Analyzing Data about TED Talks

- network visualization
- network spatial analysis
- word2vec analysis

Wong Yeung Sum_54450514

Lam Choi Fai_56556200

Cheung Tsun Ho Aaron_56563276

Angus Au-Yeung__56571891

Ho Lai Tung_56961174

Lee Yat Shun_57040150

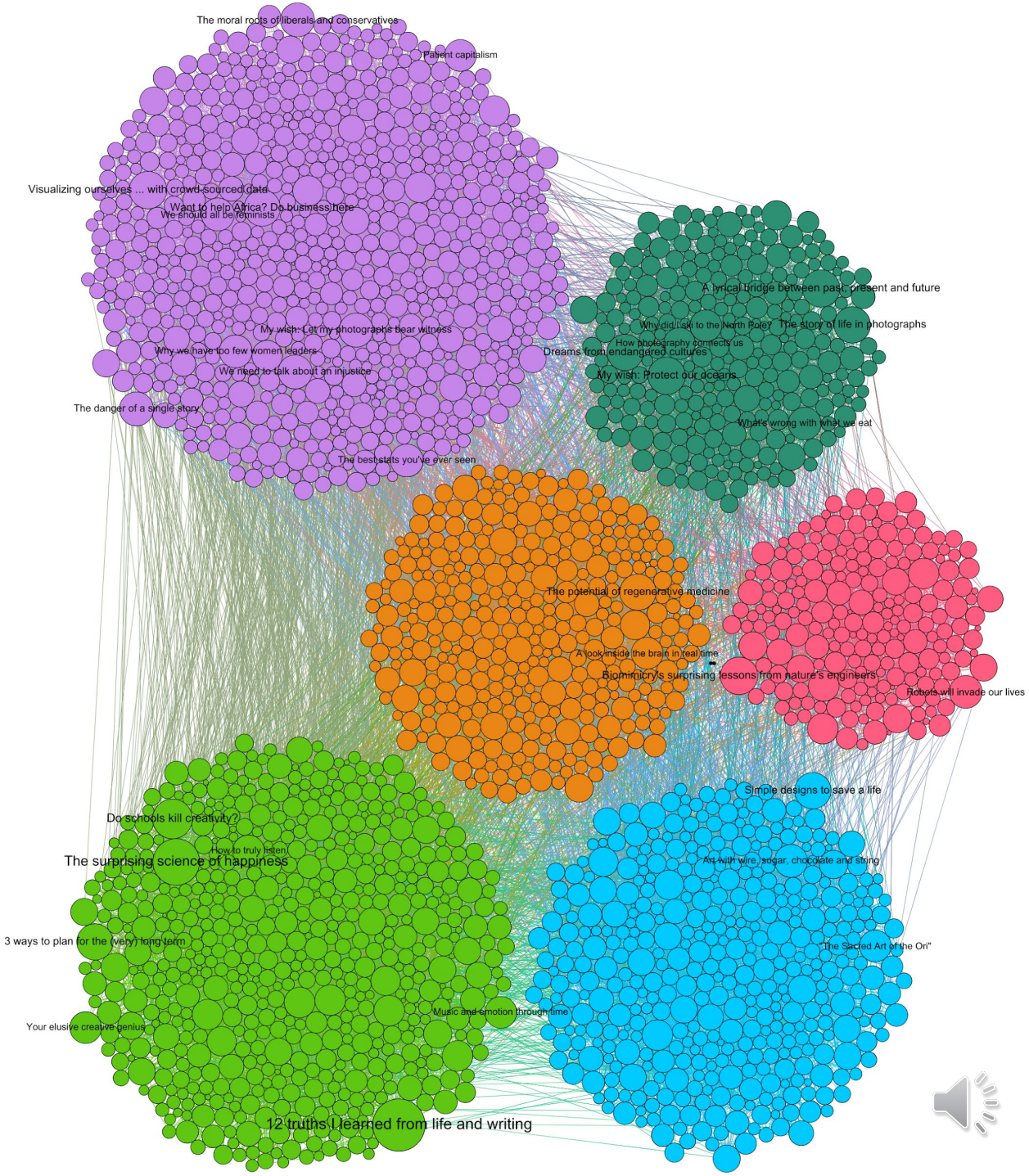
Network Visualization

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(C^{(i)}, C^{(j)}), \text{ where}$$

- A is the adjacency matrix
- $m = |E|$ is the number of edges in the network
- k_i is the degree of node i
- $C^{(i)}$ is the cluster group of node i
- δ(x,y) is the Kronecker delta function with value 1 if x=y and value 0 otherwise

Network Visualization

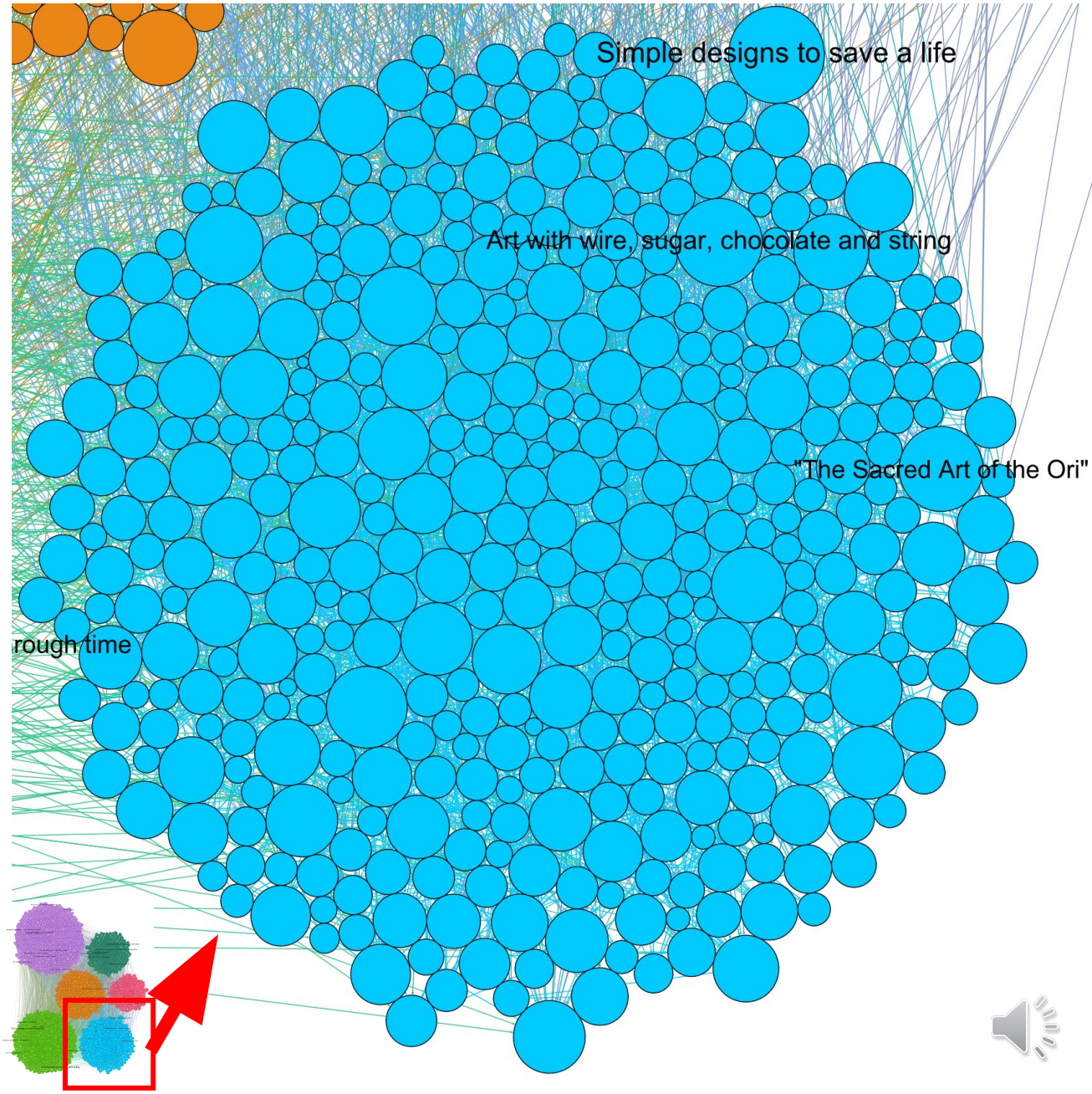
- Nodes grouped by their modularity class
- Talk title of nodes with high degree are shown
- Common theme within clusters



Network Visualization

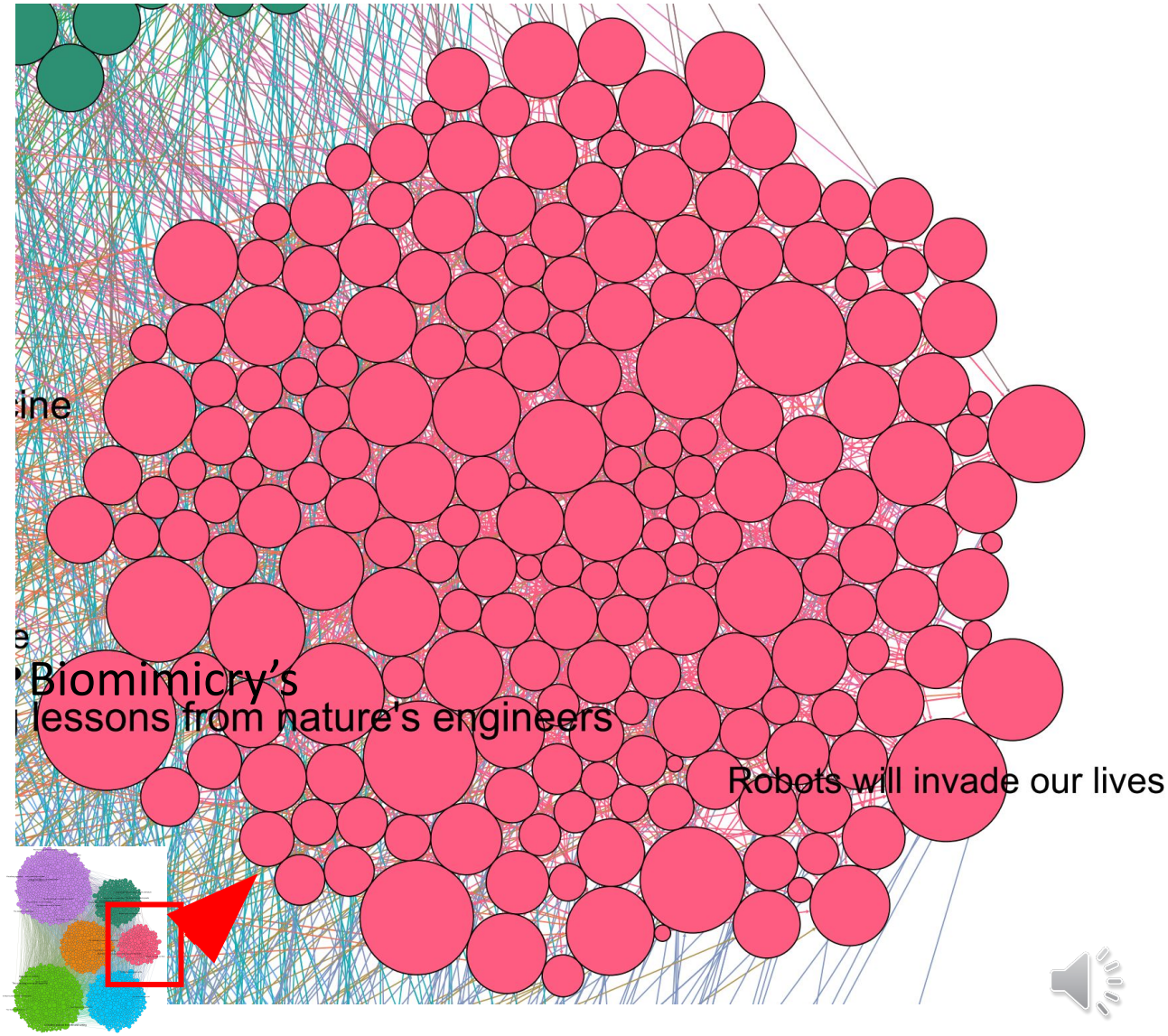
- Cyan cluster's common theme: arts and design

- 'Simple designs to save a life'
- 'Art with wire, sugar, chocolate and string'
- 'The Sacred Art of the Ori'



Network Visualization

- Red cluster's common theme: engineering and technology
- 'Biomimicry's lessons from nature's engineers'
- 'Robots will invade our lives'



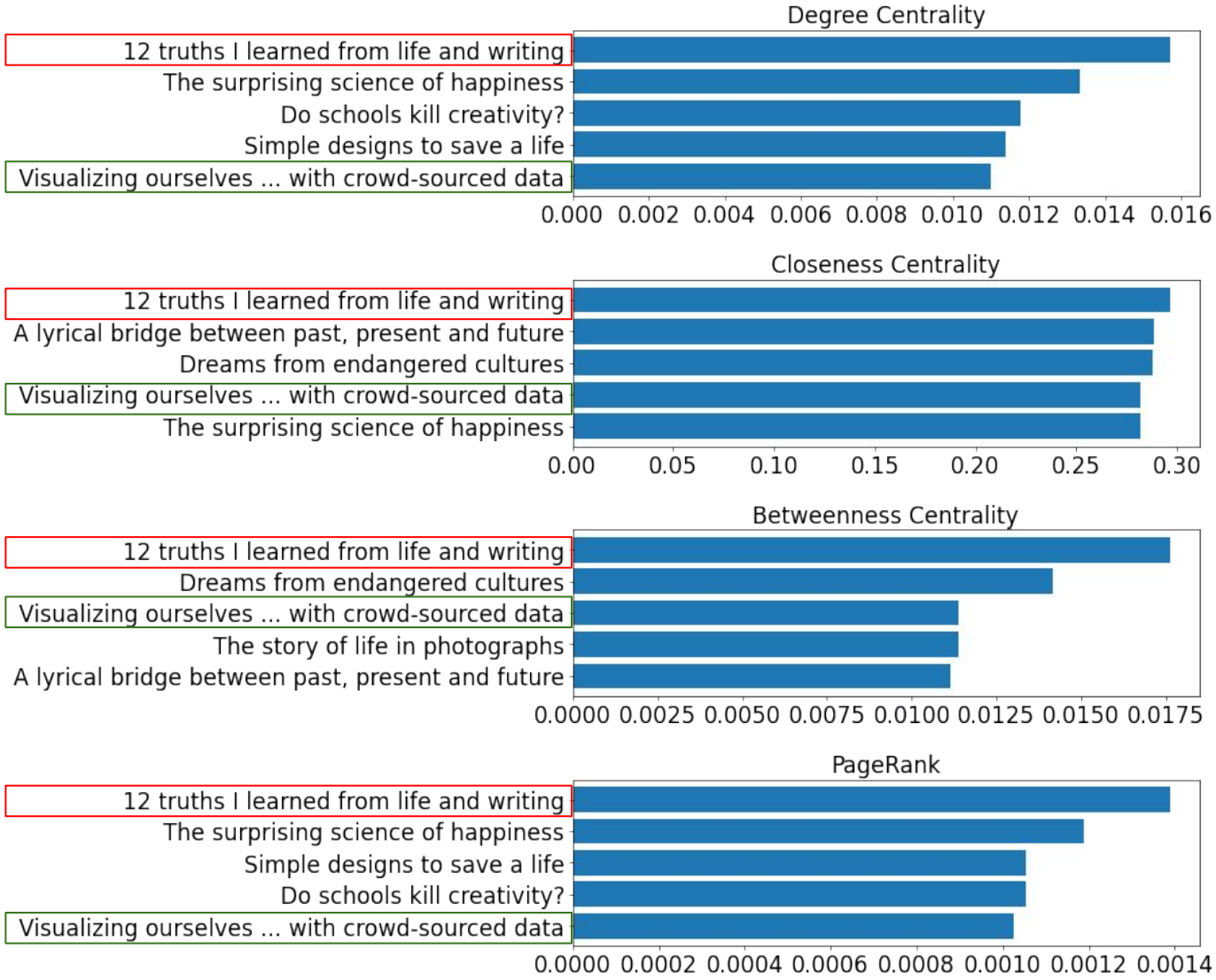
Network Analysis (Four Centralities)

From the four centralities

- Degree Centrality
- Closeness Centrality
- Betweenness Centrality
- Pagerank

"12 truths I learned from life and writing " is no.1 rank, means it has most significant impact in the network

- has greatest ability to influence other nodes
- And connect different groups
- The backlink is the highest

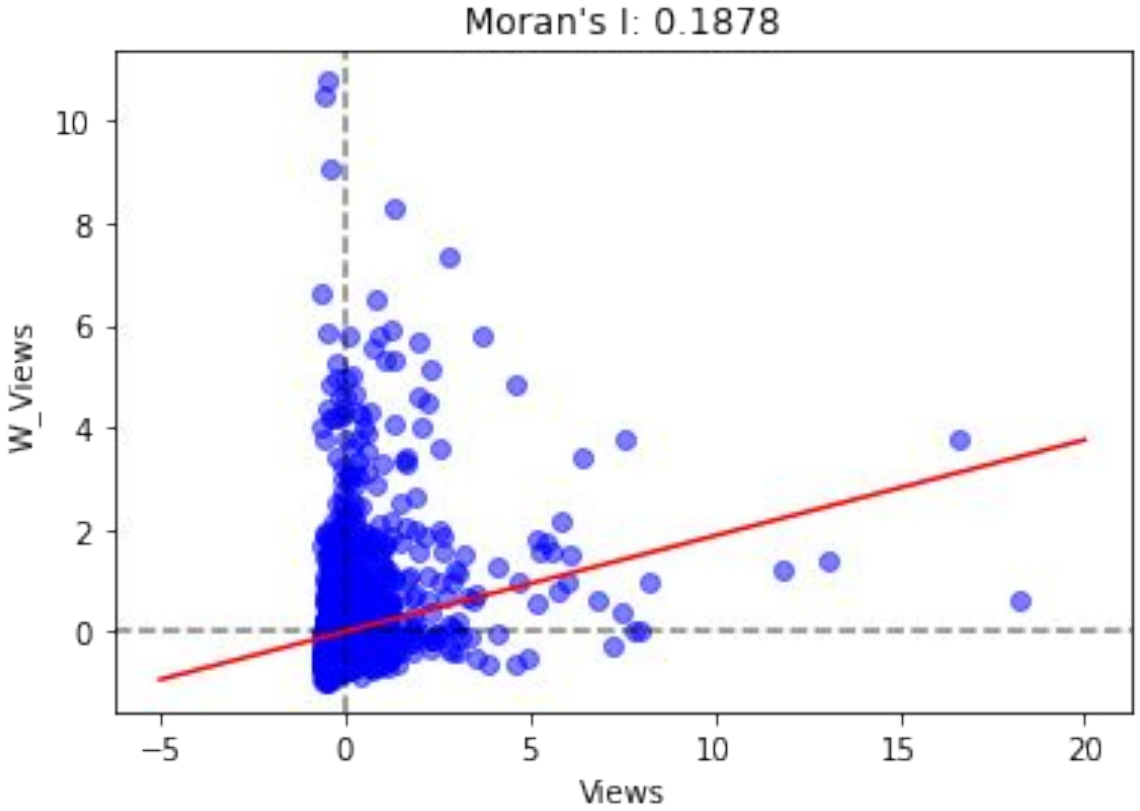


Network Spatial Analysis

How talks in recommended list affect the number of views on the talk

Correlation Analysis: Moran's I

- Converting the network into spatial adjacency matrix
  - W_{ij} = 1 if the nodes are connected by edge
- The global metrics, Moran's I shows a positive weak spatial correlation
- Moran's I = 0.1878
- Randomness hypothesis:
  - statistically significant



Moran's I Scatter Plot

Expected value E(I)	Variance VAR(I)	Z score
-3.9231e-03	8.6938e-05	20.1828

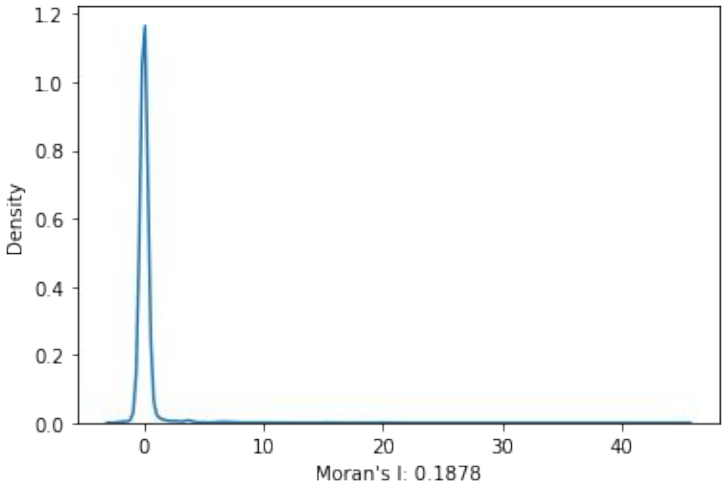
Randomness Hypothesis Testing

Clustering Analysis: LISA

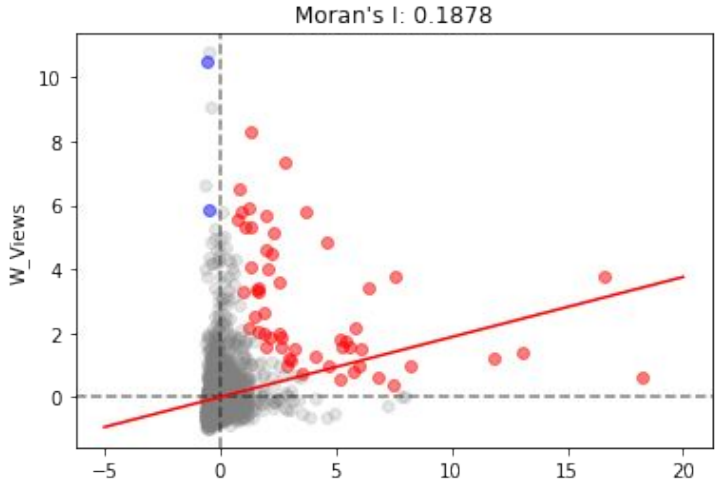
- Observing the clusters
- By computing LISA, it provides the clustering effect among the ted talks
- Using a 95 percent confidence level, red points are the ted talks with high LISA and blue points are the ted talks with low LISA.

Title	LISA	E(Ii)	VAR(Ii)	Z Score
Do schools kill creativity?	9.6962	-0.0004	0.9446	9.9771
Averting the climate crisis	-0.1395	-0.0004	0.9582	-0.1421
Simplicity sells	-0.0015	-0.0004	0.9582	-0.0011
Greening the ghetto	0.0001	-0.0004	0.9582	0.0005
The best stats you've ever seen	0.3454	-0.0004	0.9576	0.3533

Table Containing LISA of Each Data Point



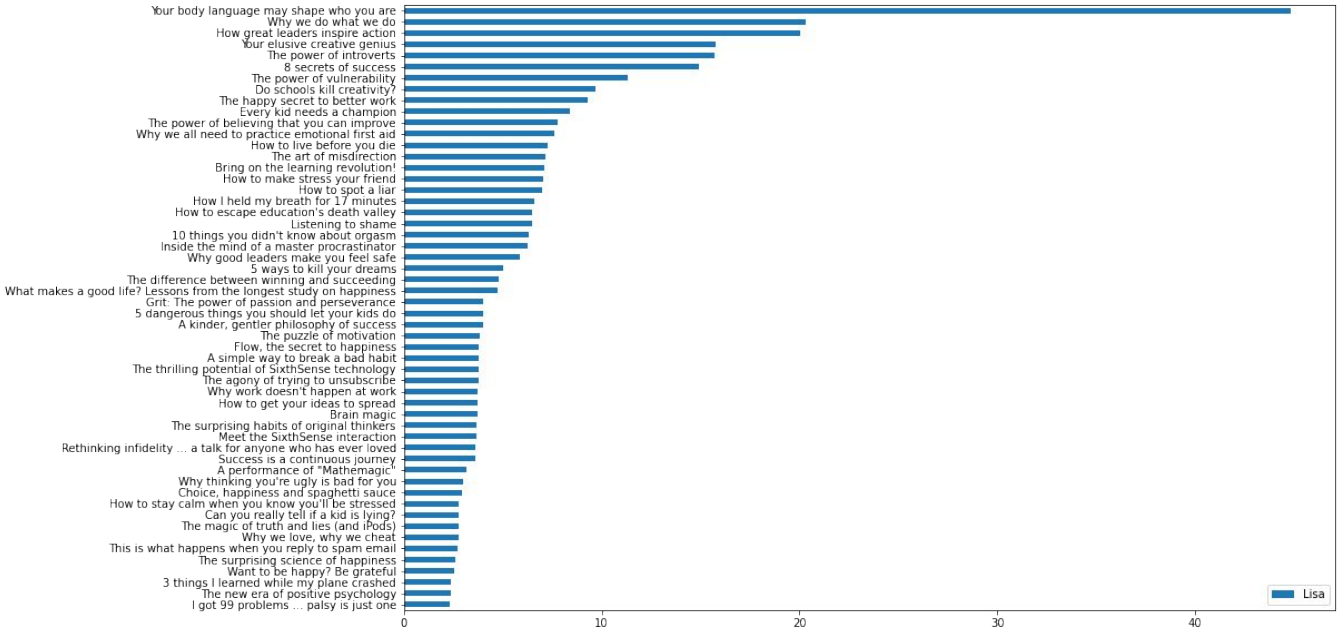
Distribution Plot of LISA



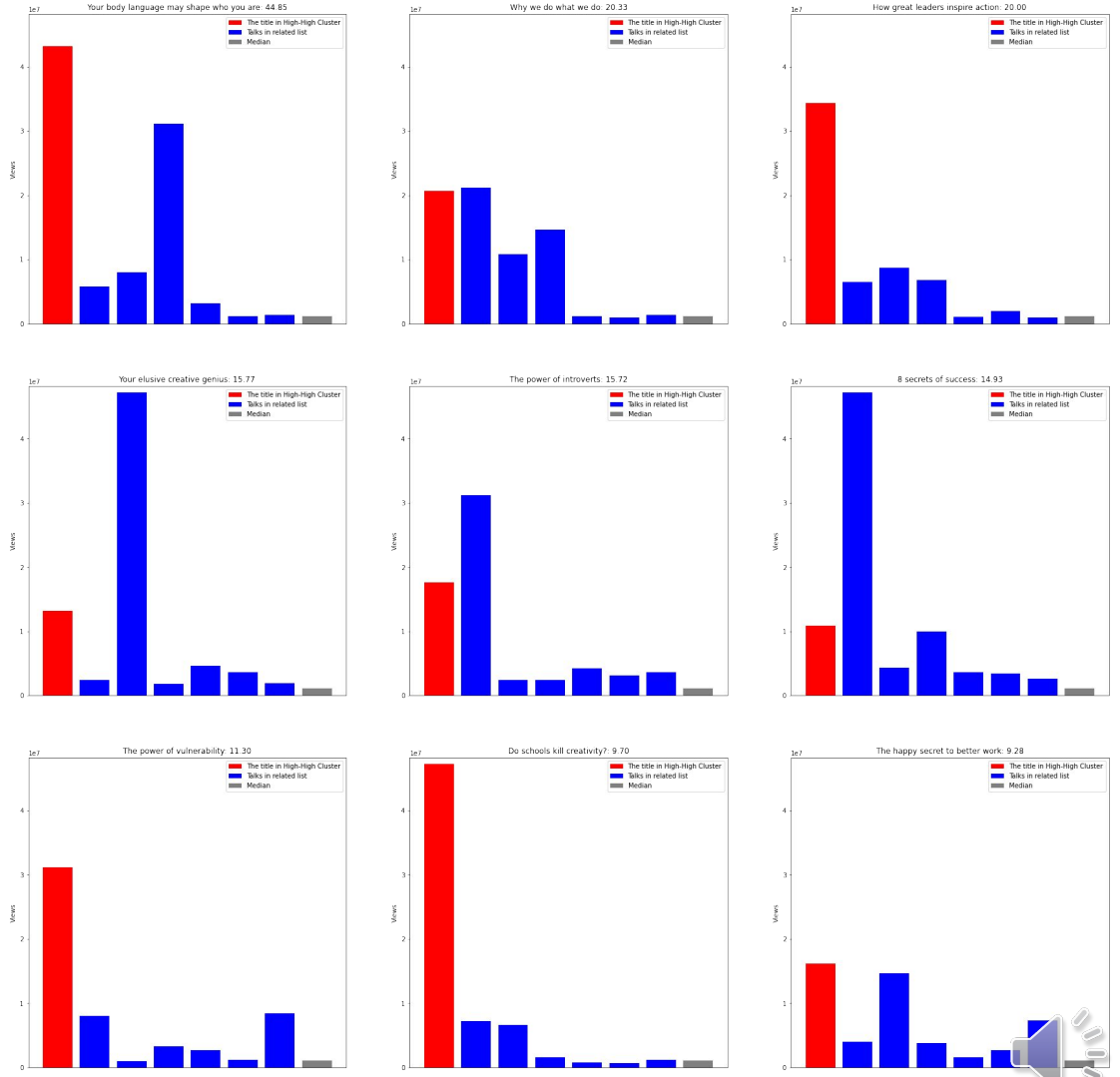
Moran's I Scatter Plot

(Red: HH cluster; Blue: outliers; Gray: insignificant)

Ted Talks in the High-High Cluster



LISA of Some Ted Talks in High-High Cluster

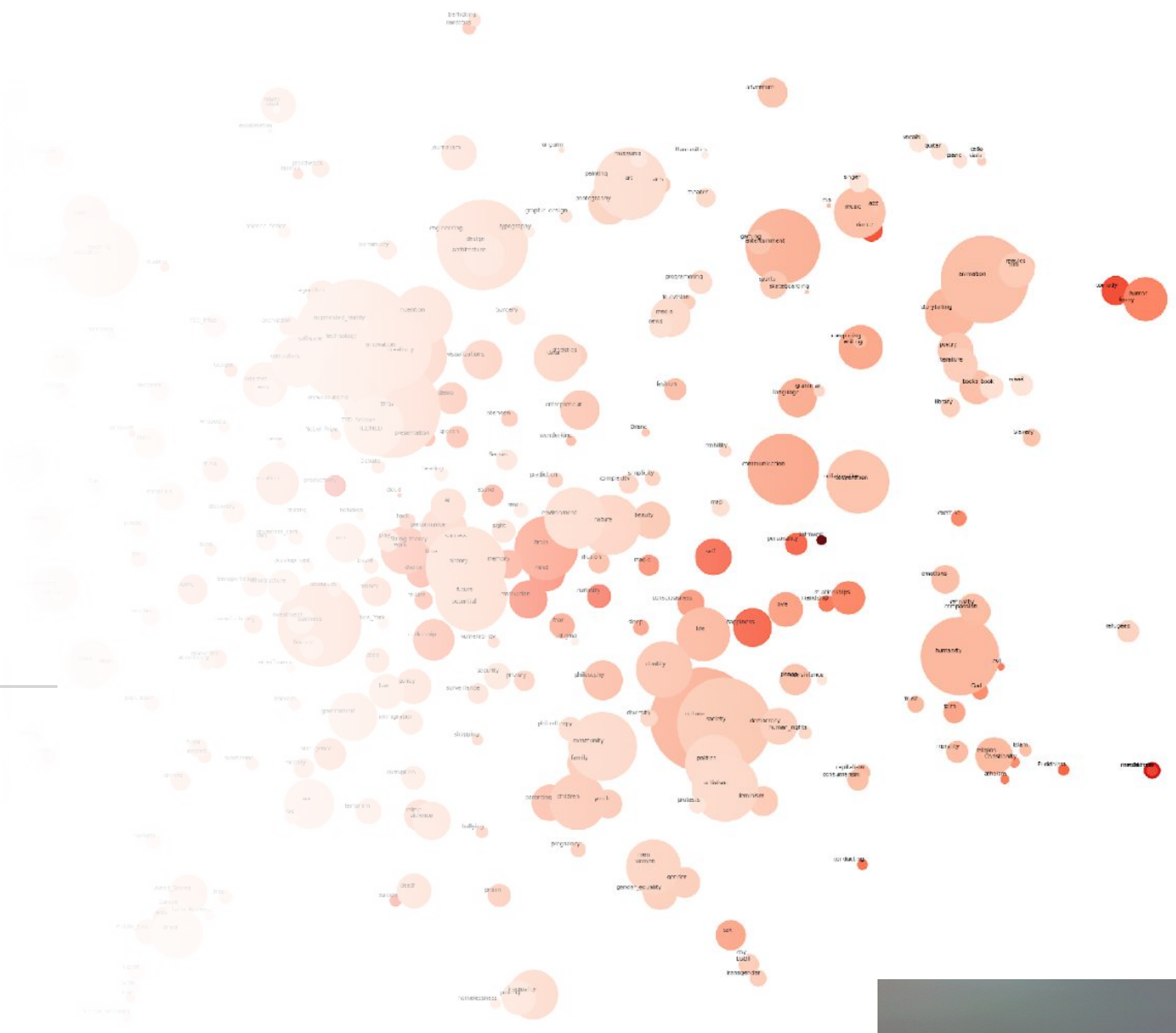


Red: High LISA Talk **Blue: Related Talks** **Gray: Median**

Summary

- The ted talks generally exhibits a weak positive correlation in the related lists.
- From the High-High cluster, the local spatial relationship can be clearly observed.
- For predicting the number of views, the topological relationship of the network has to be considered.

Similarity of key words in topics



Similarity of topics key words

- Among the 4005 talks and its unique key words, some of the key words appears very frequently while some doesn't
- A brief look into the key words in topics, 458 distinct key words are found. Output shows the key words ranked by "views" and "views_avg"
  - views: total views of the talks with the key word
  - topic_count: count of key word appears in 4005 talks
  - views_avg: "Views" divided by "topic_count"
- Word2vec analysis: a natural language processing using a neural network model to learn word associations from a large corpus of text. With the limited resource and time, an open-source training dataset "word2vec-google-news-300" is used for the word embedding and its similarity. Among the 4005 talks topics with 458 distinct keywords, 380words appeared in the google training dataset, which we believe it is justify for our project scope.
- due to the use of large size gensim word2vec api, colab is used during data processing and plot graph

In [106]:
```
df_topic_final_views_avg.sort_values("views", ascending = False)
```

Out[106]:

topic	views	topic_count	views_avg
science	2058367804	993	2072878
culture	1841213512	680	2707667
technology	1780541509	979	1818735
TEDx	1207538926	581	2078380
business	1163479109	443	2626364
...	...	...	...
exoskeleton	1797810	1	1797810
cello	1753908	4	438477
autism	570089	1	570089
testing	533106	2	266553
gay	380309	2	190154

380 rows × 3 columns

Out[103]:

topic	views	topic_count	views_avg
introvert	51452920	6	8575487
success	353490805	53	6669638
mindfulness	105759277	17	6221134
String_theory	55993294	10	5599329
productivity	148120757	28	5290027
...	...	...	...
Humanities	1914613	3	638204
autism	570089	1	570089
cello	1753908	4	438477
testing	533106	2	266553
gay	380309	2	190154

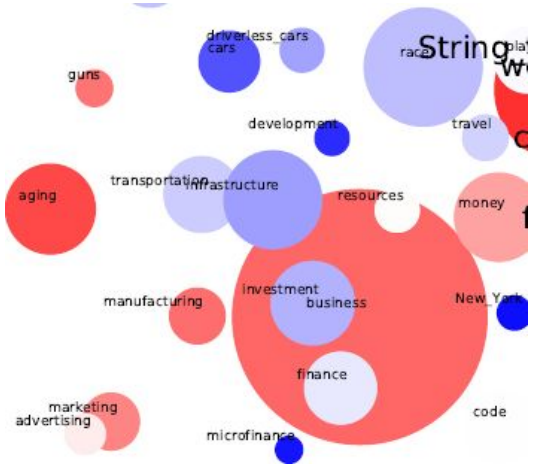
380 rows × 3 columns

Word2vec analysis

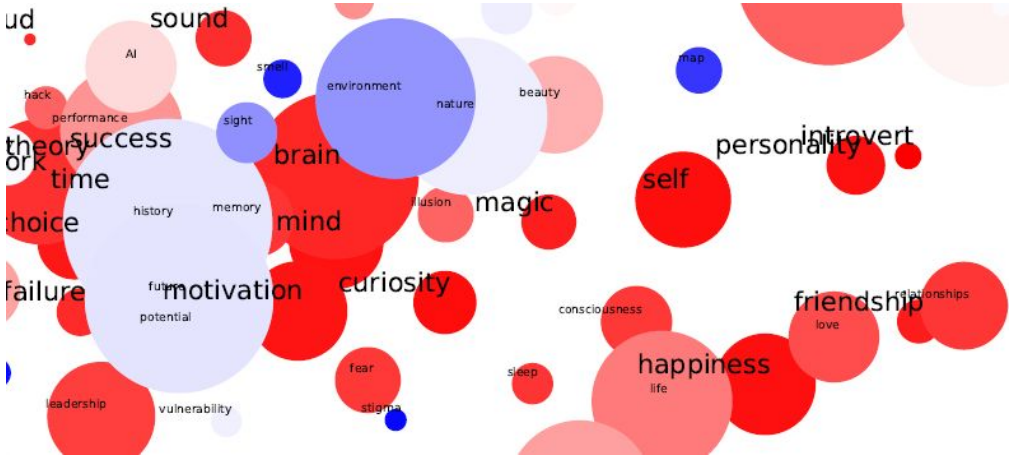
```
# define the function to plot graph
def tsne_plot(model, filename):
  labels = []
  wordvecs = []

  for word in vocab:
    wordvecs.append(model[word])
    labels.append(word)

  tsne_model = TSNE(perplexity= perplexity_lv, n_components= 2, init="pca", random_state=8)
  coordinates = tsne_model.fit_transform(wordvecs)
```

- With using “tsne.plot” and principal components analysis to plot the trained word2vec models result of the 380 key words, we can understand the clustering of the key words according to their similarity.
- The size of the nodes refers to the number of topics having the key words, and the color refers to the “views”. Top 10% word is enlarged the title
- Other than focusing on big dark nodes (existing_hot_topics), small dark nodes (upcoming_topics) should also be focused.

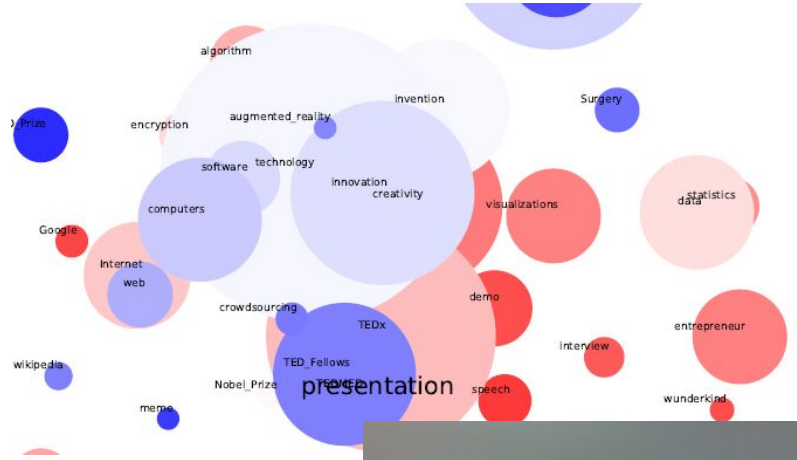
Business related



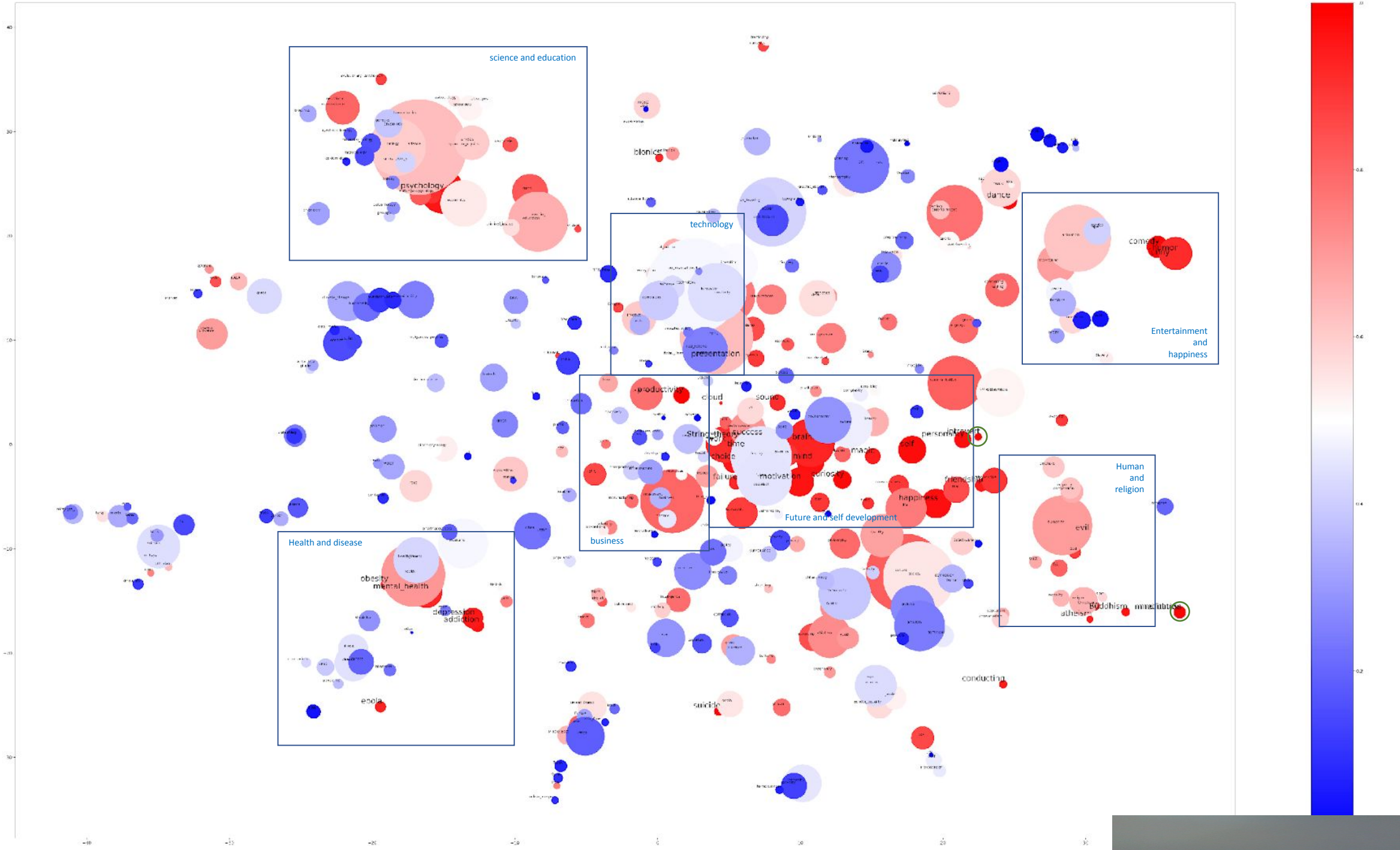
Future and Self-growing related



technology related



By views


science and education
psychology
technology
bionics
presentation
dance
comedy
humor
funny
Entertainment and happiness
productivity
cloud
sound
success
time
brain
mind
motivation
failure
curiosity
self
happiness
friendship
Future and self development
business
Human and religion
evil
Buddhism
atheism
Health and disease
obesity
mental_health
depression
addiction
ebola
suicide
conducting

Network analysis of topics key words

The analysis reveal the connectivity between the 380 key words and also its importance. Total we have 19799 edges.

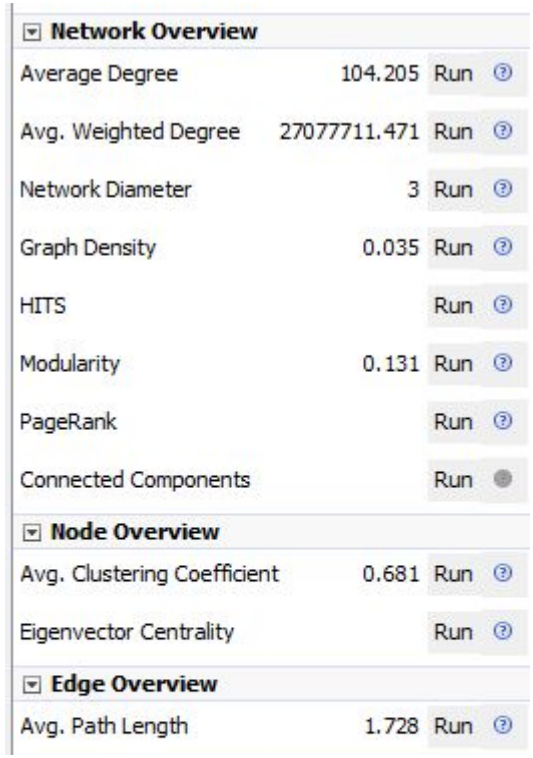
• Degree centrality: Protein with a higher degree (more connections) are more central to the structure and tend to have a greater ability to influence others. **Science** has a highest degree centrality, 993 edges, and **Technology** and **culture** also has 979 and 680 edges.

• Closeness centrality: Closeness centrality indicates the influence to the entire network that higher the closeness centrality mean closer to the other key words. **Technology** and **TEDx** have 0.9 and 0.88 closeness which is the highest in network.

• Betweenness centrality: Key words with high betweenness connect different groups. Again **technology** has the highest betweenness that 2118 connections within the network will pass through.

• PageRank: A node has high rank if the sum of the ranks of its backlinks is high. We see **science** has the highest page rank.

In summary, we found that **science** and **technology** has the highest centrality within the network.



Label	weigth
science	993
technology	979
culture	680
TEDx	581
society	557

Label	PageRank
science	0.050476
technology	0.041965
society	0.026422
culture	0.020993
TEDx	0.020277

Label	Closeness Centrality
technology	0.908873
TEDx	0.887588
science	0.885514
innovation	0.849776
society	0.847875

Label	Betweenness Centrality
technology	2118.878308
TEDx	1820.220658
science	1803.705821
culture	1502.057259
innovation	1497.984408

Network analysis of topics key words - PageRank

- Weight: (Range from 1 to 359), weights are the number of times the pairs of nodes co-exist in one ted talk, and are used for the edge size
- Filter: No filter for All_nodes.png (Edges = 19799), for others, filter weight >= 10 (Edges = 2496).
- Graph: Color = centrality measures, the higher the darker (Except All_nodes.png), Only edges with weights > 120 shown. For pagerank_v2.png, only edges with weights > 100 shown, adjusted the edge size