CLAIMS

What is claimed is:

1. A method comprising:

employing a computer system to receive a target string forming a part of an electronic communication;

employing a computer system to process the target string into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

in response to receiving the target string, employing the computer system to determine a string eligibility criterion according to the target string;

employing the computer system to pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, employing the computer system to perform a comparison between a frequency spectrum of the target signal and a frequency spectrum determined for a candidate string of the plurality of candidate strings; and

employing the computer system to determine whether the electronic communication is spam or non-spam according to a result of the comparison.

2. The method of claim 1, further comprising determining whether the electronic communication is spam or non-spam according to a comparison between a timestamp of the target string and a timestamp of the candidate string.

3. The method of claim 1, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the method further comprises, in response to performing the comparison, employing

the computer system to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.

4. The method of claim 3, further comprising determining whether the target communication is spam or non-spam according to a plurality of timestamps, each timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

5. The method of claim 3, further comprising:

in response to assigning the target string to the selected cluster, determining a count of cluster members of the selected cluster; and

determining whether the target communication is spam or non-spam according to the count of cluster members.

6. The method of claim 3, further comprising identifying the target communication as belonging to a selected spam wave according to the selected cluster.

7. The method of claim 1, wherein pre-filtering the corpus comprises:

determining a first string length of the target string and a second string length of a reference string of the corpus, and

when the first string length differs from the second string length by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

8. The method of claim 1, wherein pre-filtering the corpus comprises:

determining a first spectrum score of the target string as a function of:

$$S = \sum_i A_i$$

wherein A_i denotes an amplitude of the frequency spectrum of the target signal, determined at frequency i;

determining a second spectrum score of a reference string of the corpus; and

when the first spectrum score differs from the second spectrum score by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

9. The method of claim 1, wherein performing the comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in S} \left| A^T{}_i - A^C{}_i \right|^{\alpha},$$

wherein i indexes frequencies, S denotes a subset of frequencies of the frequency spectrum of the target signal, $A^T{}_i$ denotes an amplitude of the frequency spectrum of the target signal at frequency i, $A^C{}_i$ denotes an amplitude of the frequency spectrum determined for the candidate string at frequency i, and wherein α is a positive constant.

10. The method of claim 9, wherein the subset S consists exclusively of an index i corresponding to the zero frequency of the frequency spectrum of the target signal.

11. The method of claim 9, wherein the subset S consists exclusively of an index i corresponding to a maximum amplitude of the frequency spectrum of the target signal.

12. The method of claim 1, further comprising, in response to processing the target string into a target signal, employing the computer system to apply a digital low-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

13. The method of claim 1, further comprising, in response to processing the target string into a target signal, employing the computer system to apply a digital high-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

14. A computer system comprising at least a processor programmed to:

receive a target string forming a part of an electronic communication;

process the target string into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

in response to receiving the target string, determine a string eligibility criterion according to the target string;

pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, perform a comparison between a frequency spectrum of the target signal and a frequency spectrum determined for a candidate string of the plurality of candidate strings; and

determine whether the electronic communication is spam or non-spam according to a result of the comparison.

15. The system of claim 14, wherein the processor is further programmed to determine whether the electronic communication is spam or non-spam according to a comparison between a timestamp of the target string and a timestamp of the candidate string.

16. The system of claim 14, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the processor is further programmed, in response to performing the comparison, to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.

17. The system of claim 16, further comprising determining whether the target communication is spam or non-spam according to a plurality of timestamps, each timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

18. The system of claim 16, wherein the processor is further programmed to:

in response to assigning the target string to the selected cluster, determine a count of cluster members of the selected cluster; and

determine whether the target communication is spam or non-spam according to the count of cluster members.

19. The system of claim 16, wherein the processor is further programmed to identify the target communication as belonging to a selected spam wave according to the selected cluster.

20. The system of claim 14, wherein pre-filtering the corpus comprises:

determining a first string length of the target string and a second string length of a reference string of the corpus, and

when the first string length differs from the second string length by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

21. The system of claim 14, wherein pre-filtering the corpus comprises:

determining a first spectrum score of the target string as a function of:

$$S = \sum_i A_i$$

wherein A_i denotes an amplitude of the frequency spectrum of the target signal, determined at frequency i;

determining a second spectrum score for a reference string of the corpus; and

when the first spectrum score differs from the second spectrum score by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

22. The system of claim 14, wherein performing the comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in S} \left| A^T{}_i - A^C{}_i \right|^{\alpha},$$

wherein i indexes frequencies, S denotes a subset of frequencies of the frequency spectrum of the target signal, $A^T{}_i$ denotes an amplitude of the frequency spectrum of the target signal at frequency i, $A^C{}_i$ denotes an amplitude of the frequency spectrum determined for the candidate string at frequency i, and wherein α is a positive constant.

23. The system of claim 22, wherein the set S consists exclusively of an index i corresponding to the zero frequency of the frequency spectrum of the target signal.

24. The system of claim 22, wherein the set S consists exclusively of an index i corresponding to a maximum amplitude of the frequency spectrum of the target signal.

25. The system of claim 14, wherein the processor is further programmed, in response to processing the target string into a target signal, to apply a digital low-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

26. The system of claim 14, wherein the processor is further programmed, in response to processing the target string into a target signal, to apply a digital high-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

27. A method comprising:

employing a computer system to receive an electronic communication;

in response to receiving the electronic communication, employing the computer system to extract a target string from the electronic communication;

employing the computer system to transmit the target string to an anti-spam server; and

in response to transmitting the target string, receiving a target label indicative of whether the electronic communication is spam or non-spam, wherein the target label is determined at the anti-spam server and wherein determining the target label comprises:

employing the anti-spam server to process the target string of characters into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

employing the anti-spam server to determine an eligibility criterion according to the target string;

employing the anti-spam server to pre-filter a corpus of reference strings according to the eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, employing the anti-spam server to employing the computer system to perform a comparison between a frequency spectrum of the target signal and a frequency spectrum determined for a candidate string of the plurality of candidate strings; and

employing the anti-spam server to determine whether the electronic communication is spam or non-spam according to a result of the comparison.

28. A method comprising:

employing a computer system to receive a target string forming a part of an electronic communication;

employing a computer system to process the target string of characters into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

in response to receiving the target string, employing the computer system to determine a string eligibility criterion according to the target string;

employing the computer system to pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, employing the computer system to determine an inter-string distance separating the target string from a candidate string of the plurality of candidate strings, the inter-string distance determined according to a first amplitude of a frequency spectrum of the target signal and according to a second amplitude of a frequency spectrum determined for the candidate string; and

employing the computer system to determine whether the target communication is spam or non-spam according to the inter-string distance.

29. A method comprising:

employing a computer system to receive a target string forming a part of an electronic communication;

employing a computer system to process the target string into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

employing the computer system to determine a frequency spectrum of the target signal;

employing the computer system to perform a comparison between the frequency spectrum of the target signal and a frequency spectrum determined for a reference string selected from a set of reference strings; and

employing the computer system to determine whether the target communication is spam or non-spam according to a result of the comparison.

CLAIMS

1. A method comprising:

employing at least one processor of a computer system to receive a target string forming a part of an electronic communication;

employing the at least one processor to process the target string into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

in response to receiving the target string, employing the at least one processor to determine a string eligibility criterion according to the target string;

employing the at least one processor to pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, employing the at least one processor to perform a comparison between a frequency spectrum of the target signal and a frequency spectrum determined for a candidate string of the plurality of candidate strings, wherein determining the frequency spectrum of the target signal comprises applying a Fourier transform to the target signal to represent the target signal as a plurality of frequency components of the target signal, each frequency component having a distinct frequency and an amplitude determined for the distinct frequency; and

employing the at least one processor to determine whether the electronic communication is spam or non-spam according to a result of the comparison.

2. The method of claim 1, further comprising determining whether the electronic communication is spam or non-spam according to a comparison between a timestamp of the target string and a timestamp of the candidate string.

3. The method of claim 1, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the method further comprises, in response to performing the comparison, employing the at least one processor to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.

4. The method of claim 3, further comprising determining whether the target communication is spam or non-spam according to a plurality of timestamps, each timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

5. The method of claim 3, further comprising:

in response to assigning the target string to the selected cluster, determining a count of cluster members of the selected cluster; and

determining whether the target communication is spam or non-spam according to the count of cluster members.

6. The method of claim 3, further comprising identifying the target communication as belonging to a selected spam wave according to the selected cluster.

7. The method of claim 1, wherein pre-filtering the corpus comprises:

determining a first string length of the target string and a second string length of a reference string of the corpus, and

when the first string length differs from the second string length by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

8. The method of claim 1, wherein pre-filtering the corpus comprises:

determining a first spectrum score of the target string as a function of:

$$S = \sum_i A_i$$

wherein A_i denotes an amplitude of a frequency component of the target signal, the frequency component having frequency i;

determining a second spectrum score of a reference string of the corpus; and

when the first spectrum score differs from the second spectrum score by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

9. The method of claim 1, wherein performing the comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in S} |A_i^T - A_i^C|^\alpha,$$

wherein S denotes a subset of frequencies, A^T_i denotes an amplitude of a frequency component of the target signal with frequency i, A^C_i denotes an amplitude of a frequency component determined for the candidate string at frequency i, and wherein α is a positive constant.

10. The method of claim 9, wherein the subset S consists exclusively of a frequency equal to zero.

11. The method of claim 9, wherein the subset S consists exclusively of a frequency of a frequency component having the largest amplitude among the plurality of frequency components of the target signal.

12. The method of claim 1, further comprising, in response to processing the target string into a target signal, employing the at least one processor to apply a digital low-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

13. The method of claim 1, further comprising, in response to processing the target string into a target signal, employing the at least one processor to apply a digital high-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

14. A computer system comprising at least one processor programmed to:

receive a target string forming a part of an electronic communication;

process the target string into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

in response to receiving the target string, determine a string eligibility criterion according to the target string;

pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, perform a comparison between a frequency spectrum of the target signal and a frequency spectrum determined for a candidate string of the plurality of candidate strings, wherein determining the frequency spectrum to the target signal comprises applying a Fourier transform to the target signal to represent the target signal as a plurality of frequency components of the target signal, each frequency component having a distinct frequency and an amplitude determined for the distinct frequency; and

determine whether the electronic communication is spam or non-spam according to a result of the comparison.

15. The system of claim 14, wherein the at least one processor is further programmed to determine whether the electronic communication is spam or non-spam according to a comparison between a timestamp of the target string and a timestamp of the candidate string.

16. The system of claim 14, wherein the corpus of reference strings comprises a plurality of clusters, each cluster including a set of similar strings, wherein each candidate string of the plurality of candidate strings is representative of a distinct cluster, and wherein the at least one processor is further programmed, in response to performing the comparison, to select a cluster from the plurality of clusters and to assign the target string to the selected cluster.

17. The system of claim 16, further comprising determining whether the target communication is spam or non-spam according to a plurality of timestamps, each timestamp of the plurality of timestamps corresponding to a member of the selected cluster.

18. The system of claim 16, wherein the at least one processor is further programmed to:

in response to assigning the target string to the selected cluster, determine a count of cluster members of the selected cluster; and

determine whether the target communication is spam or non-spam according to the count of cluster members.

19. The system of claim 16, wherein the at least one processor is further programmed to identify the target communication as belonging to a selected spam wave according to the selected cluster.

20. The system of claim 14, wherein pre-filtering the corpus comprises:

determining a first string length of the target string and a second string length of a reference string of the corpus, and

when the first string length differs from the second string length by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

21. The system of claim 14, wherein pre-filtering the corpus comprises:

determining a first spectrum score of the target string as a function of:

$$S = \sum_i A_i,$$

wherein A_i denotes an amplitude of a frequency component of the target signal, the frequency component having frequency i;

determining a second spectrum score for a reference string of the corpus; and

when the first spectrum score differs from the second spectrum score by an amount smaller than a predetermined threshold, adding the reference string to the set of candidate strings.

22. The system of claim 14, wherein performing the comparison comprises determining an inter-string distance as a function of:

$$\sum_{i \in S} |A_i^T - A_i^C|^\alpha,$$

wherein S denotes a subset of frequencies, A^T_i denotes an amplitude of a frequency component of the target signal having frequency i, A^C_i denotes an amplitude of a frequency

component determined for the candidate string at frequency i, and wherein α is a positive constant.

23. The system of claim 22, wherein the subset S consists exclusively of a frequency equal to zero.

24. The system of claim 22, wherein the subset S consists exclusively of a frequency of a frequency component having the largest amplitude among the plurality of frequency components of the target signal.

25. The system of claim 14, wherein the at least one processor is further programmed, in response to processing the target string into a target signal, to apply a digital low-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

26. The system of claim 14, wherein the at least one processor is further programmed, in response to processing the target string into a target signal, to apply a digital high-pass filter to the target signal, and wherein the frequency spectrum of the target signal comprises a frequency spectrum of the filtered target signal.

27. A method comprising:

employing at least one processor of a computer system to receive an electronic communication;

in response to receiving the electronic communication, employing the at least one processor to extract a target string from the electronic communication;

employing the at least one processor to transmit the target string to an anti-spam server; and

in response to transmitting the target string, receiving a target label indicative of whether the electronic communication is spam or non-spam, wherein the target label is determined at the anti-spam server and wherein determining the target label comprises:

employing the anti-spam server to process the target string of characters into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

employing the anti-spam server to determine an eligibility criterion according to the target string;

employing the anti-spam server to pre-filter a corpus of reference strings according to the eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, employing the anti-spam server to employing the computer system to perform a comparison between a frequency spectrum of the target signal and a frequency spectrum determined for a candidate string of the plurality

of candidate strings, wherein determining the frequency spectrum of the target signal comprises applying a Fourier transform to the target signal to represent the target signal as a plurality of frequency components of the target signal, each frequency component having a distinct frequency and an amplitude determined for the distinct frequency; and

employing the anti-spam server to determine whether the electronic communication is spam or non-spam according to a result of the comparison.

28. A method comprising:

employing at least one processor of a computer system to receive a target string forming a part of an electronic communication;

employing the at least one processor to process the target string into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

in response to receiving the target string, employing the at least one processor to determine a string eligibility criterion according to the target string;

employing the at least one processor to pre-filter a corpus of reference strings according to the string eligibility criterion, to produce a plurality of candidate strings;

in response to selecting the candidate strings, employing the at least one processor to determine an inter-string distance separating the target string from a candidate string of the plurality of candidate strings, the inter-string distance determined according to a first amplitude of a frequency spectrum of the target signal and according to a second amplitude of a frequency spectrum determined for the candidate string, wherein determining the frequency spectrum of the target signal comprises applying a Fourier transform to the target signal to represent the target signal as a plurality of frequency components of the target signal, each frequency component having a distinct frequency and an amplitude determined for the distinct frequency; and

employing the at least one processor to determine whether the target communication is spam or non-spam according to the inter-string distance.

29. A method comprising:

employing at least one processor of a computer system to receive a target string forming a part of an electronic communication;

employing the at least one processor to process the target string into a target signal consisting of a sequence of numbers, wherein each character of the target string is mapped to a number of the sequence of numbers;

employing the at least one processor to perform a comparison between a frequency spectrum of the target signal and a frequency spectrum determined for a reference string

selected from a set of reference strings, wherein determining the frequency spectrum of the target signal comprises applying a Fourier transform to the target signal to represent the target signal as a plurality of frequency components of the target signal, each frequency component having a distinct frequency and an amplitude determined for the distinct frequency; and

employing the at least one processor to determine whether the target communication is spam or non-spam according to a result of the comparison.