Science Communication: Research Productivity and Data Analytics using Open Source Software

Vijay Kumar Verma

Central Library

IIT Delhi

Week: 05

Lecture 20: Zipf's Law

Dear Learners, Welcome to our NPTEL course on Science Communication, Research Productivity and Data Analytics using Open Source Software.

Now we are going to cover Zipf's Law. Zipf's Law basically is an empirical Law which describes the distribution of frequency of elements in a dataset. Zipf's Law deals basically with the distribution of frequency of elements in a dataset. If you see all around us then we find that most of the things are average. For example, if you want to find out the height of the student of a particular class, we get almost the average height, in some cases the student would be very tall, another student would be very short but usually you get the average height of the student.

But this does not happen in the case of language. Basically when we talk about Zipf's Law, Zipf's Law is the distribution of frequency of elements in a dataset. it describes the frequency distribution of words in a large text corpus. For example, if you have a large text corpus or a corpus having a million or billion of words. For example, let us see Wikipedia or any other archive of any particular newspaper X. And if you want to find out which is the word which occurred the maximum number of times, which is the word which occurred the least number of times, here we have the use of Zipf's Law. Basically Zipf's Law describes the frequency distribution of words in a large text corpus and Zipf's Law is named after linguist George Kingley Zipf who observed that a frequency of a word in a text is inversely proportional to its rank.

But Zipf observed that the frequency of a word is inversely proportional to the rank. That means the word of a large corpus which rank is first and the word which in that particular corpus having rank second, there is a relationship between the rank and frequency. And the frequency of a word in a text is inversely proportional to the rank. That means the rank multiplied by frequency in any given corpus is a constant. The frequency of the word is inversely proportional to its frequency rank. Based on his observation he found that or based on the observation of Zipf, we can say that 20% of the words filled 80% of the corpus, which means some words are used very frequently in a corpus as compared to other words.

Some words use very frequently, some words use less frequently and the words which are used frequently are short and the rare words that means the words which are used less frequently are longer. Many words in a corpus, in a particular language Corpus, are used very rarely and this type of word is known as hapax legomena. Its singular is hapax legomenon. That means the word which is used very rarely is hapax legomena. The word that is used only once is basically the hapax legomena. This famous example of hapax legomena from the work of Shakespeare is honorificabilitudinitatibus. And this particular word has been used only once in the literature of Shakespeare as per the available record.

Here I have taken an example from Wikipedia and the source I have already mentioned here. On x axis this one is x axis, this is frequency rank, the rank of the word how many times, what is their rank in Wikipedia and this y axis is basically the frequency occurrence, how many times of the particular word is occurring in Wikipedia. And we found that, the study found that 100 most frequent words from Wikipedia were taken and the graph was drawn. This is the shape of the graph. In that it was found that 'the' word, which was at number one, was used approximately 5 million times. The word 'which' was at rank 14 used 360,000 times. While the word 'being' was at the rank 100 used 50,000 times.

When we multiply the rank to the number of times the word is being used that means 1 multiplied by 5 million, that is 50 lakhs, is 50,00,000. When we multiply the rank of the word with the number of times it appeared in Wikipedia like in this case the word 'the' having rank 1 the which appeared 5 million or 50 lakhs time in Wikipedia. If we are multiplying these two we are getting 50,00,000. Similarly if we multiply rank 14 with 360000 we are getting 50,40,000. And the word 'being' which appears at the 100th place means being 50000 times. 100*50,000, we are getting 50,00,000. What we are observing here is almost constant this particular quantity, this particular 50,40,000, this is 50,00,000. They are constant. That means when we are multiplying the rank with frequency in a given corpus we are getting the same result and that's what Zipf's Law is.

Zipf says the most frequent words occur approximately twice as often as the second most frequent word. three times as often as the third most frequent word, and so on. He says that the most frequent word occurs approximately twice as compared to the second most frequent word. Similarly the first most frequent word would occur approximately three times as compared to the third most frequent word and so on means the product of rank and frequency would almost always be constant. This is the general equation of Zipf's Law in which frequency is equal to c/r^s.

$$f(r) = c/r^s$$

We have already seen where $f(r)$ is the frequency of the element of rank r, c is a constant, and s is a parameter that typically falls in the range of 1 and 2. We understood Zipf's Law.

Zipf's Law is basically the Law which talks about the occurrence of words in a given corpus of words. And it says that some words occur quite frequently some other words occur very less number of times.

What are the different applications of Zipf's Law?

Zipf's Law has various applications; the most important application of Zipf's Law is in the field of linguistics. Zipf observed that some words like ‘the’ ‘and’ ‘is’ appear very frequently. While there are some very less common words like ‘ephemeral’ are used very less number of times.

In economics there is application of Zipf's Law. We see that most of the wealth of a country is concentrated on very few people. The distribution of income we can say that means a smaller proportion of the population has a larger distribution of income. It is basically the application of Zipf's Law that we can find that in a particular society the wealth is concentrated to a very few.

Zipf's Law can also be applied in city size distribution. For example, major cities have very large populations while smaller cities and towns have very less population. And there is a distribution of population and that population is unevenly distributed in cities.

Zipf's Law can also be used in information retrieval. In information retrieval the common search terms are used frequently while more specific or niche queries are less common. While designing any information retrieval system we should keep in mind that the common search term should be used more as compared to any nick term or any new term.

In the case of biology we see the application of Zipf's Law. We see that few species are highly abundant in certain areas while the majority of the species are less common. This is basically the application of Zipf's Law.

Zipf's Law is also applicable in the internet and social networking. We see that on social media platforms a few accounts or posts often garner a very large number of likes as compared to those which have very less number of likes. What the Zipf's Law says is that a few words have the more number of occurrences as compared to some words having the less number of occurrences. The same thing happens here also in case of the internet and social network. some of the posting will have the larger number of audience which popularity is more as compared to those posting which popularity is less. Zipf's Law is also used in scientific citation. Some articles which get a very large number of very good citations. some of the articles get less citation so that here also there is an application of Zipf's Law.

Zipf's Law is also used in finances as we often see that the share of some financial institutions and companies jump very high and perform very high as compared to the

share of some other companies and institutions. Zipf's Law can also be used in the music industry as we have seen that some of the songs are very much popular as compared to songs which are less popular. Traffic flow we see that in some areas in metros or in cities the traffic is very high. In some areas the traffic is sparsely low, that is basically the application of the Zipf's Law.

Now when we see the application of Zipf's Law in the context of designing an information retrieval system, then it's very important to see how Zipf's Law can be applied in the designing of a good information storage and retrieval system. The search engine follows Zipf's Law and some of the common words which are less informative, like ‘the’, ‘and’, ‘though’, ‘in’. These are not in their search paradigm, they are in the high frequency and the information retrieval algorithm uses basically the technique TF-IDF that means term frequency, inverse document frequency to give less weight to highly frequent terms. TF-IDF is basically giving the least weight to highly frequently used terms and by using Zipf's Law those words which use maximum number of times but having less relevance can have less weightage while giving the search result.

Zipf's Law in the case of designing a good information retrieval system can be used for the query expansion. because we see that designing a good query is very much important to get good results. With the help of Zipf's Law we can find out which word should be used in our query strategy, in our search string to get the maximum number of relevant results. Zipf's Law in the case of information retrieval systems can be used for stop word removal. There are various small words. There are various sets of words which are dominant in the language, which you use very frequently. But these words are not relevant in the context of searching. By that way using Zipf's Law or information retrieval system can put those words under the stop word list. And by that way we will get more relevant results by putting our queries.

In the context of an information retrieval system Zipf's Law can be used for indexing optimization. because the search result is based on the back end index. The back end index of an information retrieval system can be such a way that those words which are not relevant should be stopped, should not be the part of the index. And by that way we will get the maximum relevant result. Zipf's Law in the case of information retrieval systems can be used for the ranking algorithm also. Zipf's Law in the context of an information retrieval system can be used for the ranking algorithm. because an information retrieval system can put the ranking of the terms. And based on the ranking the relevant result can be obtained by the users.

This can be also used for the user query suggestions. because the user can suggest the query based on the different words which can be used for retrieval of the result. And this Law can be used for optimizing document retrieval. Because with the help of Zipf's Law the system algorithm can prioritize documents containing less common terms and having

higher discriminative power potentially leading to more accurate and contextually relevant results. In case of document retrieval with the help of Zipf's Law the system can prioritize which document should be delivered, which document should be seen by the users while putting the ranking term. And by that way the user can get the relevant result. Similarly Zipf's Law can be used so the dynamic query can be expanded and the relevant result can be derived using Zipf's Law. Feedback can be provided based on the results, what we obtain. what the user obtains. And the system can study the feedback and based on the feedback what is systematic can a more reverse system can be developed the system can be improved and more relevant results can be shown by the system.

Now that we have seen the different applications of Zipf's Law we studied the application of Zipf's Law in detail, particularly in the context of the information retrieval system.

There are some limitations of Zipf's Law. Zipf's Law is basically observational in nature. it is more of a statistical pattern rather than a fundamental Law of nature. In the example we have seen that Zipf's Law is basically a Law which gives that there is a certain pattern in the occurrence of words in a given language corpus. And that varies from one corpus to another corpus. That means Zipf's Law is not the Law which is the Law of nature. This is not the Law which is universally applicable because of the different reasons. Because the language is dynamic. There are various other aspects which are included in the development of language, not only the words. Hence we can say that Zipf's Law is basically a statistical analysis pattern showing the statistical analysis pattern rather than the Law of nature.

Zipf's Law is very much data sensitive. Different sets of data exhibit variations in the extent to which they confront the law. We can take the different sets of data and we apply Zipf's Law. We find that there is variation. We take the different sets of data, we take the different data corpus and we apply Zipf's Law. we find that there is a variation in the result of what we get. That means it is very much sensitive to data. There is a data sensitivity; Zipf's Law is sensitive to data. Zipf's Law is also a scale dependency. they may not hold uniformly across different scales. This applicability can diminish when considering less common words and when examining a smaller subset of the language corpus.

Here Zipf's Law is scale dependent. Zipf's Law if you are taking a subset of data then we find some errors in Zipf's Law. Another limitation of Zipf's Law is lack of causality. It provides a statistical pattern. But does not provide insight into the underlying cause of the observed distribution. You see when Zipf studied the corpus he found the relationship between the rank of the word, occurrence, and the frequency the occurrence is constant but why this happens there was no explanation.

There was no explanation for the statistical pattern which a particular language corpus follows. It does not explain why certain words or elements follow this distribution leaving room for interpretation or hypothesis. You see, universal Law or the Law of nature is such a way that it can be universally applicable in any context but here in the case of Zipf's Law this is not the case and there are certain limitations. And this Law does not explain why some of the words follow this particular distribution in the language corpus. Another limitation of Zipf's Law is the complexity of language. You see, the language is quite complex. The development of language is a dynamic process. The language system is a quite dynamic system. Hence a single mathematical or a single statistical formula which was suggested by Zipf's cannot be applied to the language model because of its complex nature and that is one of the limitations.

There is variability across the languages, different languages having different patterns, they have the challenging idea of universal linguistic law. You see different languages having different patterns, different languages having their different development criteria. Hence there cannot be a single law which can universally be applied to a different language. So, that is one of the limitations here. Is a limited predictive power. Zipf's Law provides a statistical distribution of the word frequency. it may have limited predictive power in certain applications predicting the occurrence of a specific word or event based solely on Zipf's Law might not be highly accurate. because we have already seen that Zipf's Law has limitations. Hence predicting the result solely based on Zipf's Law would not be good. It won't give us the correct result or correct figure.

Zipf's Law over-emphasizes on the rank. Because Zipf's did he take the rank and the frequency of occurrence. So, this Law is basically over emphasizing on the rank and neglecting the other language structure and usage. That is another limitation of Zipf's Law. Thank you learners and I duly acknowledge the sources scholars and website content providers which contents and materials I have used in my presentation. Thank you and for any query you can contact me at my email address at vkverma@library.iitd.ac.in.

Thanks for your patient hearing. Thank you.