1. Techniques Used for Vectorization and Embedding

- **TF-IDF Vectorization:**
  Converts preprocessed review text into a sparse feature matrix, where each column represents a word or n-gram weighted by term frequency and inverse document frequency. This helps highlight important words while reducing the impact of common but less informative words.
- **Word Embeddings (Word2Vec):**
  Trains distributed vector representations of words using the CBOW (Continuous Bag of Words) approach on the review corpus. Each word is mapped to a dense vector in a 100-dimensional space capturing semantic similarity. Review vectors are generated by averaging their word vectors.

2. Configuration and Tuning of TF-IDF Parameters

- **N-gram Range:** (1, 2) to include both unigrams (single words) and bigrams (two consecutive words), capturing some context.
- **Max Features:** Limited to 1000 to control dimensionality and reduce sparsity.
- **Min Document Frequency (min_df=5):** Ignores terms appearing in fewer than 5 documents to remove rare words.
- **Max Document Frequency (max_df=0.8):** Removes very common terms that appear in more than 80% of documents.

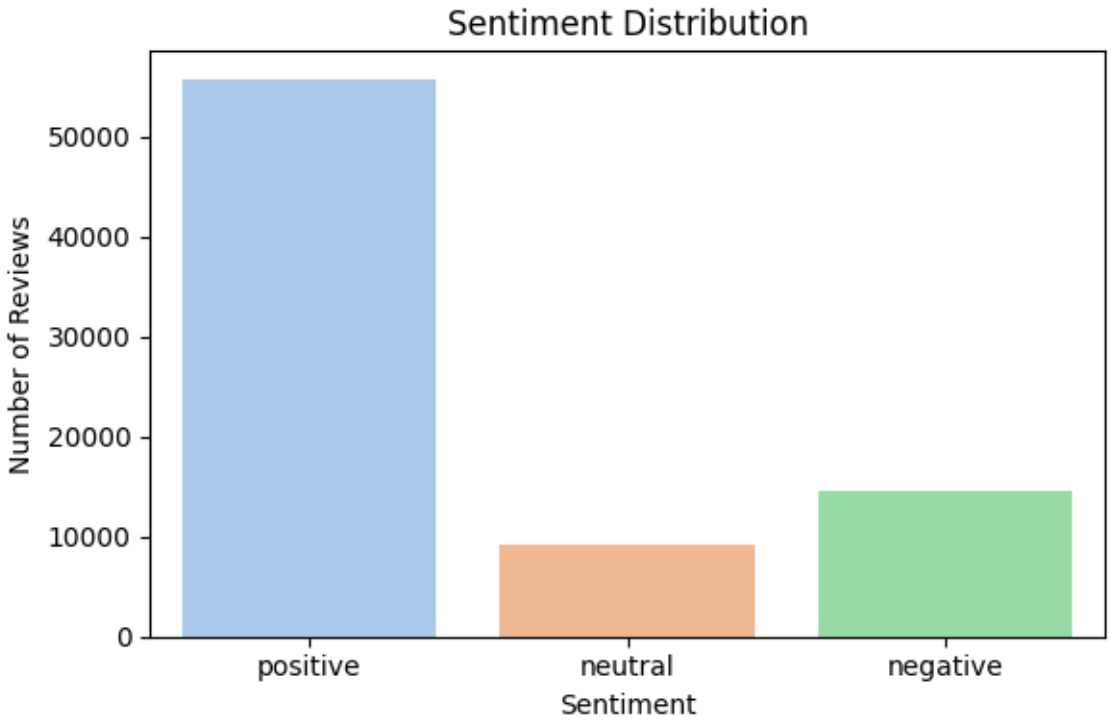
3. Comparison: TF-IDF vs. Word Embeddings

Aspect	TF-IDF	Word Embeddings (Word2Vec)
Dimensionality	Sparse, high-dimensional (e.g., 1000 features)	Dense, fixed lower-dimensional (100 dims)
Interpretability	Easy — each feature corresponds to a word/ngram	Harder — features are latent semantic dimensions
Captures Semantics	Limited — based on word frequency statistics	Better — captures semantic similarity & context
Sparsity	High sparsity (e.g., 95%)	Dense vectors (no sparsity)
Suitability	Good for interpretable, frequency-based models	Good for capturing semantic meaning, downstream ML

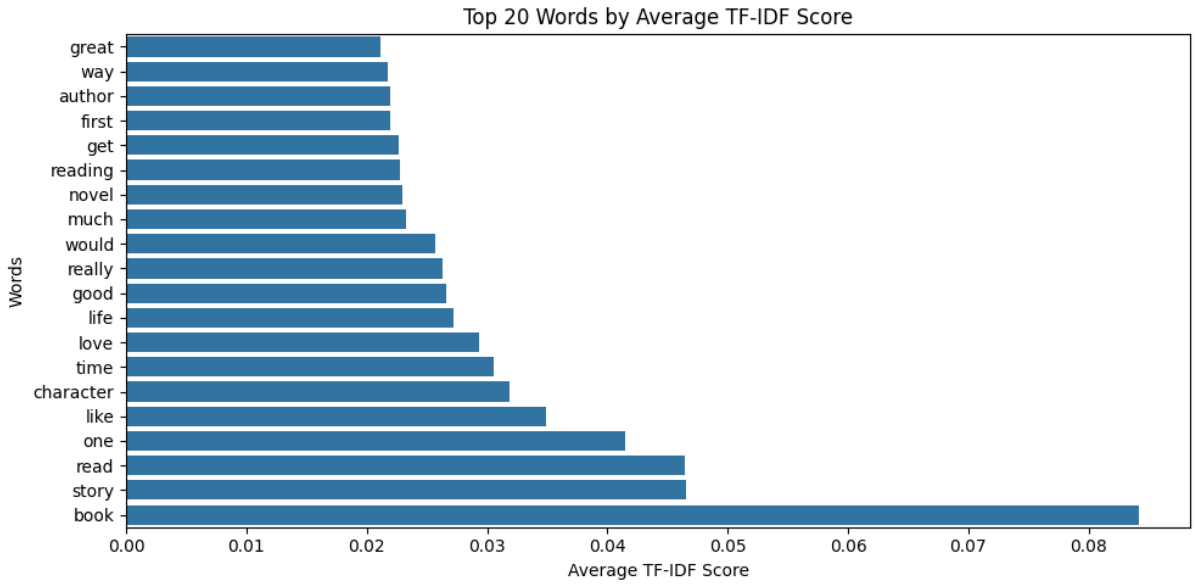
4. Sentiment Labeling Logic & Distribution

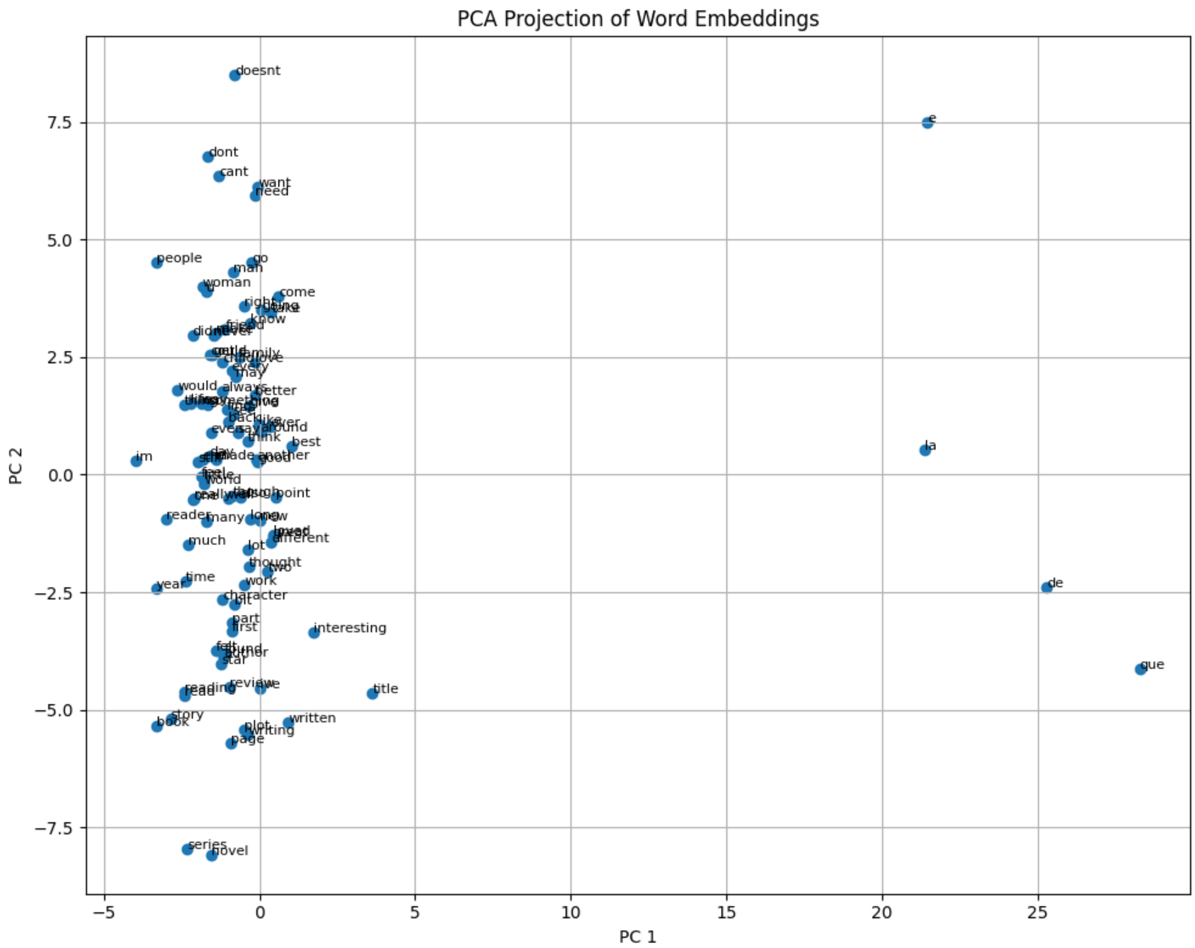
- **If Rating Exists:**
  Labels assigned based on numeric rating:
  - rating >= 4 → Positive
  - rating <= 2 → Negative
  - Others → Neutral

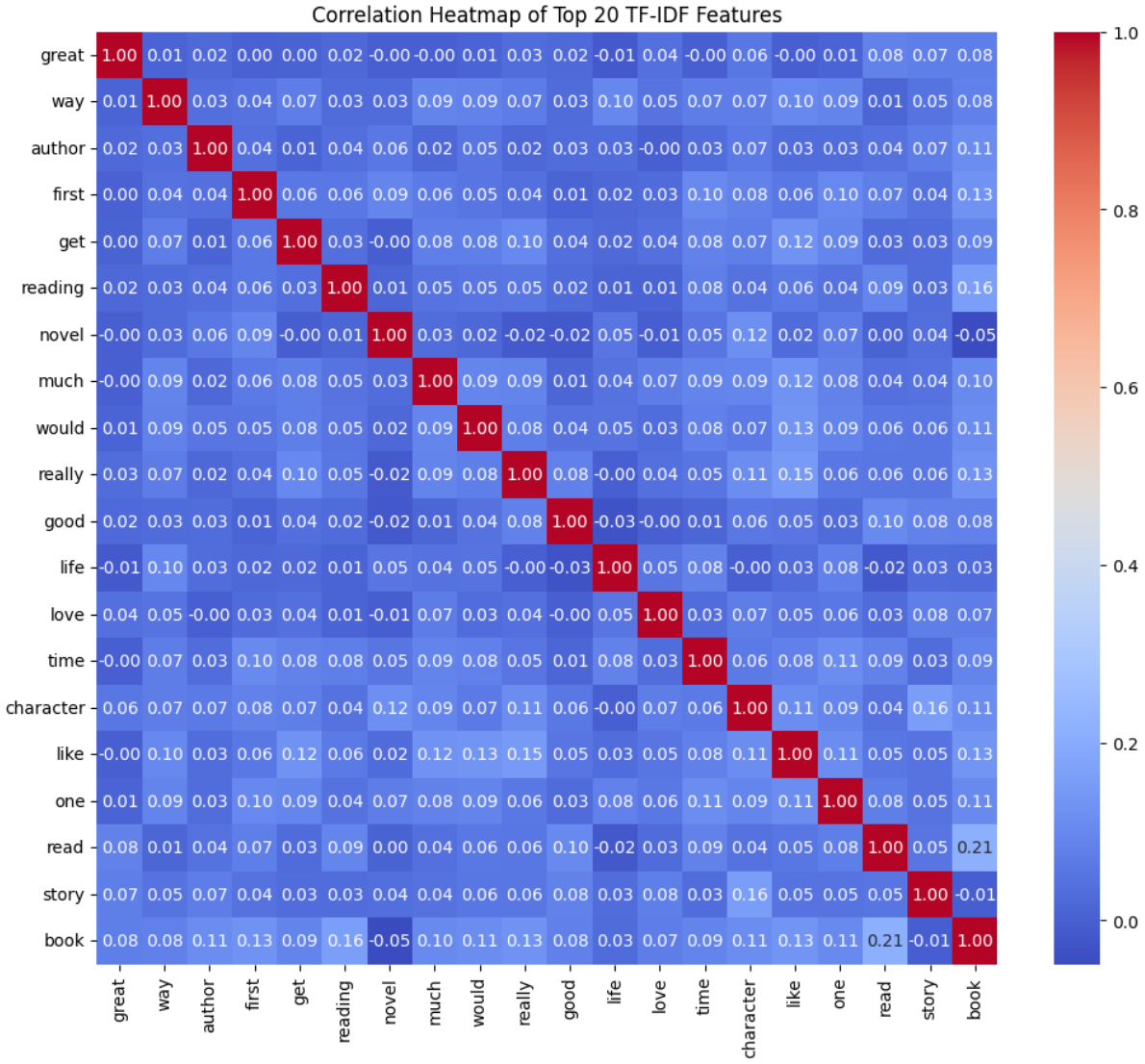
- **If No Rating:**
  Used VADER sentiment lexicon to compute polarity scores and label reviews:
  - Compound score ≥ 0.05 → Positive
  - Compound score ≤ -0.05 → Negative
  - Otherwise → Neutral
- **Label Distribution:**

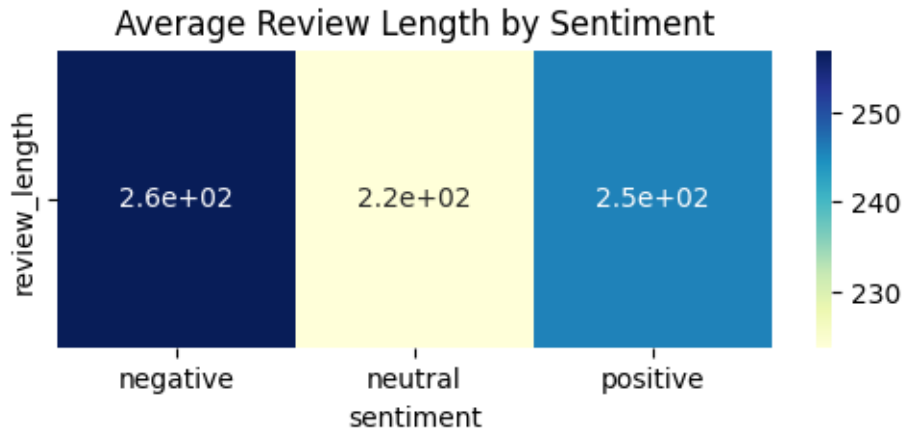


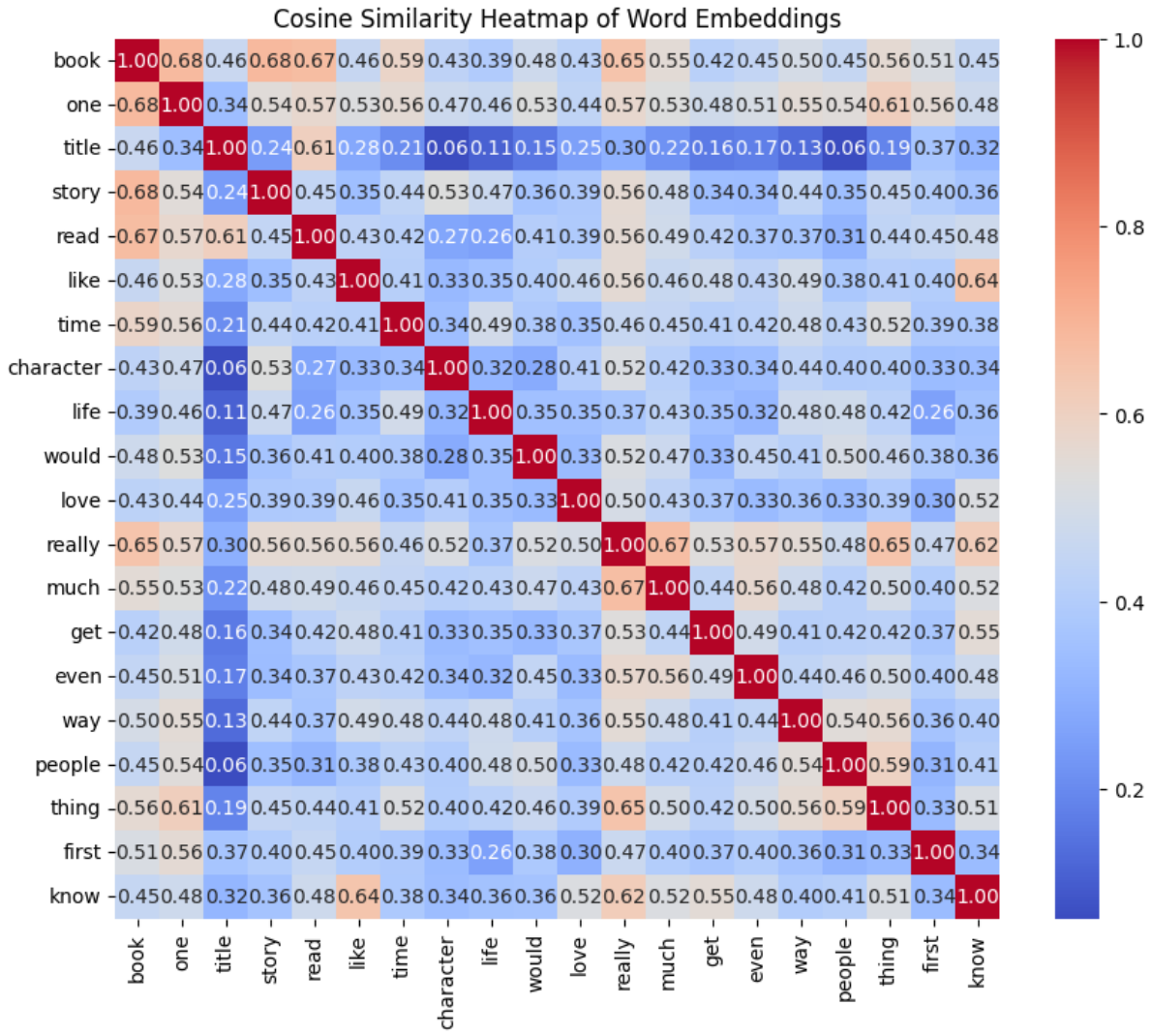
5. **Sample Visualizations**




PCA Projection of Word Embeddings
PC 1
PC 2
−5
0
5
10
15
20
25
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
doesnt
e
dont
cant
want
need
people
go
man
woman
come
know
did
la
im
best
another
good
point
reader
many
much
different
thought
time
year
work
character
de
part
first
interesting
star
que
review
title
read
story
book
plot
written
writing
page
series
novel


Correlation Heatmap of Top 20 TF-IDF Features
great 1.00 0.01 0.02 0.00 0.00 0.02 -0.00 -0.00 0.01 0.03 0.02 -0.01 0.04 -0.00 0.06 -0.00 0.01 0.08 0.07 0.08
way 0.01 1.00 0.03 0.04 0.07 0.03 0.03 0.09 0.09 0.07 0.03 0.10 0.05 0.07 0.07 0.10 0.09 0.01 0.05 0.08
author 0.02 0.03 1.00 0.04 0.01 0.04 0.06 0.02 0.05 0.02 0.03 0.03 -0.00 0.03 0.07 0.03 0.03 0.04 0.07 0.11
first 0.00 0.04 0.04 1.00 0.06 0.06 0.09 0.06 0.05 0.04 0.01 0.02 0.03 0.10 0.08 0.06 0.10 0.07 0.04 0.13
get 0.00 0.07 0.01 0.06 1.00 0.03 -0.00 0.08 0.08 0.10 0.04 0.02 0.04 0.08 0.07 0.12 0.09 0.03 0.03 0.09
reading 0.02 0.03 0.04 0.06 0.03 1.00 0.01 0.05 0.05 0.05 0.02 0.01 0.01 0.08 0.04 0.06 0.04 0.09 0.03 0.16
novel -0.00 0.03 0.06 0.09 -0.00 0.01 1.00 0.03 0.02 -0.02 -0.02 0.05 -0.01 0.05 0.12 0.02 0.07 0.00 0.04 -0.05
much -0.00 0.09 0.02 0.06 0.08 0.05 0.03 1.00 0.09 0.09 0.01 0.04 0.07 0.09 0.09 0.12 0.08 0.04 0.04 0.10
would 0.01 0.09 0.05 0.05 0.08 0.05 0.02 0.09 1.00 0.08 0.04 0.05 0.03 0.08 0.07 0.13 0.09 0.06 0.06 0.11
really 0.03 0.07 0.02 0.04 0.10 0.05 -0.02 0.09 0.08 1.00 0.08 -0.00 0.04 0.05 0.11 0.15 0.06 0.06 0.06 0.13
good 0.02 0.03 0.03 0.01 0.04 0.02 -0.02 0.01 0.04 0.08 1.00 -0.03 -0.00 0.01 0.06 0.05 0.03 0.10 0.08 0.08
life -0.01 0.10 0.03 0.02 0.02 0.01 0.05 0.04 0.05 -0.00 -0.03 1.00 0.05 0.08 -0.00 0.03 0.08 -0.02 0.03 0.03
love 0.04 0.05 -0.00 0.03 0.04 0.01 -0.01 0.07 0.03 0.04 -0.00 0.05 1.00 0.03 0.07 0.05 0.06 0.03 0.08 0.07
time -0.00 0.07 0.03 0.10 0.08 0.08 0.05 0.09 0.08 0.05 0.01 0.08 0.03 1.00 0.06 0.08 0.11 0.09 0.03 0.09
character 0.06 0.07 0.07 0.08 0.07 0.04 0.12 0.09 0.07 0.11 0.06 -0.00 0.07 0.06 1.00 0.11 0.09 0.04 0.16 0.11
like -0.00 0.10 0.03 0.06 0.12 0.06 0.02 0.12 0.13 0.15 0.05 0.03 0.05 0.08 0.11 1.00 0.11 0.05 0.05 0.13
one 0.01 0.09 0.03 0.10 0.09 0.04 0.07 0.08 0.09 0.06 0.03 0.08 0.06 0.11 0.09 0.11 1.00 0.08 0.05 0.11
read 0.08 0.01 0.04 0.07 0.03 0.09 0.00 0.04 0.06 0.06 0.10 -0.02 0.03 0.09 0.04 0.05 0.08 1.00 0.05 0.21
story 0.07 0.05 0.07 0.04 0.03 0.03 0.04 0.04 0.06 0.06 0.08 0.03 0.08 0.03 0.16 0.05 0.05 0.05 1.00 -0.01
book 0.08 0.08 0.11 0.13 0.09 0.16 -0.05 0.10 0.11 0.13 0.08 0.03 0.07 0.09 0.11 0.13 0.11 0.21 -0.01 1.00
great way author first get reading novel much would really good life love time character like one read story book
1.0
0.8
0.6
0.4
0.2
0.0


Average Review Length by Sentiment
review_length
2.6e+02
2.2e+02
2.5e+02
negative
neutral
positive
sentiment
250
240
230


Cosine Similarity Heatmap of Word Embeddings
book 1.00 0.68 0.46 0.68 0.67 0.46 0.59 0.43 0.39 0.48 0.43 0.65 0.55 0.42 0.45 0.50 0.45 0.56 0.51 0.45
one 0.68 1.00 0.34 0.54 0.57 0.53 0.56 0.47 0.46 0.53 0.44 0.57 0.53 0.48 0.51 0.55 0.54 0.61 0.56 0.48
title 0.46 0.34 1.00 0.24 0.61 0.28 0.21 0.06 0.11 0.15 0.25 0.30 0.22 0.16 0.17 0.13 0.06 0.19 0.37 0.32
story 0.68 0.54 0.24 1.00 0.45 0.35 0.44 0.53 0.47 0.36 0.39 0.56 0.48 0.34 0.34 0.44 0.35 0.45 0.40 0.36
read 0.67 0.57 0.61 0.45 1.00 0.43 0.42 0.27 0.26 0.41 0.39 0.56 0.49 0.42 0.37 0.37 0.31 0.44 0.45 0.48
like 0.46 0.53 0.28 0.35 0.43 1.00 0.41 0.33 0.35 0.40 0.46 0.56 0.46 0.48 0.43 0.49 0.38 0.41 0.40 0.64
time 0.59 0.56 0.21 0.44 0.42 0.41 1.00 0.34 0.49 0.38 0.35 0.46 0.45 0.41 0.42 0.48 0.43 0.52 0.39 0.38
character 0.43 0.47 0.06 0.53 0.27 0.33 0.34 1.00 0.32 0.28 0.41 0.52 0.42 0.33 0.34 0.44 0.40 0.40 0.33 0.34
life 0.39 0.46 0.11 0.47 0.26 0.35 0.49 0.32 1.00 0.35 0.35 0.37 0.43 0.35 0.32 0.48 0.48 0.42 0.26 0.36
would 0.48 0.53 0.15 0.36 0.41 0.40 0.38 0.28 0.35 1.00 0.33 0.52 0.47 0.33 0.45 0.41 0.50 0.46 0.38 0.36
love 0.43 0.44 0.25 0.39 0.39 0.46 0.35 0.41 0.35 0.33 1.00 0.50 0.43 0.37 0.33 0.36 0.33 0.39 0.30 0.52
really 0.65 0.57 0.30 0.56 0.56 0.56 0.46 0.52 0.37 0.52 0.50 1.00 0.67 0.53 0.57 0.55 0.48 0.65 0.47 0.62
much 0.55 0.53 0.22 0.48 0.49 0.46 0.45 0.42 0.43 0.47 0.43 0.67 1.00 0.44 0.56 0.48 0.42 0.50 0.40 0.52
get 0.42 0.48 0.16 0.34 0.42 0.48 0.41 0.33 0.35 0.33 0.37 0.53 0.44 1.00 0.49 0.41 0.42 0.42 0.37 0.55
even 0.45 0.51 0.17 0.34 0.37 0.43 0.42 0.34 0.32 0.45 0.33 0.57 0.56 0.49 1.00 0.44 0.46 0.50 0.40 0.48
way 0.50 0.55 0.13 0.44 0.37 0.49 0.48 0.44 0.48 0.41 0.36 0.55 0.48 0.41 0.44 1.00 0.54 0.56 0.36 0.40
people 0.45 0.54 0.06 0.35 0.31 0.38 0.43 0.40 0.48 0.50 0.33 0.48 0.42 0.42 0.46 0.54 1.00 0.59 0.31 0.41
thing 0.56 0.61 0.19 0.45 0.44 0.41 0.52 0.40 0.42 0.46 0.39 0.65 0.50 0.42 0.50 0.56 0.59 1.00 0.33 0.51
first 0.51 0.56 0.37 0.40 0.45 0.40 0.39 0.33 0.26 0.38 0.30 0.47 0.40 0.37 0.40 0.36 0.31 0.33 1.00 0.34
know 0.45 0.48 0.32 0.36 0.48 0.64 0.38 0.34 0.36 0.36 0.52 0.62 0.52 0.55 0.48 0.40 0.41 0.51 0.34 1.00
book one title story read like time character life would love really much get even way people thing first know
1.0 0.8 0.6 0.4 0.2