MASTER'S PROJECT in IR:
Simulation of Zipf Density

by

the Beta Distribution

2015

By Swetha B Davuluri

Department of Computer Science

and

Information Technology

Montclair State University

***Advisor*: Dr. H.M. Hubey**

Advising Committee

Dr. Jing Peng

Dr. Dajin Wang

Contents

1. Introduction

Every day large amount of information is stored in many firms to store crucial data about the staff, budget, resources in the form of documents, excel spread sheets etc. to track the performance of the company. This process of archiving and storing information is being implemented for thousands of years. With the invention of computers storing information has become very easy and efficient retrieval of information has become necessary. This is the stage where information retrieval comes into role. It was invented in mid 1950s and considerably increased over the last 40 years. One of the most efficient methods was described by H.P. Luhn in 1957, in which he proposed using words as indexing units for documents and measuring word overlap as a criterion for retrieval.

Later various methods were developed during the period of 1970s and 1980s. These methods worked effectively on short text documents available to researchers at that time. But they could not test these methods on large text documents as they are unavailable. So this remained unanswered. Later in 1992, the formation of Text Retrieval Conference (TREC) by many government agencies encouraged scientists to carry on their research on large text documents. This helped them in making advancements in the old technologies and inventions of new ones. TREC has also branched Information Retrieval into relevant but crucial areas like retrieval of spoken information, retrieval of non-English language, filtering of information, interactions of user with a retrieval system etc. The algorithms developed in IR were the first ones to be employed for searching the World Wide Web from 1996 to 1998. Every day many systems in Information Retrieval are used by many users in different departments.

2. Vector Space Model

Vector space model is defined as an algebraic model for describing text documents as vectors of identifiers such as index terms. It is used for filtering information in the large data sources, extracting information, indexing and relevancy rankings. It was first used in the SMART Information Retrieval system. All the documents and the queries required to extract those documents are represented in the form of vectors in the following way.

1) $d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$

2) $q = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$

Each dimension represents a separate term. The value of the term in the vector is a non-zero, if it occurs in the document. It gets increasing with the increase in the number of times of its occurrence in the document. Many kinds of ways of computing the values of these terms are developed by researchers which is also known as (term) weights. One of the best available schemes is tf-idf weighting.

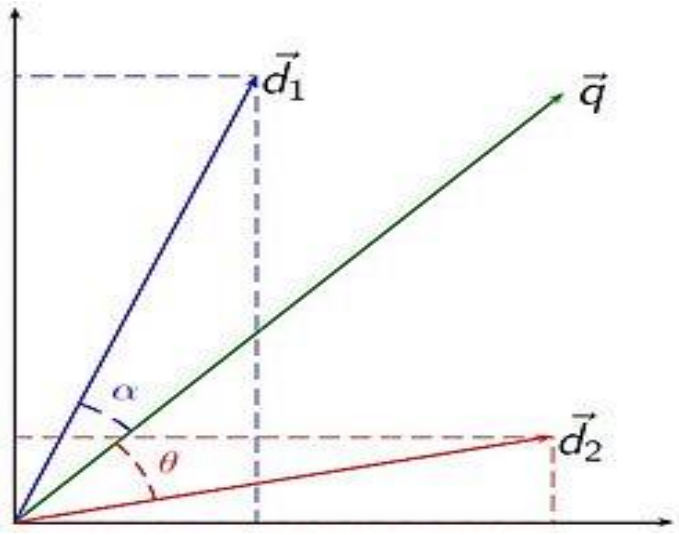


Fig 1: Vector Space Model
http://en.wikipedia.org/wiki/Vector_space_model

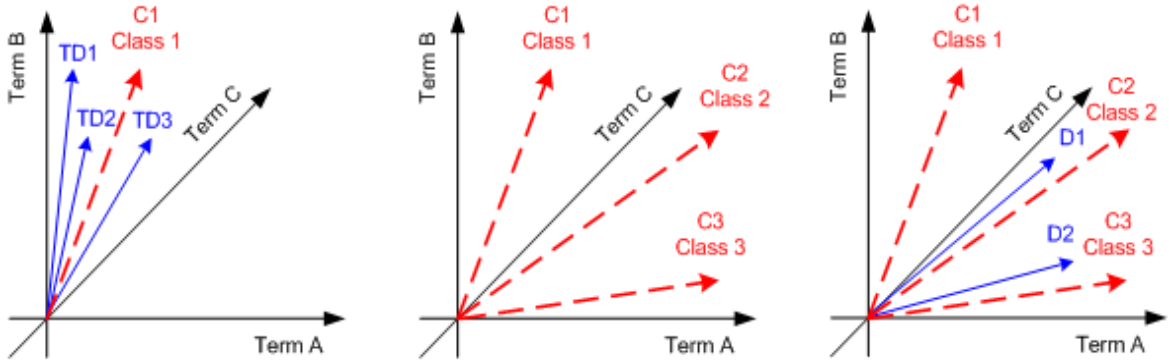
Every application has its own way of defining terms. Single words, keywords, or longer phrases are usually called terms. If the words are chosen to be the terms, the number of words in the vocabulary represents the dimensionality of the vector. Vector operations can be used to compare documents with their queries.

In the vector space model text is represented by a vector of terms. Documents are represented in the form of binary vectors of terms. The definition of a term is not innate in the model, but terms are usually phrases and words. If words are chosen as terms, then every word in the vocabulary turns as an independent dimension in a vector space. Value of each component is the co-occurrence of target words with component label.

Context can be defined based on location, syntactic roles or distribution within collection of documents. In this high dimensional space any text in a document can be represented by a vector.

A non-zero value is assigned to a term if it belongs to a text in the text-vector along the dimension with respect to the term. As any text contains a limited set of terms (the vocabulary can be millions of terms), most text vectors are scattered.

In general no term is assigned a negative value. So all these text vectors reside in the positive quadrant of vector space. Queries are also similar to documents. Length and direction of vectors act as attributes for query and document weights. The rank of the retrieved documents is derived from the vector distance measure between the query and documents.



A) Training Document Representations in Vector Space define Class Representation (Centroid)

B) Class Representation by Class Vectors (CentroidS)

C) Classification of Documents by similarity between Document Vector and Class Vector

Fig 2: Vector Space Model
http://www.iicm.tugraz.at/cguetl/courses/isr/opt/classification/Vector_Space_Model.html

During keyword search, relevant rankings of documents can be measured using the hypothesis of similarities theory on documents, by comparing the deviation of angles between the original query vector where the query is represented as the same kind of vector as the documents and each document vector. In general, instead of the angle itself, it is very easy to calculate the cosine of the angle between the vectors.

3) $$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

Where $\mathbf{d_2} \cdot \mathbf{q}$ is the intersection (i.e. the dot product) of the document (d2 in the figure to the right) and the query (q in the figure) vectors, $\|\mathbf{d_2}\|$ is the norm of vector d2, and $\|\mathbf{q}\|$ is the norm of vector q. The norm of a vector is calculated as such:

4) $$\|\mathbf{q}\| = \sqrt{\sum_{i=1}^{n} q_i^2}$$

There are three stages in the vector space model procedure. The first stage is indexing the document where the crucial words containing content are extracted from the text of document. The second stage is the weighting of the indexed terms according to their repetition in the text document to enhance retrieval of required document relevant to the user. The last stage prioritizes the document in the pool of documents with respect to the query according to a similarity measure.

a. Document Indexing

It is common that many of the words in a document do not describe the content, but are just used to support the important words like *the*, *is*, *can* etc. All these non-significant words are removed by using automatic document indexing from the document vector. So the document vector will only be formed by words containing content. This indexing can be based on term frequency.

These terms have both low and high frequency within a document which are considered to be function words. But in practice, it is difficult to implement the term frequency in automatic indexing. Instead stop list is used to remove high frequency words (stop words). This stop list holds common words and this makes the indexing method language dependent. Usually with the help of a stop list around 40-50% of the total number of words in a document is removed.

Methods of non linguistic for indexing have also been executed. The hypothesis that there is some statistical difference in the distribution of content bearing words and function words acts as basis for probabilistic indexing. Probabilistic indexing is used to rank the terms in the collection of words with respect to the term frequency in the entire collection. The function words are formulated by applying Poisson distribution over all documents, as terms bearing content cannot be modeled. Bernoulli model is an extension of the use of Poisson model. Now-a-days, a new method which uses serial clustering of words in text has been developed for automatic indexing. If the word is content bearing, the value of such clustering is an indicator.

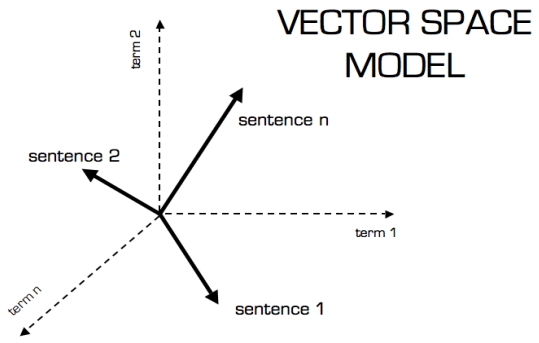


Fig3:Vector Space Model
http://pyevolve.sourceforge.net/wordpress/?p=2497

b. Weighting of indexed terms

Weighting of indexed terms can be explained by regulating the exhaustivity and specificity of the keyword search, where the exhaustivity is related to recall and specificity to precision. Single term statistics is the factor which is completely used as a base for the term weighting for vector space model. Term weighting has three main factors which are collection frequency factor, term frequency factor and length normalization factor. Multiplication of these three factors makes the resulting term weight.

As stated by Luhn, the frequency of occurrence of terms is generally used as a weighting scheme for terms within a text document. The term frequency is generally used as the root of a weighted document vector and is kind of content descriptive for the documents. Term weighting can also be done using binary document vector. But the results are more accurate with term frequency when using the vector space model.

Many weighting schemes have been developed to distinguish one document from the other. This factor is called collection frequency document. Most of them for example the inverse document frequency assume that the importance of a term is proportional with the number of documents the term appears in. Many Experiments have proved that this document discrimination factors lead to a more effective retrieval which is an improvement in precision and recall.

The third possible weighting factor is a document length normalization factor. When compared with the short documents, usually long documents have larger set of terms. This increases the probability of extracting long documents than the shorter ones. Many types of weight schemes have been researched and more precise results with respect to recall and precision are obtained using term frequency with inverse document frequency and length normalization.

c. Similarity Coefficients

Using associative coefficients based on the inner product of the document vector and query vector the similarity in vector space models is determined. Here word overlap indicates similarity. Usually the normalization of inner product is done. The most established similarity measure is the cosine coefficient. It is the angle between the query vector and the document vector. For example Jaccard and Dice coefficients are the other examples.

3. Precision and Recall

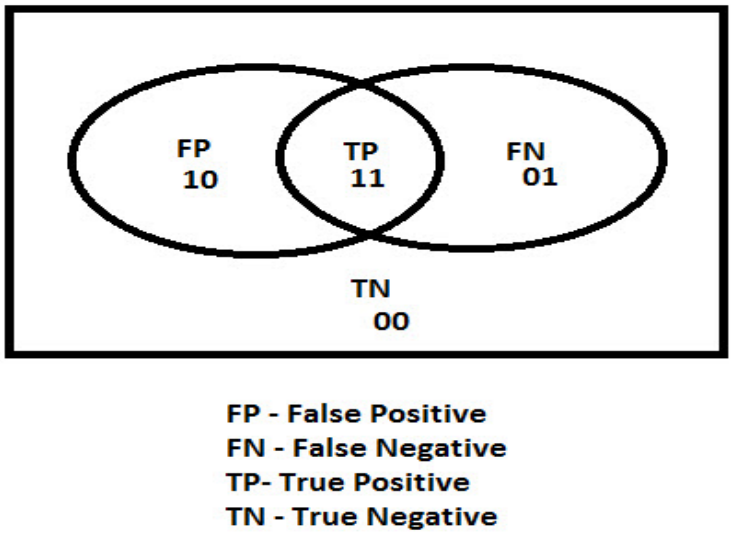


Fig 4: Vector Space Model

Precision is the fraction of retrieved documents that are relevant to the need of user's information.

5) $$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Precision usually takes all documents that are retrieved into consideration. It can also be derived by using a cut-off rank by extracting only the topmost priority results returned by the system. If n is the cut-off rank, it is represented as precision at n or P@n.

For example for searching a text on a group of documents, precision is the number of correct documents divided by the number of all returned documents. Precision is also used with recall, the percent of all related documents which are returned by the search of text in a pool of documents. To generate a single measurement for a system these two measures are sometimes used together in the F1 Score (or f-measure).

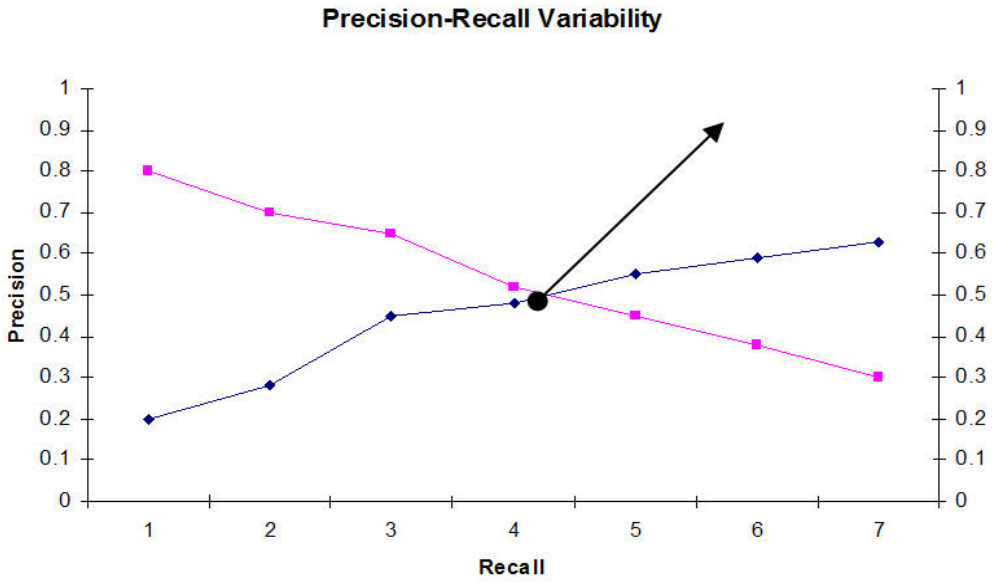


Fig 5: Precision - Recall
http://acl2014.org/acl2014/P14-1/xml/P14-1045.xhtml

Recall is defined as the fraction of the documents which are retrieved successfully relevant to the implied query.

6)
$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

For example for text search on a set of documents recall is the number of correct documents divided by the number of documents that should have been retrieved. Recall is called sensitivity in binary classification. Hence it can be defined as the probability that a relevant document is retrieved by applying the query. It is unimportant to achieve 100% recall by extracting all documents in return to any query. Therefore only recall is not enough. We should also measure the number of non-relevant documents which can be done by calculating the precision.

Precision and recall are defined with the following equations.

7) $$\text{Precision} = \frac{tp}{tp + fp}$$

8) $$\text{Recall} = \frac{tp}{tp + fn}$$

4. TF-IDF Model

An information retrieval system faces difficulty to anticipate the query terms (words) users will issue in order to extract the relevant documents. Sometimes almost every word in the documents can be a possible query term. As in many Web search engines the best solution is to index all the words appearing in the documents.

After indexing all the terms in a document, the next step is, if given a query (contain multiple terms), how to calculate a relevance score for every documents in a pool of documents. The simplest calculation will be to count the original number of words the query has in common with a document. This number is called co-ordination level.

9) For a term i in the document

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

$$\text{RelScore}_d = \sum_{t \in q \cap d} w_{t_d}, w_{t_d} \equiv TF_{t,d} \cdot IDF_t$$

Relevance score does not depend on the query term frequency in a document. Usually the document containing more number of query terms is more relevant. We can also use Term Frequency which is the frequency of occurrence as a weight. The summation of the frequencies of all query terms in a document is the relative score. However, completely depending on frequency of term is still constrained because query terms vary in their capability to discriminate documents. If a query term occurs in many documents, it is not a good discriminator. It should be given less weight than the term occurring in few documents. For examples, when querying the

keyword “information retrieval”, it is very rare that the documents containing “information” will be more relevant than documents containing “retrieval”. In 1972, Sparck Jones introduced Inverse Document Frequency (IDF) which is a measure of term specificity (discriminative power).

IDF depends on counting the number of documents in the pool of documents. A term has less discriminative power, if it occurs in many documents. Sparck Jones illustrated that IDF exceeds Co-ordination level matching. IDF almost occurs in every term weighting scheme, coupled with Term frequency which is the frequency of occurrence of a term t in document d. This scheme of weighting terms is usually called TF-IDF. Heuristic study is the basis for Inverse document frequency (IDF). Term Frequency and Inverse Document Frequency is represented by Tf-Idf. Typically, the tf-idf weight is composed of two terms: the first computes the normalized Term Frequency (TF) which is the number of times a word appears in a document. This is divided by the total number of words in that document. Inverse Document Frequency (IDF) is the second term mentioned for calculating tf-idf weight. This is computed by taking the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

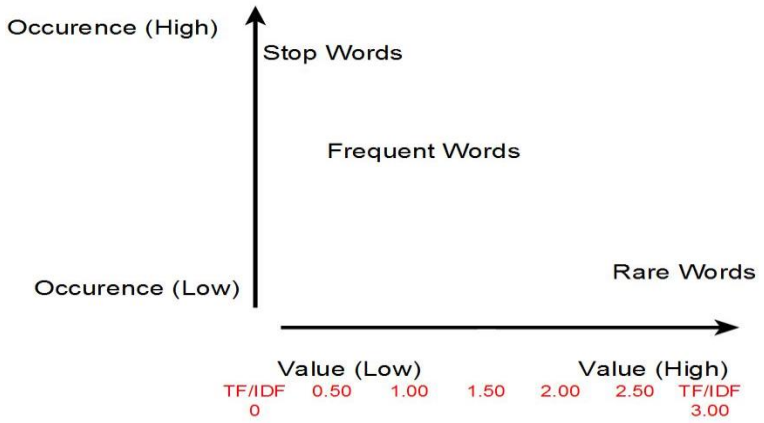


Fig 6: TF-IDF Model
http://trimc-nlp.blogspot.com/2013/04/tfidf-with-google-n-grams-and-pos-tags.html

Text mining and information retrieval is done using the values of *T*f-idf weights. This measure weight is a statistical evaluation to estimate how crucial a term is to a document in a collection of documents or corpus. The importance of terms grows proportional to the number of times a term appears in the document but is offset by the frequency of the term in the corpus. Tf-idf weighting scheme variations are usually performed by search engines. It acts as a central tool in ranking and scoring the relativity of document given a query by user. One of the simplest ranking functions is computed by summing the tf-idf for each query term; many more sophisticated

ranking functions are variants of this simple model. Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

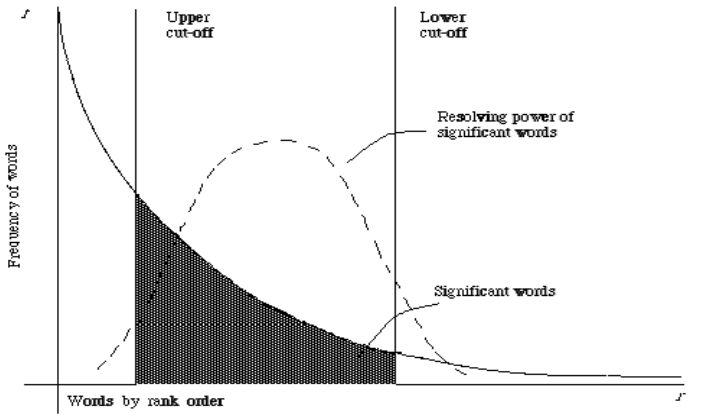


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

Fig 7: TF-IDF Model
http://www.cs.sfu.ca/~cameron/Teaching/D-Lib/IR.html

Term Frequency (TF):

It measures how many times a word appears in a document. As we have different documents which have different lengths, there is a possibility that a term can appear more frequently in long documents when compared with the smaller ones. Then normalization is done on the term frequency by dividing with the length of the document.

Equation for Term Frequency (TF) :
TF(t) = (Number of times term t occurs in a document) / (Total number of terms in the document).

Inverse Document Frequency (IDF):

It measures how crucial a word is in the document. While calculating Term Frequency (TF), all the words in document are considered equally important. But we have certain words, such as "is", "that","and" and "of". These words will appear a lot of times in the document but are of little important. Thus we need to weigh down the more frequent terms and scale up the rare terms by computing in the following way.

Equation for Inverse Document Frequency (IDF):
IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

Let us illustrate the concept of Term frequency (TF) and Inverse document frequency (IDF) with an example. For example, consider a document containing 150 words where in the word cat appears 5 times. The term frequency (i.e., tf) for cat is then (5 / 150) = 0.033. Now, assume we have 10 million documents and the word cat appears in ten thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 10,000) = 3. Thus, the Tf-idf weight is the product of these quantities: 0.033 * 3 = 0.099.

5. Rejection Sampling

Rejection sampling is one of the popular methods which is used to produce random samples from an arbitrary target probability distributions. It requests the outline of a suitable proposal probability density function (pdf) from which applicant specimens can be drawn. These examples are either acknowledged or rejected relying upon a test including the proportion of the target and proposition densities. Rejection sampling system is an effective calculation to test from a log-curved target density that achieves high acknowledgement rates by enhancing the proposition density whenever a sample is rejected. In this venture, the Rejection sampling method that can be connected with a wide class of target probability distributions, potentially non-log curved and displaying different modes. The proposed strategy yields an arrangement of proposition densities that merge toward the target pdf, therefore accomplishing high acceptance rates.

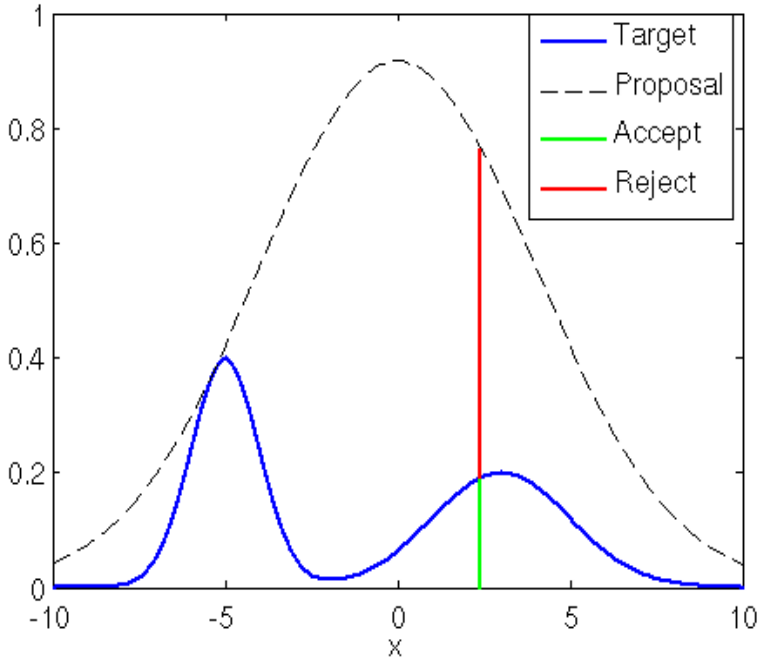


Figure 8: Rejection Sampling

https://theclevermachine.wordpress.com/2012/09/10/rejection-sampling/

Generating arbitrary numbers is a computational charge. It has numerous difficulties. A decent approach to create random numbers in computational insights includes studying different distributions utilizing computational techniques. Subsequently, the probability distribution of every practical number can be uniform or pseudo-arbitrary.

Generating random number is based upon Rejection Sampling. The principle thought is the point at which you produce a number in the certain range, yield that number quickly. In the event that the number is out of the certain extent, reject it and re-test once more. As every number in the expected reach has the same probability of being picked, a uniform appropriation is created.

Assume that we need to test from a target distribution f(x) that is troublesome or difficult to test from directly. Assume additionally that we have a proposition distribution g(x) from which we have a sensible technique for examining (e.g. the uniform dissemination).

Step 1: Generate X,Y, on uniform density [0,1].

Step 2: Reject if Y> f(x)

The idea of rejection sampling is, envision the large rectangular board, in that dashes are tossing on the board to charting the thickness capacity of a random variable. Imagine that the darts are consistently distributed around the board. Now, reject the darts which are tossing outside of the limit. The remaining darts will be distributed consistently inside the territory under the limit, and the x-positions of these darts will be distributed by random variable's thickness. This is because there is the most space for the darts to land where the bend is most and therefore the probability density is most.

The general method of rejection sampling expect that the board is not so much rectangular yet it can be of any shape as per the distribution strategy that we know how to test from and which is in any event as high at each point as the distribution we need to sample from, so that the previous totally encloses the latter.

The central figure of value of a rejection sampler is the mean acceptance rate, which implies the imagined number of acknowledged samples over the aggregate number of proposed ones.

6. Beta Distribution

The Beta distribution is a probability distribution which is continuous and is defined in the interval [0, 1]. This distribution depends mainly on two positive parameters denoted by α and β. These parameters are responsible to control the shape of the beta distribution curve. The Beta distribution is an excellent way to represent outcomes like probabilities or proportions. This beta distribution is normally used to design the pattern of random variables when restricted to a finite interval in a variety of day to day applications. One of its most common uses of beta distribution is to model one's uncertainty about the probability of success of an experiment.

Assume a probabilistic experiment can have just two results, either success, with probability X, or failure, with 1-X probability. Assume that X is obscure and all its chances are considered just as equally likely. This uncertainty can be depicted by allotting to X a uniform distribution on the interval [0, 1]. This is proper because of the fact that X, being a probability, can take just values between 0 and 1; moreover, the uniform distribution gives the same probability to all the points which are present in the given interval, which reflects the way that no possible value of X is, from the earlier, regarded more probable than all the others.

Now, assume that we perform n independent repetitions of the test and we watch k successes and n-k failures. After performing the analysis, we commonly need to know how we ought to change the distribution at first assigned to X, to properly consider the data gave by the observed results.

As such, we need to calculate the conditional probability of X, conditional on the quantity of successes and failures we have watched. The outcomes of this calculation are a Beta distribution. Specifically, the conditional distribution of X, conditional on having observed k successes out of n trials, is a Beta distribution with parameters k+1 and n-k+1.

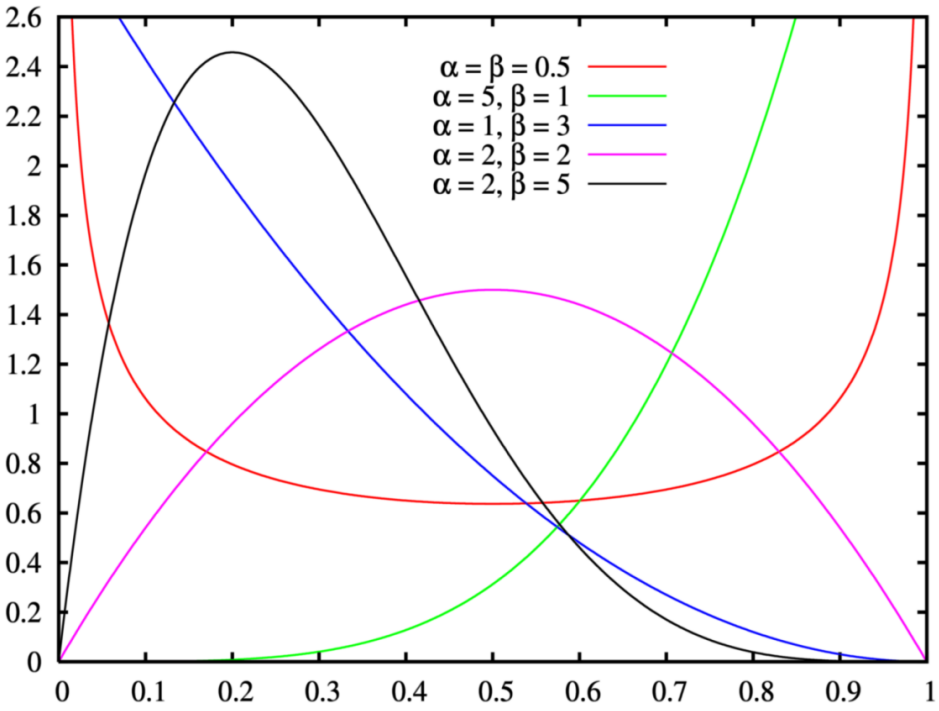


Figure 9: Beta Distribution

http://stats.stackexchange.com/questions/114004/when-is-beta-distribution-bell-shaped-or-concave

Beta Distribution is defined on the continuum between 0 and 1.The parameters α and β mainly influences the shape of the distribution curves. For example, if $\alpha < 1$ and $\beta < 1$, the shape of graph will be a "U" (see the red plot on the picture above). If $\alpha = 1$ and $\beta = 2$, the graph is a straight line. If $\alpha = 1$ and $\beta = 3$ the shape of graph is almost a straight line (blue line).

Beta distribution is a distribution that models events which are constrained to take place within an interval defined by a minimum and maximum value. The Beta function B in the denominator plays the role of a "normalizing constant" which assures that the total area under the density curve equals 1. The Beta function is equal to a ratio of Gamma functions. Keeping in mind that for integers, $\Gamma(k) = (k-1)!$

Beta Distribution

Parameters: $\alpha > 0$ and $\beta > 0$

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)} x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0 \le x \le 1$$

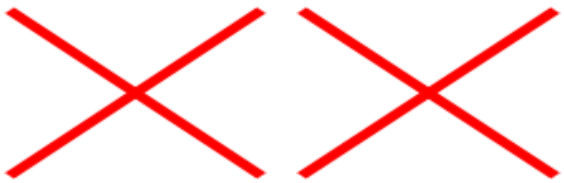
$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

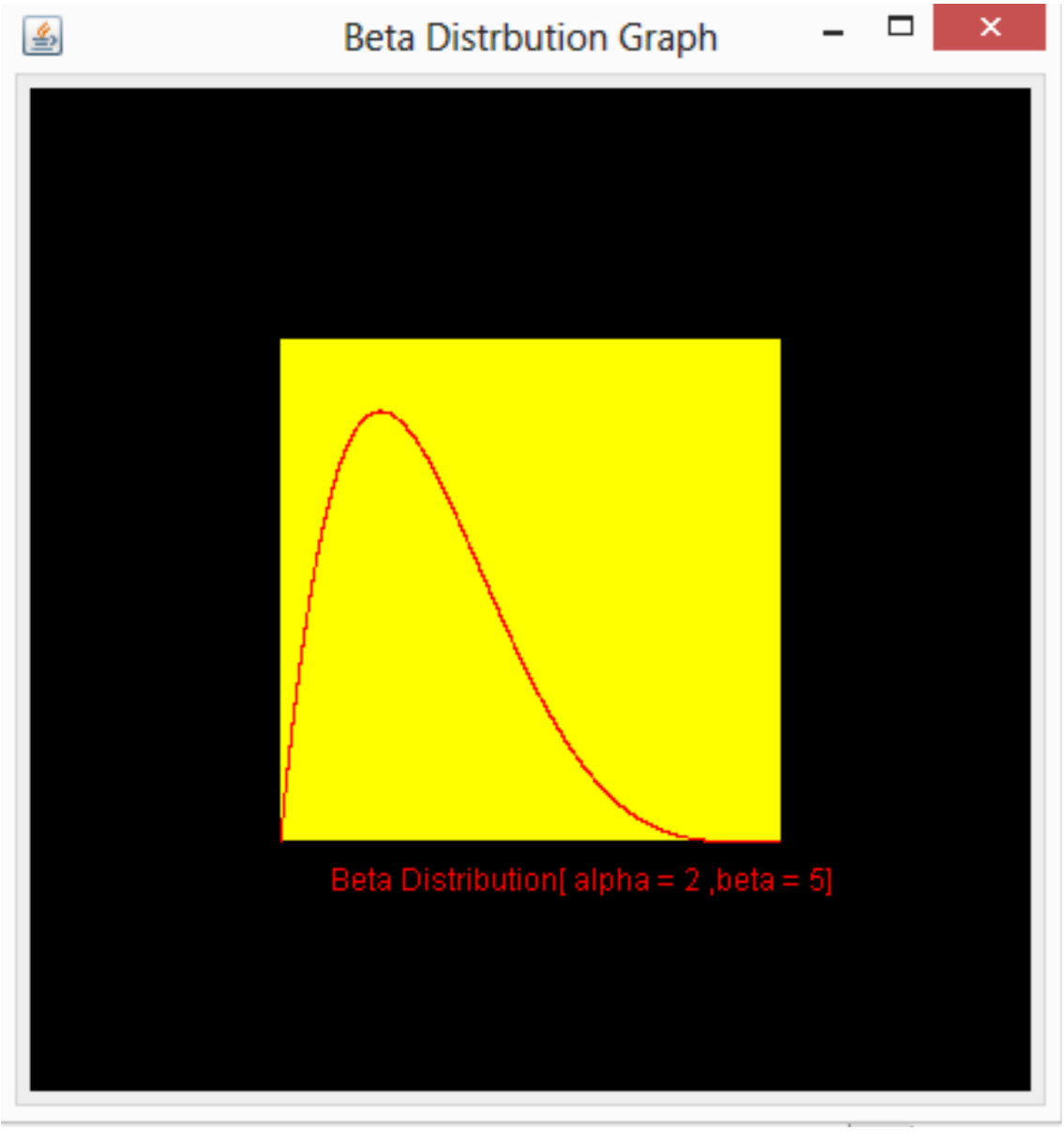
$$x_{\text{mode}} = \frac{\alpha-1}{\alpha+\beta-2}$$

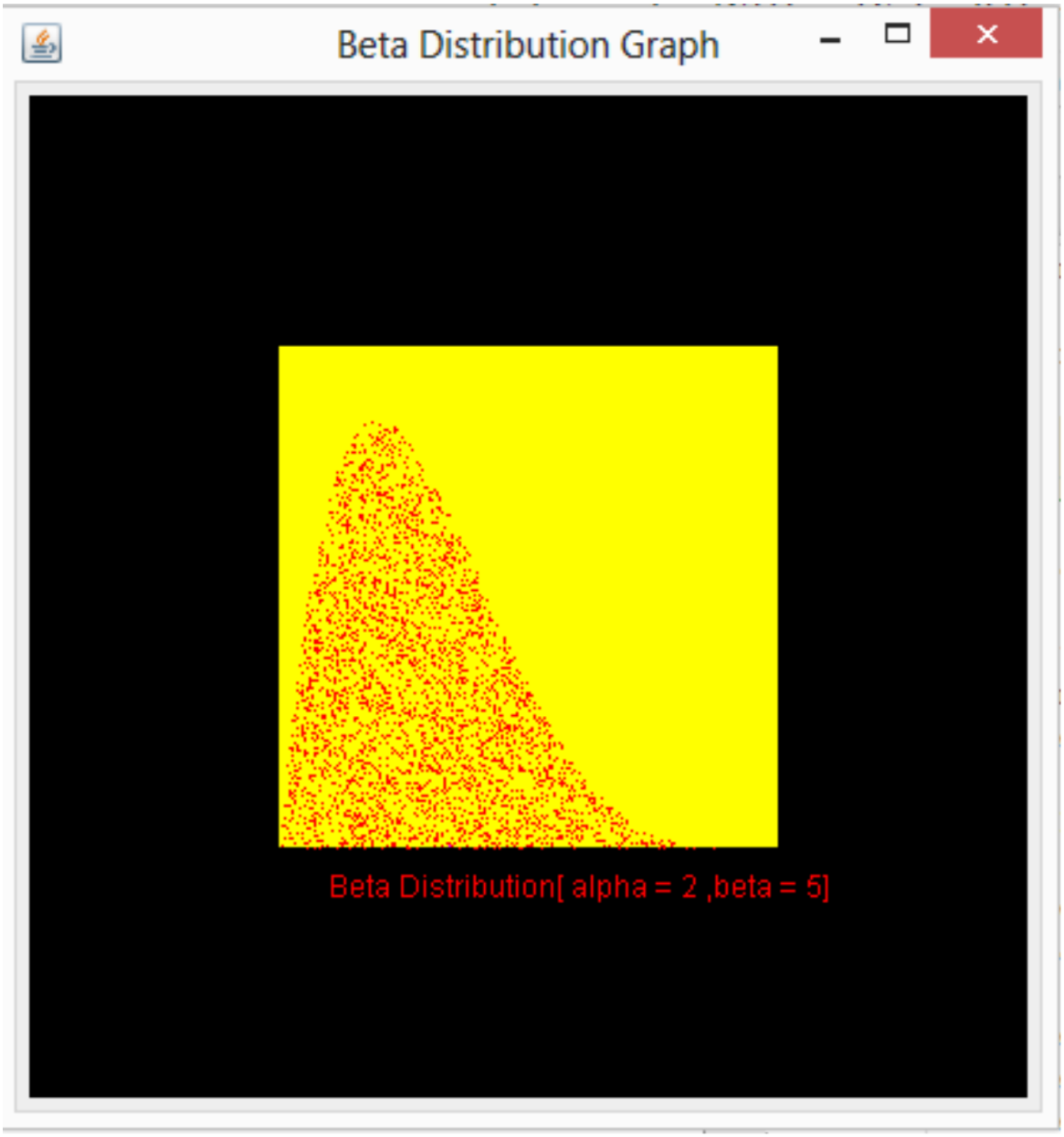
7. Simulation

Program



8. Outputs




Beta Distribution Graph
Beta Distribution[alpha = 2 ,beta = 5]

9. Conclusion

Information Retrieval plays a key role for the advancement of business in many organizations. Different concepts which are used in extracting information is discussed in this paper. It includes brief account on the different steps implemented in Vector space model, concepts of precision and recall, Receiver operator characteristics, Tf-Idf models. Information retrieval is a fastest growing technology in many industries to collect data in a required format, process the information and helps the analyzers in making effective business decisions.

10. References

1. http://en.wikipedia.org/wiki/Sensitivity_(tests)
2. http://www.tfidf.com/
3. http://web4.cs.ucl.ac.uk/staff/jun.wang/blog/2009/07/08/tf-idf/
4. http://cogsys.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html
5. http://en.wikipedia.org/wiki/Vector_space_model
6. http://nlp.stanford.edu/IR-book/
7. http://comminfo.rutgers.edu/~aspoerri/InfoCrystal/Ch_2.html
8. http://www.mathworks.com/help/stats/beta-distribution.html
9. http://en.wikipedia.org/wiki/Beta_distribution
10. http://mathworld.wolfram.com/BetaDistribution.html