

CLAIMS

1. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor applying the stacked regression method, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories.

2. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor applying the bias formula method, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories.

3. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor applying the noise elimination method, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories.

4. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor ensemble method consisting of a plurality of algorithmic methods the results of which are averaged, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories.

5. The computer-implemented method of claim 4, wherein the ensemble includes a stacked regression method.

6. The computer-implemented method of claim 4, wherein the ensemble includes a bias formula method.

7. The computer-implemented method of claim 4, wherein the ensemble includes a noise elimination method.

8. A system for calculating category proportions for a population set, comprising:

a training module comprising a training processor coupled to memory, the memory including software instructions causing the training processor to:

receive training data, the training data comprising a number of training text elements;

present to at least some of the training text elements to a user, receive from the user an indication of categories to which the training text elements belong, and tag the training text elements with the indicated category; and

store the category tagged training text elements; and

a population set category estimating module comprising an estimating processor coupled to memory, the memory including software instructions causing the estimating processor to:

receive population data, the population data comprising a number of population text elements;

calculate a content profile for each category indicated in the training set;

calculate a content profile for the population set; and

calculate, based upon the content profiles associated with the categories indicated for the elements in the training set and the content profile of the population set, the proportions of the population set that belong in each category;

wherein calculating the proportions includes applying at least one of a stacked regression method, a bias formula method, and a noise elimination method.

9. The system of claim 8, wherein calculating the proportions includes applying a plurality of calculating methods in an ensemble and averaging the results of the different methods.
10. The system of claim 8, wherein calculating the proportions includes the stacked regression method.
11. The system of claim 8, wherein calculating the proportions includes the bias formula method.
12. The system of claim 8, wherein calculating the proportions includes the noise elimination method.

Claims

1. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor applying a stacked regression method, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories.

2. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor applying a bias formula method, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories.

3. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor applying a noise elimination method, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories.

4. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor ensemble method consisting of a plurality of algorithmic methods the results of which are averaged, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories;

wherein the ensemble includes a stacked regression method.

5. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor ensemble method consisting of a plurality of algorithmic methods the results of which are averaged, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories;

wherein the ensemble includes a bias formula method.

6. A computer-implemented method performed by a computer processor, comprising:

(a) receiving by the computer processor a training set of elements, each element in the training set being assigned to one of a plurality of categories and having one of a plurality of content profiles associated therewith;

(b) receiving by the computer processor a population set of elements, each element in the population set having one of the plurality of content profiles associated therewith; and

(c) calculating by the computer processor ensemble method consisting of a plurality of algorithmic methods the results of which are averaged, based on the content profiles associated with and the categories assigned to elements in the training set and the content profiles associated with the elements of the population set, a distribution of elements of the population set over the categories;

wherein the ensemble includes a noise elimination method.

7. A system for calculating category proportions for a population set, comprising:

a training module comprising a training processor coupled to memory, the memory including software instructions causing the training processor to:

receive training data, the training data comprising a number of training text elements;

present the at least some of the training text elements to a user, receive from the user an indications of categories to which the training text elements belong, and tag the training text elements with the indicated category; and

store the category tagged training text elements; and

a population set category estimating module comprising an estimating processor coupled to memory, the memory including software instructions causing the estimating processor to:

receive population data, the population data comprising a number of population text elements;

calculate a content profile for each category indicated in the training set;

calculate a content profile for the population set; and

calculate, based upon the content profiles associated with the categories indicated for the elements in the training set and the content profile of the population set, the proportions of the population set that belong in each category;

wherein calculating the proportions includes applying at least one of a stacked regression method, a bias formula method, and a noise elimination method.

8. The system of claim 7, wherein calculating the proportions includes applying a plurality of calculating methods in an ensemble and averaging the results of the different methods.

9. The system of claim 7, wherein calculating the proportions includes the stacked regression method.

10. The system of claim 7, wherein calculating the proportions includes the bias formula method.

11. The system of claim 7, wherein calculating the proportions includes the noise elimination method.