Best Topic Estimation

The Best Topic Estimation tool

Our 'Shape of Stories' Algorithms all consider one specific crucial parameter: the number of clusters. The 'Shape of Stories' Algorithms are performing two main steps:

1. Sorting texts into clusters using the sentiment scores patterns throughout each text.
2. Creating Visualization of the sentiment score 'story' for each cluster on the texts belonging to those clusters.

The defining parameter of this process is the number of clusters. In order to get optimized results and avoid overfitting or underfitting issues, an optimal number of clusters will allow us to accurately rank texts according to the shape of the sentiment scores throughout a text.

The Best Number of Topics k Algorithm

The algorithm to find the best number of clusters or topics (k) is using Non-Negative Matrix Factorization (NMF) as our technique to separate texts in clusters.

To learn more about Non-Negative Matrix Factorization, read the TIPS_NLP_NMF.

The basic concept of NMF is that it breaks down a matrix A of size mxn in two lower dimensions matrices W of size mxk and H of size kxn such that:

$$A = WH$$

where k is the number of topics.

To find the optimal k, the Best Topic Estimation script iteratively tests different values for k and computes the coefficient of dispersion and average number of empty clusters for each value of k.

The **coefficient of dispersion (COD)** is the average difference a group of numbers has from the median. The value is reported as a percentage of the median. The COD score ranges from 0 to 1 with optimal value reached at the convergence point.

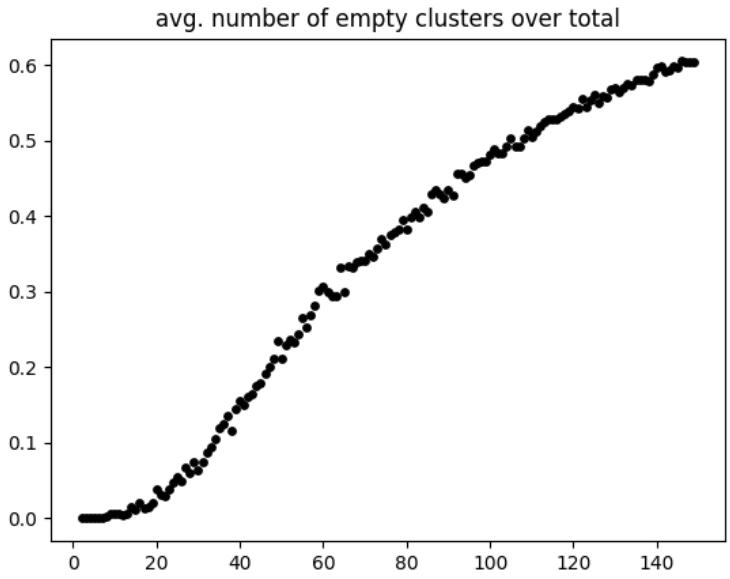
The **Average Number of Empty Topics** is computed using the number of topics left blank due to a k too large for our dataset.

Outputs

Using the two coefficients described above, the script creates two scatterplots and one CSV file outputting results.

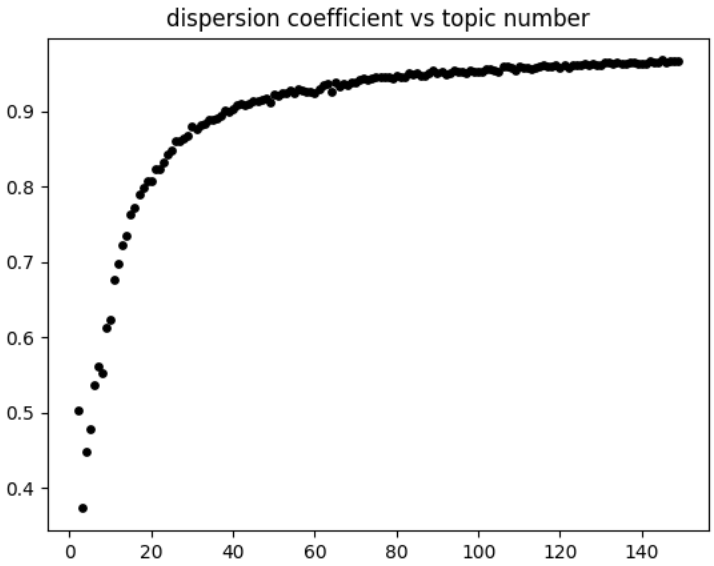
Average Number of Blank Topics Plot

Below is the graph showing the average number of clusters for each value of k (the number of clusters). The *x axis* is the value of k and the *y axis* is the percentage of blank clusters.



Dispersion Coefficient Plot

Below is the graph showing the dispersion coefficient for each value of k (the number of clusters). The *x axis* is the value of k and the *y axis* is the dispersion coefficient in percentage.



CSV File

Below is a short part of the table displaying the results from the Best Topic Estimation script.

topic number	dispersion coefficient	avg. number of empty clusters over total
2	0.502160284	0
3	0.373047185	0
4	0.44861575	0
5	0.478331372	0
6	0.536203081	0
7	0.561405013	0
8	0.551988915	0.0025
9	0.611973198	0.006666667
10	0.624080503	0.006
11	0.676924712	0.005454545
12	0.698391177	0.005
13	0.722266524	0.006153846
14	0.73413476	0.014285714
15	0.763544178	0.010666667
16	0.77278677	0.02
17	0.789129735	0.012941176
18	0.798926887	0.015555556
19	0.808050287	0.021052632
20	0.808015624	0.038

The table contains 3 categories:

- The number of topics or clusters k
- The dispersion coefficient
- The Average number of Empty Clusters

References

Burrows, J.F. 1987. "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style." *Literary and Linguistic Computing*, Vol. 2, No. 2, pp. 61-70.

Franzosi, Roberto. 2004. *From Words to Numbers*. Cambridge, UK: Cambridge University Press.

Franzosi, Roberto. 2010. *Quantitative Narrative Analysis*. Thousand Oaks, CA: Sage.

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The Emotional Arcs of Stories Are Dominated by Six Basic Shapes". *EPJ Data Science*, Vol. 5, No. 31, pp. 1-12.

Vonnegut, Kurt. https://www.youtube.com/watch?v=oP3c1h8v2ZQ

Vonnegut, Kurt. https://www.youtube.com/watch?v=GOGru_4z1Vc

Vonnegut, Kurt. 2005. "Here is a Lesson in Creative Writing." In: pp. 23-28, Kurt Vonnegut, *A Man Without a Country. Edited by Daniel Simon.* New York: Seven Stories Press.