Data Mining using the Topic Model Latent Dirichlet Allocation

Topics in Data Science
Department of Computer Science
Ryerson University
Toronto, Canada

Jorge López
jlopez@ryerson.ca

***Abstract*— The aim of this paper is to show how Probabilistic Topic Models, in particular the Latent Dirichlet Allocation (LDA), can assist IT practitioners to uncover the hidden topic structure of large collections of unstructured data. We will focus on Questions and Answers (Q&A) forums, like the ones publicly available in *stackoverflow.com*, which contains several thousands of entries that are questions about an specific software product posed and responded by the public.**

The problem that this tool may help to resolve is as follows: For the IT practitioner who ”owns” the product and is concerned with making it better, he or she would be very interested in knowing what the user is talking about the product. There won't be enough man power to sift through all vast amount of entries, neither have the necessary time for doing this with the resources available, and it would be extremely slow.

Here we will show how mining a Q&A forum data-set using the Latent Dirichlet Allocation model can help the IT practitioner to extract the information that may serve to improve the quality of a given product. We are going to determine what the best probability distribution is for our exemplified fits in order to find a model of probability of appearance of words in our corpus by determining what the best distribution for the words and topics is.

***Index Terms*—LDA, bag-of-words, generative Model, Dirichlet Distribution, Text Mining, Probabilistic Topic Model, Weibull distribution.**

I. Introduction

NOWADAYS the amount of information that we have access to, that is produced by electronic means such as the web is overwhelming, therefore we need tools that help us make sense of it. Humans are incapable of synthesizing such amount of information because it would require a very large number of resources[1].

This is where *Machine Learning* steps in, where researchers have developed Probabilistic Topic Models, including the Latent Dirichlet Allocation model (LDA). We would like to start by providing a synopsis of Topic Models and LDA. We continue by describing the data-set that we are analyzing from *stackoverflow.com* and that we will be mining using the LDA model. As explained before, this data-set contains Q&A of a given product in which an IT practitioner is interested in knowing what the public is talking more about what topic. This knowledge can be translated as a relevant issue that if addressed might conduct to improving the product. After this, we present the methodology used for getting the results, we present them and analyze them including finding the best distribution for all the five topics found in our research and for each of them, and finally we will briefly talk about some other applications of LDA-like models. Topic models are based on the idea that documents represent a mixture of topics, and in this setting, a topic is a probability distribution over words [4].

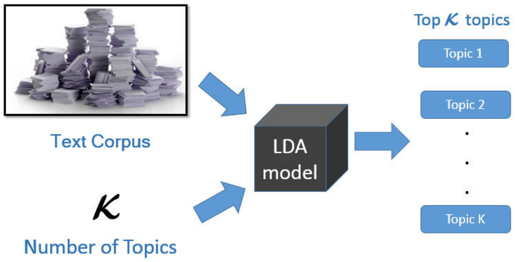


Fig. 1. High level functioning of LDA model

A. Probabilistic Topic Models

Topic Models (TM) are algorithms to uncover the topic structure that fills a large document archive. A TM is a *generative model* for documents as it defines a probabilistic method to derive documents. These documents are produced according to latent (random) variables. We aim to discover the most effective fitting of a generative model that encompass the set of random variables that explain the observed data [4].

One assumption of these processes, is the one referred as the *bag-of-words* assumption, that basically considers the number of times that the words are produced, as the only provided information that matters to the model [4].

B. Latent Dirichlet Allocation model

LDA is a generative statistical model intended for collections of discrete data like a text corpus. It falls under the category of unsupervised algorithms and the only inputs that it requires are the observed variables (words) of the documents and the k number of topics that we would like to discover.

The output of the LDA is that for each generated topic we get a list of n words associated with that topic. The words are listed in order from the most probable occurrence to the least probable. The topic occurrence is also ordered in the same way as the words[1] (Figure 1).

The *Dirichlet distribution (DD)* is a distribution over Multinomial Distributions (MD) In order to visualize how the DD basically works, suppose that we have only three topics named t_1, t_2 and t_3, We can think of MD as living in the triangular space (simplex), Each of the edges of the triangle represent the three distributions over words where each has$probabiity = 1$, The middle point between edges gives $p = 0.5$ to two of the words, and the centroid means the uniform distribution over the three words We can visualize this by looking at the simplex directly from above, the circles represent volume. Now in order to strengthen the concept we can observe a 3-D of the 2-D simplex in.

Next we are going to present the graphical model of LDA [1]. Here the nodes represent random variables, the edges are possible dependence, the observed variable is shaded, and the plates denote replicated structure. The letter θ represents per-document topic proportions, the letter z is the per-word topic assignment, w is the observed word, N is the number of word and D is the number of documents in the corpus.

II. Q&A data-set description

We have used the data-set database administrators (dba) Q&A forum, of *stackoverflow.com* to apply the LDA model to perform the analysis. We can access this web site in *http://dba.stackexchange.com/questions*, that at the present time has 51,027 questions included. In this site there are Q&A forums for different kind of products that has information for different periods of time. The data of our interest was found archived in the web site *https://archive.org/download/stackexchange*, and it was identified as the file *dba.stackexchange.com.7z*,

The raw data is originally in XML format. As our selected data-set incorporates several years of information, we have chosen to extract only the year 2014 entries.

Basically the data-set layout is as follows: creation time stamp of the question, id, title of the question, body of the question and the answers for that question.

III. Methodology

We have used two programming languages for processing the data-sets: R and $Perl$. The former known for its open-source straightforward implementation of the LDA algorithm through a package [2], the latter because its efficiency manipulating strings of characters. In order to produce the expected results, the data has been processed by doing the following: extracting the text corpus, cleansing the text and converting it from XML to CSV format. Afterwards we run our R program to obtain the keywords (words) per topic, for doing this we need to supply the text corpus and the k number of topics that we want to discover. For determining k, there is no analytic way to do it, rather the one in use is empirical figure(2). The LDA algorithm implemented in the CRAN R package basically performs the following:

```
LDA algorithm [1]
Step 1. Select randomly a distribution over topics.
Step 2. For each document $w$ in a corpus $D$ do:
Step 2.1 Select a topic from the distribution over topics in a random way.
Step 2.2 Select in random way a word from the distribution over the vocabulary.
```

The above algorithm will cluster the top k topics and their corresponding words with the highest probability of occurrence in the text corpus.

IV. Results

We have run the LDA for a fit of 5, 10 and 15 top topics (k) and 20 keywords each (t), therefore we are going to get as a result in these cases, matrices of $5x20$, $10x20$ and $15x20$ respectively. This is $M_{k,t}$ where $p(k)>p(k_{i+1})$ and $p(t_i) > p(t_{i+1})$ figure(3).

Next we would like to perform some analysis by producing various plots only for topic 1 for $k = 5$: the probability per term figure(4), the Document Frequency figure(5) the Inverse Document Frequency (idf) figure(7) and finally the probability by idf, note that the size of each term is proportional to the probability in the topic, in this case The algorithm found 329 (N) documents for topic 1 and the keywords associated in descending order of probability and corresponding inverse document frequency (idf), where the closer this value is from 0, the more the widespread, and vice versa. The arrangement of the output keywords, suggest that the main issue expressed by thousands of users that posted in this forum is related to how can they connect to the DBMS figure(7). This knowledge may lead to product improvements action-ed by the IT practitioner,

A. Best Distribution fit

We have used the *CurveExpert* software to help us determine what the best distribution for our data is. We have found that the Weibull distribution offers a good fit of $R^2 = 0.9843$. The fit for all the topics and unique words is shown in figure 8. The regression equation for this distribution is:

$$y = a - be^{-cx^d}$$

with parameters $a = 437.389, b = 353.654, c = 0.001$ and $d = 1.938$.

The plots for topics 1 to 5 are included in the appendix of this paper, and they are the figures A1, A2, A3, A4 and A5. Moreover in order to insure that the distribution found (Weibull) holds for many values of k, we have included distribution evaluation for other values of k, $k = 10$ (Fig. A6, A7) and $k = 15$ (Fig. A8, A9), in both cases the best distribution found was Weibull.

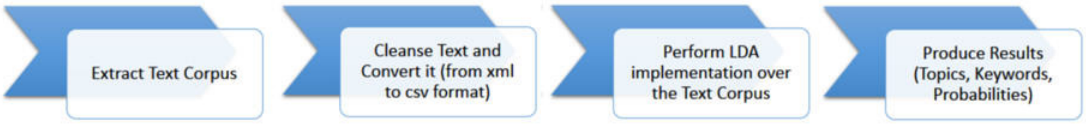

Extract Text Corpus
Cleanse Text and Convert it (from xml to csv format)
Perform LDA implementation over the Text Corpus
Produce Results (Topics, Keywords, Probabilities)

Fig. 2. General Process for fitting our Q&A forum with LDA

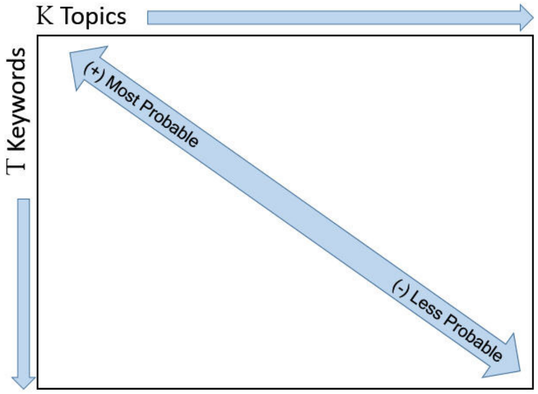


Fig. 3. Output of LDA Topic Extraction

V. Other Applications

Topic Models can be applied to a variety of areas, including but not limited to the following areas: Population Genetics. Computer Vision, Topic Evolution. Software Analysis and Social Networks.

VI. Conclusions and Further Work

Topic Models are a collection of algorithms that give a statistical solution to understanding the more than ever growing amount of digitized data. We found a model of probability of appearance of words in our corpus by determining the best distribution for the keywords and topics in it. As a future work, we would like to study the possibility of approximating the input parameter k (number of topics) for LDA by deriving a formula from an empirical analysis.

Acknowledgments

We would like to thank Dr. Andriy Miranskyy and Dr. Abdolreza Abhari for their motivation and feedback to undertake and improve this work.

References

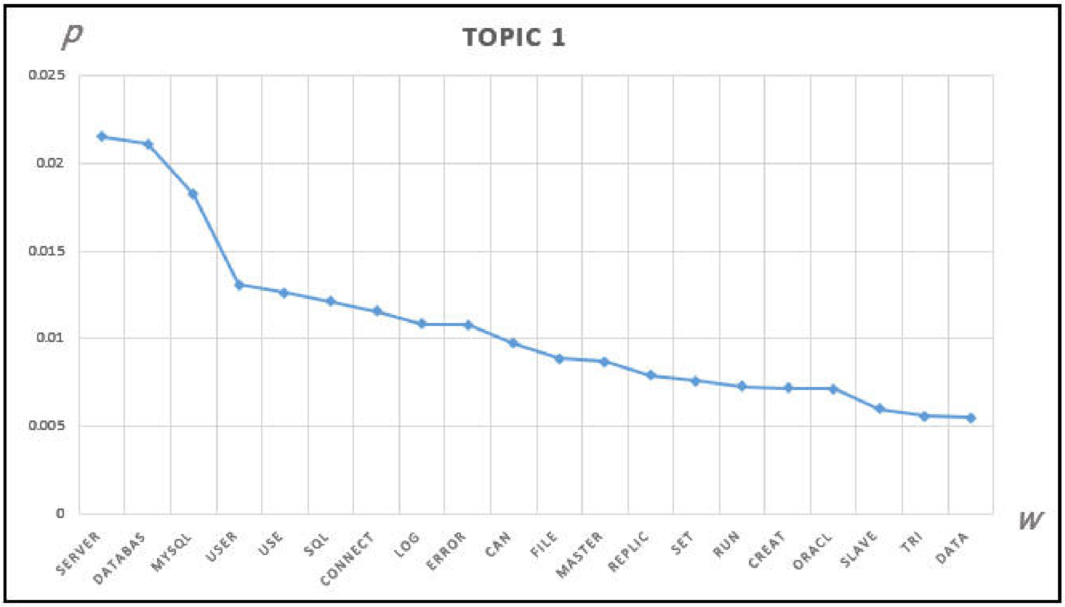
[1] D. M. Blei, *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3, pp. 993-1022, 2003.

[2] K. H. Bettina Grun, *Topic Models: An R package for fitting Topic Models*. Journal of Statistical Software, pp. 1-20, 2011.

[3] G. W. A. S. Ayushi Aggarwal, *Mining Issue Tracking Systems using Topic Models for Trend Analysis*. Corpus Exploration, and Understanding," Communications of the ACM, pp. 52-58, 2014.

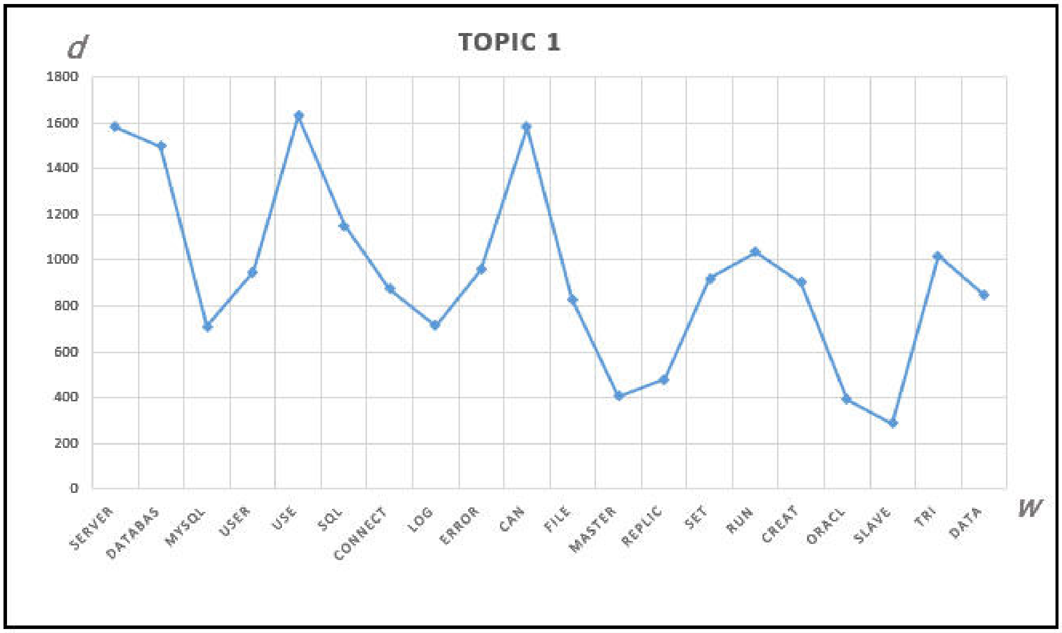
[4] M. Steyvers, *Probabilistic Topic Models*. University of California, Irvine, pp. 1-15, 2011.

[5] S. P. J. C. Ashwinkumar Ganesan, *LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation*. Communications of the ACM, p. 7, 2015.



p = probability, *w* = keyword

Fig. 4. PDF per term for Topic 1



d = number of documents that belong to Topic 1, where the keyword appears
w = keyword

Fig. 5. Document frequency for Topic 1

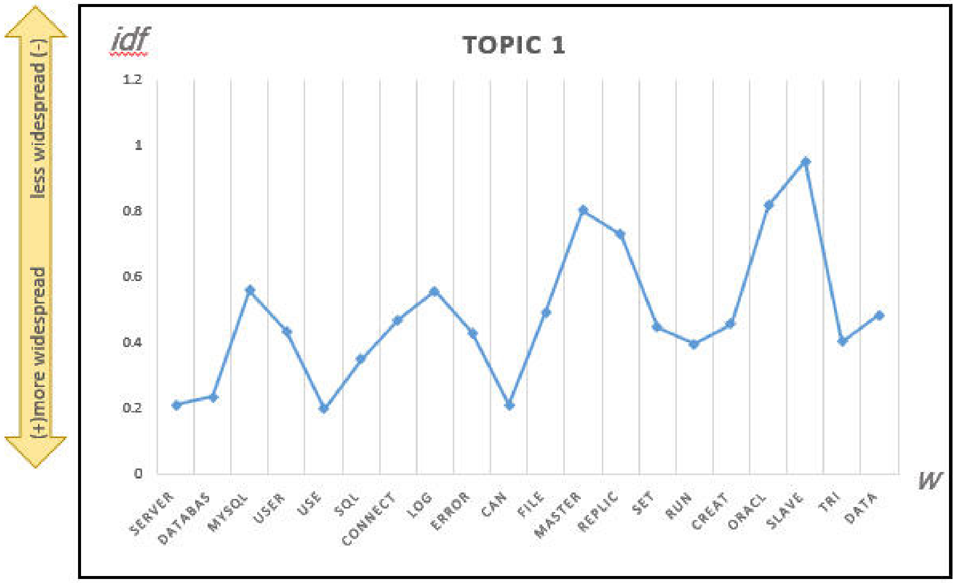


Fig. 6. Inverse Document frequency for Topic 1

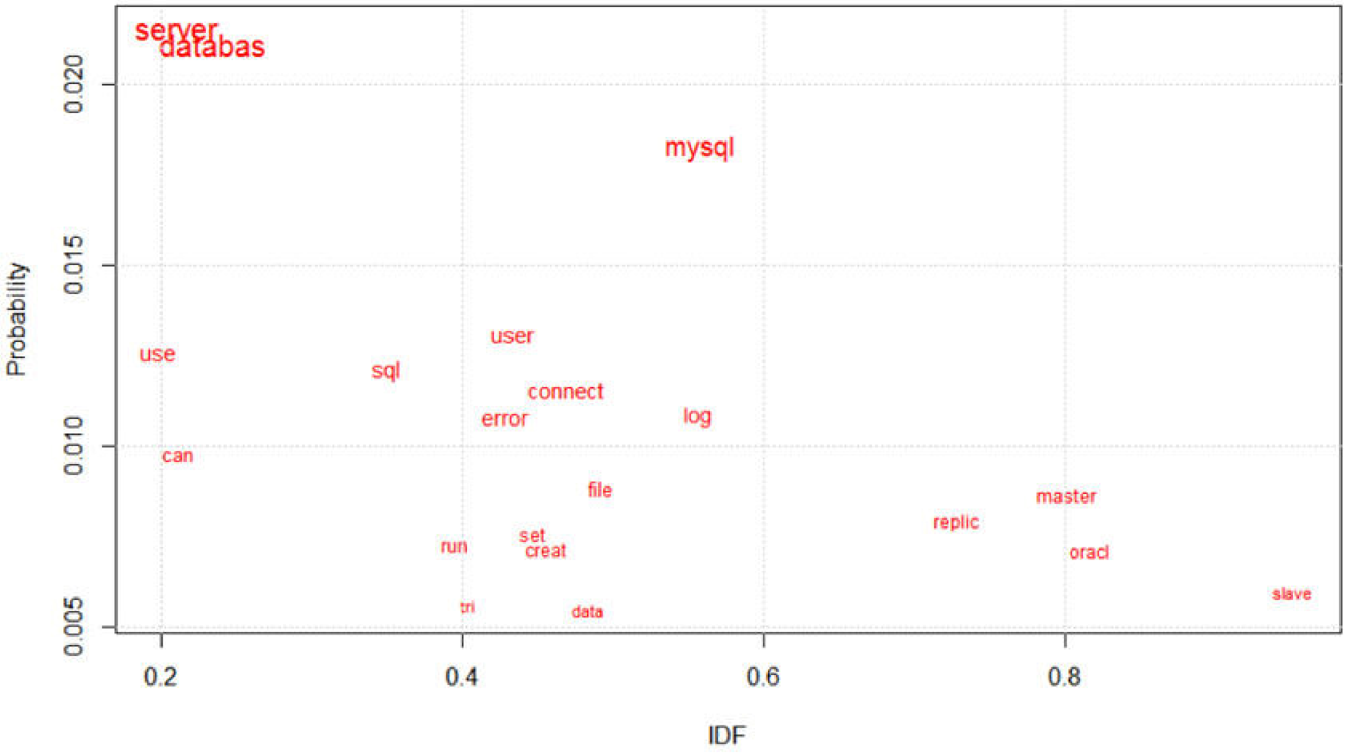


Fig. 7. Probability by idf for Topic 1

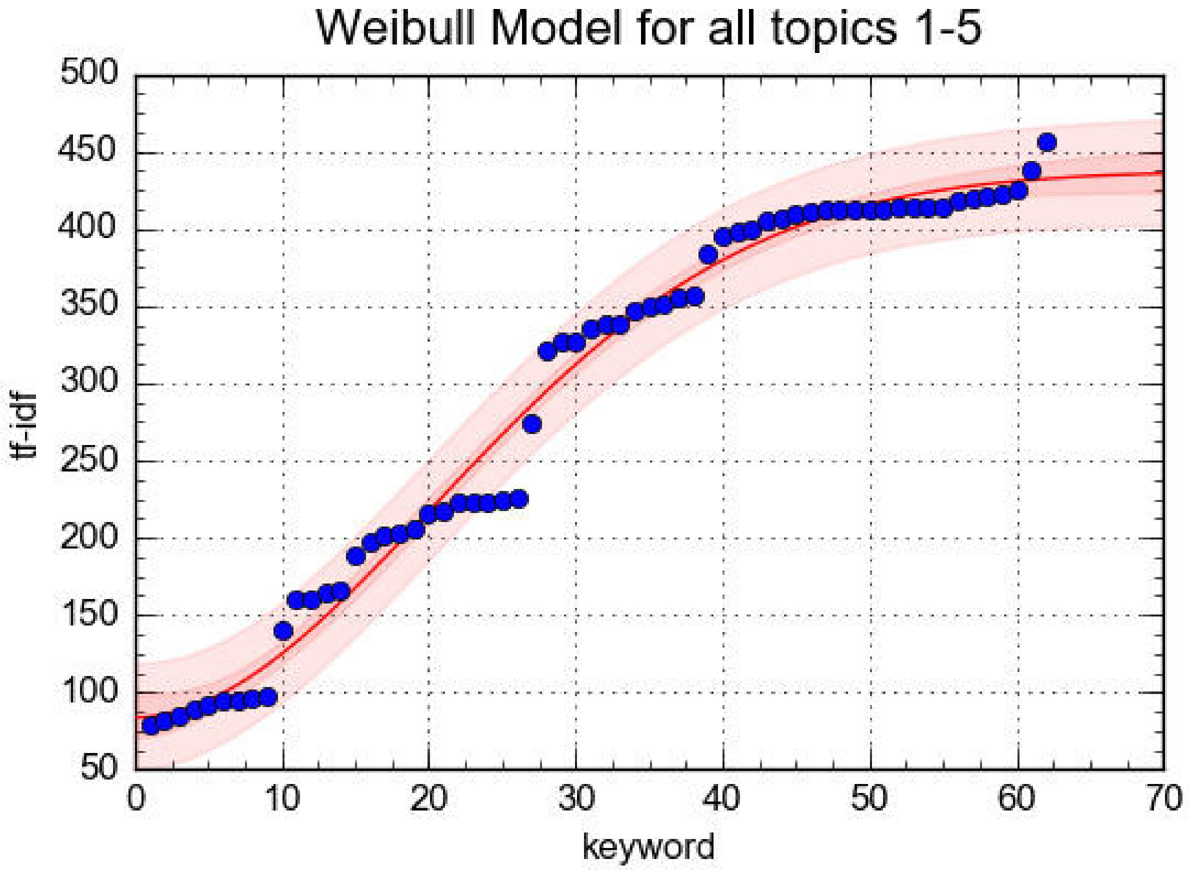
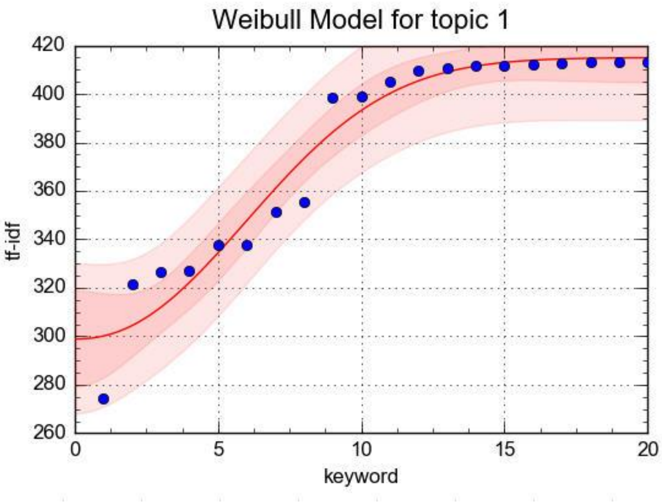


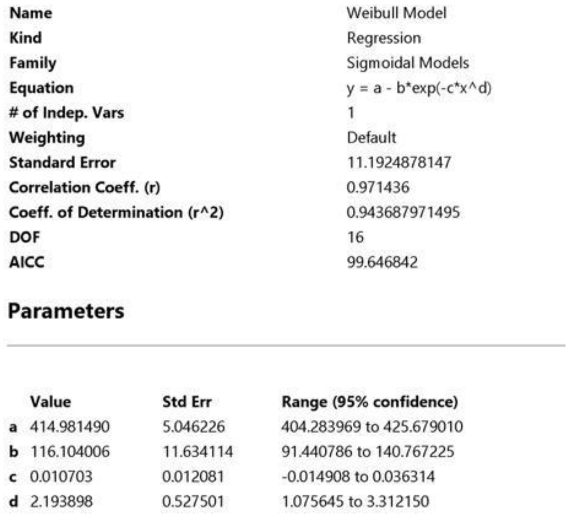
Fig. 8. Weibull Distribution for all topics

Appendix

We are including here the plots for the distributions found by *CurveExpert* for *k=1...5*, *k=10* and *k=15*.

Topic 1:





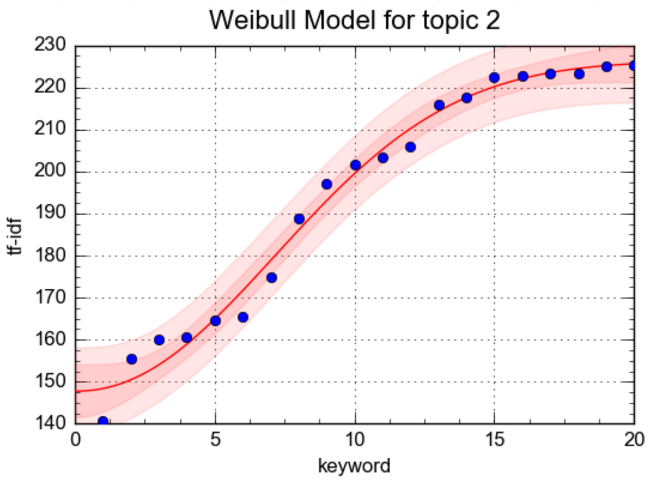
Name	Weibull Model
Kind	Regression
Family	Sigmoidal Models
Equation	y = a - b*exp(-c*x^d)
# of Indep. Vars	1
Weighting	Default
Standard Error	11.1924878147
Correlation Coeff. (r)	0.971436
Coeff. of Determination (r^2)	0.943687971495
DOF	16
AICC	99.646842

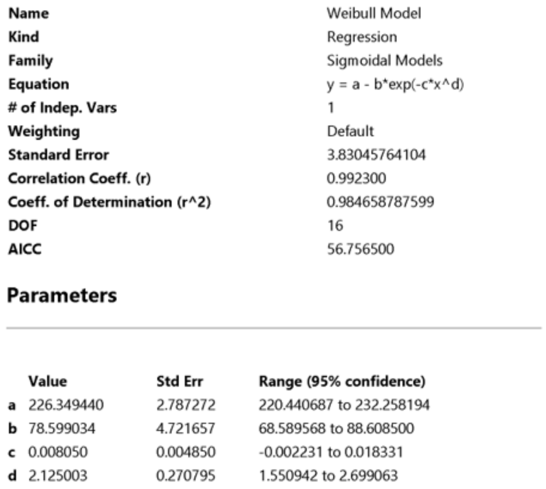
Parameters

	Value	Std Err	Range (95% confidence)
a	414.981490	5.046226	404.283969 to 425.679010
b	116.104006	11.634114	91.440786 to 140.767225
c	0.010703	0.012081	-0.014908 to 0.036314
d	2.193898	0.527501	1.075645 to 3.312150

Fig. A1. Topic 1 Distribution

Topic 2.





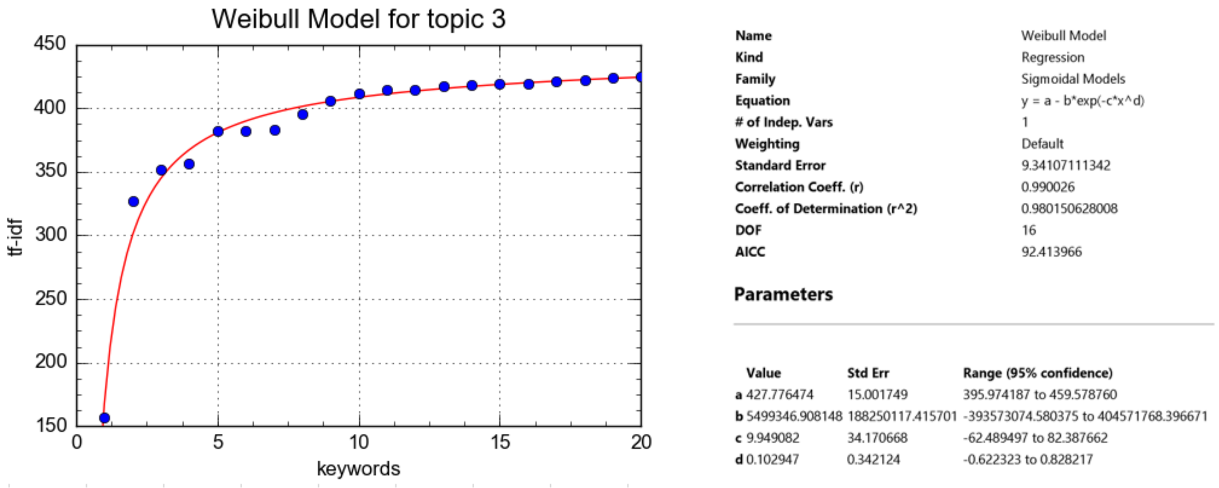
Name	Weibull Model
Kind	Regression
Family	Sigmoidal Models
Equation	y = a - b*exp(-c*x^d)
# of Indep. Vars	1
Weighting	Default
Standard Error	3.83045764104
Correlation Coeff. (r)	0.992300
Coeff. of Determination (r^2)	0.984658787599
DOF	16
AICC	56.756500

Parameters

	Value	Std Err	Range (95% confidence)
a	226.349440	2.787272	220.440687 to 232.258194
b	78.599034	4.721657	68.589568 to 88.608500
c	0.008050	0.004850	-0.002231 to 0.018331
d	2.125003	0.270795	1.550942 to 2.699063

Fig. A2. Topic 2 Distribution

Topic 3.



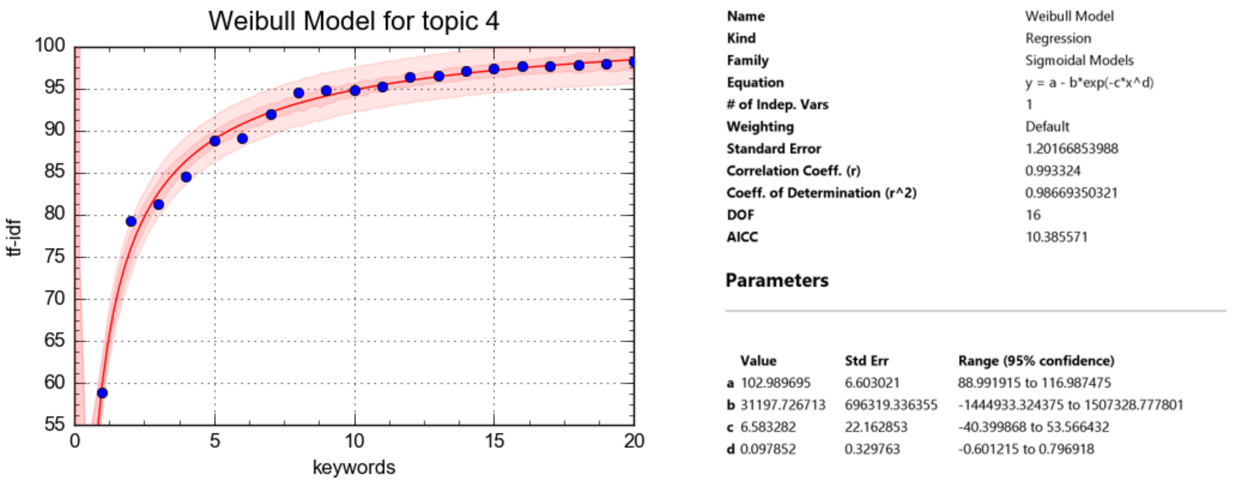
Name	Weibull Model
Kind	Regression
Family	Sigmoidal Models
Equation	y = a - b*exp(-c*x^d)
# of Indep. Vars	1
Weighting	Default
Standard Error	9.34107111342
Correlation Coeff. (r)	0.990026
Coeff. of Determination (r^2)	0.980150628008
DOF	16
AICC	92.413966

Parameters

	Value	Std Err	Range (95% confidence)
a	427.776474	15.001749	395.974187 to 459.578760
b	5499346.908148	188250117.415701	-393573074.580375 to 404571768.396671
c	9.949082	34.170668	-62.489497 to 82.387662
d	0.102947	0.342124	-0.622323 to 0.828217

Fig. A3. Topic 3 Distribution

Topic 4.



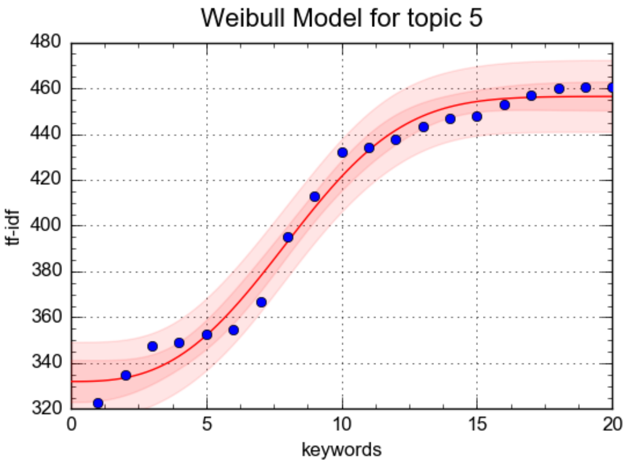
Name	Weibull Model
Kind	Regression
Family	Sigmoidal Models
Equation	y = a - b*exp(-c*x^d)
# of Indep. Vars	1
Weighting	Default
Standard Error	1.20166853988
Correlation Coeff. (r)	0.993324
Coeff. of Determination (r^2)	0.98669350321
DOF	16
AICC	10.385571

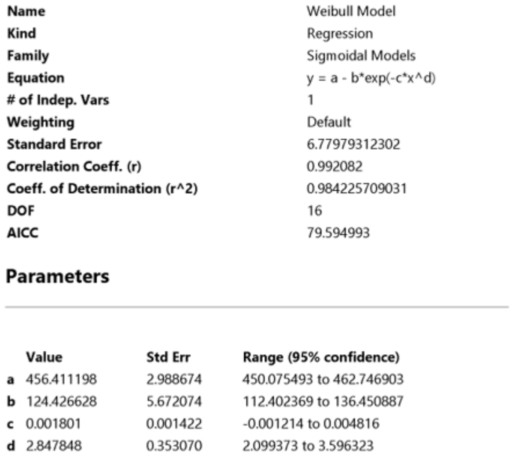
Parameters

	Value	Std Err	Range (95% confidence)
a	102.989695	6.603021	88.991915 to 116.987475
b	31197.726713	696319.336355	-1444933.324375 to 1507328.777801
c	6.583282	22.162853	-40.399868 to 53.566432
d	0.097852	0.329763	-0.601215 to 0.796918

Fig. A4. Topic 4 Distribution

Topic 5.





Name	Weibull Model
Kind	Regression
Family	Sigmoidal Models
Equation	y = a - b*exp(-c*x^d)
# of Indep. Vars	1
Weighting	Default
Standard Error	6.77979312302
Correlation Coeff. (r)	0.992082
Coeff. of Determination (r^2)	0.984225709031
DOF	16
AICC	79.594993

Parameters

	Value	Std Err	Range (95% confidence)
a	456.411198	2.988674	450.075493 to 462.746903
b	124.426628	5.672074	112.402369 to 136.450887
c	0.001801	0.001422	-0.001214 to 0.004816
d	2.847848	0.353070	2.099373 to 3.596323

Fig. A5 Topic 5 Distribution

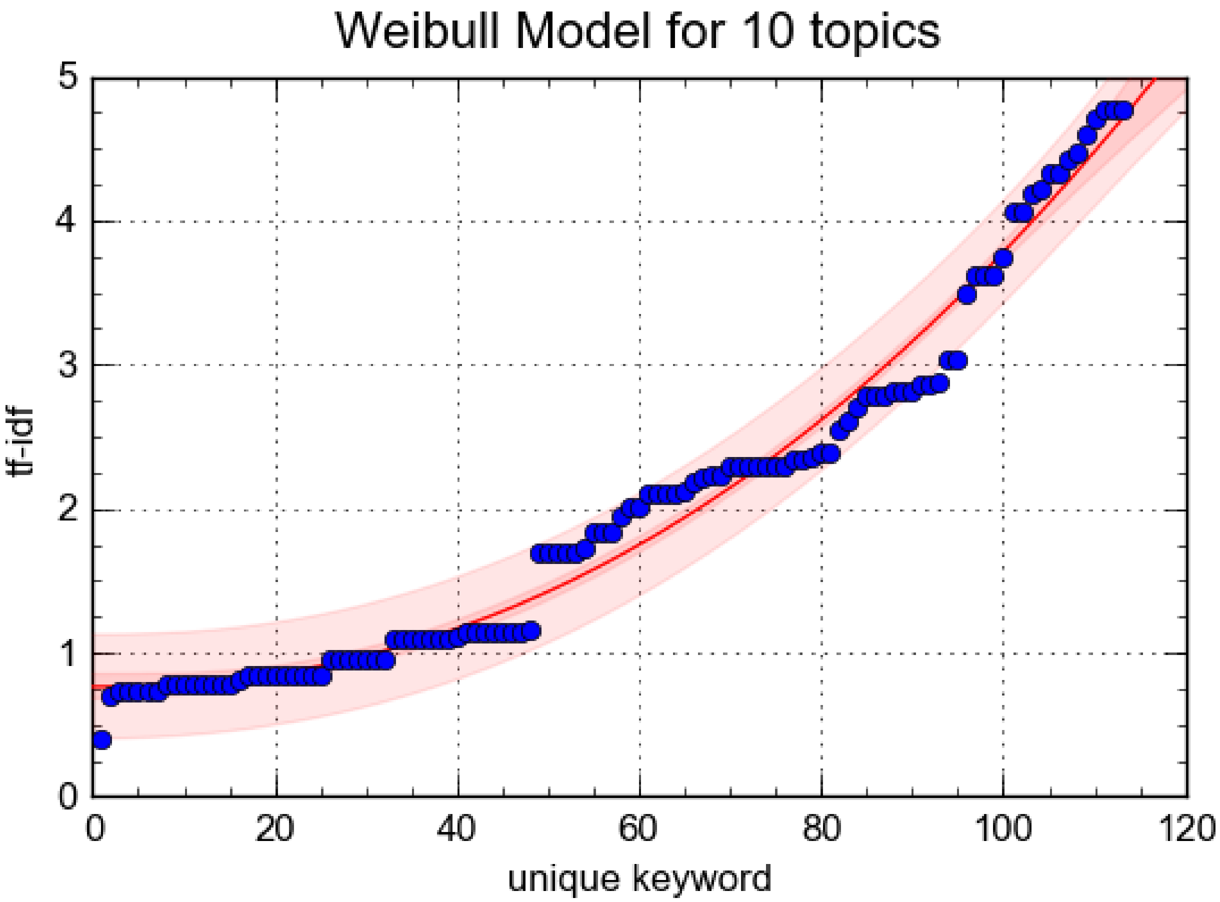
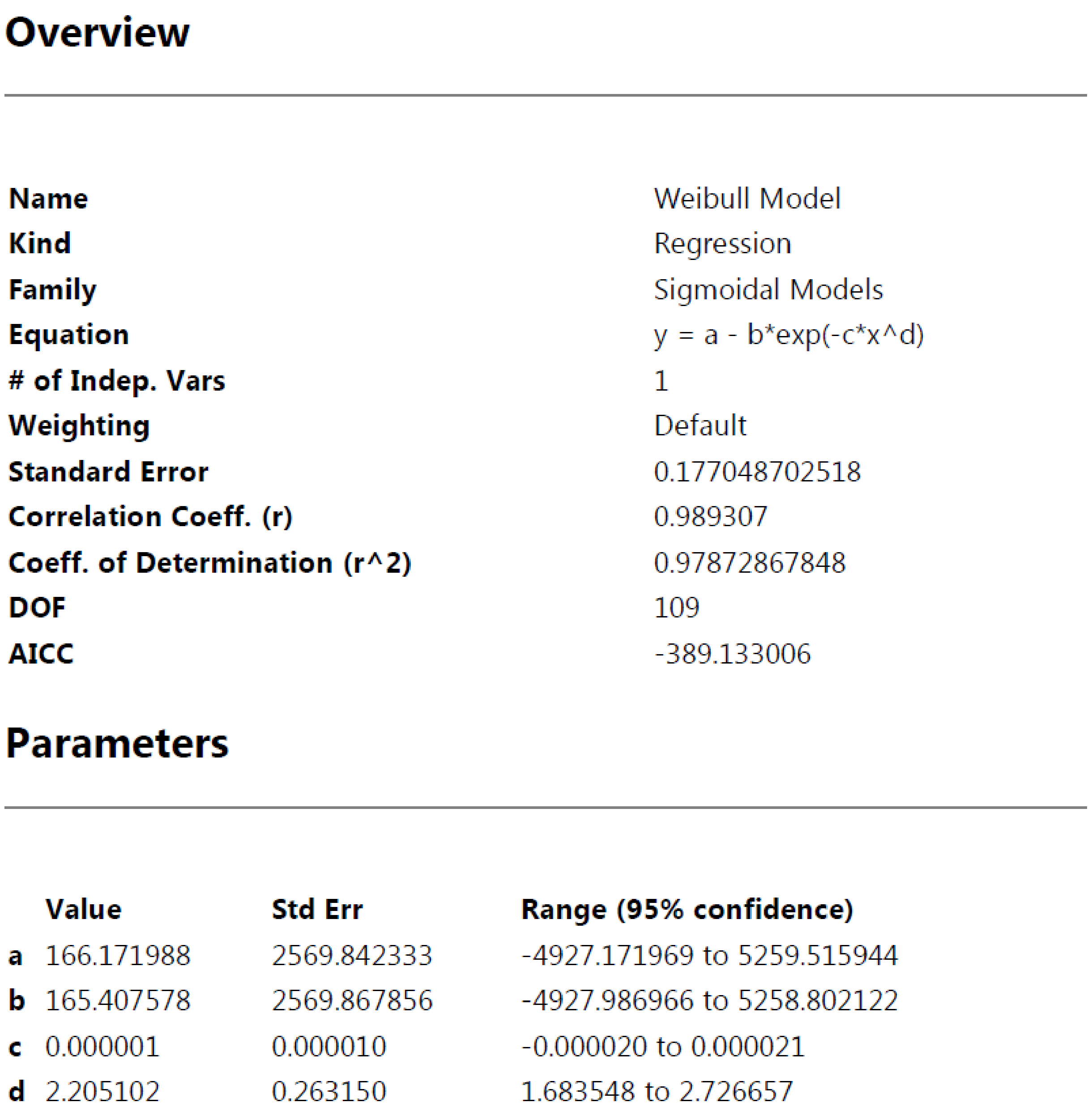


Figure A6 Distribution for 10 topics



Overview

Name	Weibull Model
Kind	Regression
Family	Sigmoidal Models
Equation	y = a - b*exp(-c*x^d)
# of Indep. Vars	1
Weighting	Default
Standard Error	0.177048702518
Correlation Coeff. (r)	0.989307
Coeff. of Determination (r^2)	0.97872867848
DOF	109
AICC	-389.133006

Parameters

	Value	Std Err	Range (95% confidence)
a	166.171988	2569.842333	-4927.171969 to 5259.515944
b	165.407578	2569.867856	-4927.986966 to 5258.802122
c	0.000001	0.000010	-0.000020 to 0.000021
d	2.205102	0.263150	1.683548 to 2.726657

Figure A7 Details for Distribution with 10 topics

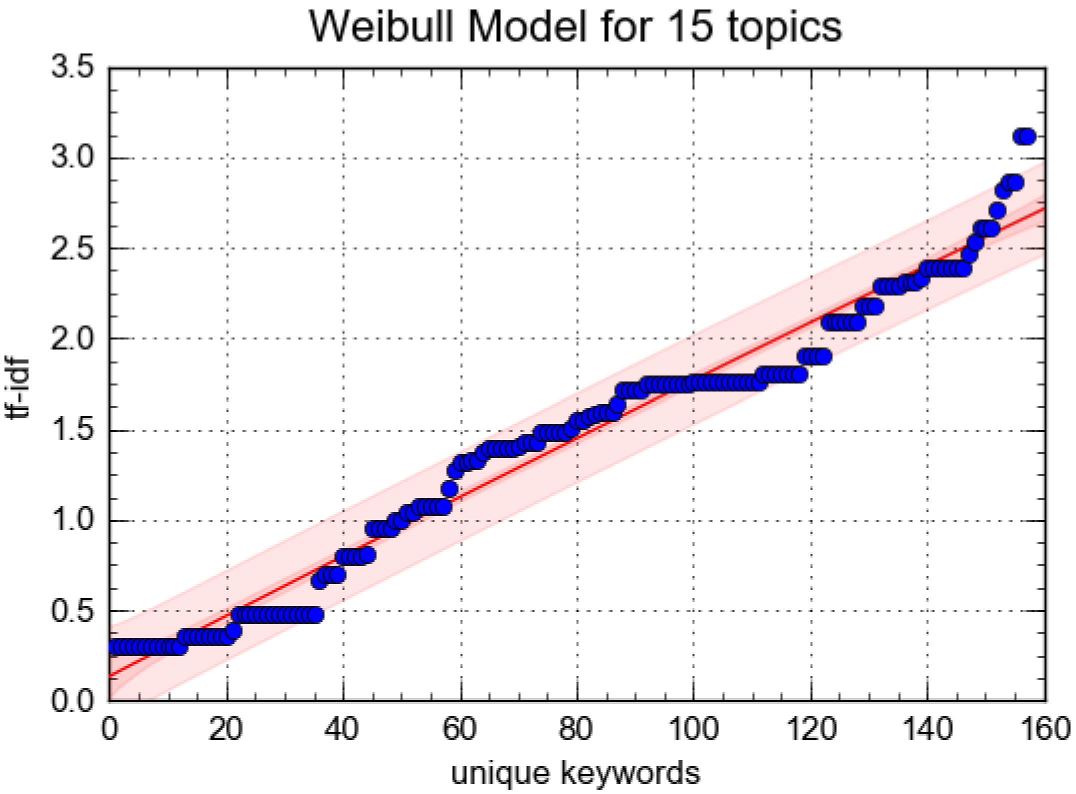
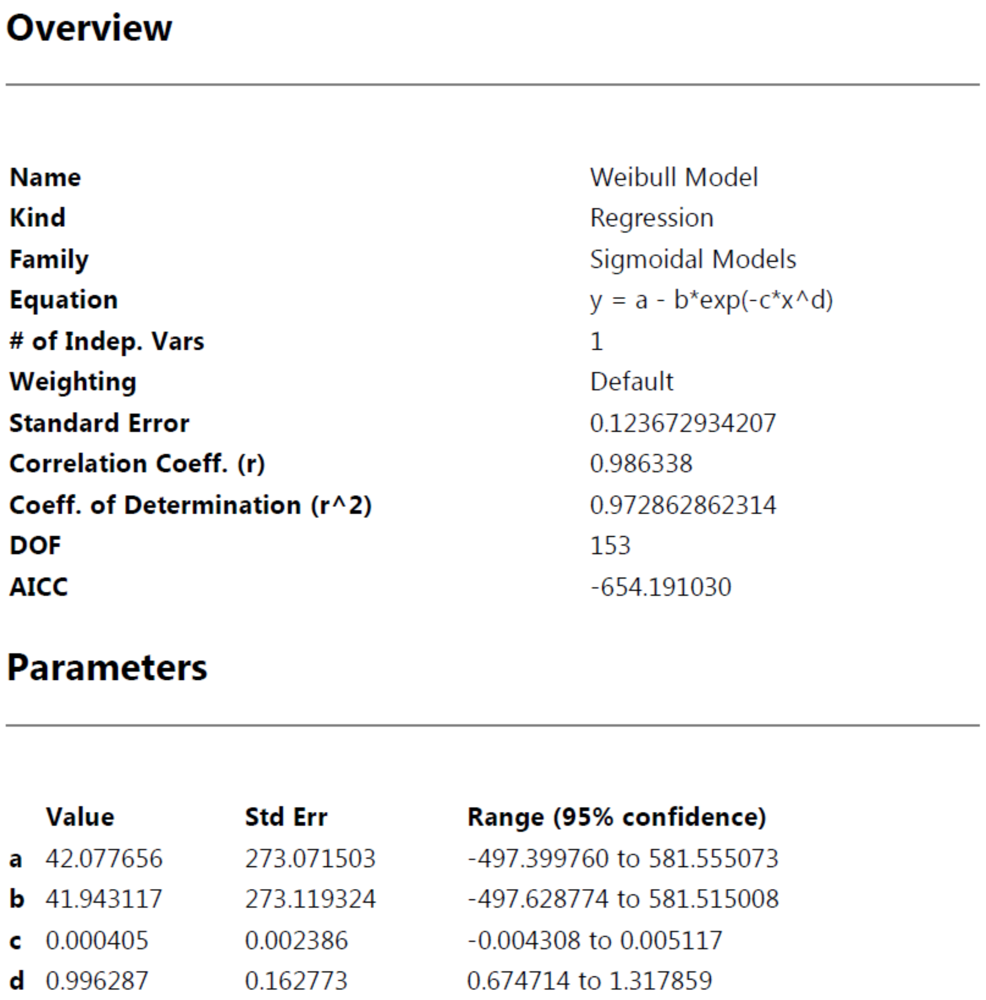


Fig. A8 Distribution for 15 topics



Overview

Name	Weibull Model
Kind	Regression
Family	Sigmoidal Models
Equation	y = a - b*exp(-c*x^d)
# of Indep. Vars	1
Weighting	Default
Standard Error	0.123672934207
Correlation Coeff. (r)	0.986338
Coeff. of Determination (r^2)	0.972862862314
DOF	153
AICC	-654.191030

Parameters

	Value	**Std Err**	**Range (95% confidence)**
a	42.077656	273.071503	-497.399760 to 581.555073
b	41.943117	273.119324	-497.628774 to 581.515008
c	0.000405	0.002386	-0.004308 to 0.005117
d	0.996287	0.162773	0.674714 to 1.317859

Fig. A9 Details for Distribution with 15 topics