

CLAIMS

1. A computer-implemented method, comprising:
- determining whether a site-specific script for extracting financial data from a particular financial institution website is available;
 - in response to determining that a site-specific script for extracting financial data from the particular financial institution website is not available, generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based on at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website;
 - generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website; and
 - extracting, for one or more users, financial data from the particular financial institution website using the generated site-specific script.
2. The method of claim 1, wherein generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based on at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website comprises:
- crawling the financial institution website to identify one or more web pages in the financial institution website and one or more web page segments in the one or more web pages in the financial institution website;
 - determining, based on the statistical analysis, respective categorizations for the one or more identified web pages; and
 - determining, based on the statistical analysis, respective categorizations for the one or more web page segments in the identified one or more web pages.

3. The method of claim 2, wherein determining, based on the statistical analysis, respective categorizations for the one or more identified web pages comprises:

determining a respective plurality of scores for each web page in the one or more identified web pages, each score in the plurality of scores indicating a confidence that the web page corresponds to a particular category;

determining, for each web page in the one or more identified web pages, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page satisfies a threshold, associating the web page in the one or more web pages with a particular category corresponding to the score.

4. The method of claim 3, further comprising:

in response to determining that no score in the respective plurality of scores for a web page satisfy a threshold, providing the web page to a user for manual categorization; and

associating the web page with a category specified by the user.

5. The method of claim 2, wherein determining, based on the statistical analysis, respective categorizations of one or more web page segments in the one or more web pages in the financial institution website comprises:

determining a respective plurality of scores for each web page segment in the one or more web pages, each score in the plurality of scores indicating a confidence that a web page segment corresponds to a particular category;

determining, for each web page segment, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page segment satisfies a threshold, associating the web page segment with a particular category corresponding to the score.

6. The method of claim 5, further comprising:
in response to determining that no score in the respective plurality of scores for a web
page segment satisfy a threshold, providing the web page segment to a user for manual
categorization; and
5 associating the web page segment with a category specified by the user.

7. The method of claim 1, wherein the statistical analysis of (i) web pages and (ii) web
page segments comprises:
generating, based on one or more machine learning techniques, a first data model that
10 models a relationship between a serialized DOM of web pages to respective categories of the
web pages;
generating, based on one or more machine learning techniques, a second data model
that models a relationship between a serialized DOM of web page segments to respective
categories of the web page segments; and
15 using the first and second data models to respectively categorize web pages and web
page segments in the financial institution website.

8. The method of claim 1, wherein generating, based on the site map of the financial
institution website, a site-specific script for extracting financial data from the financial
20 institution website does not require human input.

9. A computer storage medium encoded with a computer program, the program comprising instructions that when executed by data processing apparatus cause the data processing apparatus to perform operations comprising:

determining whether a site-specific script for extracting financial data from a particular financial institution website is available;

in response to determining that a site-specific script for extracting financial data from the particular financial institution website is not available, generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based on at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website;

generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website; and

extracting, for one or more users, financial data from the particular financial institution website using the generated site-specific script.

10. The medium of claim 9, wherein generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based on at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not
5 in the financial institution website comprises:

crawling the financial institution website to identify one or more web pages in the financial institution website and one or more web page segments in the one or more web pages in the financial institution website;

determining, based on the statistical analysis, respective categorizations for the one or
10 more identified web pages; and

determining, based on the statistical analysis, respective categorizations for the one or more web page segments in the identified one or more web pages.

11. The medium of claim 10, wherein determining, based on the statistical analysis, respective categorizations for the one or more identified web pages comprises:

determining a respective plurality of scores for each web page in the one or more identified web pages, each score in the plurality of scores indicating a confidence that the web page corresponds to a particular category;

determining, for each web page in the one or more identified web pages, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page satisfies a threshold, associating the web page in the one or more web pages with a particular category corresponding to the score.

12. The medium of claim 11, further comprising:

in response to determining that no score in the respective plurality of scores for a web page satisfy a threshold, providing the web page to a user for manual categorization; and

associating the web page with a category specified by the user.

13. The medium of claim 10, wherein determining, based on the statistical analysis, respective categorizations of one or more web page segments in the one or more web pages in the financial institution website comprises:

determining a respective plurality of scores for each web page segment in the one or more web pages, each score in the plurality of scores indicating a confidence that a web page segment corresponds to a particular category;

determining, for each web page segment, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page segment satisfies a threshold, associating the web page segment with a particular category corresponding to the score.

14. The medium of claim 13, further comprising:

in response to determining that no score in the respective plurality of scores for a web page segment satisfy a threshold, providing the web page segment to a user for manual categorization; and

5 associating the web page segment with a category specified by the user.

15. The medium of claim 9, wherein the statistical analysis of (i) web pages and (ii) web page segments comprises:

generating, based on one or more machine learning techniques, a first data model that
10 models a relationship between a serialized DOM of web pages to respective categories of the web pages;

generating, based on one or more machine learning techniques, a second data model that models a relationship between a serialized DOM of web page segments to respective categories of the web page segments; and

15 using the first and second data models to respectively categorize web pages and web page segments in the financial institution website.

16. The medium of claim 9, wherein generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial
20 institution website does not require human input.

17. A system comprising one or more computers programmed to perform operations comprising:

determining whether a site-specific script for extracting financial data from a particular financial institution website is available;

in response to determining that a site-specific script for extracting financial data from the particular financial institution website is not available, generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based on at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website;

generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website; and

extracting, for one or more users, financial data from the particular financial institution website using the generated site-specific script.

18. The system of claim 17, wherein generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based on at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website comprises:

crawling the financial institution website to identify one or more web pages in the financial institution website and one or more web page segments in the one or more web pages in the financial institution website;

determining, based on the statistical analysis, respective categorizations for the one or more identified web pages; and

determining, based on the statistical analysis, respective categorizations for the one or more web page segments in the identified one or more web pages.

19. The system of claim 18, wherein determining, based on the statistical analysis, respective categorizations for the one or more identified web pages comprises:

determining a respective plurality of scores for each web page in the one or more identified web pages, each score in the plurality of scores indicating a confidence that the web page corresponds to a particular category;

determining, for each web page in the one or more identified web pages, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page satisfies a threshold, associating the web page in the one or more web pages with a particular category corresponding to the score.

20. The system of claim 19, further comprising:

in response to determining that no score in the respective plurality of scores for a web page satisfy a threshold, providing the web page to a user for manual categorization; and

associating the web page with a category specified by the user.

21. The system of claim 18, wherein determining, based on the statistical analysis, respective categorizations of one or more web page segments in the one or more web pages in the financial institution website comprises:

determining a respective plurality of scores for each web page segment in the one or more web pages, each score in the plurality of scores indicating a confidence that a web page segment corresponds to a particular category;

determining, for each web page segment, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page segment satisfies a threshold, associating the web page segment with a particular category corresponding to the score.

22. The system of claim 21, further comprising:
in response to determining that no score in the respective plurality of scores for a web
page segment satisfy a threshold, providing the web page segment to a user for manual
categorization; and
5 associating the web page segment with a category specified by the user.

23. The system of claim 17, wherein the statistical analysis of (i) web pages and (ii) web
page segments comprises:

generating, based on one or more machine learning techniques, a first data model that
10 models a relationship between a serialized DOM of web pages to respective categories of the
web pages;

generating, based on one or more machine learning techniques, a second data model
that models a relationship between a serialized DOM of web page segments to respective
categories of the web page segments; and

15 using the first and second data models to respectively categorize web pages and web
page segments in the financial institution website.

24. The system of claim 17, wherein generating, based on the site map of the financial
institution website, a site-specific script for extracting financial data from the financial
20 institution website does not require human input.

CLAIMS

1. A computer-implemented method, comprising:
 - determining, in a computer, whether a site-specific script for extracting financial data from a particular financial institution website is available;
 - in response to determining that a site-specific script for extracting financial data from the particular financial institution website is not available, generating, in the computer, a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website;
 - generating, in the computer, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website; and
 - extracting, in the computer, for one or more users, financial data from the particular financial institution website using the generated site-specific script.
2. The method of claim 1, wherein generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website comprises:
 - crawling the financial institution website to identify one or more web pages in the financial institution website and one or more web page segments in the one or more web pages in the financial institution website;
 - determining, based on the statistical analysis, respective categorizations for the one or more identified web pages; and
 - determining, based on the statistical analysis, respective categorizations for the one or more web page segments in the identified one or more web pages.

3. The method of claim 2, wherein determining, based on the statistical analysis, respective categorizations for the one or more identified web pages comprises:

determining a respective plurality of scores for each web page in the one or more identified web pages, each score in the plurality of scores indicating a confidence that the web page corresponds to a particular category;

determining, for each web page in the one or more identified web pages, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page satisfies a threshold, associating the web page in the one or more web pages with a particular category corresponding to the score.

4. The method of claim 3, further comprising:

in response to determining that no score in the respective plurality of scores for a web page satisfy a threshold, providing the web page to a user for manual categorization; and

associating the web page with a category specified by the user.

5. The method of claim 2, wherein determining, based on the statistical analysis, respective categorizations of one or more web page segments in the one or more web pages in the financial institution website comprises:

determining a respective plurality of scores for each web page segment in the one or more web pages, each score in the plurality of scores indicating a confidence that a web page segment corresponds to a particular category;

determining, for each web page segment, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page segment satisfies a threshold, associating the web page segment with a particular category corresponding to the score.

6. The method of claim 5, further comprising:
 - in response to determining that no score in the respective plurality of scores for a web page segment satisfy a threshold, providing the web page segment to a user for manual categorization; and
 - associating the web page segment with a category specified by the user.
7. The method of claim 1, wherein the statistical analysis of (i) web pages and (ii) web page segments comprises:
 - generating, based on one or more machine learning techniques, a first data model that models a relationship between a serialized DOM of web pages to respective categories of the web pages;
 - generating, based on one or more machine learning techniques, a second data model that models a relationship between a serialized DOM of web page segments to respective categories of the web page segments; and
 - using the first and second data models to respectively categorize web pages and web page segments in the financial institution website.
8. The method of claim 1, wherein generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website does not require human input.

9. A non-transitory computer storage medium encoded with a computer program, the program comprising instructions that when executed by data processing apparatus cause the data processing apparatus to perform operations comprising:

determining whether a site-specific script for extracting financial data from a particular financial institution website is available;

in response to determining that a site-specific script for extracting financial data from the particular financial institution website is not available, generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website;

generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website; and

extracting, for one or more users, financial data from the particular financial institution website using the generated site-specific script.

10. The medium of claim 9, wherein generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website comprises:

crawling the financial institution website to identify one or more web pages in the financial institution website and one or more web page segments in the one or more web pages in the financial institution website;

determining, based on the statistical analysis, respective categorizations for the one or more identified web pages; and

determining, based on the statistical analysis, respective categorizations for the one or more web page segments in the identified one or more web pages.

11. The medium of claim 10, wherein determining, based on the statistical analysis, respective categorizations for the one or more identified web pages comprises:

determining a respective plurality of scores for each web page in the one or more identified web pages, each score in the plurality of scores indicating a confidence that the web page corresponds to a particular category;

determining, for each web page in the one or more identified web pages, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page satisfies a threshold, associating the web page in the one or more web pages with a particular category corresponding to the score.

12. The medium of claim 11, further comprising:

in response to determining that no score in the respective plurality of scores for a web page satisfy a threshold, providing the web page to a user for manual categorization; and

associating the web page with a category specified by the user.

13. The medium of claim 10, wherein determining, based on the statistical analysis, respective categorizations of one or more web page segments in the one or more web pages in the financial institution website comprises:

determining a respective plurality of scores for each web page segment in the one or more web pages, each score in the plurality of scores indicating a confidence that a web page segment corresponds to a particular category;

determining, for each web page segment, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page segment satisfies a threshold, associating the web page segment with a particular category corresponding to the score.

14. The medium of claim 13, further comprising:
 - in response to determining that no score in the respective plurality of scores for a web page segment satisfy a threshold, providing the web page segment to a user for manual categorization; and
 - associating the web page segment with a category specified by the user.
15. The medium of claim 9, wherein the statistical analysis of (i) web pages and (ii) web page segments comprises:
 - generating, based on one or more machine learning techniques, a first data model that models a relationship between a serialized DOM of web pages to respective categories of the web pages;
 - generating, based on one or more machine learning techniques, a second data model that models a relationship between a serialized DOM of web page segments to respective categories of the web page segments; and
 - using the first and second data models to respectively categorize web pages and web page segments in the financial institution website.
16. The medium of claim 9, wherein generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website does not require human input.

17. A system comprising one or more computers programmed to perform operations comprising:

determining whether a site-specific script for extracting financial data from a particular financial institution website is available;

in response to determining that a site-specific script for extracting financial data from the particular financial institution website is not available, generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website;

generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website; and

extracting, for one or more users, financial data from the particular financial institution website using the generated site-specific script.

18. The system of claim 17, wherein generating a site map of web pages and web page segments in the financial institution website, wherein the site map is generated based at least in part on a statistical analysis of (i) web pages and (ii) web page segments that are not in the financial institution website comprises:

crawling the financial institution website to identify one or more web pages in the financial institution website and one or more web page segments in the one or more web pages in the financial institution website;

determining, based on the statistical analysis, respective categorizations for the one or more identified web pages; and

determining, based on the statistical analysis, respective categorizations for the one or more web page segments in the identified one or more web pages.

19. The system of claim 18, wherein determining, based on the statistical analysis, respective categorizations for the one or more identified web pages comprises:

determining a respective plurality of scores for each web page in the one or more identified web pages, each score in the plurality of scores indicating a confidence that the web page corresponds to a particular category;

determining, for each web page in the one or more identified web pages, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page satisfies a threshold, associating the web page in the one or more web pages with a particular category corresponding to the score.

20. The system of claim 19, further comprising:

in response to determining that no score in the respective plurality of scores for a web page satisfy a threshold, providing the web page to a user for manual categorization; and

associating the web page with a category specified by the user.

21. The system of claim 18, wherein determining, based on the statistical analysis, respective categorizations of one or more web page segments in the one or more web pages in the financial institution website comprises:

determining a respective plurality of scores for each web page segment in the one or more web pages, each score in the plurality of scores indicating a confidence that a web page segment corresponds to a particular category;

determining, for each web page segment, whether a score in the respective plurality of scores satisfies a threshold; and

in response to determining that a score in the respective plurality of scores for a web page segment satisfies a threshold, associating the web page segment with a particular category corresponding to the score.

22. The system of claim 21, further comprising:
in response to determining that no score in the respective plurality of scores for a web page segment satisfy a threshold, providing the web page segment to a user for manual categorization; and
associating the web page segment with a category specified by the user.
23. The system of claim 17, wherein the statistical analysis of (i) web pages and (ii) web page segments comprises:
generating, based on one or more machine learning techniques, a first data model that models a relationship between a serialized DOM of web pages to respective categories of the web pages;
generating, based on one or more machine learning techniques, a second data model that models a relationship between a serialized DOM of web page segments to respective categories of the web page segments; and
using the first and second data models to respectively categorize web pages and web page segments in the financial institution website.
24. The system of claim 17, wherein generating, based on the site map of the financial institution website, a site-specific script for extracting financial data from the financial institution website does not require human input.