DATA MINING

RevRank: A Fully Unsupervised Algorithm for Selecting the Most Helpful Book Reviews

Group Members :
Shruti Badadhe
Shruti Bhatambre
Sushmitha P
M. Krishna kanth

WHY ????

People’s decision of buying books is significantly influenced by reviews of products.

Large number of reviews

Early bird bias

winner circle bias

Imbalance vote bias

Project Description

We present an algorithm for automatically ranking user generated book reviews according to review helpfulness.

It is fully unsupervised.

It is better than user votes.

It finds hidden helpful reviews.

Given a collection of reviews, our REVRANK algorithm identifies a lexicon of dominant terms that constitutes the core of a virtual optimal review.

This lexicon defines a feature vector representation.

Reviews are then converted to this representation and ranked according to their distance from a ‘virtual core’ review vector.

Our experiments show that RevRank clearly outperforms a baseline imitating the Amazon user vote review ranking system.

Main Idea - ‘Dominant Concepts’

Automatic detection of dominant concepts.

The dominant concepts deal with type of lexicon :

- Frequent words
- examples are book, author (dan brown), category (thriller, frictional, historical), Christianity , great.
- Infrequent but informative/interesting/book specific words
- examples are council of nicea, foucault’s pendulum).

About term Dominance

$$D_{R_p}(t) = f_{R_p}(t) \cdot c \cdot \frac{1}{\log B(t)}$$

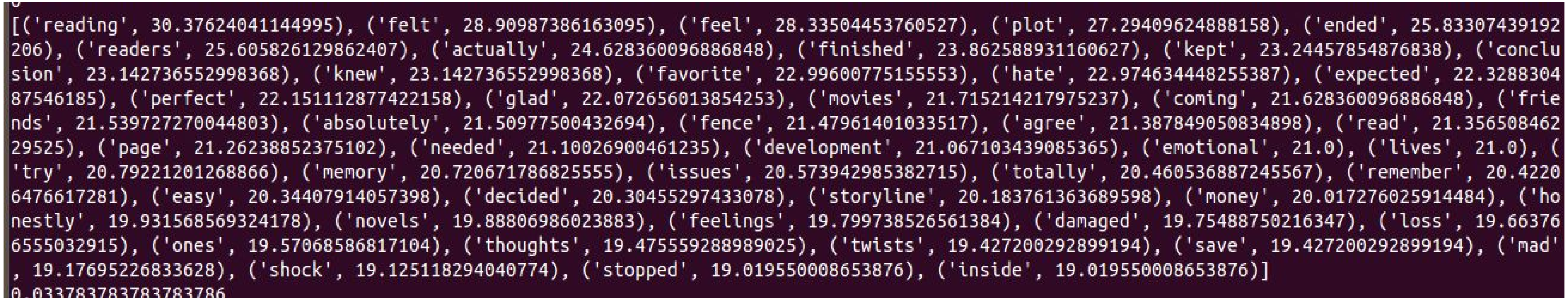
DRp(t) - dominance score of term t in R

Rp - the collection of reviews for book b(for given corpus Rp)

fRp(t) - observed frequency of term t in R

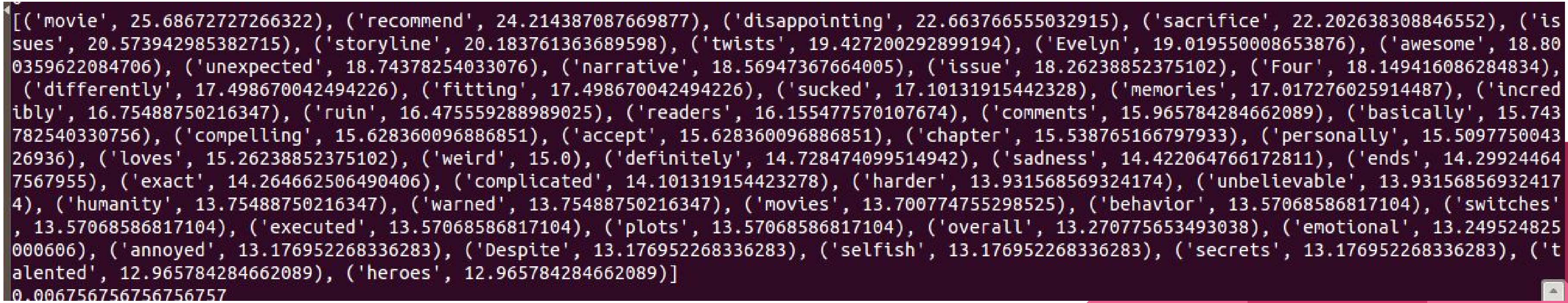
B(t) - frequency of t in a balanced corpus

Different result with different corpus



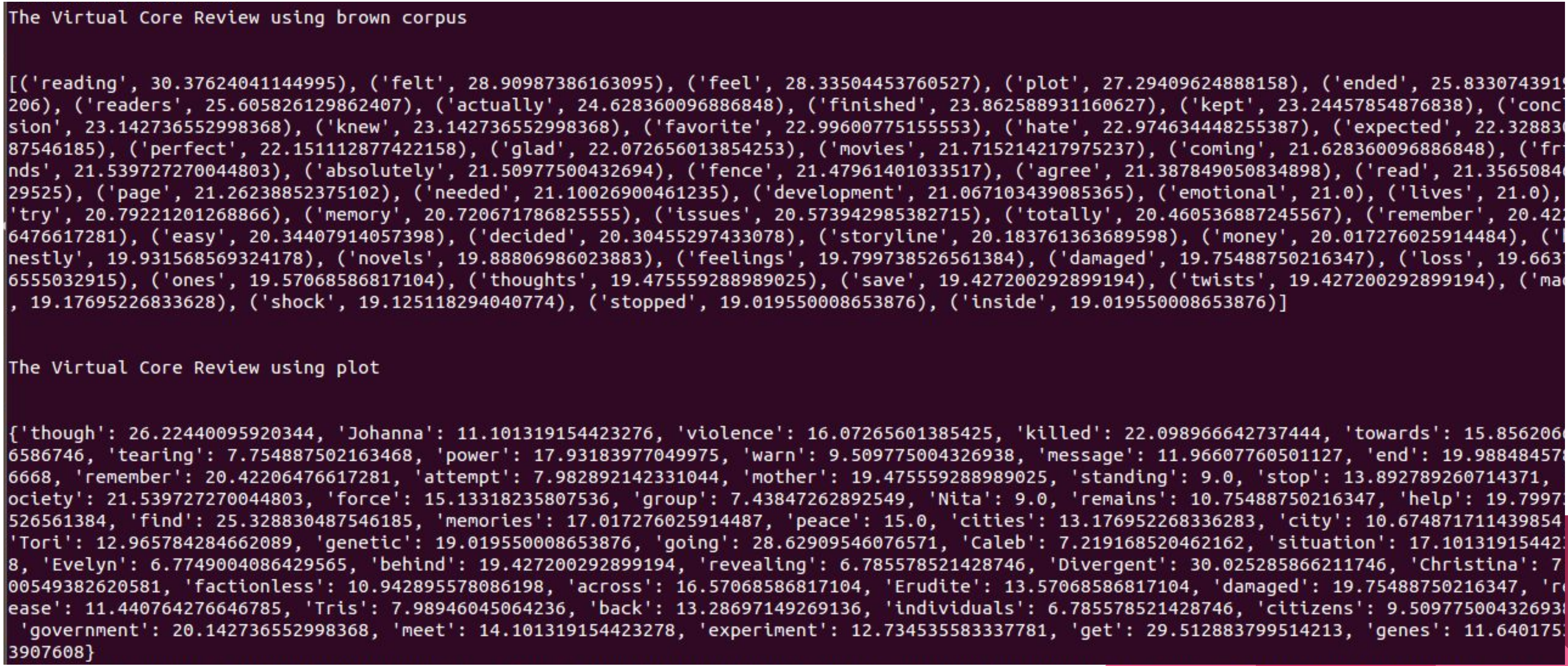
```
[('reading', 30.37624041144995), ('felt', 28.90987386163095), ('feel', 28.33504453760527), ('plot', 27.29409624888158), ('ended', 25.83307439192
206), ('readers', 25.605826129862407), ('actually', 24.628360096886848), ('finished', 23.862588931160627), ('kept', 23.24457854876838), ('conclu
sion', 23.142736552998368), ('knew', 23.142736552998368), ('favorite', 22.99600775155553), ('hate', 22.974634448255387), ('expected', 22.3288304
87546185), ('perfect', 22.151112877422158), ('glad', 22.072656013854253), ('movies', 21.715214217975237), ('coming', 21.628360096886848), ('frie
nds', 21.539727270044803), ('absolutely', 21.50977500432694), ('fence', 21.47961401033517), ('agree', 21.387849050834898), ('read', 21.356508462
29525), ('page', 21.26238852375102), ('needed', 21.10026900461235), ('development', 21.067103439085365), ('emotional', 21.0), ('lives', 21.0), (
'try', 20.79221201268866), ('memory', 20.720671786825555), ('issues', 20.573942985382715), ('totally', 20.460536887245567), ('remember', 20.4220
6476617281), ('easy', 20.34407914057398), ('decided', 20.30455297433078), ('storyline', 20.183761363689598), ('money', 20.017276025914484), ('ho
nestly', 19.931568569324178), ('novels', 19.88806986023883), ('feelings', 19.799738526561384), ('damaged', 19.75488750216347), ('loss', 19.66376
6555032915), ('ones', 19.57068586817104), ('thoughts', 19.475559288989025), ('twists', 19.427200292899194), ('save', 19.427200292899194), ('mad'
, 19.17695226833628), ('shock', 19.125118294040774), ('stopped', 19.019550008653876), ('inside', 19.019550008653876)]
0.0337837837837837786
```

Book specific ::



```
[('movie', 25.68672727266322), ('recommend', 24.214387087669877), ('disappointing', 22.663766555032915), ('sacrifice', 22.202638308846552), ('is
sues', 20.573942985382715), ('storyline', 20.183761363689598), ('twists', 19.427200292899194), ('Evelyn', 19.019550008653876), ('awesome', 18.80
0359622084706), ('unexpected', 18.74378254033076), ('narrative', 18.56947367664005), ('issue', 18.26238852375102), ('Four', 18.149416086284834),
 ('differently', 17.498670042494226), ('fitting', 17.498670042494226), ('sucked', 17.10131915442328), ('memories', 17.017276025914487), ('incred
ibly', 16.75488750216347), ('ruin', 16.475559288989025), ('readers', 16.155477570107674), ('comments', 15.965784284662089), ('basically', 15.743
782540330756), ('compelling', 15.628360096886851), ('accept', 15.628360096886851), ('chapter', 15.538765166797933), ('personally', 15.5097750043
26936), ('loves', 15.26238852375102), ('weird', 15.0), ('definitely', 14.728474099514942), ('sadness', 14.422064766172811), ('ends', 14.29924464
7567955), ('exact', 14.264662506490406), ('complicated', 14.101319154423278), ('harder', 13.931568569324174), ('unbelievable', 13.93156856932417
4), ('humanity', 13.75488750216347), ('warned', 13.75488750216347), ('movies', 13.700774755298525), ('behavior', 13.57068586817104), ('switches'
, 13.57068586817104), ('executed', 13.57068586817104), ('plots', 13.57068586817104), ('overall', 13.270775653493038), ('emotional', 13.249524825
000606), ('annoyed', 13.176952268336283), ('Despite', 13.176952268336283), ('selfish', 13.176952268336283), ('secrets', 13.176952268336283), ('t
alented', 12.965784284662089), ('heroes', 12.965784284662089)]
0.006756756756756757
```

Virtual core using wiki plot



```
The Virtual Core Review using brown corpus


[('reading', 30.37624041144995), ('felt', 28.90987386163095), ('feel', 28.33504453760527), ('plot', 27.29409624888158), ('ended', 25.833074391
206), ('readers', 25.605826129862407), ('actually', 24.628360096886848), ('finished', 23.862588931160627), ('kept', 23.24457854876838), ('conc
sion', 23.142736552998368), ('knew', 23.142736552998368), ('favorite', 22.99600775155553), ('hate', 22.974634448255387), ('expected', 22.32883
87546185), ('perfect', 22.151112877422158), ('glad', 22.072656013854253), ('movies', 21.715214217975237), ('coming', 21.628360096886848), ('fr
nds', 21.539727270044803), ('absolutely', 21.50977500432694), ('fence', 21.47961401033517), ('agree', 21.387849050834898), ('read', 21.356508
29525), ('page', 21.26238852375102), ('needed', 21.10026900461235), ('development', 21.067103439085365), ('emotional', 21.0), ('lives', 21.0),
'try', 20.79221201268866), ('memory', 20.720671786825555), ('issues', 20.573942985382715), ('totally', 20.460536887245567), ('remember', 20.42
6476617281), ('easy', 20.34407914057398), ('decided', 20.30455297433078), ('storyline', 20.183761363689598), ('money', 20.017276025914484), ('
nestly', 19.931568569324178), ('novels', 19.88806986023883), ('feelings', 19.799738526561384), ('damaged', 19.75488750216347), ('loss', 19.663
6555032915), ('ones', 19.57068586817104), ('thoughts', 19.475559288989025), ('save', 19.427200292899194), ('twists', 19.427200292899194), ('ma
, 19.17695226833628), ('shock', 19.125118294040774), ('stopped', 19.019550008653876), ('inside', 19.019550008653876)]


The Virtual Core Review using plot


{'though': 26.22440095920344, 'Johanna': 11.101319154423276, 'violence': 16.07265601385425, 'killed': 22.098966642737444, 'towards': 15.856206
6586746, 'tearing': 7.754887502163468, 'power': 17.93183977049975, 'warn': 9.509775004326938, 'message': 11.96607760501127, 'end': 19.98848457
6668, 'remember': 20.42206476617281, 'attempt': 7.982892142331044, 'mother': 19.475559288989025, 'standing': 9.0, 'stop': 13.892789260714371,
ociety': 21.539727270044803, 'force': 15.13318235807536, 'group': 7.43847262892549, 'Nita': 9.0, 'remains': 10.75488750216347, 'help': 19.7997
526561384, 'find': 25.328830487546185, 'memories': 17.017276025914487, 'peace': 15.0, 'cities': 13.176952268336283, 'city': 10.674871711439854
'Tori': 12.965784284662089, 'genetic': 19.019550008653876, 'going': 28.62909546076571, 'Caleb': 7.219168520462162, 'situation': 17.10131915442
8, 'Evelyn': 6.7749004086429565, 'behind': 19.427200292899194, 'revealing': 6.785578521428746, 'Divergent': 30.025285866211746, 'Christina': 7
00549382620581, 'factionless': 10.942895578086198, 'across': 16.57068586817104, 'Erudite': 13.57068586817104, 'damaged': 19.75488750216347, 'r
ease': 11.440764276646785, 'Tris': 7.98946045064236, 'back': 13.28697149269136, 'individuals': 6.785578521428746, 'citizens': 9.50977500432693
 'government': 20.142736552998368, 'meet': 14.101319154423278, 'experiment': 12.734535583337781, 'get': 29.512883799514213, 'genes': 11.640175
3907608}
```

As u can see that using wiki plot we get very book specific reviews but that need not be always helpful from a purchase point of view.

Some might be very detailed explaination of the book which might turn off the user from that review

Review Representation and the Virtual Core Review

Virtual core feature vector (VCFV) is vector having 1 in all of its coordinates for the dominant obtained in the above step.

Each review r ∈ R p is mapped to v(r) in this representation such that a coordinate k is frequency in text depending on whether or not the review r contains the kth dominant term

Review Ranking

Review Score :

$$S(r) = \frac{1}{p(|r|)} \cdot \frac{d_r}{|r|}$$

Where

$$d_r = v_r \cdot VCFV$$

Punishment Factor :

$$p(|r|) = \begin{cases} c & |r| < |\bar{r}| \quad \text{or} \quad |r| < l \\ 1 & \text{otherwise} \end{cases}$$

Punish Factor is needed to punish reviews which are too short or too long.

C was set to 20 to purposely create high threshold for reviews that are too short.

We can change the value of c according to the type of reviews needed .

Similarly p() can be adjusted to punish or favour long reviews .

The denominator |r| is already used for punishing very long length review

Score got using RevRank

1 :: 0.020270270270270270 27

2 :: 0.022641509433962263

3 :: 0.0006944444444444445

4 :: 0.019736842105263157

5 :: 0.03819444444444445

6 :: 0.0018181818181818182

7 :: 0.011904761904761904

8 :: 0.0010869565217391304

9 :: 0.0196078431372549

10 :: 0.02127659574468085

Evaluation Procedure

We compare RevRank (RVR), User Vote (UV) and Amazon Helpfulness score.

Amazon helpfulness Score ::

Sort the reviews by the upvote/total votes.

We set a baseline of about 15 total number of votes so that review arent biased.

Evaluation Procedure

User Vote ::

An assessor can read 10 reviews for each book and give upvote if found helpful

For each book :

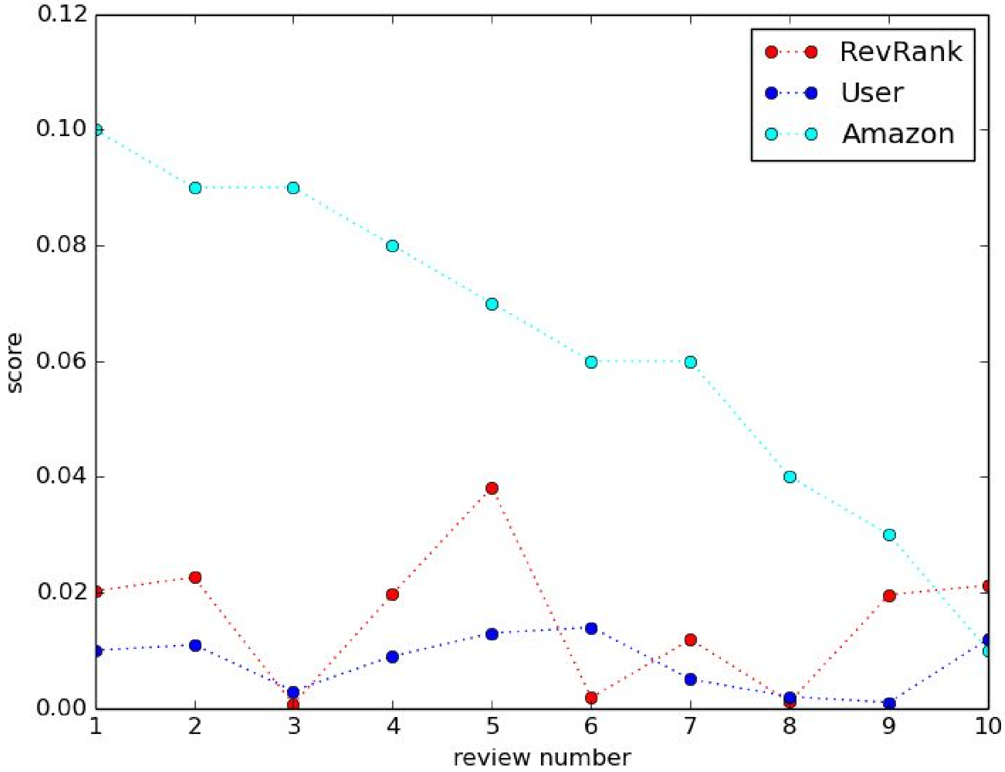
- 1 batch of 10 reviews for each book
- the batch is evaluated by 6 assessors.
- Overall : 10 reviews and 60 human evaluations.

Graph from analysis using brown corpus

By RevRank ::
5 > 2 > 10 > 1 > 4 > 9 > 7 > 6 > 8 > 3

By user :
6>5>10>2>1>4>9>7>3>8

By Amazon ::
1>2> 3>4>5>6>7>8>9>10



Some more ranks according to the wiki plot corpus

Wiki plot + paper way of scoring

4<3<9<10<6<5<8<2<1<7

Wiki plot + cosine

4<3<6<5<10<9<2<8<1<7

Word2vec

Used to produce word embeddings and is shallow model.

2 layer neural network trained to reconstruct from linguistic context of words

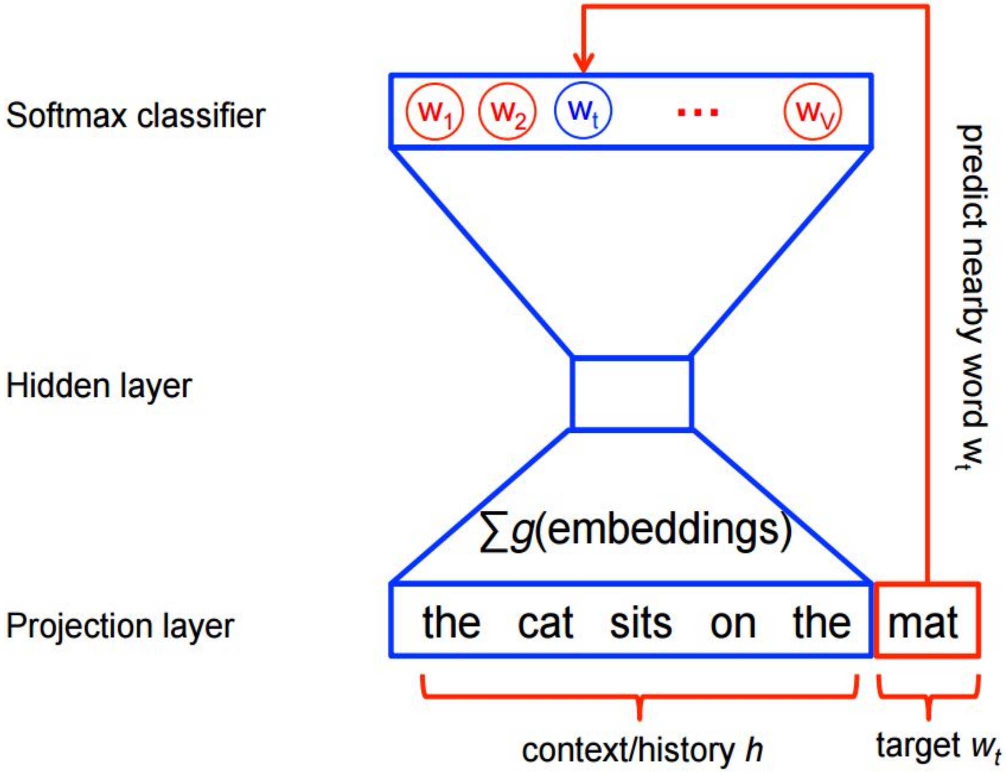
Takes a large corpus as input and produces a vector space of multiple dimension.

Has a vector for each unique word.

Word vectors are positioned in vector space such that words that share common contexts in corpus are located in close proximity to one another in space.

Can use 2 types of architecture - continuous bag-of-words, continuous skip-gram

Results can be sensitive to parametrization. Eg : Dimensionality, sub sampling, context window.



Ways to find similarity between Reviews

- Using TFIDF - The lexicon direct references to the book and the plot, references to similar books or to other books by the same author, and other important contextual aspects. This lexicon identified in an unsupervised and efficient manner, using a measure motivated by tf-idf with the use of an external balanced reference corpus.(as implemented in paper)

- Using word2vec - Reviews are converted to form list of sentences which in turn forms list of words. Found score for these words by finding their distance from words of best review which is formed using dominant words.
- Using cosine similarity - Reviews are converted into vectors and distance between these vectors found using cosine formula.

Additional Work

- Weights (tried making weights using word2vec by finding similarity score between words and the virtual core)
- Different book specific corpus(used the wiki plot text)
- Different style of analysis (using cosine etc for scoring)

Thank you