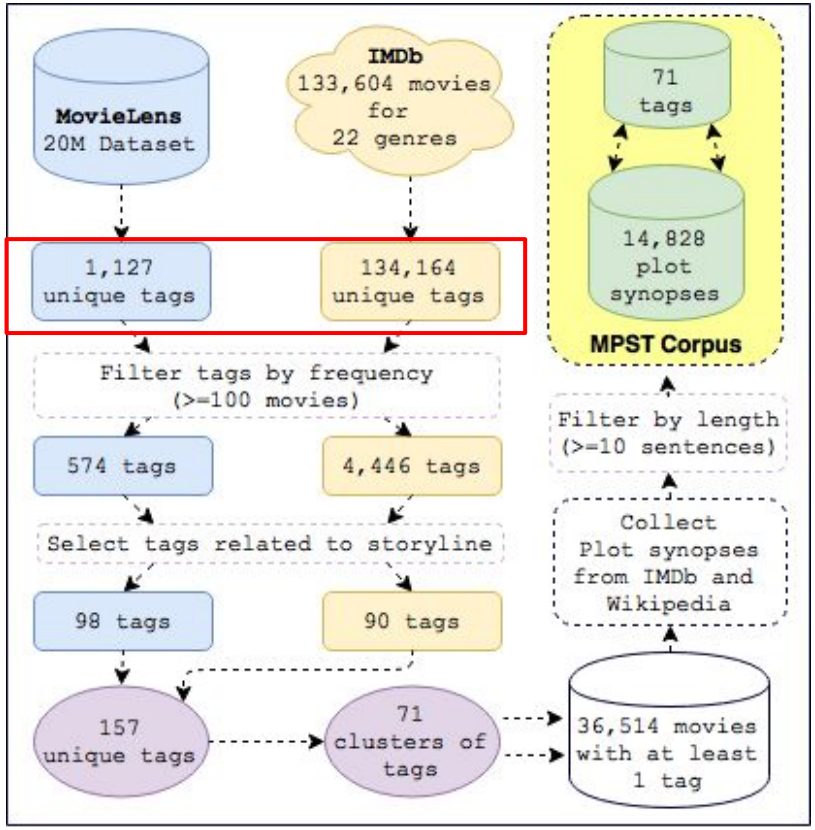


Figure 1: Overview of the data collection process

Data Collection

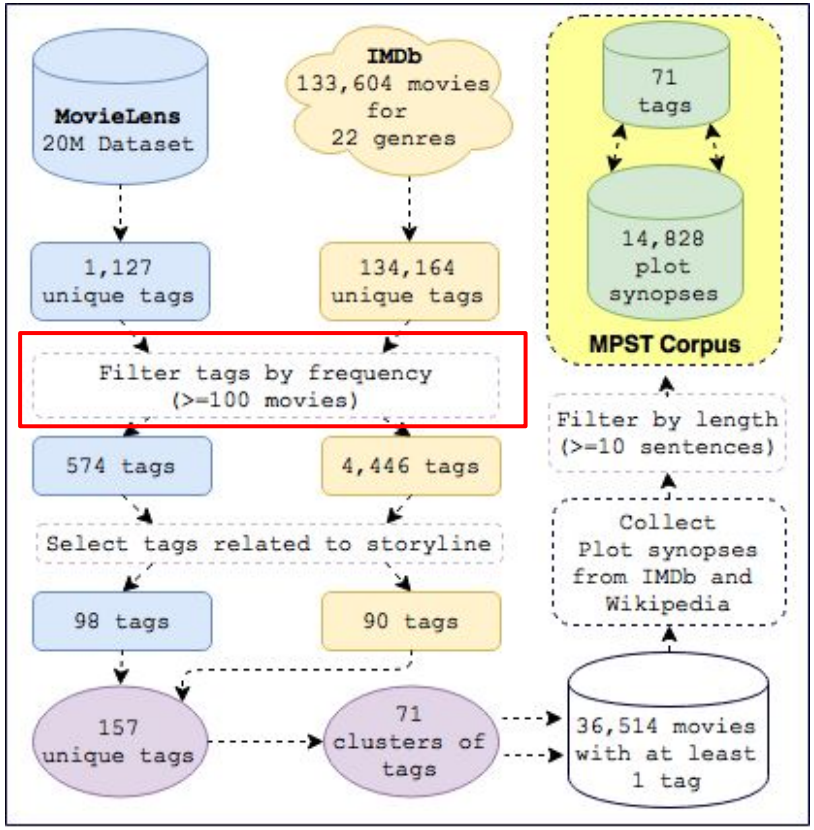


1,127
unique tags

~134K
unique tags

Figure 1: Overview of the data collection process

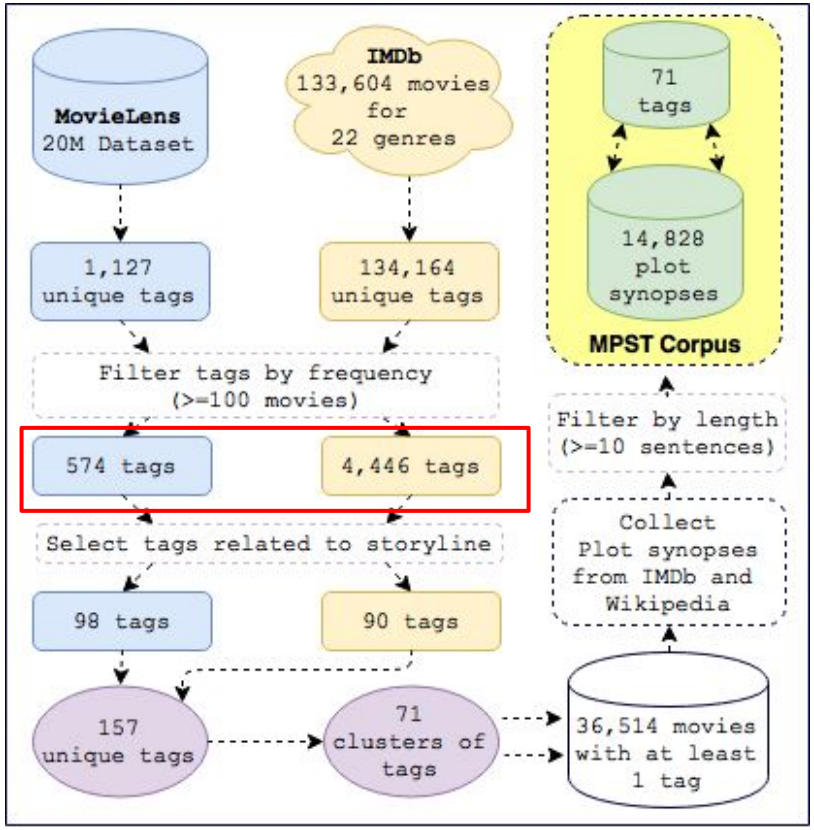
Data Collection



Filter tags by frequency
(>=100 movies)

Figure 1: Overview of the data collection process

Data Collection

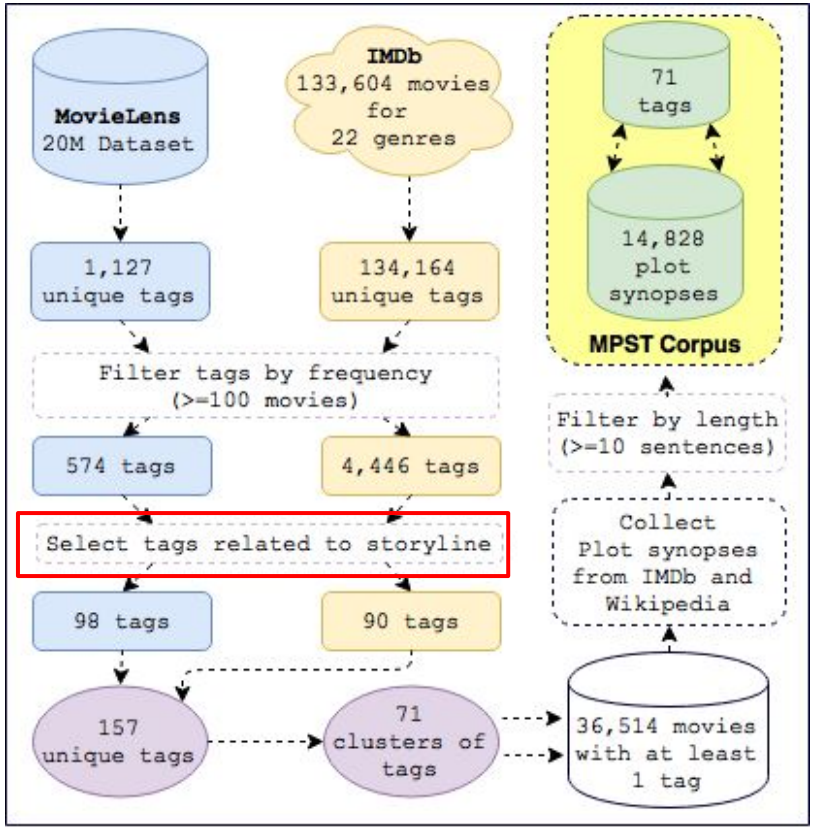


574 tags

~4K tags

Figure 1: Overview of the data collection process

Data Collection



Select tags related to storyline

Figure 1: Overview of the data collection process

Data Collection

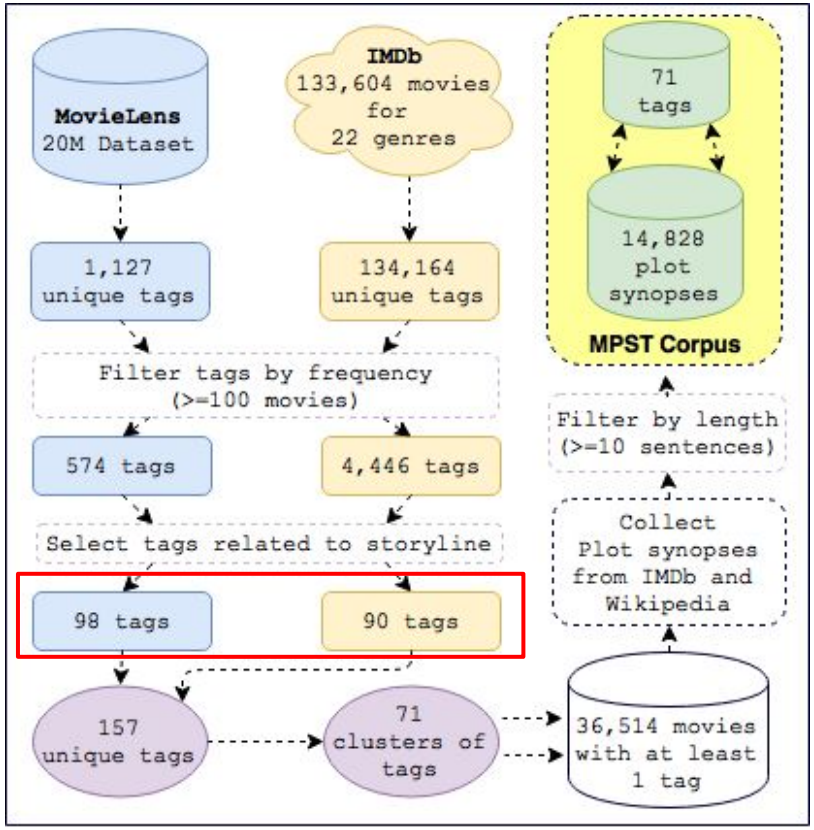


Figure 1: Overview of the data collection process

Data Collection

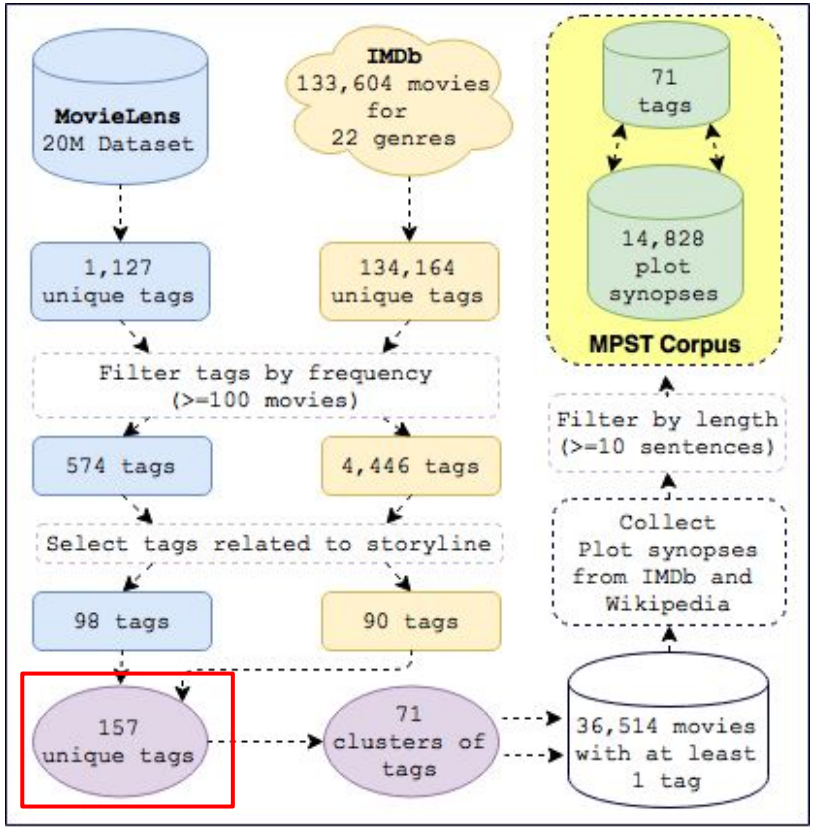


Figure 1: Overview of the data collection process

Data Collection

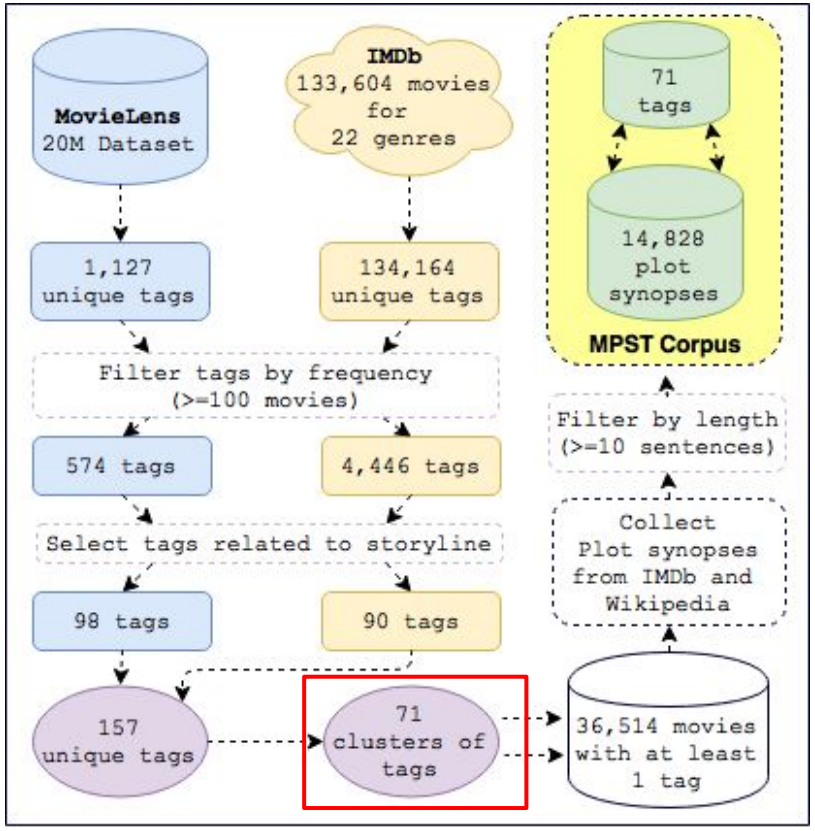
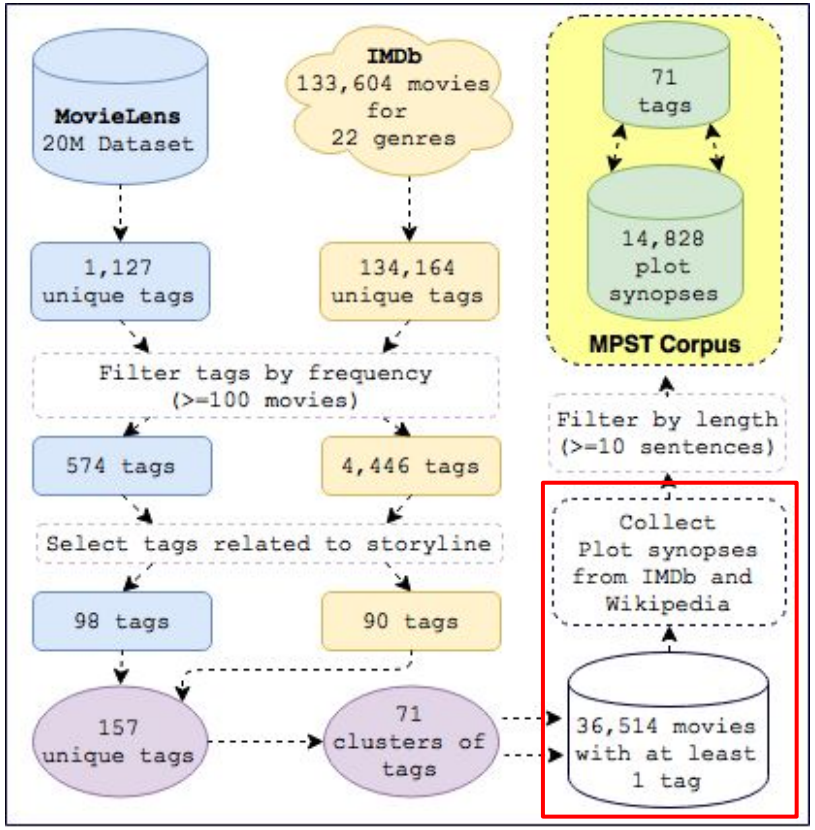


Figure 1: Overview of the data collection process

Data Collection



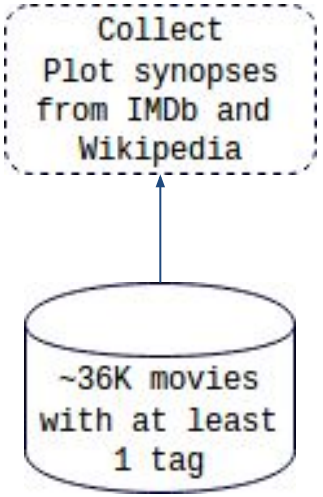
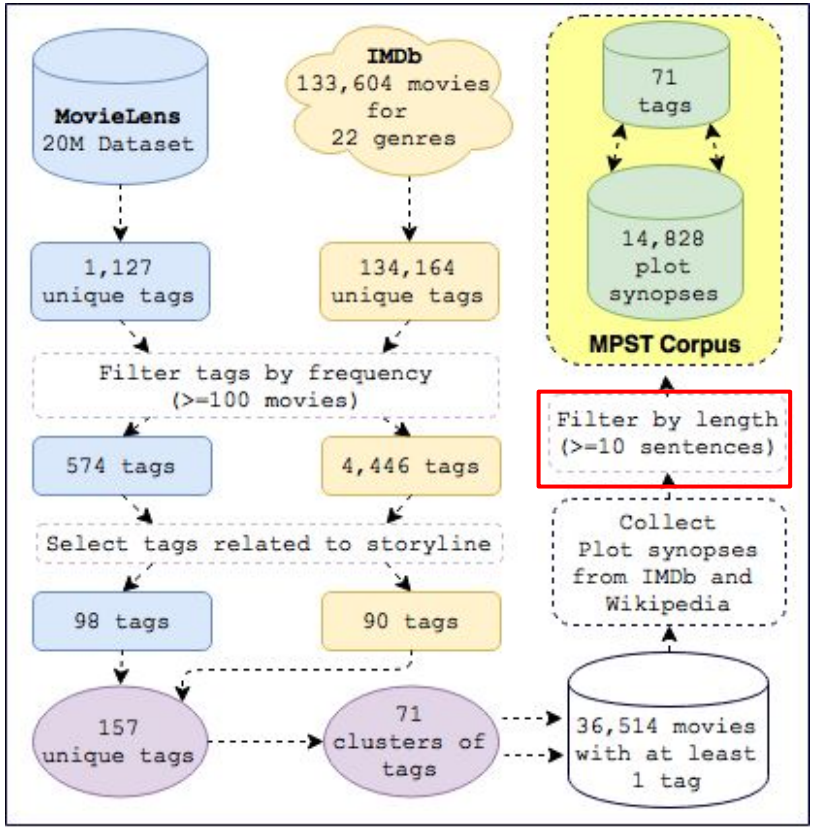


Figure 1: Overview of the data collection process

Data Collection



Filter by length
(>=10 sentences)

Figure 1: Overview of the data collection process

Finally...

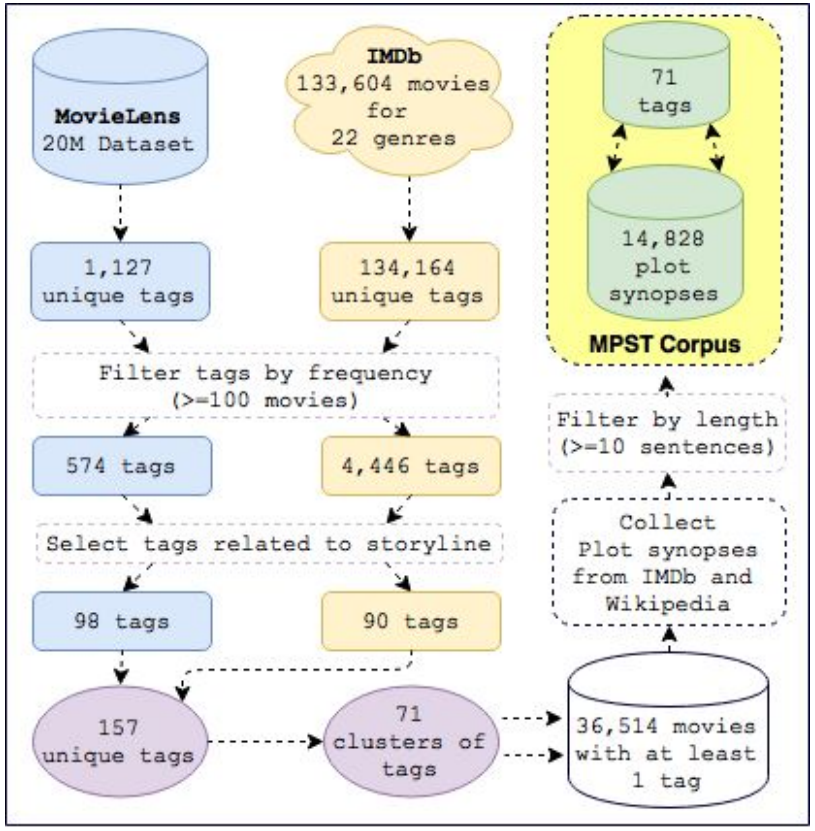


Figure 1: Overview of the data collection process

- Plot synopses of 14,828 movies
- 71 fine-grained tags
- One or more tags assigned to each movie

Fine-grained Tags

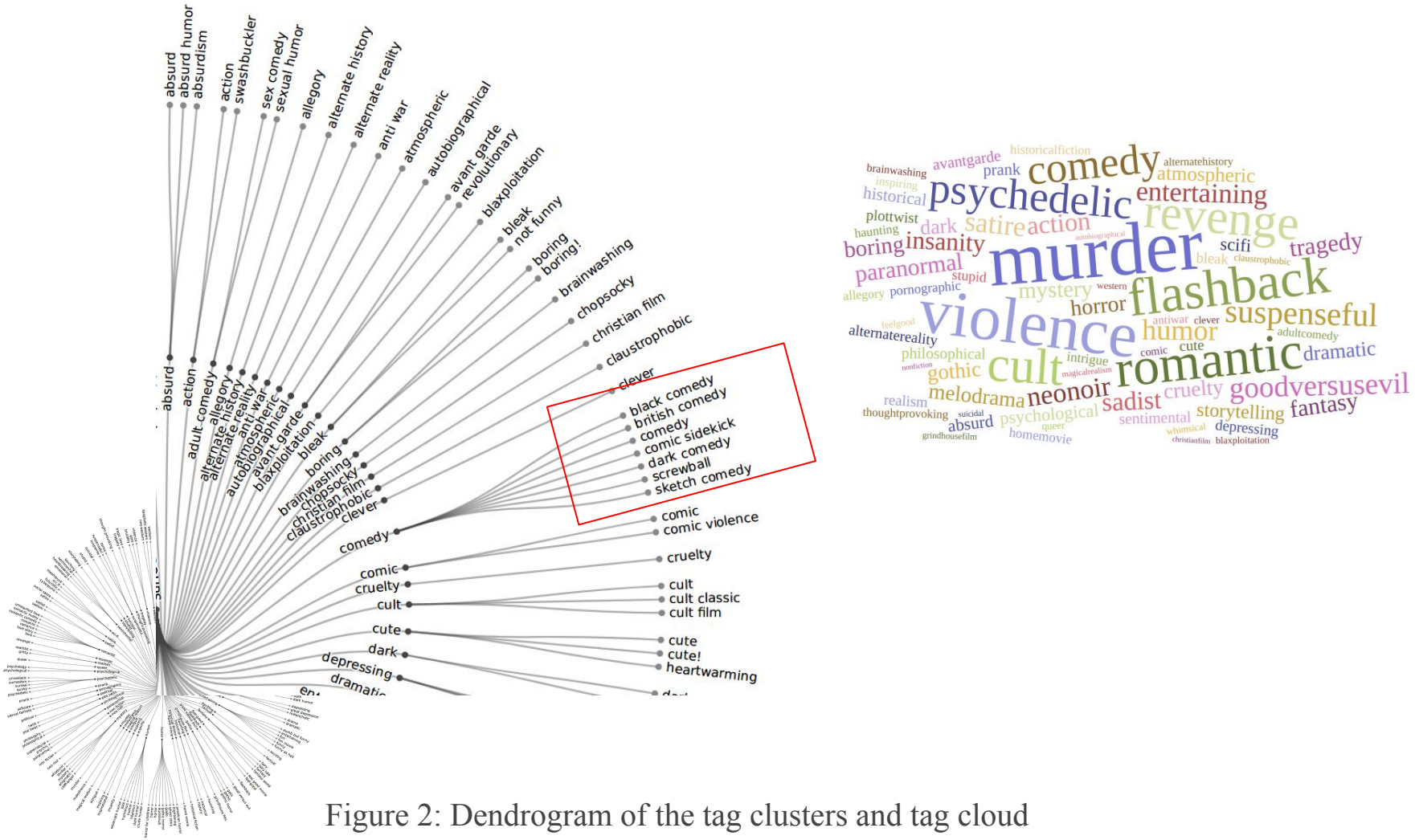


Figure 2: Dendrogram of the tag clusters and tag cloud

Overview of the Data

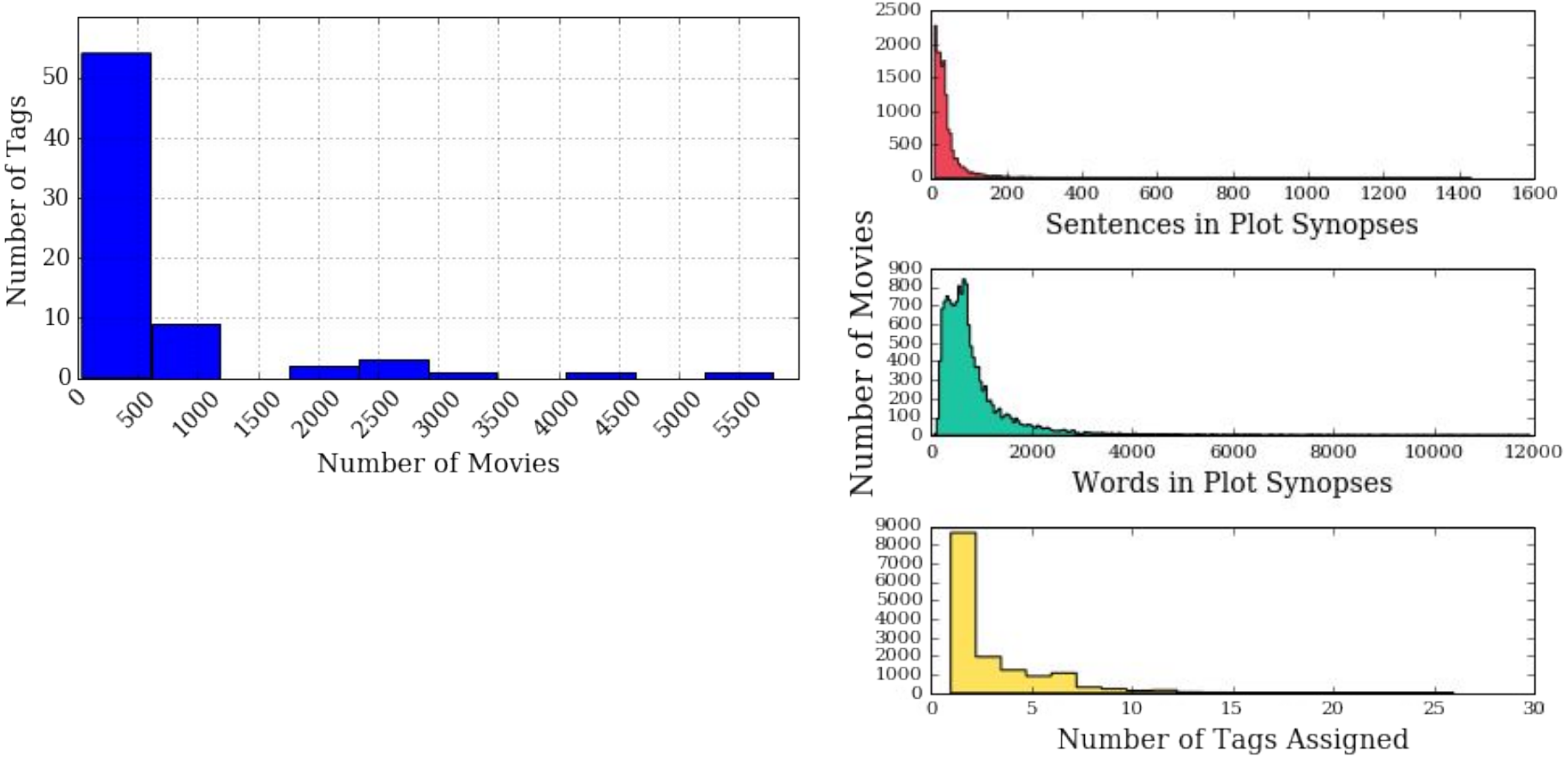


Figure 3: Number of movies for each tag and brief statistics of the data

Correlation Between Tags

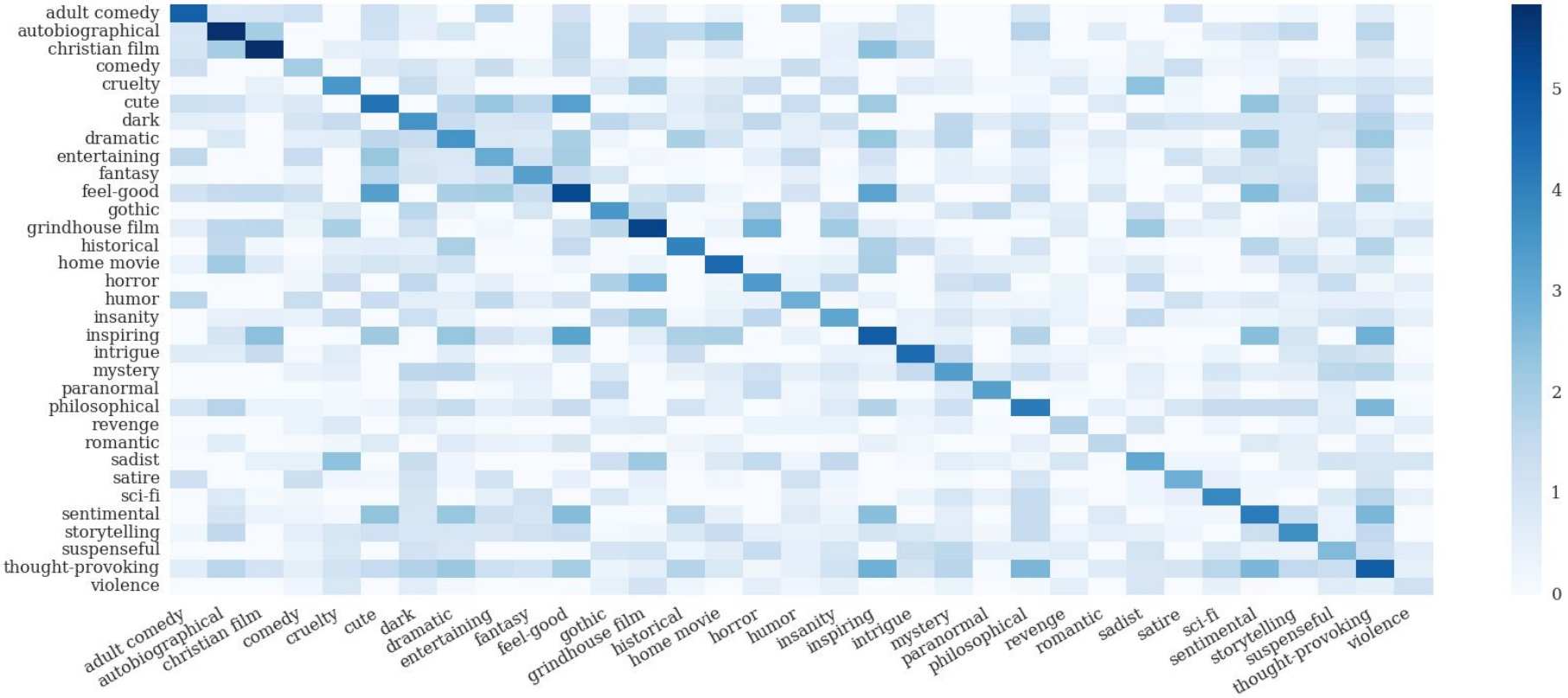


Figure 4: Heatmap of Positive Pointwise Mutual Information (PPMI) between the tags. Dark blue squares represent high PPMI, and white squares represent low PPMI.

Correlation Between Tags

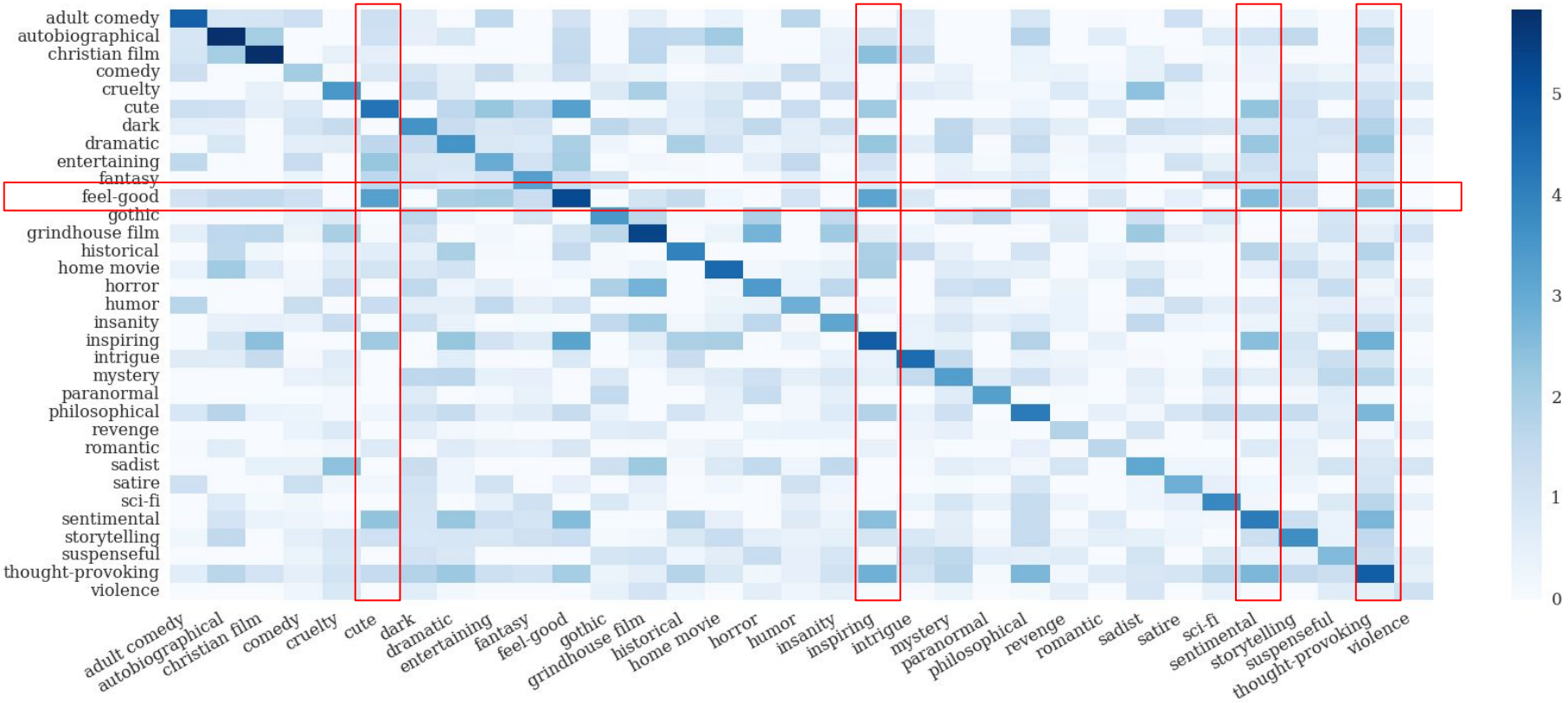


Figure 4: Heatmap of Positive Pointwise Mutual Information (PPMI) between the tags. Dark blue squares represent high PPMI, and white squares represent low PPMI.

Correlation Between Tags

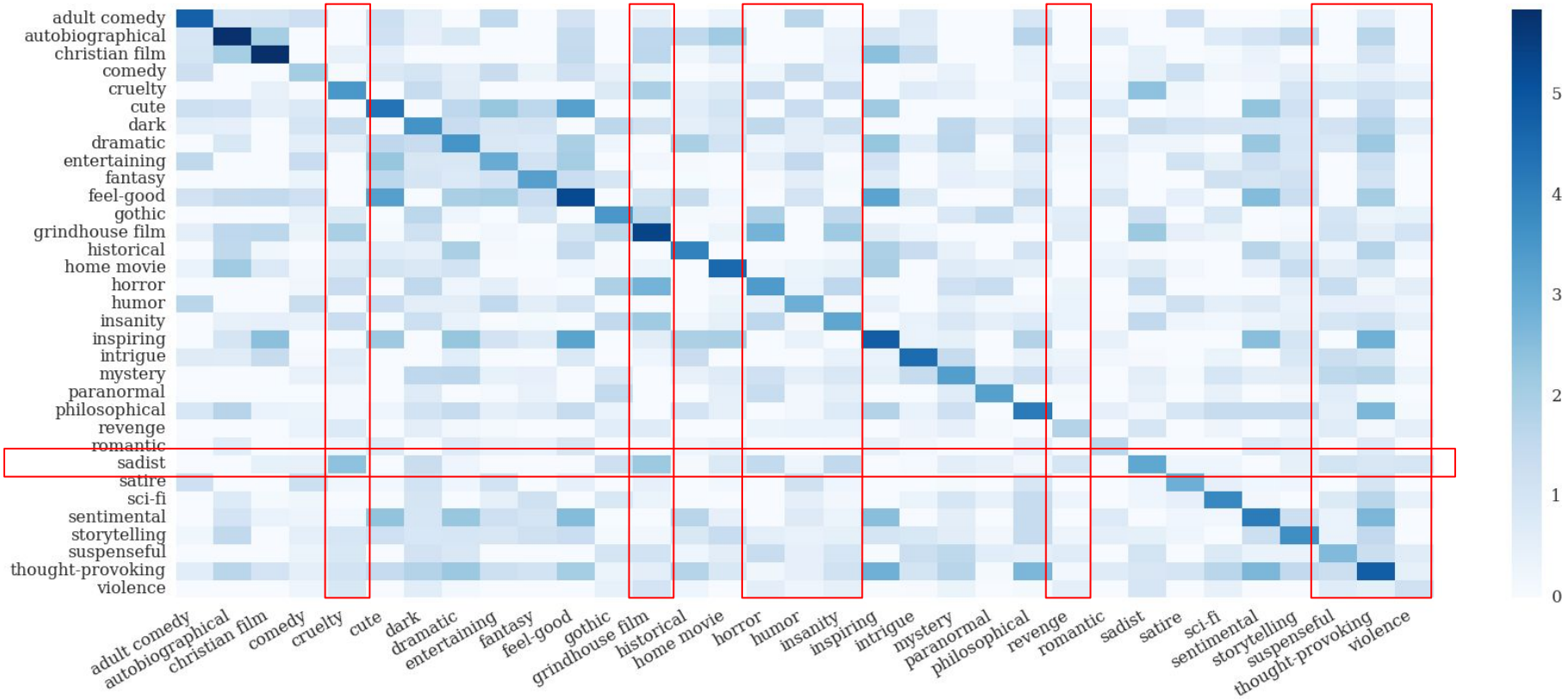
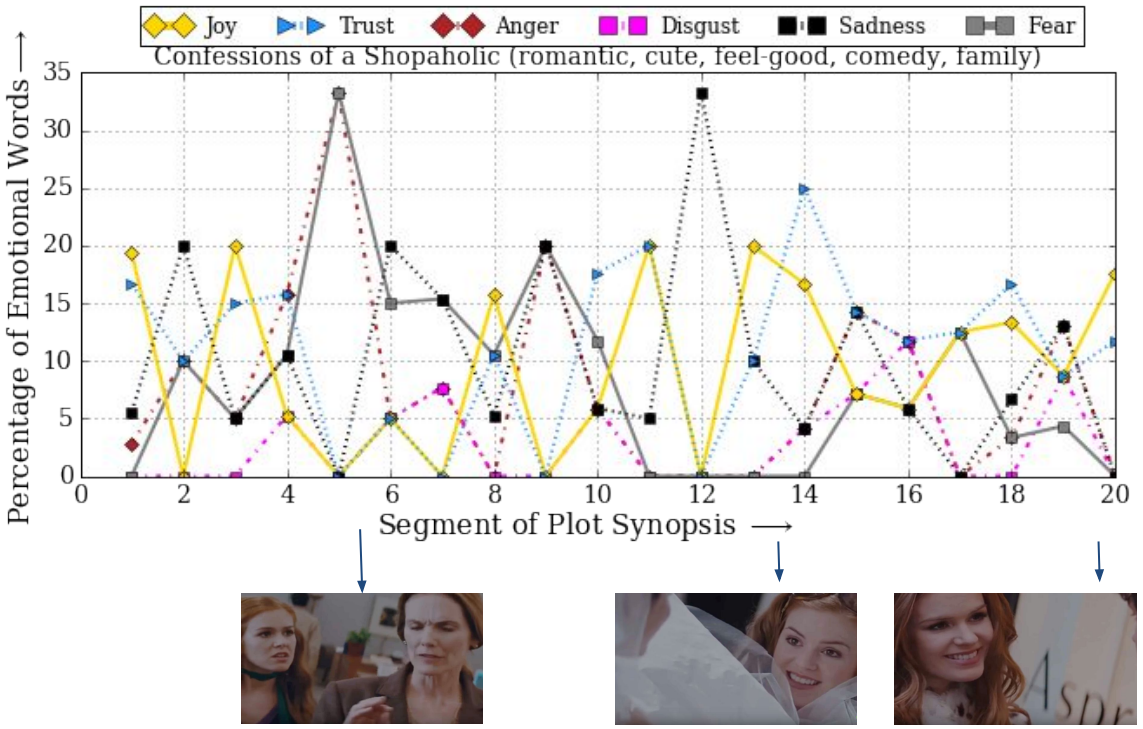


Figure 4: Heatmap of Positive Pointwise Mutual Information (PPMI) between the tags. Dark blue squares represent high PPMI, and white squares represent low PPMI.

Emotion Flow in the Synopses



Joy
winner, love, bridesmaid, kiss, shopping
Trust
wonderful, perfect, real, mother, inspired, defended
Sadness
homeless, lie, debt, upset,
Anger
Stolen, hot, horrible, insulting
Fear
nervous, avoid, homeless, war, danger
Disgust
sick, horrible, lie, thrift, furious, homeless

Figure 5: Tracking flow of emotions in the synopses. Synopsis was divided into equally sized 20 segments based on the words and percentage of the emotions for each segment were calculated using NRC emotion lexicons*. The y axis represents the percentage of emotions in each segment; whereas, the x axis represents the segments.

* Mohammad, Saif and Turney, Peter D. (2010). *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. Association for Computational Linguistics.

Outline

- Introduction
- The MPST Corpus
- Predicting Tags from Synopses
- Experiments and Results
- Conclusion and Future Works

Hand-crafted Features

- Lexical
  - Word n-grams (n=1, 2, 3), char n-grams (n=3, 4, 5), two skip n-grams (n=2, 3)
  - Experimented on minimum document frequency on the training set
- Sentiment and Emotions
  - Bag of Concepts using concept parser
    - e.g. a_lot_of,
  - Affective dimension scores with polarity
    Averaged the scores for synopses
  - Computed scores for three chunks (discussed later)
- Semantic Frames
  - SEMAFOR frame semantic parser to Bag of Framenet frames
- Word Embeddings
  - Averaged for the full synopses
- Agent Verbs and Patient Verbs[1]
  - Mary **shoots** John
  - John was **killed** by Mary
  - 100, **500***, 1000, 1500 clusters of verbs using word embeddings

1. Bamman, D., O'Connor, B., and Smith, N. A. (2013). Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria, August. Association for Computational Linguistics.

Experimental Setup

- Multi-label Classification: OneVsRest approach
- Base classifier: Logistic regression
- Fixed number of tags (3 and 5) for each movie
- Test set: 20% of data (Stratified split)
- Stratified five fold cross validation on the training set to evaluate features
- Tuned regularization parameter *C* on the best combination of features

Experimental Setup (Baselines)

- Majority Baseline
  - Assign the most frequent 3/5 tags (murder, violence, flashback, romantic, cult)
- Random Baseline
  - Assign randomly selected 3/5 tags

Experimental Setup (Evaluation Metrics)

- Multi-label performance are difficult to evaluate!
  - *Which mistake is more significant?*[1]
    "one instance with three incorrect labels **vs.** *three instances each with one incorrect label".*
  - *Less frequent classes could be underrepresented*

- Several evaluation metrics
  - *Hamming loss, average precision, ranking loss, one-error, coverage, micro and macro averaged F1 and AUC scores* [2,3,4,5,6]
  - *Mean per label recall and labels with recall > 0* [7,8,9]

- We use Micro-F1, Tag recall, and Tags learned

1. Wu, X. and Zhou, Z. (2016). A unified view of multi-label performance measures. *CoRR*, abs/1609.00288.
2. Schapire, R. E. and Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2-3):135–168.
3. Fu ̈rnkranz, J., Hu ̈llermeier, E., LozaMenc´ıa, E., and Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, Nov.
4. Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining multi-label data. In *In Data Mining and Knowledge Discovery Handbook*, pages 667–685.
5. Tsoumakas, G., Katakis, I., and Vlahavas, I. (2011). Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089.
6. Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzel, R. C. (2015). Learning to diagnose with LSTM recurrent neural networks. *CoRR*, abs/1511.03677.
7. Lavrenko, V., Manmatha, R., and Jeon, J. (2003). A model for learning the semantics of pictures. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 553–560, Cambridge, MA, USA. MIT Press.
8. Wang, C., Yan, S., Zhang, L., and Zhang, H.-J. (2009). Multi-label sparse coding for automatic image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1643–1650, June.

Results (Training Set)

	Top 3			Top 5		
	F1	**TR**	**TL**	**F1**	**TR**	**TL**
Baseline: Most Frequent	29.7	4.225	3	31.5	7.042	5
Baseline: Random	4.20	4.328	**71**	5.40	7.281	**71**
Unigram (U)	37.6	7.883	22.6	37.1	11.945	27.4
Bigram (B)	36.5	7.216	19.6	36.1	10.808	24.8
Trigram (T)	31.3	5.204	15.4	32.4	8.461	21
Char 3-gram (C3)	37.0	7.419	22	36.6	11.264	27.4
Char 4-gram (C4)	37.7	7.799	22.6	37.0	11.582	27.2
2 skip 2 gram (2S2)	34.2	6.289	19.4	34.5	9.875	25.2
2 skip 3 gram (2S3)	30.8	4.951	12.8	32.1	8.109	18.2
Bag of Concepts (BoC)	35.7	7.984	29	35.9	12.473	34.8
Concepts Scores (CS)	31.1	4.662	7.8	32.4	7.512	8.2
Word Embeddings	36.8	6.744	13.2	36.1	10.074	17.8
Semantic Frame	33.4	5.551	13.4	33.9	8.394	15.2
Agent Verbs	32.9	5.050	7.2	33.2	7.714	8
Patient Verbs	33.1	5.134	7.4	33.5	7.843	8
U+B+T	37.2	8.732	30	36.8	13.576	36.8
C3+C4	**37.8**	8.662	28.8	**37.4**	13.395	33.6
U+B+T+C3+C4	37.1	9.991	36.8	36.8	15.871	45.8
All lexical	36.7	10.046	37.6	36.5	15.838	46.4
BoC + CS	35.7	8.165	29.4	36.0	12.754	35.4
All features	36.9	**10.364**	39.6	36.8	**16.271**	47.8

Table 1: Performance of the hand-crafted features using 5-fold cross-validation on the training data.
We use three metrics to evaluate the features.
F1: micro averaged F1
TR: tag recall
TL: tags learned

Results (Test Set)

	Top 3			Top 5		
	F1	**TR**	**TL**	**F1**	**TR**	**TL**
Baseline: Most Frequent	29.7	4.23	3	28.4	14.08	5
Baseline: Random	4.20	4.21	**71**	6.36	15.04	**71**
System	**37.3**	**10.52**	47	**37.3**	**16.77**	52

Table 2: Performance of the model on test set.

Chunk Based Sentiment Representation

- Motivation is to capture the ups and downs
- *n* chunks of texts based on words
- Sentiment features for each chunk

Chunks	Top 3			Top 5		
	F1	TR	TL	F1	TR	TL
1	35.2	6.550	18.2	35.1	9.928	23.4
2	35.0	7.031	23.0	35.2	10.68	26.8
3	**35.7**	8.165	29.4	**36.0**	**12.754**	35.4
4	35.1	8.153	30.6	35.4	12.723	**36.8**
5	34.8	**8.185**	30.4	35.1	12.553	**36.8**
6	34.3	7.976	**31.2**	34.9	12.725	36.0

Table 3: Experimental results obtained by 5-fold cross-validation using chunk-based sentiment representations. Chunk-based sentiment features were combined with the other features

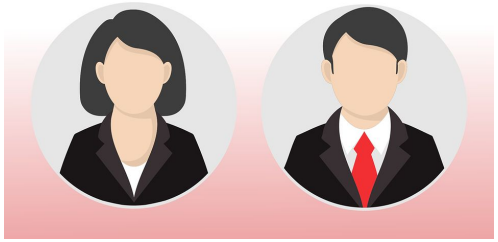
Conclusions

- Task of predicting tags for narrative texts like movie plots
- New dataset for the problem
- Feature based approach to predict tags

Future Work

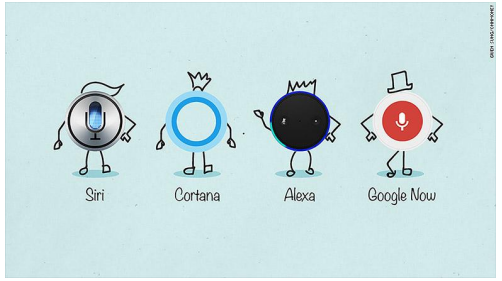
- Tackling incompleteness problem
- Better evaluation methodologies
- More sophisticated models to represent synopses

The Future is Interesting!!!



Hey Alexa/ Siri/ Google/ Cortana, I want to watch **The Nightmare at Elm Street**. How's the movie?

It's a **horror fantasy**. Please, keep your children away.
It shows **violence** and **murder**



ありがとうございました

Thank You

Find MPST Corpus @ http://ritual.uh.edu/mpst-2018/

ありがとうございました

Thank You

Find MPST Corpus @ http://ritual.uh.edu/mpst-2018/

Hand-crafted Features

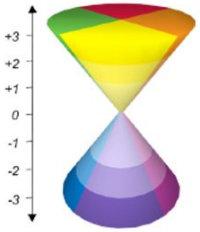
- Lexical
  - Word n-grams (n=1, 2, 3), char n-grams (n=3, 4, 5), two skip n-grams (n=2, 3)
  - Experimented on minimum document frequency on the training set

Feature	Size	Minimum Document Frequency
Unigrams	4900	100
Bigrams	5902	100
Trigrams	473	100
Char 3-grams	5598	150
Char 4-grams	19801	150
2 skip 2 grams	1382	100
2 skip 2 grams	338	100

Hand-crafted Features

- Sentiments and Emotions
  - Bag of Concepts using concept parser[1]
    - e.g. a_lot_of,
  - Affective dimension scores with polarity
    - Attention
    - Aptitude
    - Pleasantness
    - Sensitivity

    Averaged the scores for synopses
  - Computed scores for three chunks (discussed later)



	Pleasantness	Attention	Sensitivity	Aptitude
+3	ecstasy	vigilance	rage	admiration
+2	joy	anticipation	anger	trust
+1	serenity	interest	annoyance	acceptance
0	—	—	—	—
-1	pensiveness	distraction	apprehension	boredom
-2	sadness	surprise	fear	disgust
-3	grief	amazement	terror	loathing

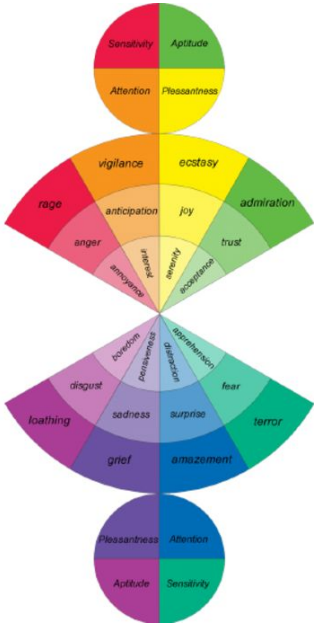
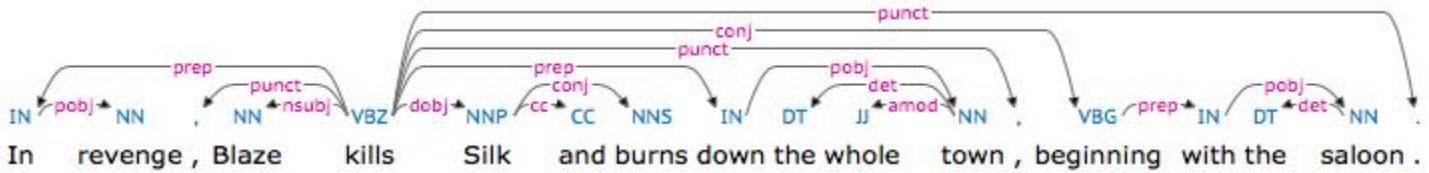


Figure 6: The hourglass of emotions[2]

1. Rajagopal, D., Cambria, E., Olsher, D., and Kwok, K. (2013). A graph-based approach to commonsense concept extraction and semantic similarity detection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 565–570. ACM.
2. Cambria, E. (2013). An introduction to concept-level sentiment analysis. In *Mexican International Conference on Artificial Intelligence*, pages 478–483. Springer.

Hand-crafted Features

- Semantic Frames
  - SEMAFOR[1] frame semantic parser to parse Framenet frames[2]
  - Modeled as bag of frames weighted by normalized frequency
  - >1200 unique frames



1. http://www.cs.cmu.edu/~ark/SEMAFOR
2. Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics- Volume 1*, pages 86–90. Association for Computational Linguistics.

Hand-crafted Features

- Word Embeddings
  - Experimented with 300 dimensional word vectors from Word2Vec (Google News) and FastText (Wikipedia)*
  - Averaged for the full synopses
- Agent Verbs and Patient Verbs[1]
  - Mary **shoots** John
  - John was **killed** by Mary
  - CoreNLP[3] to parse dependencies
  - Agent Verbs (*nsubj* or *agent*) ~23K
  - Patient Verbs (*dobj, nsubjpass, iobj)* ~20K
  - 100, **500***, 1000, 1500 clusters of verbs using word embeddings

1. Bamman, D., O'Connor, B., and Smith, N. A. (2013). Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria, August. Association for Computational Linguistics.