An Analysis of Word Formation and Evolution in Computer Science

**姓名：杨仕博**

**学号：2000013058**

**院系：信息科学技术学院**

**邮箱：2000013058@stu.pku.edu.cn**

**手机：18733155835**

**An Analysis of Word Formation and Evolution in Computer Science**

1. **Introduction**

In recent years, computer technology has experienced explosive development, and the number of the articles in related fields are increasing day by day. However, when we read computer science papers, we do not feel that there is too much unfamiliar, specialized vocabulary to hinder our reading. Is it possible that scientists have deliberately avoided the use of unfamiliar words in the development of computer science, or is it possible that many of the terms used in computer science somehow evolve from general words? As a discipline close to real life and production, does computer science only use a large number of common words that belong to a certain range, rather than using a large number of seemingly more professional words? If so, we could use our vocabulary knowledge to quickly memorize a large number of computer words - after all, they are not much different from ordinary words. By studying the knowledge of word roots, we hope to analyze the words used in computer papers from the perspective of word roots, and then verify or disprove our conjecture. In this article, we analyze computer science papers from different fields as well as ordinary articles from everyday life, and assert that computer science does not actually use too many extreme words, but prefers to use some common words that belong to certain fields, and makes them polysemous by giving them new meanings in computer science.

1. **Literature review**

In previous studies of terminology, it has been asserted that "science of terminology has developed from practical issues such as guidelines and recommendations in order to remedy communicational deficits passing phases of intensified theorisation"[1]. Specifically, in computer science, "IT terminology reflects the cutting-edge innovations in the field, providing a framework for discussing and understanding these developments. IT terminology is the backbone of effective communication within the technology sector and across various industries"[2], which means that computer science terminology is the vehicle for a lot of communication. Therefore, according to scholars, these terms must "promote clarity, efficiency, and collaboration while ensuring that the IT field remains adaptable to rapid technological changes."[2] That means our term is "Precise" and "well-defined". Thus, we can infer that overly complex, unintelligible vocabulary does not occur very often in computer science.

In "Computer Jargon Explained"[3], the author even covers the main computer terminology in only 68 words, which meant that there were not many unfamiliar words in computer science at that time -- but unfortunately, this book was written in the last century. As far as we know, there is no good paper on the current state of computer science vocabulary.

1. **Research purpose**

The purpose of this experiment is to verify our previous claims about the characteristics of words used in computer science. That is, in order to adapt to the development of technology, the vocabulary used in computer science is mostly accurate and concise, and it often involves a large number of words in a special field which are not uncommon and are often given a new meaning in Computer Science.

1. **Results and discussion**
   1. **Results**
      1. **Subjects**

To test our conjecture, we need to obtain the lexical distribution of papers in the field of computer science and the lexical distribution of everyday articles, so we find three papers from different fields of computer science and use ChatGpt to generate several daily articles. Then, we cut them all down to about 27,000 characters. Next, we use technology to count the roots involved in these words, and use these statistics to replace the root distribution used in computer papers and everyday articles, respectively.

Having prepared the data, we can proceed to analyze them.

First, we compare these distributions to see if they're similar. If not, we can verify that computer science articles tend to use words that are specific to a particular field, as opposed to words that are commonly used in everyday fields.

Next, we analyze the most frequent roots to see if they are rare roots. If not, we can conclude that most words in computer science are not special at all.

Finally, we pick out the specialized words involved in these roots, study the formation of these words, and see how they are generated, so as to analyze whether the generation of computer words is complex. If their composition is not complicated, they are easy to understand and easy to communicate.

* + 1. **Method**

For data, we select three papers from the algorithmic field, the engineering field and the artificial intelligence field of computer science. We then use automatic recognition technology to extract the English words used in these articles. We hope that the amount of data used will not vary too much, so we limit the number of characters used to around 27,000, and the excess will be deleted.

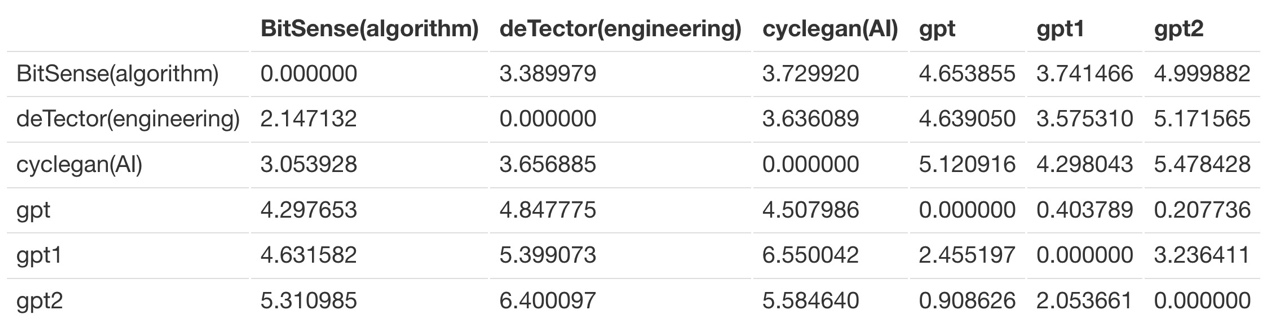
Then we refer to the roots that we need to memorize in class to form a set of roots. After some engineering work we obtained the frequency of these roots in the four texts, which are the distributions of the roots. To simplify the article, details of engineering processing are given in Appendix. Next we check whether these distributions are similar. This is simple, we just need to calculate the KL divergences(i.e., the degree of dissimilarity) between every two distributions. All we need to know is that the lower the KL divergences, the closer the two distributions are. In order to define what distributions are close, we randomly divide ChatGpt-generated essays into two classes, and define the calculated KL divergences for these two classes of essays to represent the difference value of "close" essays.

Next, we focus on the few roots that occur the most in each distribution, and empirically judge whether these roots are rare or not. If most of these roots are common, then the words corresponding to the text are not uncommon.

Finally, we select the specialized words related to the most frequently occurring roots, analyze their formation, check if they are polysemous words, reason how their meanings in computer science are generated, and use them as a representative to speculate whether most computer science words are complex constructions, and then demonstrate our previous points.

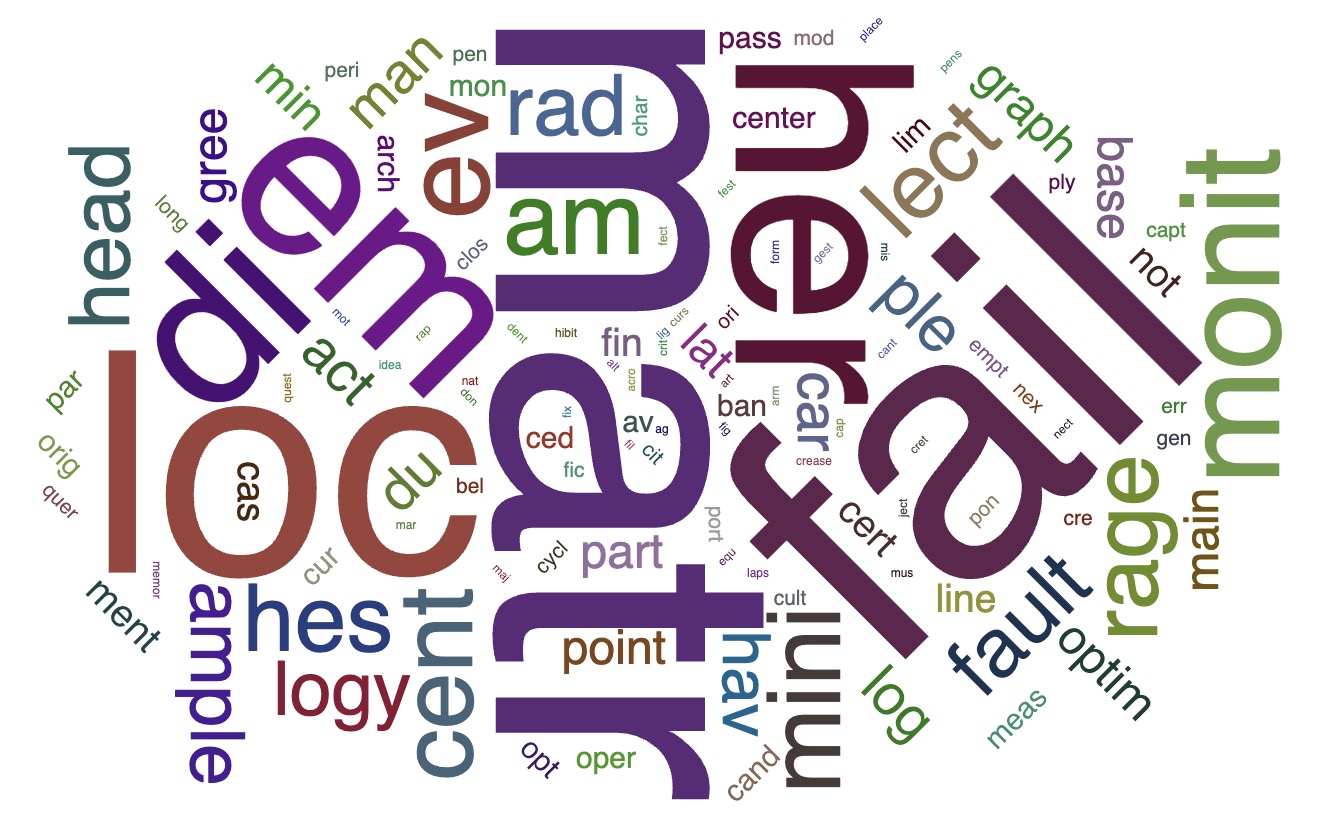
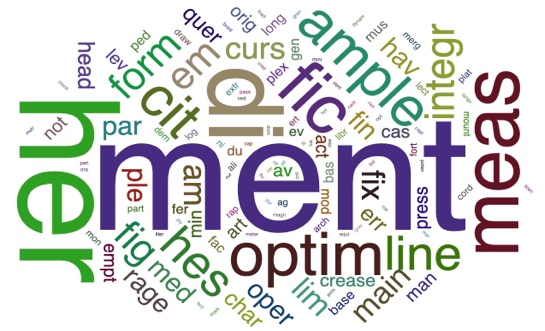
* + 1. **Result**

First we will show the KL divergences(i.e., the degree of dissimilarity) of the root distributions corresponding to several articles. The string in the table represents the name and/or kind of the paper or article. It should be mentioned that the KL divergences of the two parts of the article generated by ChatGpt after being randomly separated are 2.053661 and 3.236411, and each value in the table represents the KL divergence of the distribution above with respect to the distribution on the left.

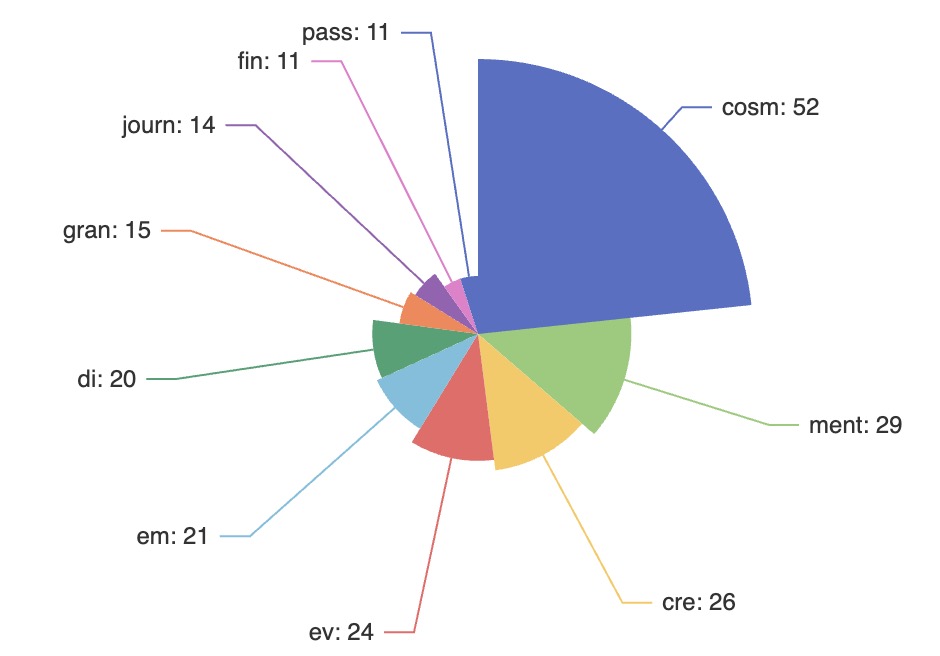
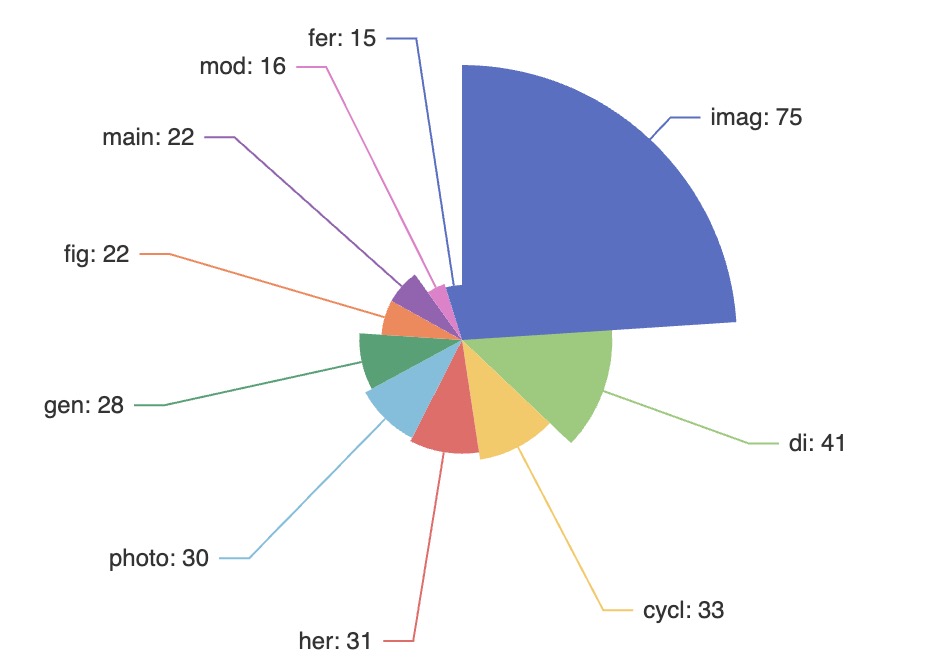
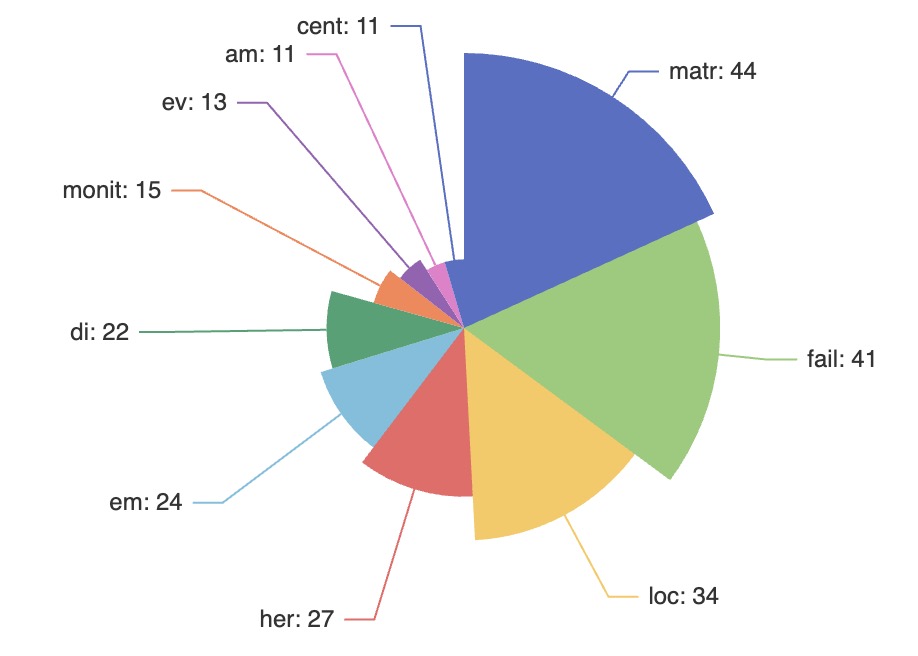
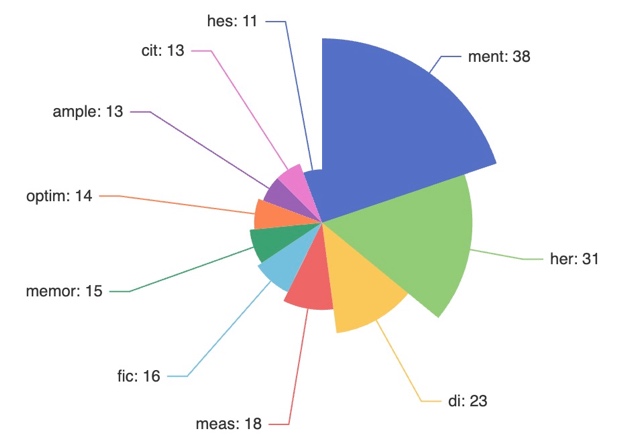


*Table1: The degree of dissimilarity(KL divergences) of root distributions involved in different texts*

Next we will show a few of the most common roots for each paper/article.

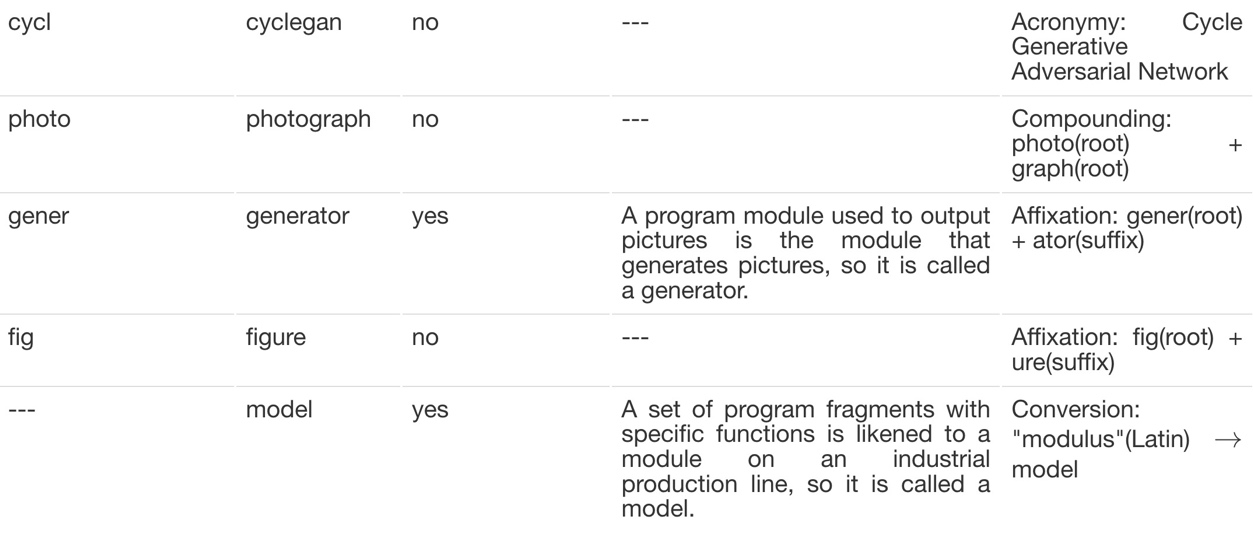
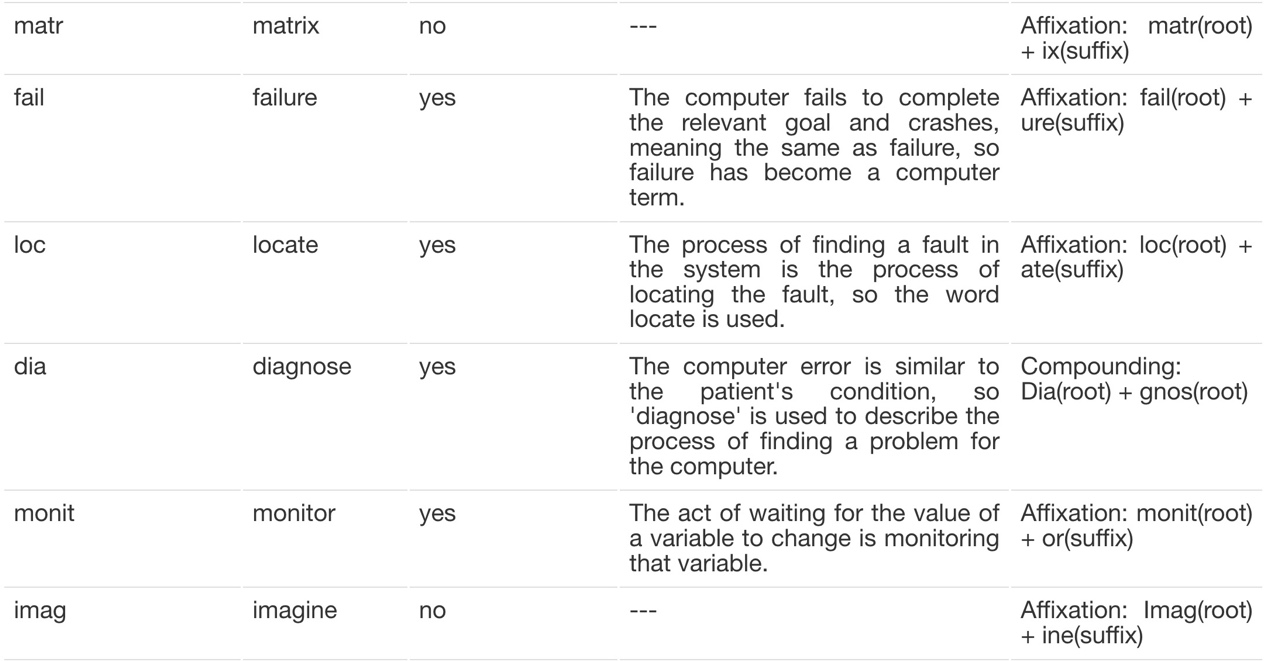
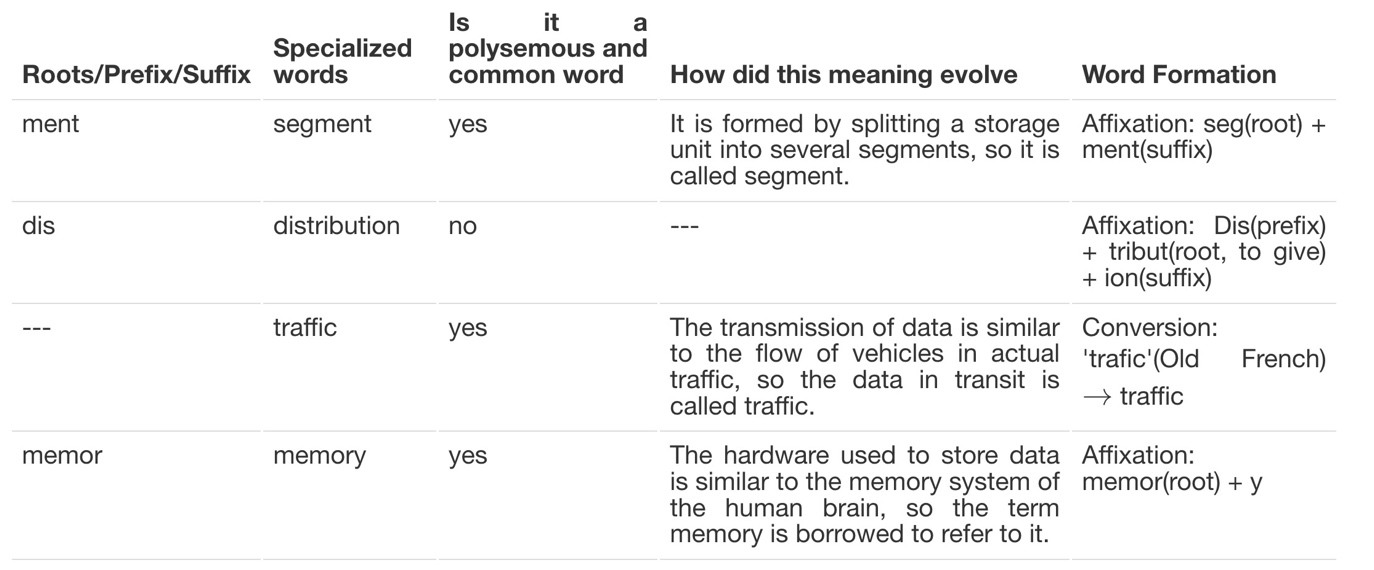


*Figure1: The word cloud map generated by the statistical root. The larger the root is, the more frequently it appears in the text. The four pictures from left to right from top to bottom belong to BitSense(algorithm), deTector(engineering), cyclegan(AI), and ChatGpt respectively.*



*Figure2: The pie chart generated by the statistical root. The graph indicates the number of times the most common root appears in the passage. The four pictures from left to right from top to bottom belong to BitSense(algorithm), deTector(engineering), cyclegan(AI), and ChatGpt respectively.*

Finally, we show the specialized words that correspond to these roots and how they are formed. It is worth noting that we find most of these terms are polysemous words whose meanings in technical terms are synchronously derived from their original meanings. We will explain how these polysemous words come to mean something in computers.



* 1. **Discussion**
     1. **Consider the characteristics of computer vocabulary as a whole**

As for the similarity between various articles, we can see that the KL divergences (i.e., the degree of dissimilarity) between daily articles range from about 2 to 3. However, the KL divergences between computer papers and daily articles are greater than 4, and most exceed 4.5. At the same time, the KL divergence between articles in different fields of computing is also partially over 3.5. So instead of simply using everyday words, computer papers use those words in a different sense and focus on different words for different areas. This also supports Nicholas Enticknap's point: "The technology and the industry are changing so rapidly it is very hard even for professionals to keep up to date."[3]

* + 1. **Consider the characteristics of computer vocabulary in terms of specific roots and vocabulary**

Considering the most frequently used roots we have extracted, in my experience, the only roots that appear very frequently but are less commonly used are "matr" and "optim". These two come up so often because the words "matrix" and "optimize" are used so often, but both of them come from mathematics. Which means, with the exception of mathematical words, the rest of the words in computer papers are likely to be relatively common words.

By examining the computer terms involved in the articles, we find that most of them are polysemous words that generated by developing a new meaning from the daily vocabulary. Computer terms often use words with similar meanings to refer to corresponding things. For example, the hardware that records information is simply replaced by memory, and the word traffic is also found for the data transmitted in the network. Even things that are not common in everyday life, such as storage units consisting of several bits, are vividly represented as segments. This shows that computer terms are indeed deliberately and simply manufactured for ease of understanding and dissemination. The way these terms are generated makes them, as what Liliya Karmazina says, "promote clarity, efficiency, and collaboration while ensuring that the IT field remains adaptable to rapid technological changes". [2]

1. **Conclusion**

By analyzing several computer papers in specific fields, we get some general understanding of computer terminology. Firstly, a large number of computer terms are derived from everyday vocabulary, except for some specialized words that have to be borrowed from mathematics. Secondly, computer terms often use the specific meanings derived from these everyday words in the compute-context, rather than the meanings in everyday life. Finally, we believe that this way of generating terms is deliberate and, based on previous literature, we believe that it was designed to accommodate the explosion of computer technology.

These results mean that we can quickly transfer our lexicology knowledge to the specialized vocabulary of the computer world - because most of these words are directly derived from everyday words.

1. **Reference**

[1] Picht H. The science of terminology: History and evolution[J]. Terminologija, 2011 (18): 6-26.

[2] Karmazina L. HISTORICAL EVOLUTION OF IT TERMINOLOGY AND ITS FURTHER DEVELOPMENT[J]. Scientific Journal of Polonia University, 2023, 59(4): 30-35.

[3] Enticknap N. Computer jargon explained[M]. Elsevier, 2014.

**Appendix**

**Appendix(A): How do we extract root distributions by engineering?**

As we know, root extraction is a complex natural language processing problem. In engineering, it is impossible for us to complete accurate root extraction in a short time. Therefore, the root extraction in this paper takes some approximate methods. To be specific, we remove some of the roots provided in the class because using these roots causes a lot of error (for example, a lot of words 'bit' correspond to root 'bi', but the two are unrelated). In the experiment, we only check whether the prefix and suffix of each word are in the root set, and if one of them is, we add the word to the list of words corresponding to the root. The reason for checking only prefixes and suffixes is that if a root is the same as a substring (which is not a prefix or a suffix) of that word, it's probably just a coincidence -- for example, the root "hal" appears in "challenges", the root "art" appears in "heartbeat", and "ni" appears in "awakening" -- obviously these words have nothing to do with these roots. To consider all the possible substrings for each word in the code would introduce a huge amount of error. However, this error can be greatly reduced if only prefixes and suffixes are detected. In this way, by counting the number of occurrences of each root, we get a general distribution of roots for all articles.