

Word-repetition patterns and keyword detection

Alvaro Corral
Centre de Recerca Matemàtica, UAB
<http://einstein.uab.es/acorralc>

Abstract:

Guided by the analogy with earthquake occurrence, we study the distribution of distances between consecutive repetitions of words in a text. We will see how this distribution shows two main characteristics:

1. Scaling. The shape of the distribution is the same for different words, but in a different scale depending on the frequency of the word.
2. Clustering: The shape of the distribution indicates an attraction between occurrences of a word, that is, short distances show an enlarged probability.
3. These properties hold for different authors and languages; however, there seems to exist a dependence with the degree of relevance of the word in the text: relevant words present higher clustering than the rest of words. We will explore if these ideas can be useful for automatic retrieval of keywords.