CLAIMS

1. A method for providing rare topic detection using hierarchical topic modeling by a processor, comprising:

executing machine learning logic to learn a hierarchical topic model from one or more data sources;

in conjunction with learning the hierarchical topic model, executing the machine learning logic to train the hierarchical topic model by iteratively removing one or more dominant words in a selected cluster using the hierarchical topic model during a progressive drilldown operation through a plurality of hierarchical topic modelling executions, wherein, at each iteration of the plurality of hierarchical topic modelling executions, the progressive drilldown operation removes those of the one or more dominant words identified during a previous iteration which are no longer discriminatory for a next execution of the plurality of hierarchical topic modelling executions, and wherein the dominant words comprise words which are used in a higher percentage of conversations, relative to a lower percentage of the conversations in which other non-dominant words are used, as identified in the one or more data sources, and the dominant words relate to one or more primary topics of the cluster, and

executing the machine learning logic to further train the learned hierarchical topic model by seeding the learned hierarchical topic model with one or more words, n-grams, phrases, text snippets to evolve the hierarchical topic model, wherein the removed dominant words are reinstated upon completion of the seeding, and wherein each of the dominant words removed from each iteration and reinstated upon completion of the seeding are together used to form a natural language explanation, provided to a user by way of a user interface, of each of the one or more primary topics resulting from the hierarchical topic model within a corpus of the one or more data sources.

2. The method of claim 1, further including generating one or more word vectors and scoring each of the one or more word vectors.

3. The method of claim 2, further including generating a plurality of clusters from the one or more word vectors, wherein the selected cluster is identified from the plurality of clusters and is a king cluster, wherein the king cluster is a largest cluster from the plurality of clusters.

4. The method of claim 1, further including:

splitting the selected cluster into a plurality of clusters at each iteration;

identifying an alternative selected cluster from the plurality of clusters while iteratively removing one or more dominant words in the alternative selected cluster, wherein the alternative selected cluster is a king cluster and the king cluster is a largest cluster from the plurality of clusters.

5. The method of claim 1, further including seeding the hierarchical topic model with an existing topic model.

6. The method of claim 1, further including seeding each of a plurality of clusters according to one or more cluster models.

7. The method of claim 1, further including identifying one or more differences between each of the plurality of clusters while iteratively removing the one or more dominant words in the selected cluster at each iteration.

8. A system for providing rare topic detection using hierarchical topic modeling in a computing environment, comprising:

one or more computers with executable instructions that when executed cause the system to:

execute machine learning logic to learn a hierarchical topic model from one or more data sources;

in conjunction with learning the hierarchical topic model, executing the machine learning logic to train the hierarchical topic model by iteratively remove one or more dominant words in a selected cluster using the hierarchical topic model during a progressive drilldown operation through a plurality of hierarchical topic modelling executions, wherein, at each iteration of the plurality of hierarchical topic modelling executions, the progressive drilldown operation removes those of the one or more dominant words identified during a previous iteration which are no longer discriminatory for a next execution of the plurality of hierarchical topic modelling executions, and wherein the dominant words comprise words which are used in a higher percentage of conversations, relative to a lower percentage of the conversations in which other non-dominant words are used, as identified in the one or more data sources, and the dominant words relate to one or more primary topics of the cluster; and

execute the machine learning logic to further train the learned hierarchical topic model by seed the learned hierarchical topic model with one or more words, n-grams, phrases, text snippets to evolve the hierarchical topic model, wherein the removed dominant words are reinstated upon completion of the seeding, and wherein each of the dominant words removed from each iteration and reinstated upon completion of the seeding are together used to form a natural language explanation, provided to a user by way of a user interface, of each of the one or more primary topics resulting from the hierarchical topic model within a corpus of the one or more data sources.

9. The system of claim 8, wherein the executable instructions when executed cause the system to generate one or more word vectors and scoring each of the one or more word vectors.

10. The system of claim 9, wherein the executable instructions when executed cause the system to generate a plurality of clusters from the one or more word vectors, wherein the selected cluster is identified from the plurality of clusters and is a king cluster, wherein the king cluster is a largest cluster from the plurality of clusters.

11. The system of claim 8, wherein the executable instructions when executed cause the system to:
split the selected cluster into a plurality of clusters at each iteration; and

identify an alternative selected cluster from the plurality of clusters while iteratively removing one or more dominant words in the alternative selected cluster, wherein the alternative selected cluster is a king cluster and the king cluster is a largest cluster from the plurality of clusters.

12. The system of claim 8, wherein the executable instructions when executed cause the system to seed the hierarchical topic model with an existing topic model.

13. The system of claim 8, wherein the executable instructions when executed cause the system to seed each of a plurality of clusters according to one or more cluster models.

14. The system of claim 8, wherein the executable instructions when executed cause the system to identify one or more differences between each of the plurality of clusters while iteratively removing one or more dominant words in the selected cluster at each iteration.

15. A computer program product for, by a processor, providing rare topic detection using hierarchical topic modeling, the computer program product comprising a non-transitory computer-readable storage medium having computer-readable program code portions stored therein, the computer-readable program code portions comprising:

an executable portion that executes machine learning logic to learn a hierarchical topic model from one or more data sources;

an executable portion that, in conjunction with learning the hierarchical topic model, executes the machine learning logic to train the hierarchical topic model by iteratively removing one or more dominant words in a selected cluster using the hierarchical topic model during a progressive drilldown operation through a plurality of hierarchical topic modelling executions, wherein, at each iteration of the plurality of the plurality of hierarchical topic modelling executions, the progressive drilldown operation removes those of the one or more dominant words identified during a previous iteration which are no longer discriminatory for a next execution of the plurality of hierarchical topic modelling executions and wherein the dominant words comprise words which are used in a higher percentage of conversations, relative to a lower percentage of the conversations in which other non-dominant words are used, as identified in the one or more data sources, and the dominant words relate to one or more primary topics of the cluster; and

an executable portion that executes the machine learning logic to further train the learned hierarchical topic model by seeding the learned hierarchical topic model with one or more words, n-grams, phrases, text snippets to evolve the hierarchical topic model, wherein the removed dominant words are reinstated upon completion of the seeding, and wherein each of the dominant words removed from each iteration and reinstated upon completion of the seeding are together used to form a natural language explanation, provided to a user by way of a user interface, of each of the one or more primary topics resulting from the hierarchical topic model within a corpus of the one or more data sources.

16. The computer program product of claim 15, further including an executable portion that generates one or more word vectors and scoring each of the one or more word vectors.

17. The computer program product of claim 16, further including an executable portion that generate a plurality of clusters from the one or more word vectors, wherein the selected cluster is identified from the plurality of clusters and is a king cluster, wherein the king cluster is a largest cluster from the plurality of clusters.

18. The computer program product of claim 15, further including an executable portion that:
splits the selected cluster into a plurality of clusters at each iteration; and
identifies an alternative selected cluster from the plurality of clusters while iteratively removing one or more dominant words in the alternative selected cluster, wherein the alternative selected cluster is a king cluster and the king cluster is a largest cluster from the plurality of clusters.

19. The computer program product of claim 15, further including an executable portion that:
seeds the hierarchical topic model with an existing topic model; or
seeds each of a plurality of clusters according to one or more cluster models.

20. The computer program product of claim 15, further including an executable portion that identifies one or more differences between each of the plurality of clusters while iteratively removing the one or more dominant words in the selected cluster at each iteration.