Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Between the Originality and Popularity: How to Create Successful Digital Content
Principal Investigator:	Minjee Sun
Prepared by (IIAI):	Yanan Liu
Other investigators:	
Date:	
Were specific aims fulfilled:	N
Readiness for extramural proposal?	N
If yes … Planned submission date	
Funding agency	
Grant mechanism	
If no … Why not? What went wrong?	PI plans to use the pipeline developed in this pilot project to extend the number of related analyses to investigate the first aim of this project, prior to any grant submission

Brief summary of accomplished results:

All content and popularity ranking of 5025 novels were collected and the LDA pipeline is developed and implemented for topics analysis.

Research report:

Aims (provided by PI):

Aim 1: Explore the optimal proportion of topics between popularity and originality

Aim 2: Predict the success of a new novel with its key topics

Data:

A cross-sectional dataset: complete content and associated popularity rankings of 5025 novels on https://www.scribblehub.com were collected.

AI/ML Approach:

In this study, topic modeling using Latent Dirichlet Allocation (LDA) was implemented in Python.

Experimental methods, validation approach:

Data Selection

Since different novels contain different chapter numbers, a histogram of chapter number is plotted.

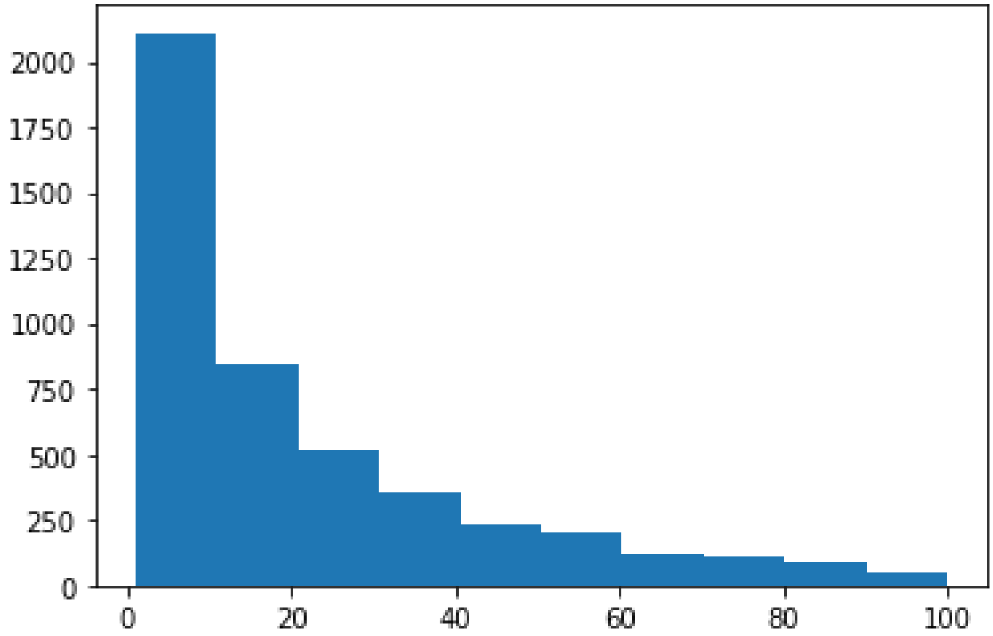


Figure 1. Histogram of chapter numbers.

The histogram clearly shows that many analyzed novels contain less than 10 chapters. Therefore, we chose the first 10 chapters for LDA to build the pipeline.

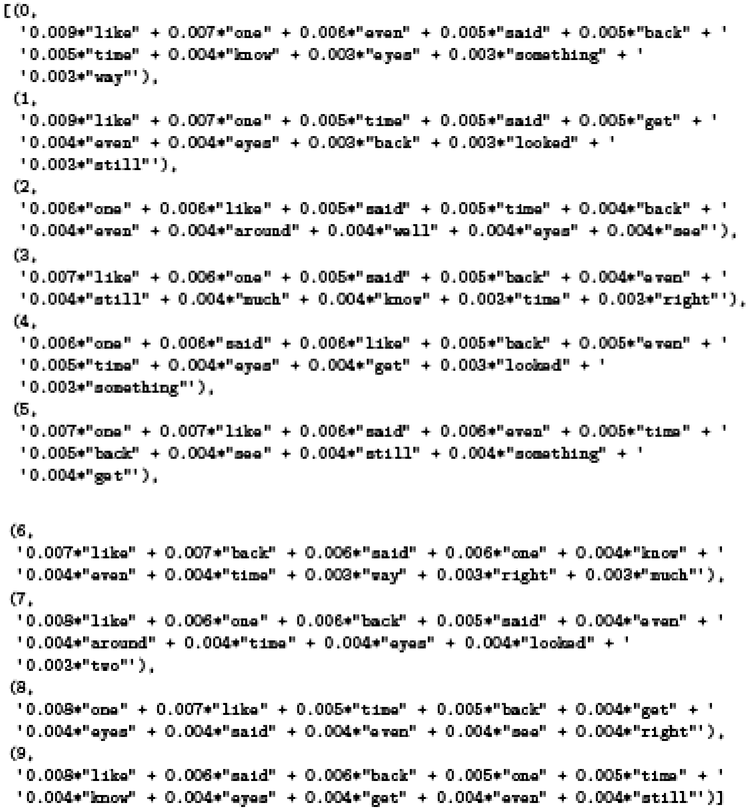
3037 novels that contain more than 10 chapters were selected for LDA analysis.

Data preparation for LDA analysis

Punctuation was removed for all the content of the first 10 chapters of each novel. All text was changed to lowercase. In order to prepare data for LDA, we started by tokenizing the text and removing stopwords. We then converted the tokenized object into a corpus and dictionary.

LDA model training

For the pipeline, we built a model with 10 topics where each topic is a combination of keywords, and each keyword contributes a certain weightage to the topic. The result is shown in Figure 2.



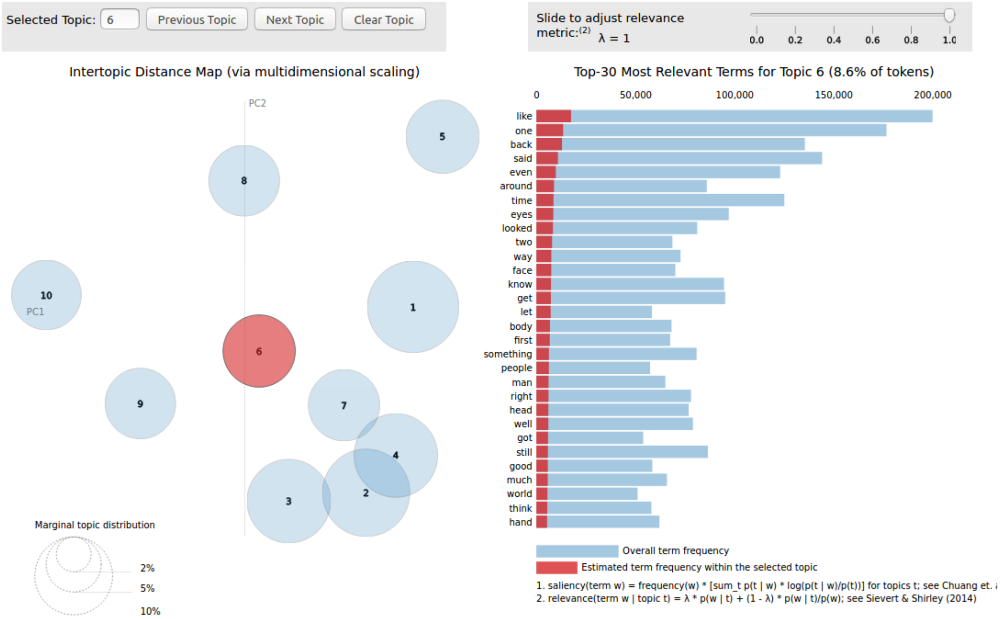
```
[(0,
  '0.009*"like" + 0.007*"one" + 0.006*"even" + 0.005*"said" + 0.005*"back" + '
  '0.005*"time" + 0.004*"know" + 0.003*"eyes" + 0.003*"something" + '
  '0.003*"way"'),
 (1,
  '0.009*"like" + 0.007*"one" + 0.005*"time" + 0.005*"said" + 0.005*"get" + '
  '0.004*"even" + 0.004*"eyes" + 0.003*"back" + 0.003*"looked" + '
  '0.003*"still"'),
 (2,
  '0.006*"one" + 0.006*"like" + 0.005*"said" + 0.005*"time" + 0.004*"back" + '
  '0.004*"even" + 0.004*"around" + 0.004*"well" + 0.004*"eyes" + 0.004*"see"'),
 (3,
  '0.007*"like" + 0.006*"one" + 0.005*"said" + 0.005*"back" + 0.004*"even" + '
  '0.004*"still" + 0.004*"much" + 0.004*"know" + 0.003*"time" + 0.003*"right"'),
 (4,
  '0.006*"one" + 0.006*"said" + 0.006*"like" + 0.005*"back" + 0.005*"even" + '
  '0.005*"time" + 0.004*"eyes" + 0.004*"get" + 0.003*"looked" + '
  '0.003*"something"'),
 (5,
  '0.007*"one" + 0.007*"like" + 0.006*"said" + 0.006*"even" + 0.005*"time" + '
  '0.005*"back" + 0.004*"see" + 0.004*"still" + 0.004*"something" + '
  '0.004*"get"'),


 (6,
  '0.007*"like" + 0.007*"back" + 0.006*"said" + 0.006*"one" + 0.004*"know" + '
  '0.004*"even" + 0.004*"time" + 0.003*"way" + 0.003*"right" + 0.003*"much"'),
 (7,
  '0.008*"like" + 0.006*"one" + 0.006*"back" + 0.005*"said" + 0.004*"even" + '
  '0.004*"around" + 0.004*"time" + 0.004*"eyes" + 0.004*"looked" + '
  '0.003*"two"'),
 (8,
  '0.008*"one" + 0.007*"like" + 0.005*"time" + 0.005*"back" + 0.004*"get" + '
  '0.004*"eyes" + 0.004*"said" + 0.004*"even" + 0.004*"see" + 0.004*"right"'),
 (9,
  '0.008*"like" + 0.006*"said" + 0.006*"back" + 0.005*"one" + 0.005*"time" + '
  '0.004*"know" + 0.004*"eyes" + 0.004*"get" + 0.004*"even" + 0.004*"still"')]
```

Figure 2. 10 topics of LDA analysis

<u>LDA model results visualization</u>

A popular visualization package, pyLDAvis was used to help interactively with a better understanding and interpreting individual topics, and a better understanding the relationships between the topics.


Selected Topic: 6

Slide to adjust relevance metric:[2] λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 6 (8.6% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. a
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. LDA visualization

Validation

With the initial input, some adjustments for parameters and stopwords seem necessary to get a meaningful result. The PI is working on it after this cooperation.

Results:

The desired pipelines were developed and provided to the PI who will employ them directly to fulfill the stated aims.

Ideas/aims for future extramural project:

PI plans to use the pipeline developed in this pilot project to extend the number of related analyses to investigate the first aim of this project, prior to any grant submission

Publications resulting from project:

None