Predicting Success of Literary Writings

Mohit Raman, Raj Kansagra, Ajinkya Mandhare, Sri Harsha Anand Pushkala
{ mohitram | rkansagr | mandhare | anandpus } @usc.edu

Introduction

- Examine the quantitative connection between writing style and success of a novel
- Perform a comparative study of 4 NLP Techniques
- Perform statistical analysis to identify the NLP technique best suited for a genre

Who will benefit from this?

- Authors, publication houses, script writers, literary agents

Data Annotation

- Constructed a Web Scraper that mined the Gutenberg pages
- Extracted the meta-data for each book
- Meta-data was used to annotate the training data as successful or not

- Computing Threshold:
  0.7 * (No. of Downloads/Recency of File) + 0.3 * (No. of Votes)

Approach I: n-Grams/ Syntactic n-Grams

- Evaluate the role of commonly occurring patterns(nGrams) in Novel success
- Improve by considering the neighbours taken by following path in syntactic trees
- Syntactic nGrams provide better results by identifying related patterns
  eg: “The strongest rain ever recorded”

Approach II: Sentiment Distribution

- Examine if Sentiment Distribution has any correlation with success of Literary work
- Key chapters of a Novel: First few, Middle few and last few
- Features: Sentiments from the key chapters

Success Neu_1 Pos_2 Neg_3….Neg_45 Pos_46….Neg_98 Pos_99
Failure Neu_1 Neu_2 Neg_3….Pos_45 Pos_46….Neg_98 Neu_99
….
Success Neu_1 Pos_2 Neg_3….Neg_45 Pos_46….Neg_98 Pos_99

Approach III: PCFG

- Approach I:
- Created features from CFG rules for each sentence for classification
- Sparse feature space resulted in bad results

- Approach II:
- Used lexical parsing to get PCFGs (G_{succ},$G_{nonsucc}$) from training set for each genre
- Use PCFGs for classification
- Decent results for relatively new books of famous genres

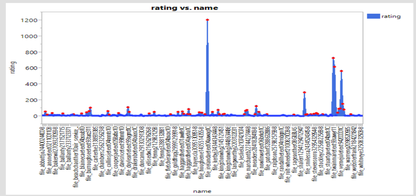
Approach IV: Artificial Neural Networks

Constructed a 5 layer Artificial Neural Network to analyse some statistical data related to the books. Some of the features extracted are:

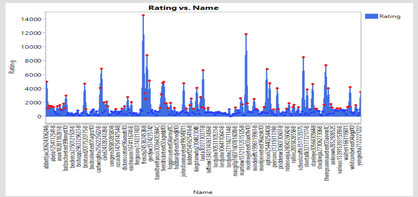
- The richness of the text
- Distribution of verbal clauses
- Ease of reading
- Average length of the line
- Number of Downloads

Threshold plot for Data Annotation

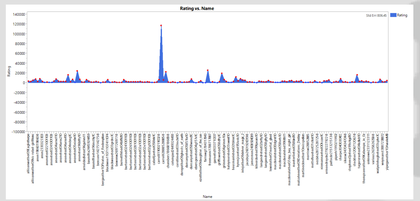
Fiction



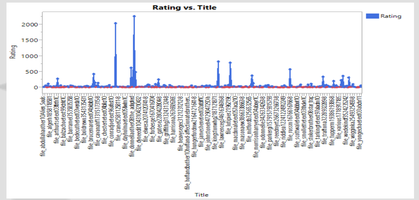
Drama



Adventure

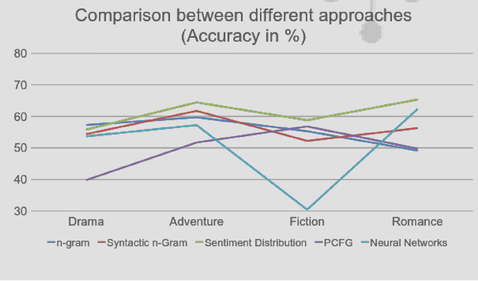


Romance



Comparison between the different approaches (Accuracy in %)

APPROACH	GENRE				Avg
	Adventure	Fiction	Drama	Romance	
n-Gram	59.7	55.3	57.26	49.18	55.36
Syntactic N-Grams	61.73	52.24	54.38	56.27	56.15
Sentiment Distribution	64.44	58.77	55.8	65.33	61.08
PCFG	51.73	56.81	39.8	49.78	49.39
Neural Networks	57.23	30.45	53.69	62.33	50.93



Conclusion

- There exists distinct linguistic patterns shared among successful literature
- Statistical stylometry can be effective in discriminating successful literature
- Interesting relation observed between sentiment distribution and literary success
- Comparative insights between the different approaches

Future Work

- Genre-wise sentiment annotated data can increase accuracy of sentiment distribution
- Combine features from different approaches and conduct further comparative studies
- Dynamically selecting best technique depending on genre

References

- Vikas Ganjigunte Ashok, Song Feng and Yejin Choi. *S uccess with Style: Using Writing Style to Predict the Success of Novels*