



US 20160162456A1

(19) **United States**

(12) **Patent Application Publication**
Munro et al.

(10) **Pub. No.: US 2016/0162456 A1**
(43) **Pub. Date: Jun. 9, 2016**

(54) **METHODS FOR GENERATING NATURAL LANGUAGE PROCESSING SYSTEMS**

(71) Applicants: **Robert J. Munro**, San Francisco, CA (US); **Schuyler D. Erle**, San Francisco, CA (US); **Christopher Walker**, San Francisco, CA (US); **Sarah K. Luger**, San Francisco, CA (US); **Jason Brenier**, Oakland, CA (US); **Gary C. King**, Los Altos, CA (US); **Paul A. Tepper**, San Francisco, CA (US); **Ross Mechanic**, San Francisco, CA (US); **Andrew Gilchrist-Scott**, Berkeley, CA (US); **Jessica D. Long**, San Francisco, CA (US); **James B. Robinson**, San Francisco, CA (US); **Brendan D. Callahan**, Philadelphia, PA (US); **Michelle Casbon**, San Antonio, TX (US); **Ujjwal Sarin**, San Francisco, CA (US); **Aneesh Nair**, Fremont, CA (US); **Veena Basavaraj**, San Francisco, CA (US); **Tripti Saxena**, Cupertino, CA (US); **Edgar Nunez**, Union City, CA (US); **Martha G. Hinrichs**, San Francisco, CA (US); **Haley Most**, San Francisco, CA (US); **Tyler J. Schnoebelen**, San Francisco, CA (US)

(72) Inventors: **Robert J. Munro**, San Francisco, CA (US); **Schuyler D. Erle**, San Francisco, CA (US); **Christopher Walker**, San Francisco, CA (US); **Sarah K. Luger**, San Francisco, CA (US); **Jason Brenier**, Oakland, CA (US); **Gary C. King**, Los Altos, CA (US); **Paul A. Tepper**, San Francisco, CA (US); **Ross Mechanic**, San Francisco, CA (US); **Andrew Gilchrist-Scott**, Berkeley, CA (US); **Jessica D. Long**, San Francisco, CA (US); **James B. Robinson**, San Francisco, CA (US); **Brendan D. Callahan**, Philadelphia, PA (US); **Michelle Casbon**, San Antonio, TX (US); **Ujjwal Sarin**, San Francisco, CA

(US); **Aneesh Nair**, Fremont, CA (US); **Veena Basavaraj**, San Francisco, CA (US); **Tripti Saxena**, Cupertino, CA (US); **Edgar Nunez**, Union City, CA (US); **Martha G. Hinrichs**, San Francisco, CA (US); **Haley Most**, San Francisco, CA (US); **Tyler J. Schnoebelen**, San Francisco, CA (US)

(73) Assignee: **Idibon, Inc.**, San Francisco, CA (US)

(21) Appl. No.: **14/964,517**

(22) Filed: **Dec. 9, 2015**

Related U.S. Application Data

(60) Provisional application No. 62/089,736, filed on Dec. 9, 2014, provisional application No. 62/089,742, filed on Dec. 9, 2014, provisional application No. 62/089,745, filed on Dec. 9, 2014, provisional application No. 62/089,747, filed on Dec. 9, 2014.

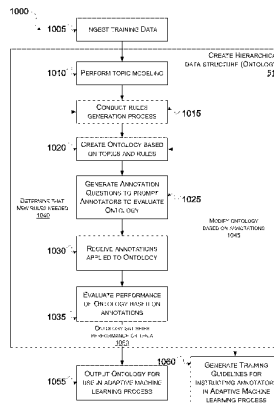
Publication Classification

(51) **Int. Cl.**
G06F 17/24 (2006.01)
G06F 17/22 (2006.01)
G06F 17/28 (2006.01)

(52) **U.S. Cl.**
CPC **G06F 17/241** (2013.01); **G06F 17/28** (2013.01); **G06F 17/2241** (2013.01)

(57) **ABSTRACT**

Methods are presented for generating a natural language model. The method may comprise: ingesting training data representative of documents to be analyzed by the natural language model, generating a hierarchical data structure comprising at least two topical nodes within which the training data is to be subdivided into by the natural language model, selecting a plurality of documents among the training data to be annotated, generating an annotation prompt for each document configured to elicit an annotation about said document indicating which node among the at least two topical nodes said document is to be classified into, receiving the



annotation based on the annotation prompt; and generating the natural language model using an adaptive machine learning process configured to determine patterns among the anno-

tations for how the documents in the training data are to be subdivided according to the at least two topical nodes of the hierarchical data structure.

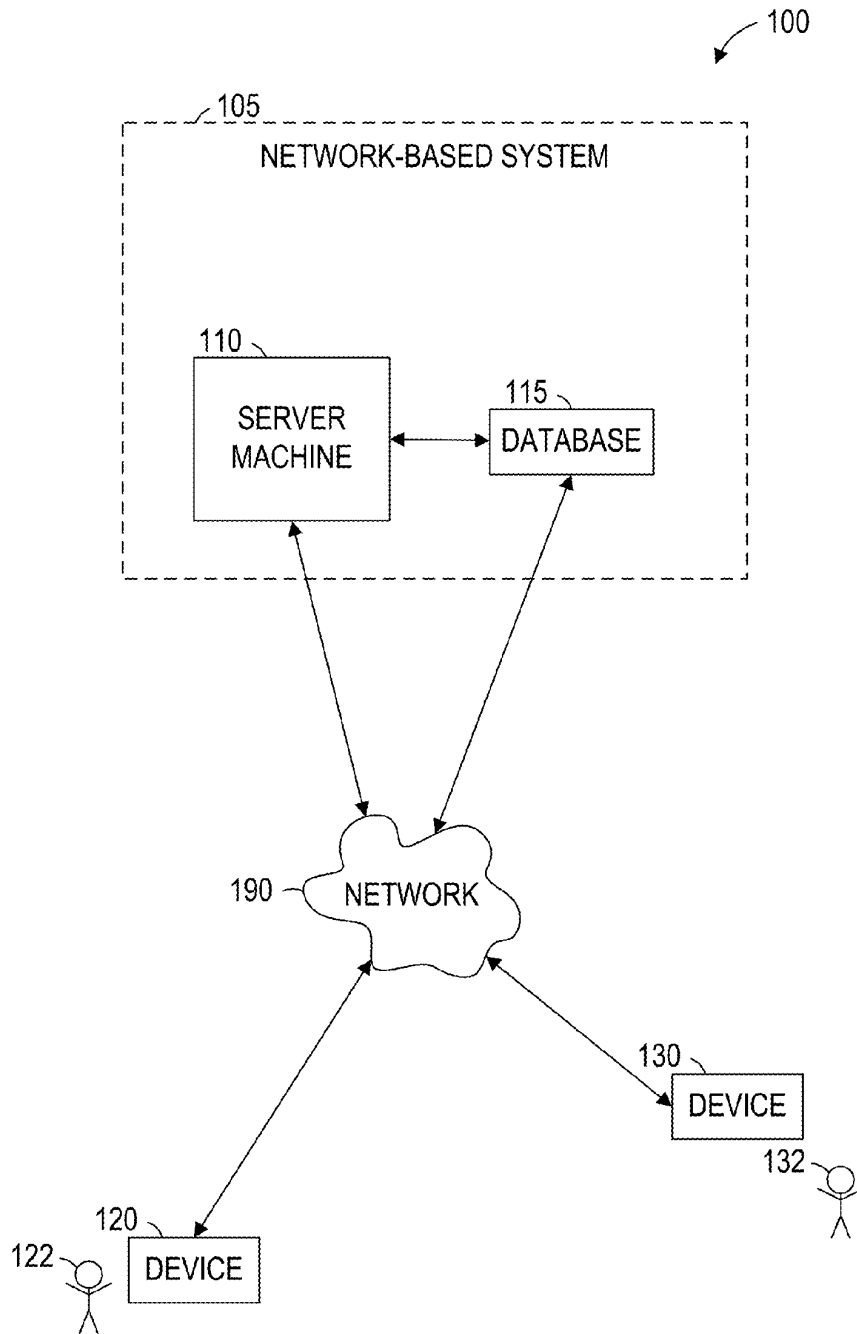


FIG. 1A

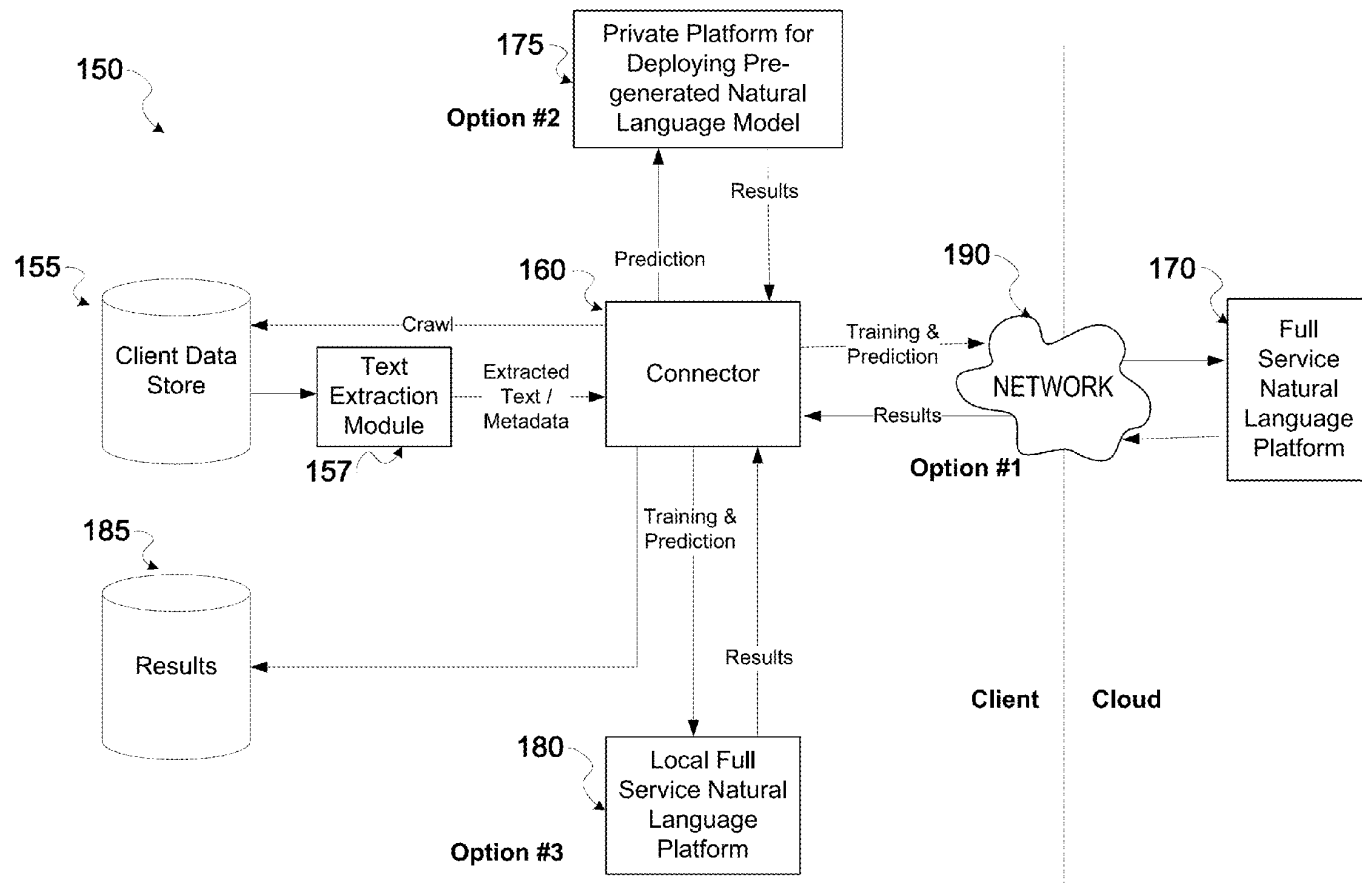


FIG. 1B

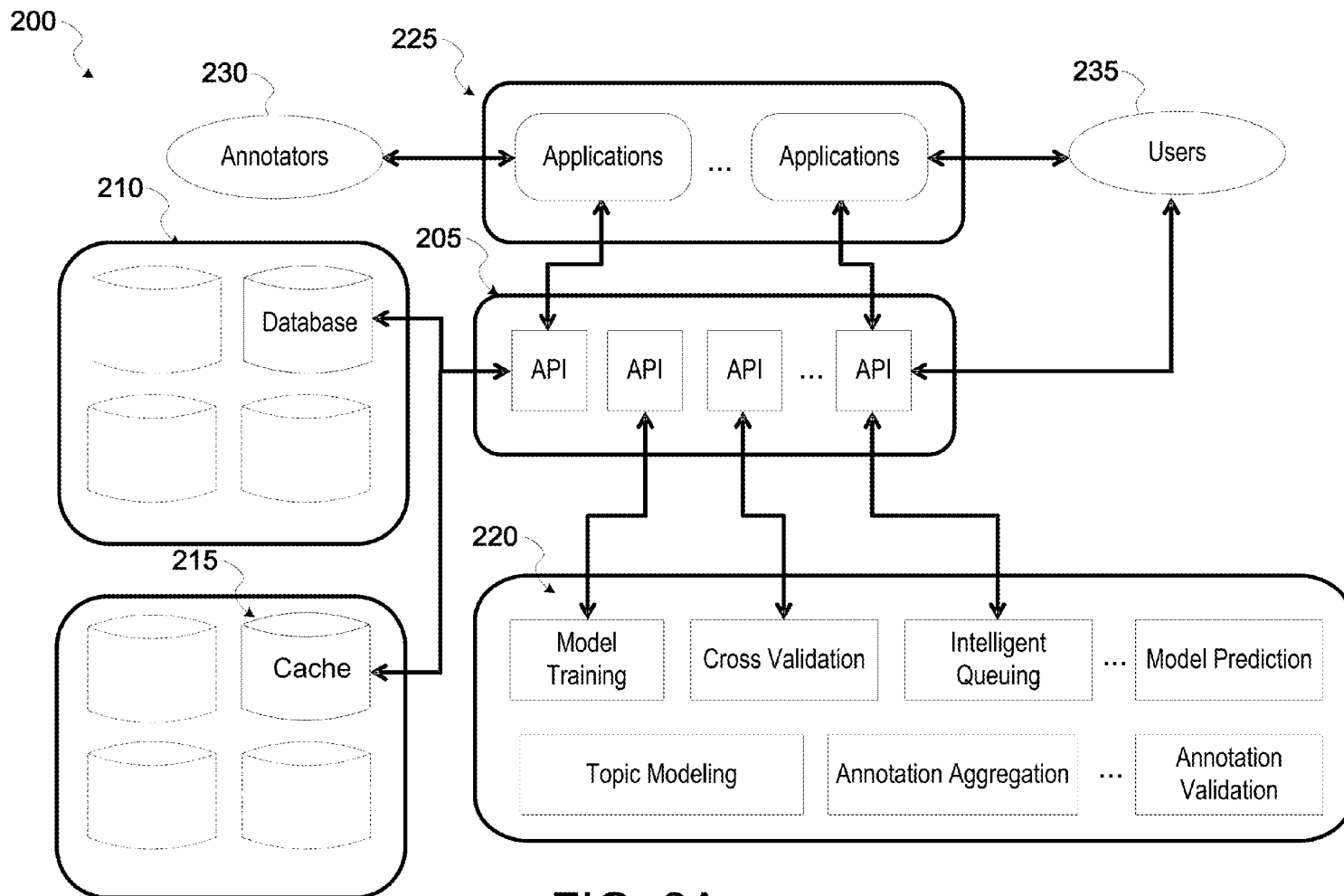


FIG. 2A

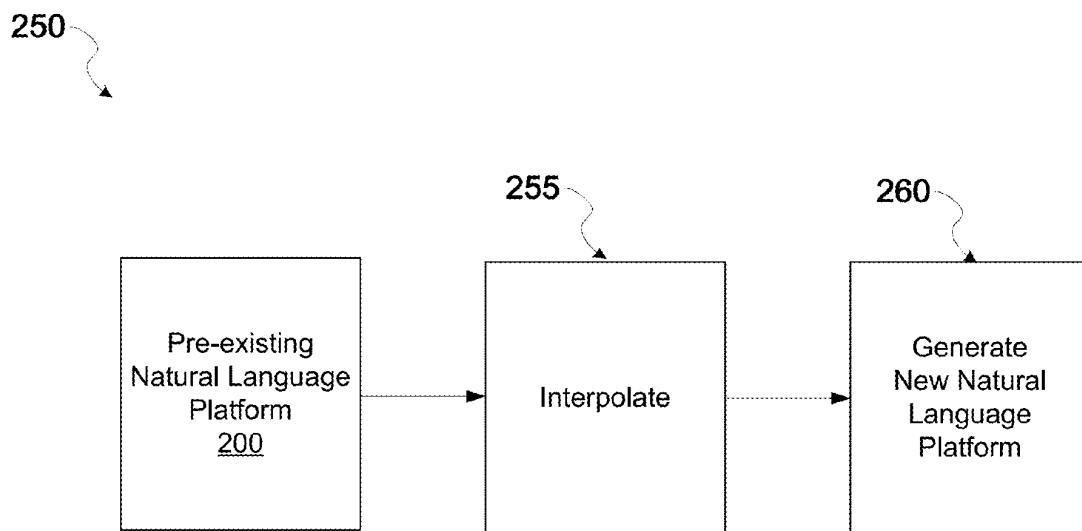


FIG. 2B

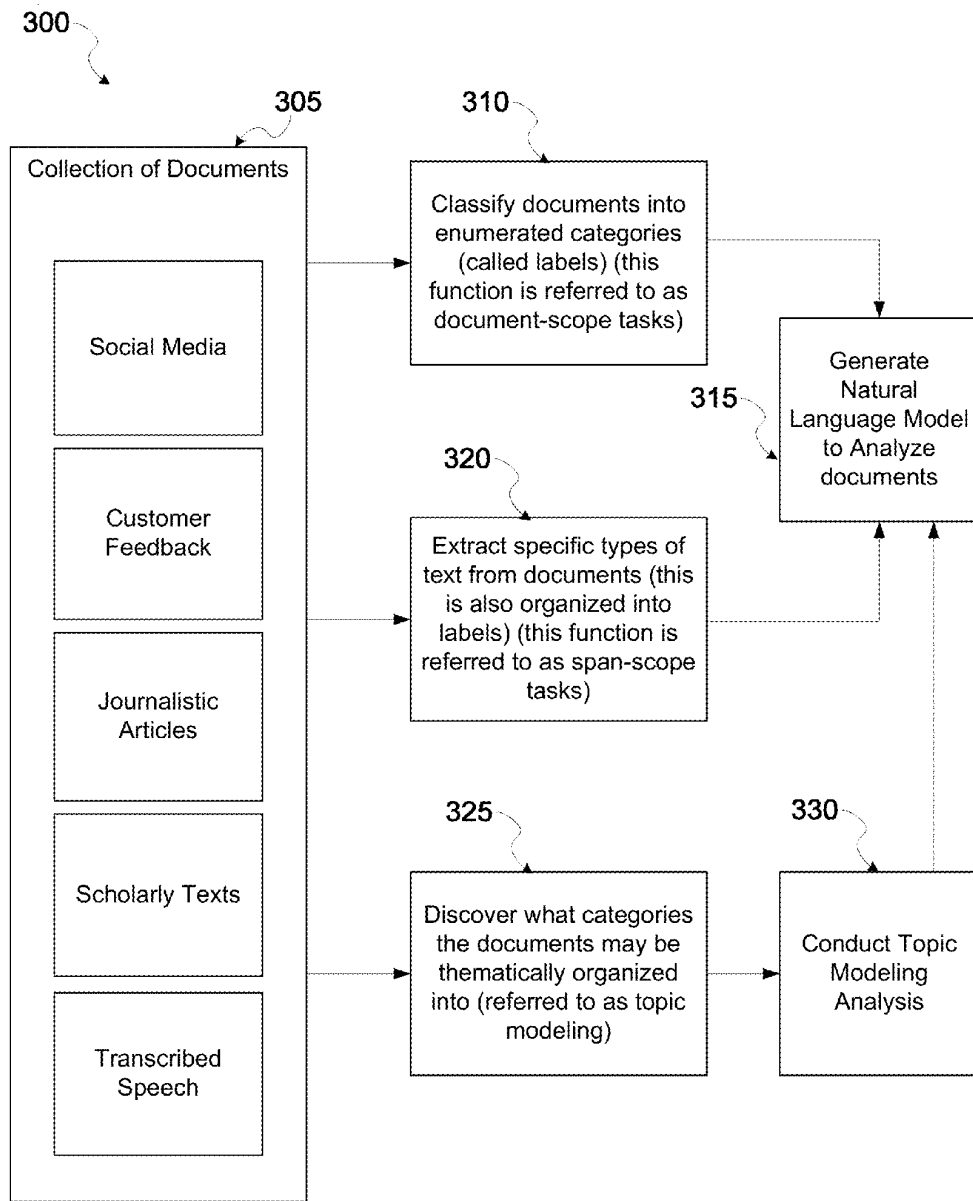


FIG. 3

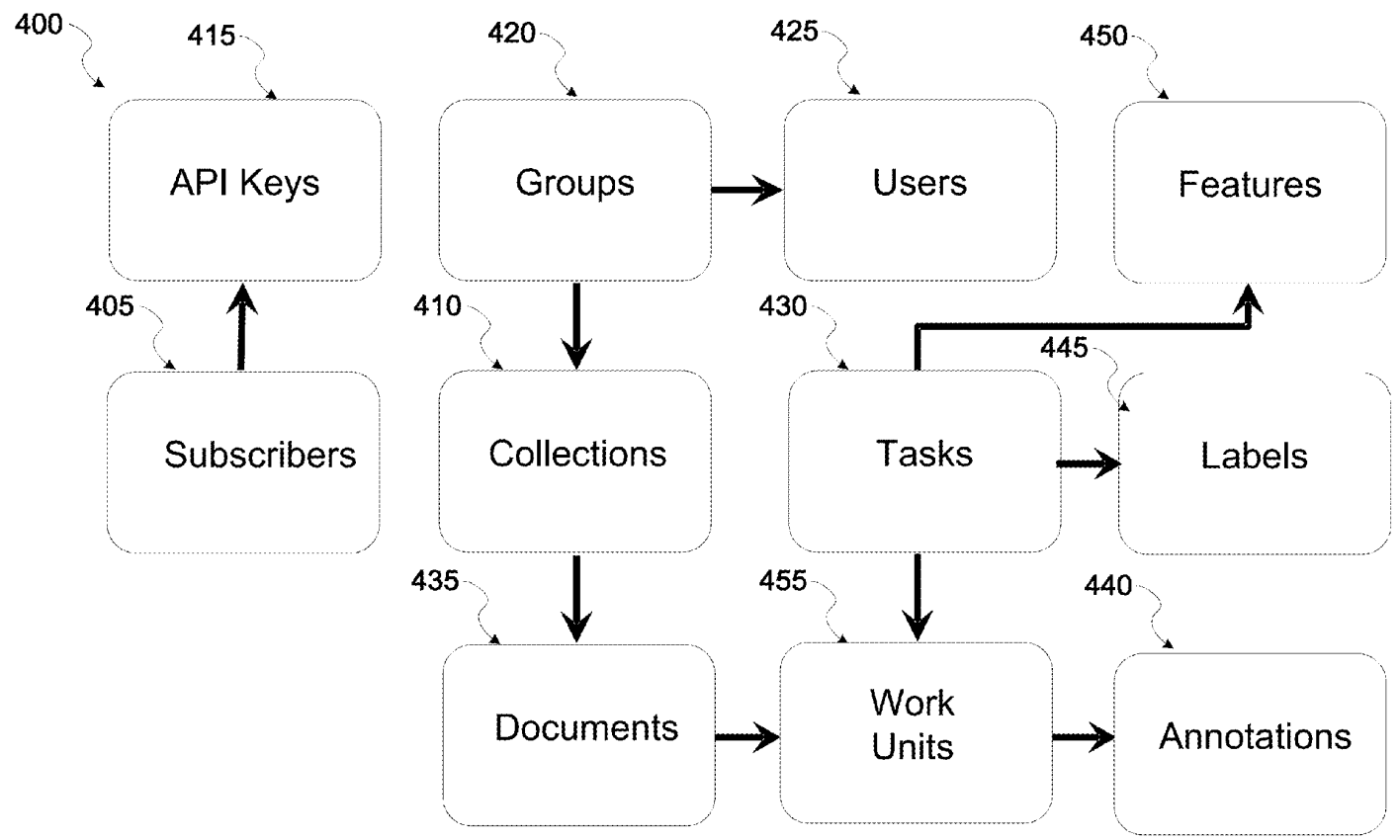


FIG. 4

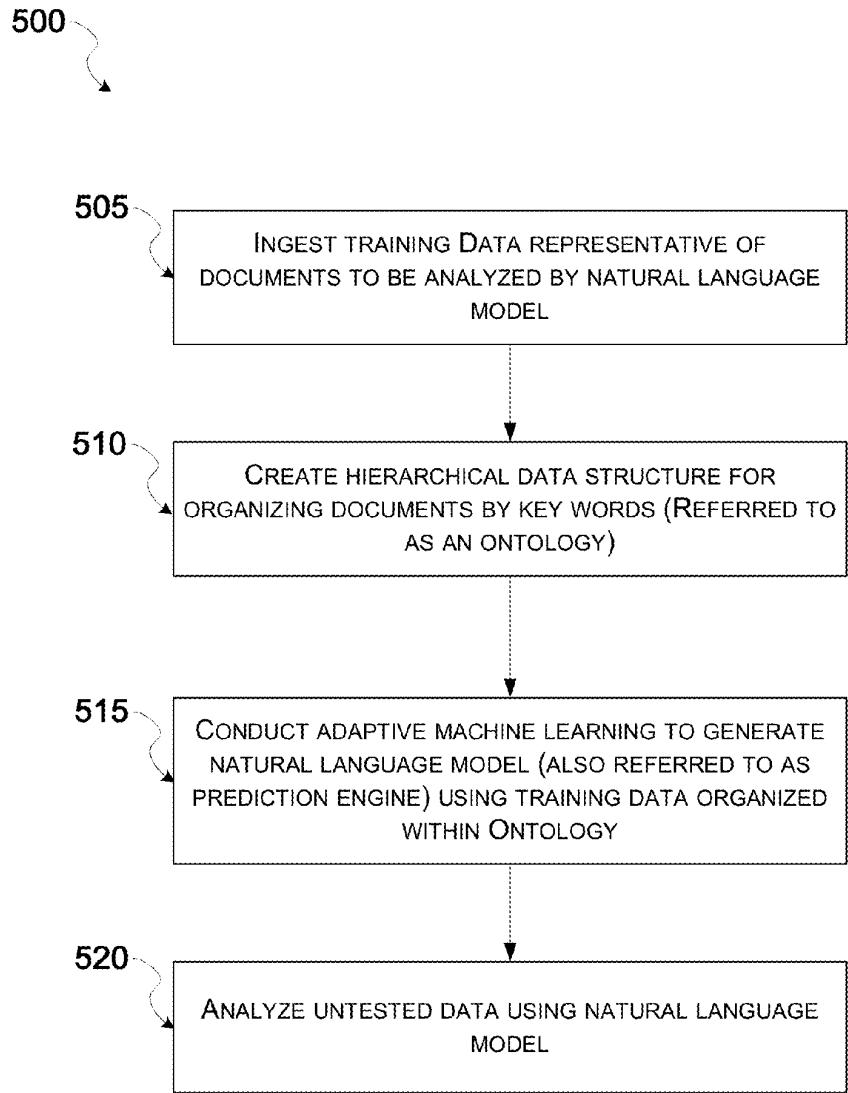


FIG. 5

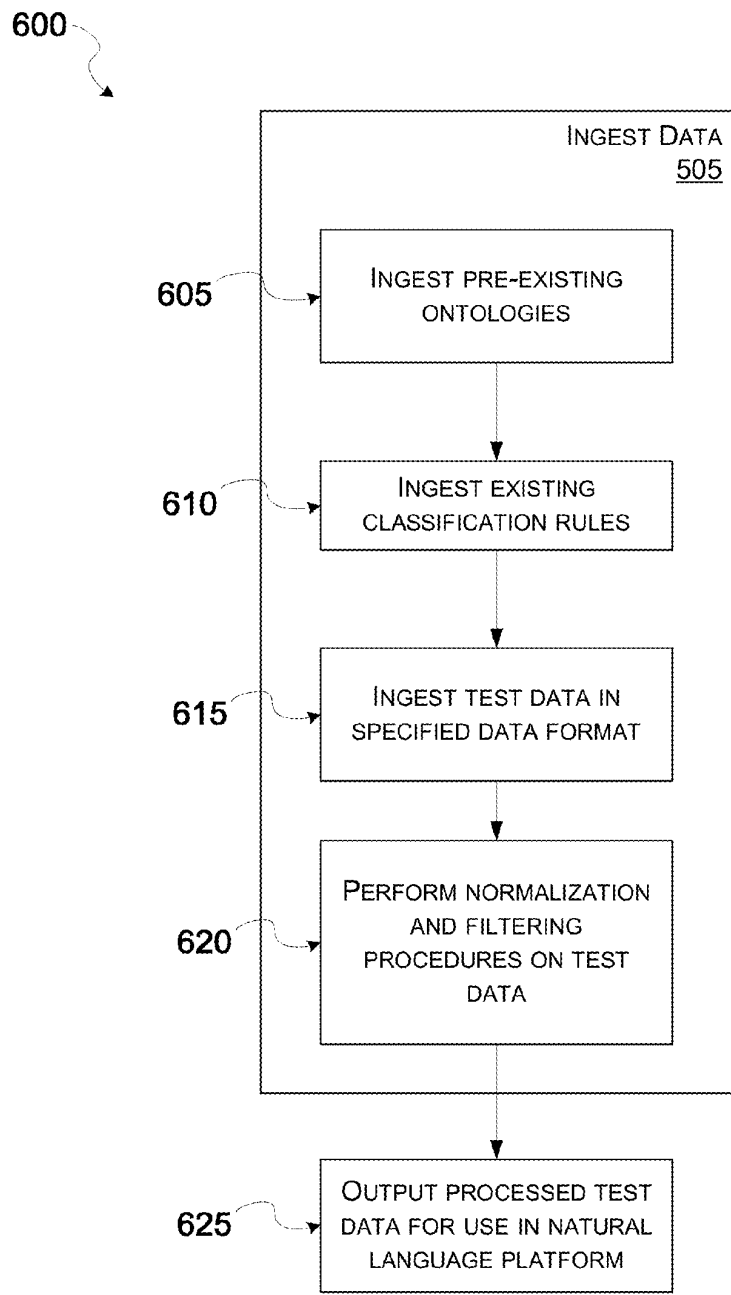


FIG. 6

700

Task	Label	Subtask
Sentiment	Positive	
Sentiment	Negative	
Sentiment	Neutral	
Relevant	Relevant	
Relevant	Irrelevant	

710

Task, Label, Subtask
Sentiment, Positive,
Sentiment, Negative,
Sentiment, Neutral,
Relevant, Relevant,
Relevant, Irrelevant,

720

- ▼ Relevant
- Irrelevant
- Relevant
- ▼ Sentiment
- Negative
- Neutral
- Positive

FIG. 7

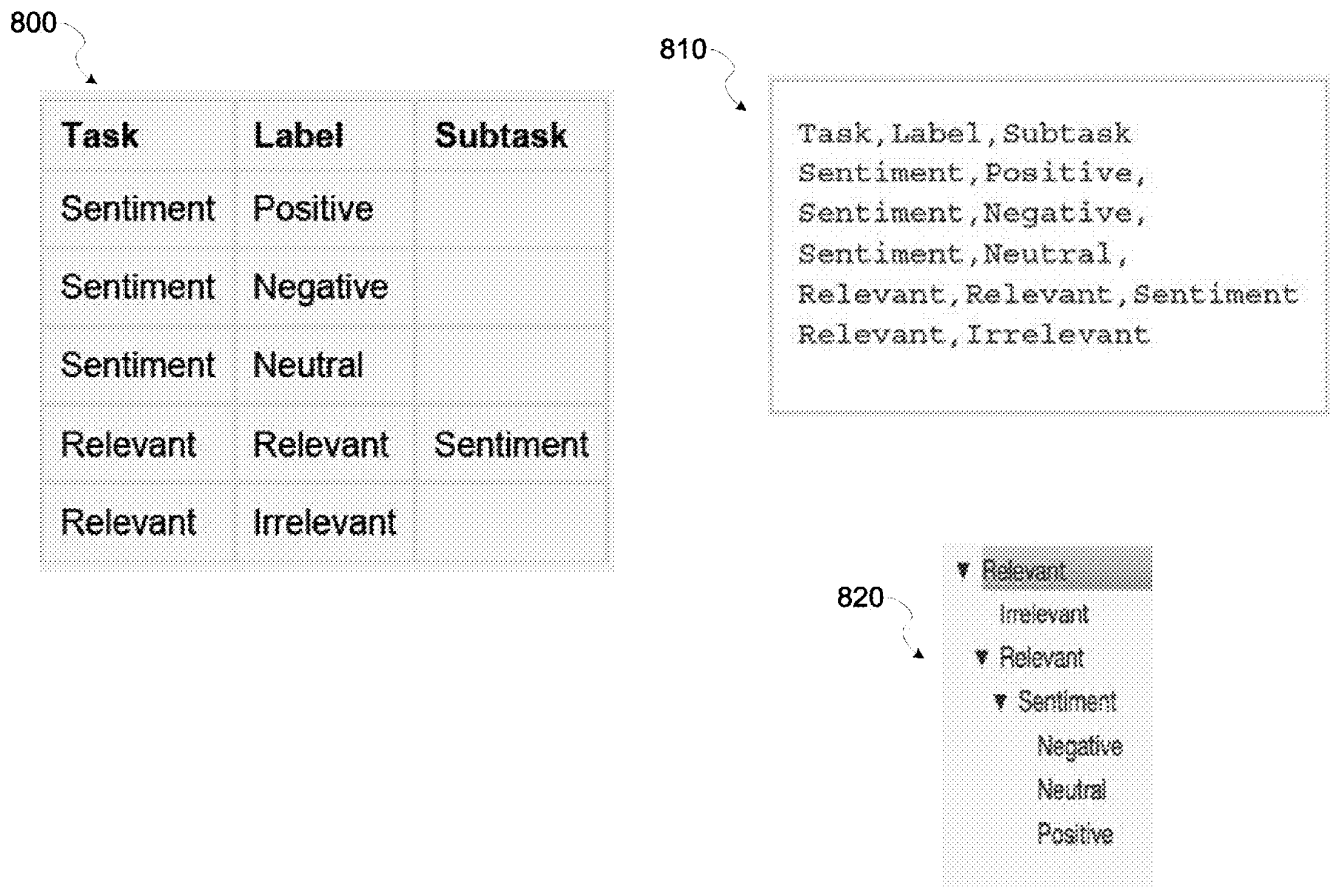


FIG. 8

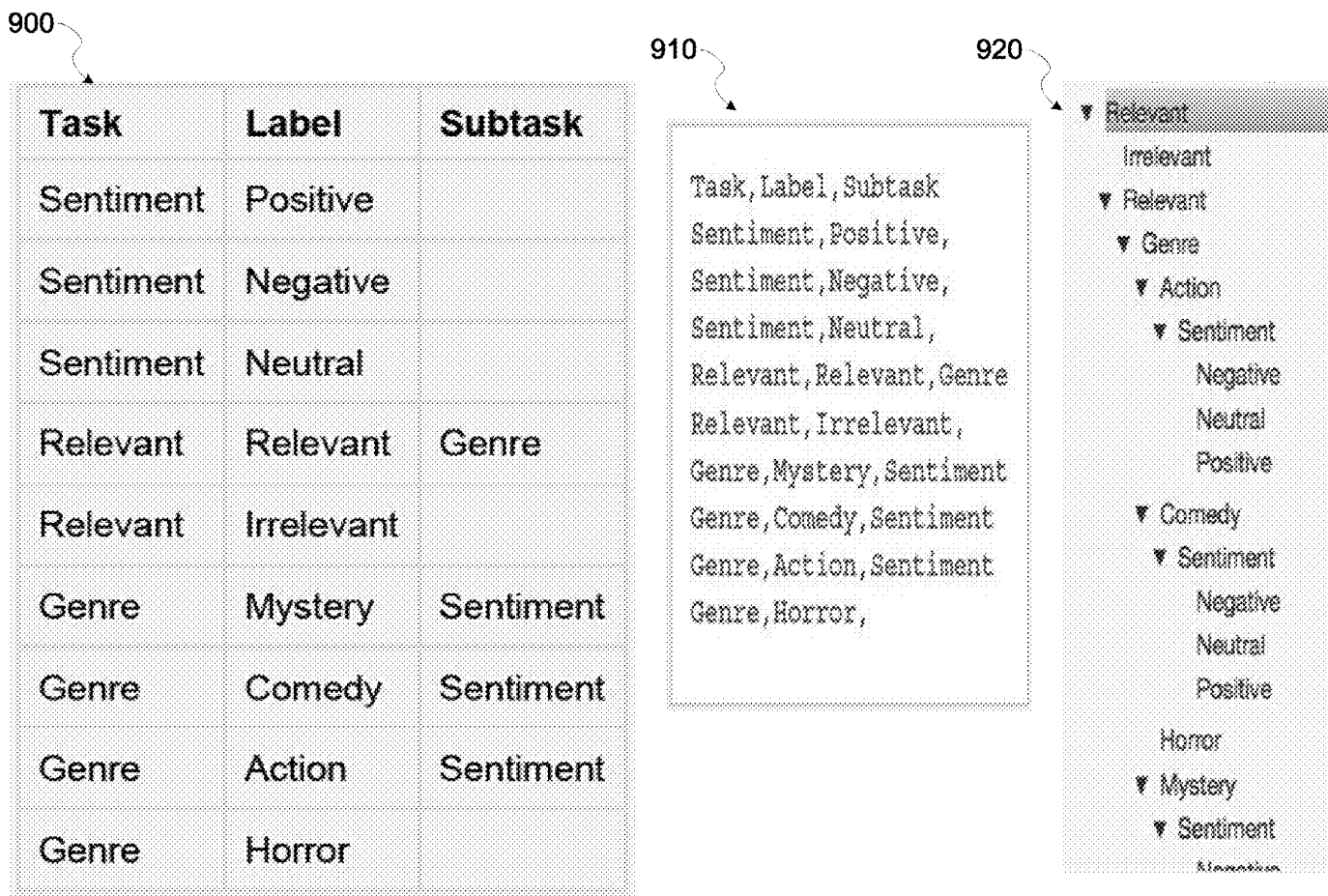


FIG. 9A

930

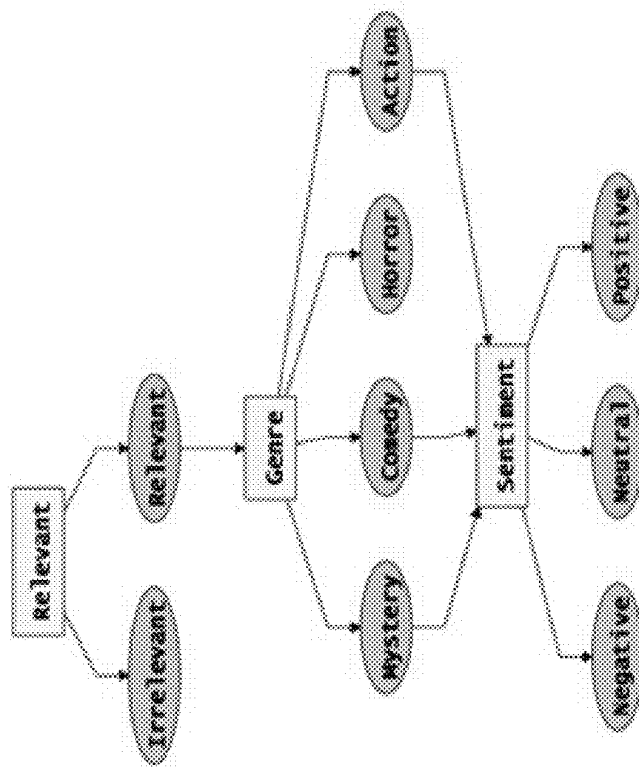


FIG. 9B

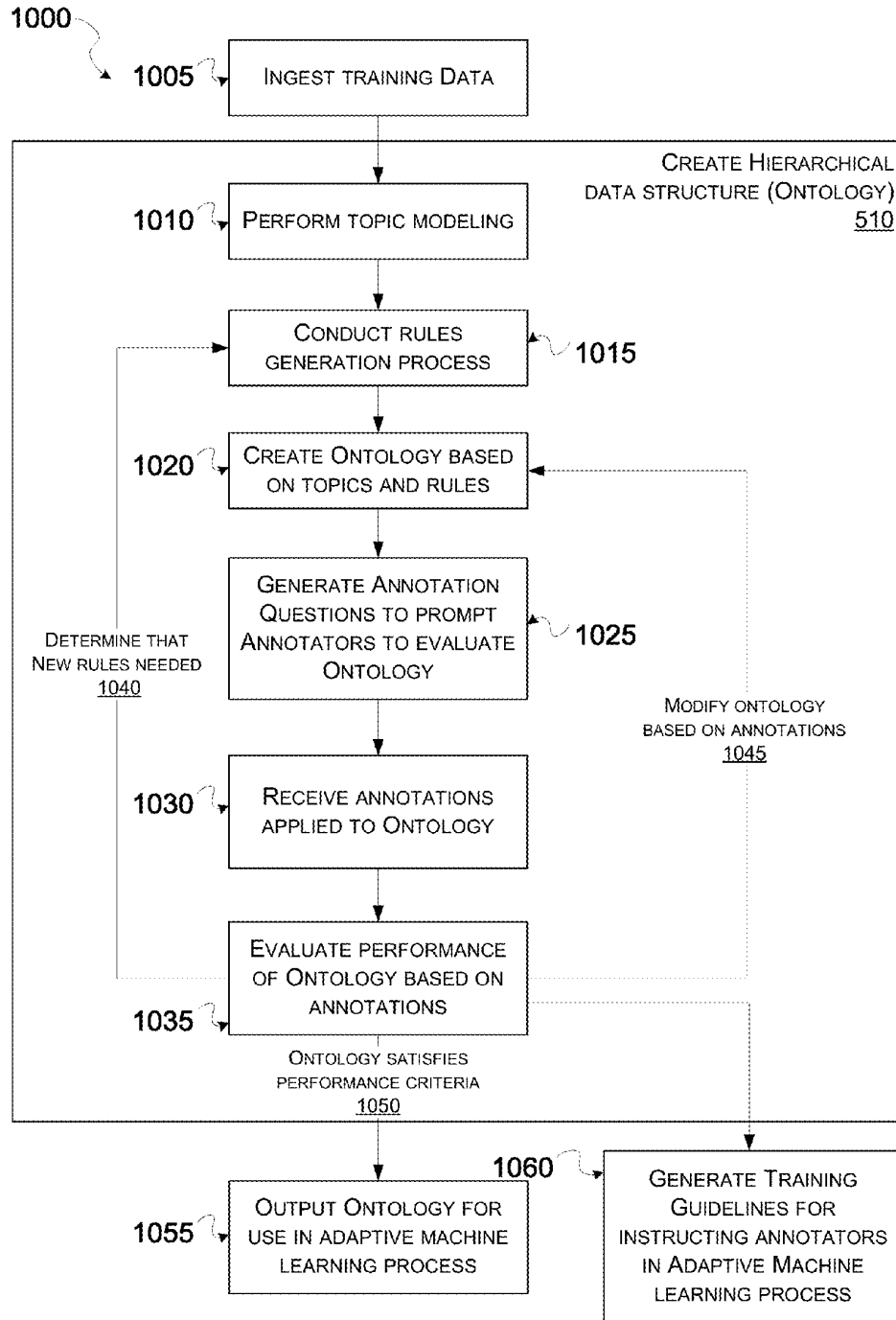


FIG. 10A

1027

➤ Annotate

colleen_jrnc_demo_step_2
JP Morgan Chase issued media audit, ready for initial presentation

The screenshot displays a document annotation interface. On the left, a sidebar titled "Document Labels for corporate..." contains a list of categories: Customer Profile, Executives, JPMC Financials, Legal, Other, Recommendations, and Security Issues. Each category has a radio button next to it. Below the list is a "NEXT DOCUMENT" button. The main content area shows a news article snippet with a photo of a woman. The article text includes: "JP Morgan to close 200 bank branches by 2018", "By Sara Flaherty @SARAFLAHERTY", and "JP Morgan Chase executives made their case Tuesday for why the bank should remain whole, even as it continues to pare down some branches." Below the article is a "View on web" link and social media icons for Facebook, Twitter, and LinkedIn. At the bottom of the interface is a "Done annotating" button.

FIG. 10B

1037

Agreement: 0.741

Total number of documents: 688
 Number of documents used for computation: 588

Results table agreement

Last calculated: October 16, 2015 11:29:48

Suggested labels to collapse

Labels collapsed	Number of documents affected by collapse	Updated agreement for collection
["Legal", "Other"]	205	0.773
["Legal", "Security", "Social"]	264	0.730
["Other", "Business", "Other"]	372	0.716
["Business", "Other"]	196	0.708
["Other", "Security", "Social"]	223	0.703

Per-analyst/label agreement

Analyst email	Number of documents annotated	Agreement value
anna@proton.com	150	0.794
anna@proton.com	150	0.794
anna@proton.com	150	0.771
anna@proton.com	150	0.800

Agreement per label

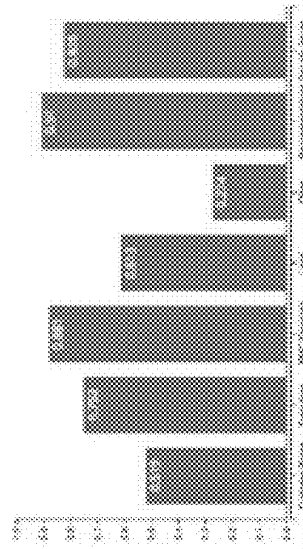


FIG. 10C

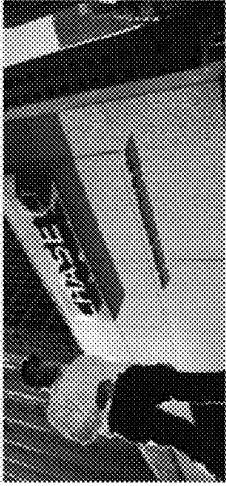
1100

Work Unit #1

Top World News
@TopNewsA1

JPMorgan faces scrutiny over Asia hiring practices: JPMorgan Chase's hiring of the son of a Chinese comme... cwb.cw/1xFD0nt

#CNC
11:36 PM · 6 Feb 2015
CNBC



JPMorgan faces fresh scrutiny over Asia hiring practices

JPMorgan Chase's hiring of the son of a Chinese commerce minister is being scrutinized by investigators looking at whether the bank improperly hired relatives of government officials

Line of Business

What is the best label for this document? (Check all that apply)

- Analyst Reports
- Branch Network
- Consumer
- Legal & Compliance
- SMB
- Corporate News
- Electronic Banking
- Mortgage & Housing

Submit None are good labels

FIG. 11A

1150

Work Unit #1

February 26: NorthStar Realty Finance Announces Plan to Spin-off European Real Estate Business into a Separate Publicly-Traded REIT

February 25: Northstar Realty Finance: Chatham Lodging Trust Caps Record Year with Strong Fourth Quarter

To view Conference Call-transcripts, click https://www.buySellSignals.net/BuySellSignals/report/Users/Stock/Daily/Link/1079_SSP_ConferenceCalltranscripts.html

27 Feb, 2015: Northstar Realty Finance's (NRF) CEO David Hamamoto on Q4 2014 Results - Earnings Call Transcript

27 Feb, 2015: Northstar Realty Finance (NRF) Q4 2014 Results - Earnings Call Webcast

ISIN: US66704R1005

N:NRF: NRFN

Source: www.BuySellSignals.com

Relevance

What is the best label for this document?

Highly_relevant (Highly_relevant)

Semi-relevant (Semi-relevant)

Totally_irrelevant (Totally_irrelevant)

Submit

FIG. 11B

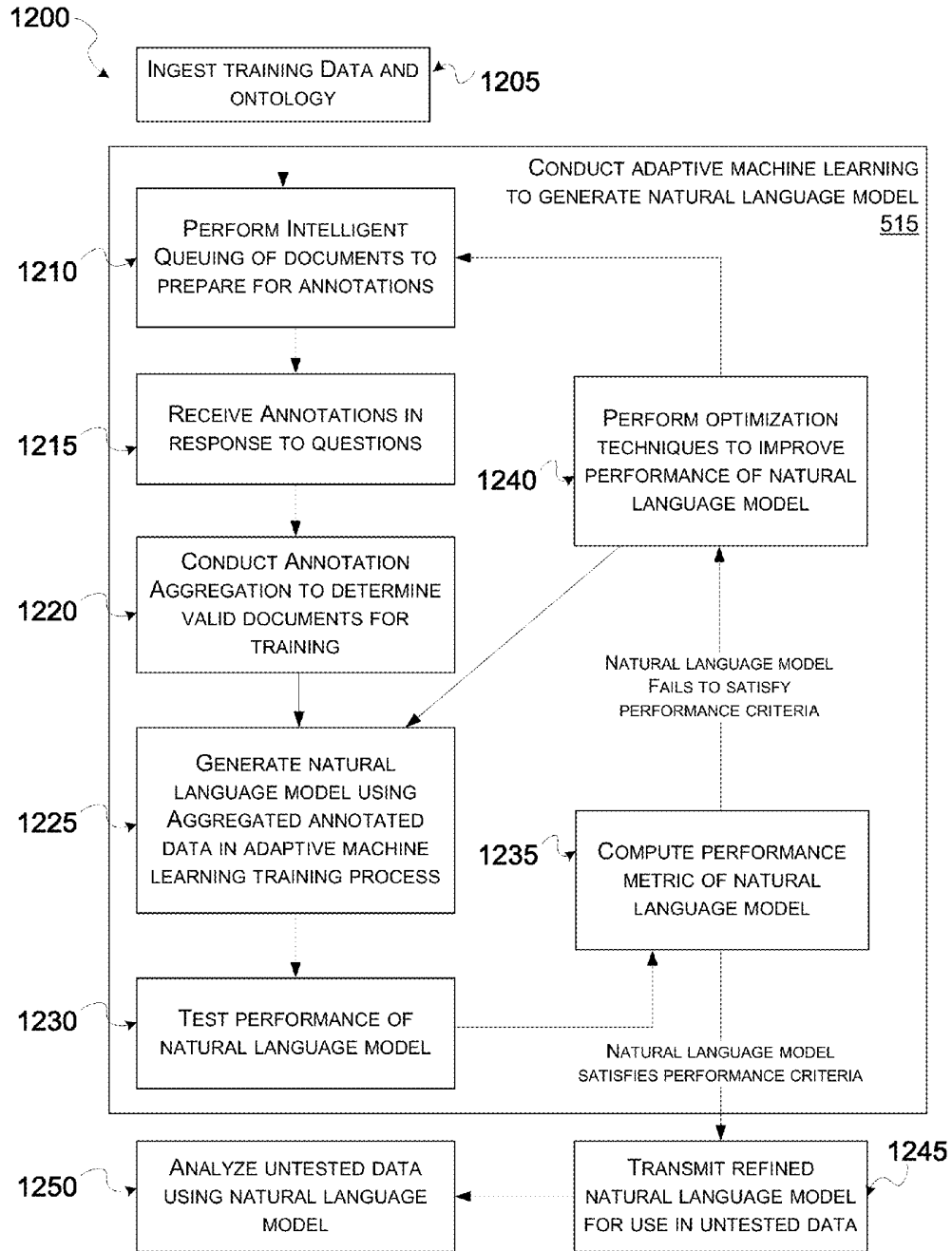


FIG. 12

1300

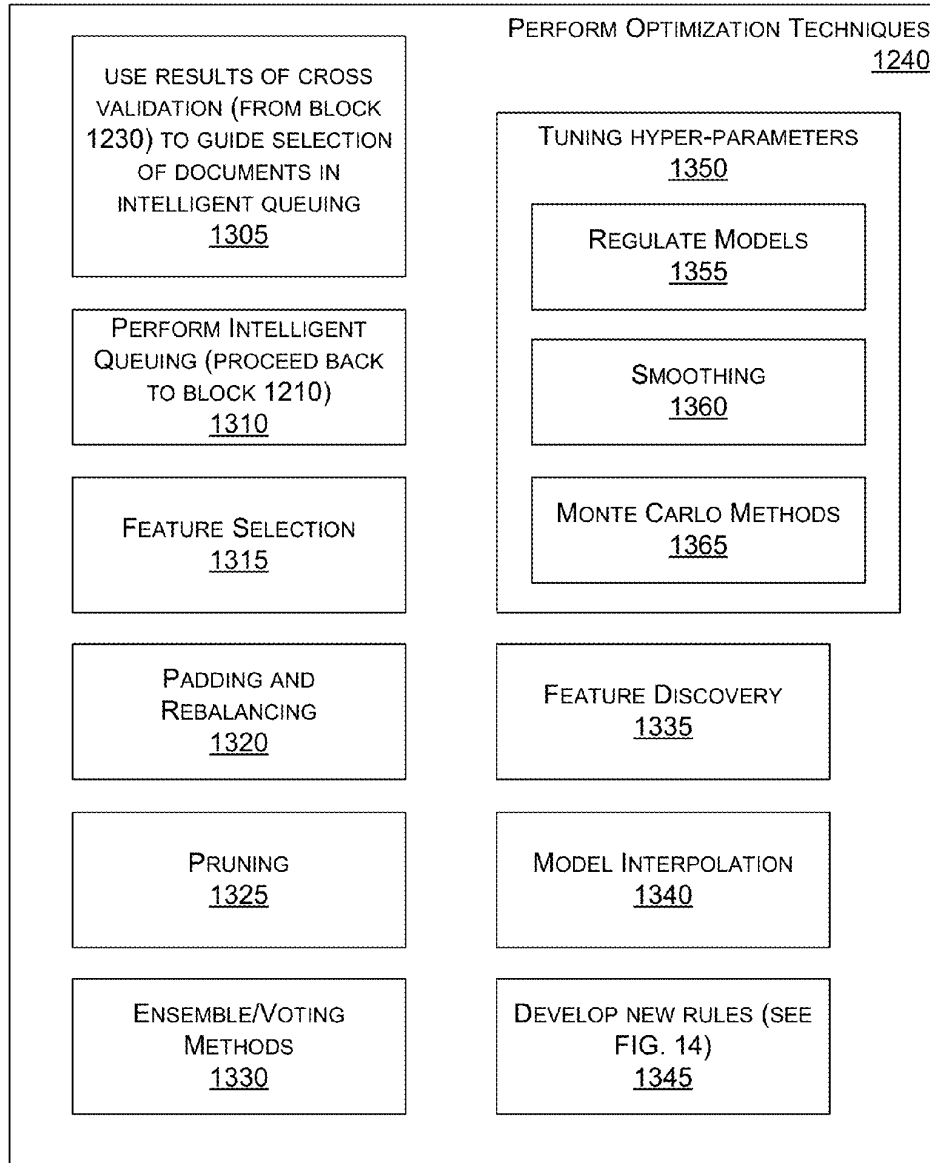


FIG. 13

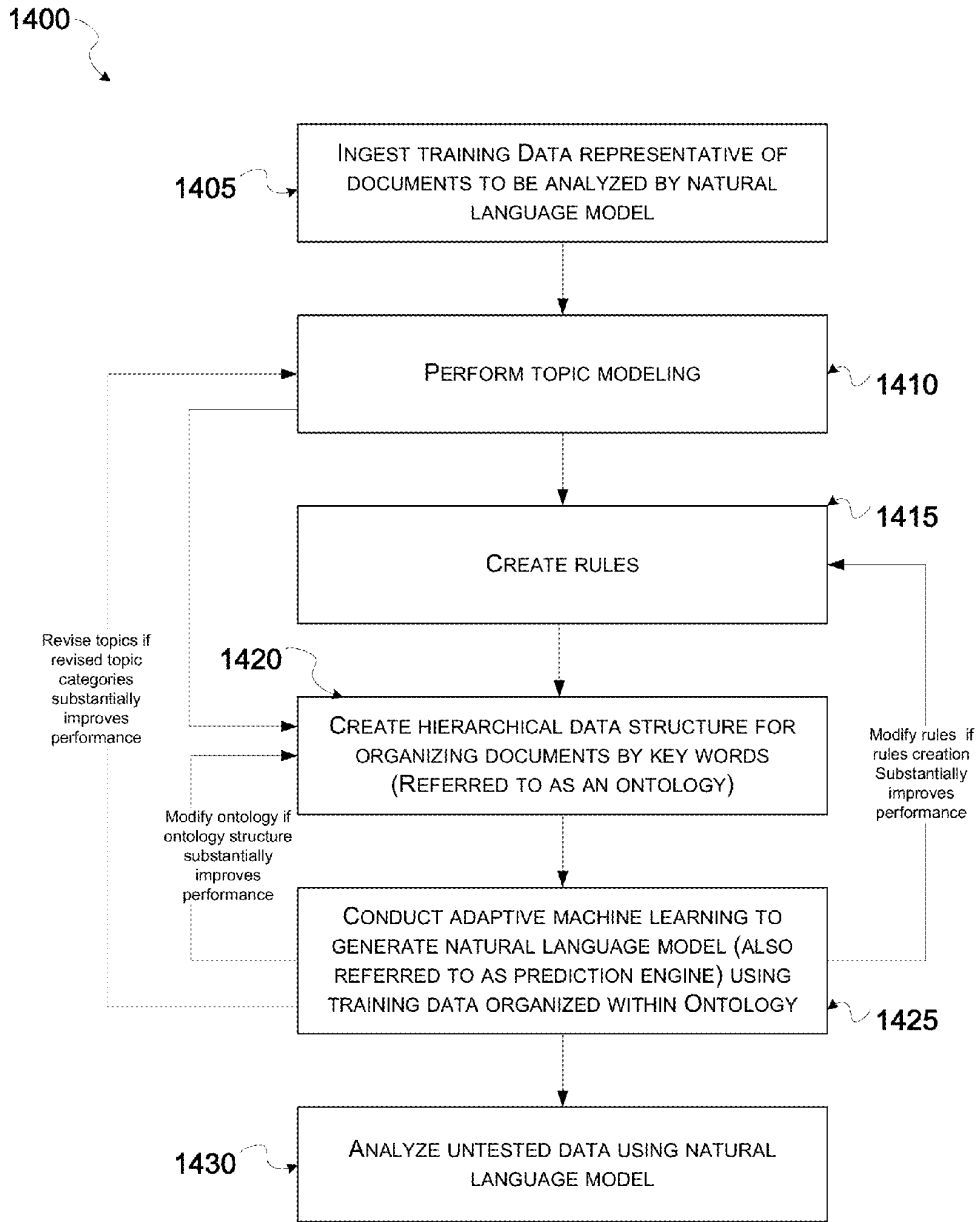


FIG. 14

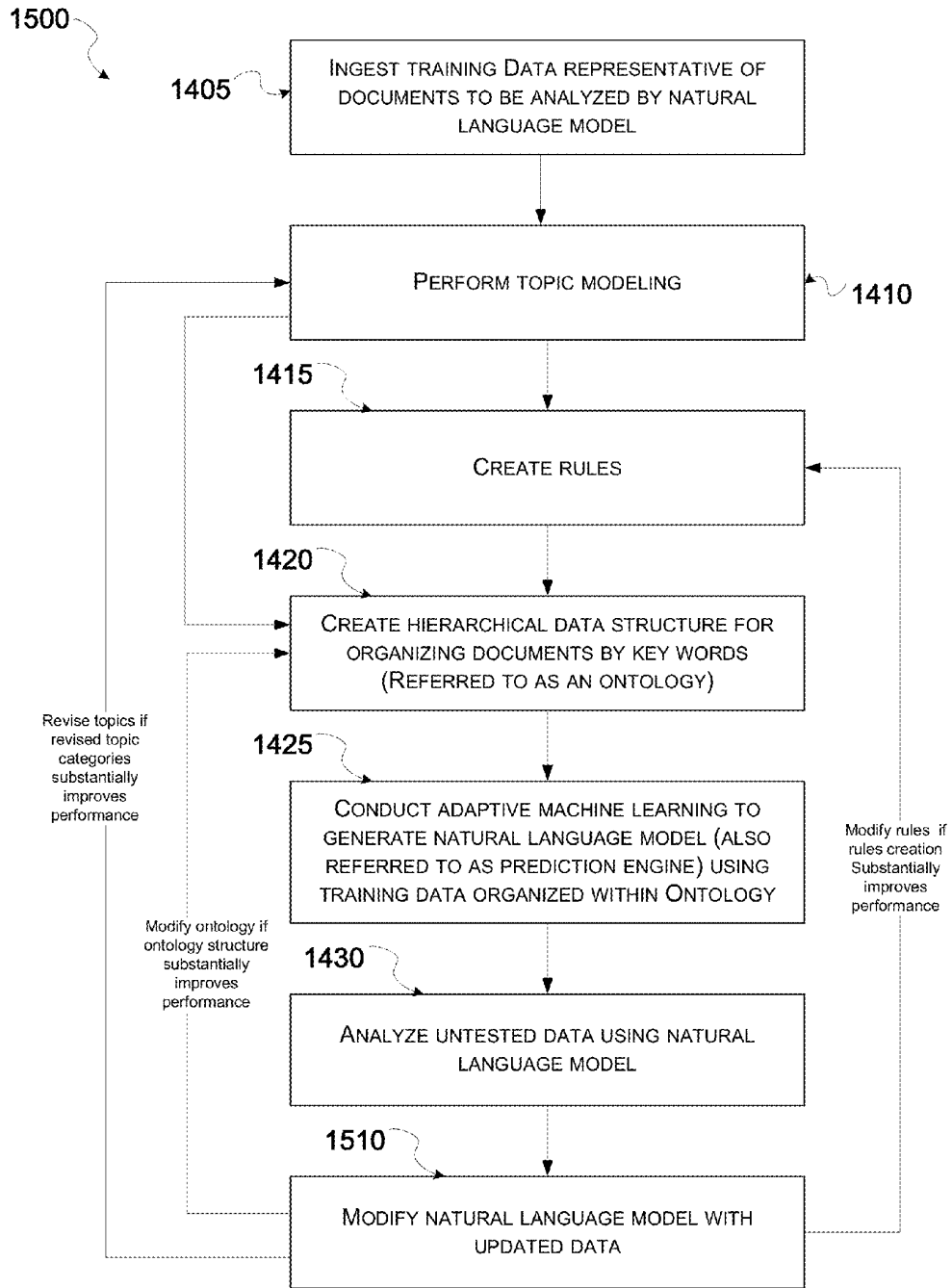


FIG. 15

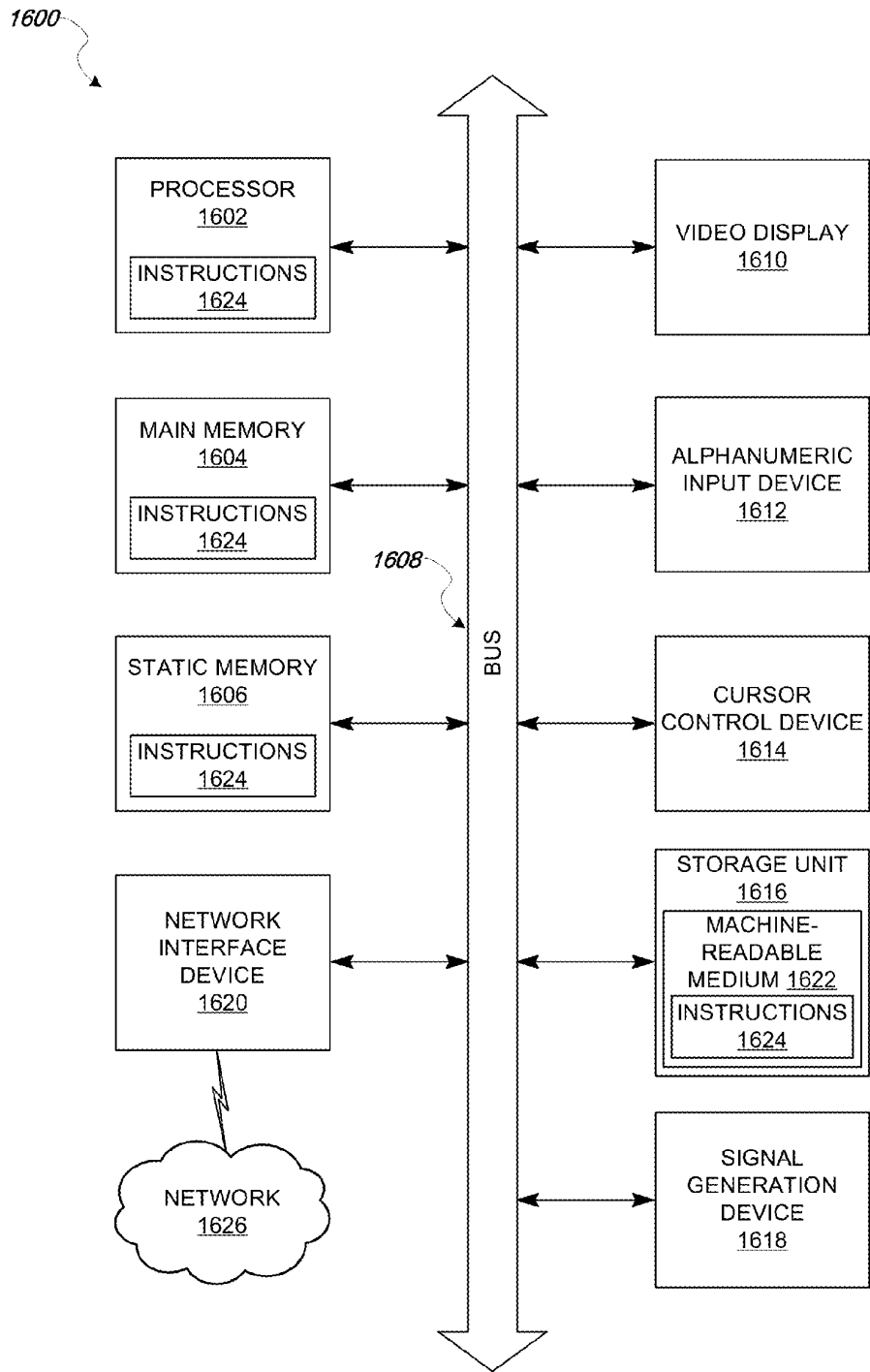


FIG. 16

METHODS FOR GENERATING NATURAL LANGUAGE PROCESSING SYSTEMS

CROSS REFERENCES TO RELATED APPLICATIONS

[0001] This application claims the benefits of U.S. Provisional Application 62/089,736, filed Dec. 9, 2014, and titled, "METHODS AND SYSTEMS FOR ANNOTATING NATURAL LANGUAGE PROCESSING," U.S. Provisional Application 62/089,742, filed Dec. 9, 2014, and titled, "METHODS AND SYSTEMS FOR IMPROVING MACHINE PERFORMANCE IN NATURAL LANGUAGE PROCESSING," U.S. Provisional Application 62/089,745, filed Dec. 9, 2014, and titled, "METHODS AND SYSTEMS FOR IMPROVING FUNCTIONALITY IN NATURAL LANGUAGE PROCESSING," and U.S. Provisional Application 62/089,747, filed Dec. 9, 2014, and titled, "METHODS AND SYSTEMS FOR SUPPORTING NATURAL LANGUAGE PROCESSING," the disclosures of which are incorporated herein by reference in their entireties and for all purposes.

[0002] This application is also related to US non provisional applications (Attorney Docket No. 1402805.00007_IDB007), titled "ARCHITECTURES FOR NATURAL LANGUAGE PROCESSING," (Attorney Docket No. 1402805.00012_IDB012), titled "OPTIMIZATION TECHNIQUES FOR ARTIFICIAL INTELLIGENCE," (Attorney Docket No. 1402805.00013_IDB013), titled "GRAPHICAL SYSTEMS AND METHODS FOR HUMAN-IN-THE-LOOP MACHINE INTELLIGENCE," (Attorney Docket No. 1402805.00014_IDB014), titled "METHODS AND SYSTEMS FOR IMPROVING MACHINE LEARNING PERFORMANCE," (Attorney Docket No. 1402805.00015_IDB015), titled "METHODS AND SYSTEMS FOR MODELING COMPLEX TAXONOMIES WITH NATURAL LANGUAGE UNDERSTANDING," (Attorney Docket No. 1402805.00016_IDB016), titled "AN INTELLIGENT SYSTEM THAT DYNAMICALLY IMPROVES ITS KNOWLEDGE AND CODE-BASE FOR NATURAL LANGUAGE UNDERSTANDING," (Attorney Docket No. 1402805.00017_IDB017), titled "METHODS AND SYSTEMS FOR LANGUAGE-AGNOSTIC MACHINE LEARNING IN NATURAL LANGUAGE PROCESSING USING FEATURE EXTRACTION," (Attorney Docket No. 1402805.00018_IDB018), titled "METHODS AND SYSTEMS FOR PROVIDING UNIVERSAL PORTABILITY IN MACHINE LEARNING," and (Attorney Docket No. 1402805.00019_IDB019), titled "TECHNIQUES FOR COMBINING HUMAN AND MACHINE LEARNING IN NATURAL LANGUAGE PROCESSING," each of which are filed concurrently herewith, and the entire contents and substance of all of which are hereby incorporated in total by reference in their entireties and for all purposes.

TECHNICAL FIELD

[0003] The subject matter disclosed herein generally relates to processing data. In some example embodiments, the present disclosures relate to methods for generating natural language models.

BACKGROUND

[0004] It has long been a goal to program machines to process human-readable language, sometimes in part as an effort to generate artificial intelligence. However, program-

ming computers to process human-readable language has proven to be far more difficult than imagined, particularly as languages continue to change and evolve, and the meaning of words and phrases are more ambiguous and nuanced than assumed. A number of techniques are available for processing natural language by computers, but the methods for generating these models either are inaccurate and imprecise or require months of refinement and programming to accurately model specific subject areas of language. It is desirable, therefore, to develop improved methods for generating natural language models that are accurate and quick while also reducing human time spent generating the models.

BRIEF SUMMARY

[0005] In some embodiments, a method for generating a natural language model is presented. The method may include: ingesting, by a natural language platform comprising at least one processor coupled to at least one memory, training data representative of documents to be analyzed by the natural language model; generating, by the natural language platform and based on topical content within the training data, a hierarchical data structure, the hierarchical data structure comprising at least two topical nodes, wherein at least two topical nodes represent partitions organized by two or more topical themes among the topical content of the training data within which the training data is to be subdivided into; selecting among the training data, by the natural language platform, a plurality of documents to be annotated; generating, by the natural language platform, at least one annotation prompt for each document among the plurality of documents to be annotated, said annotation prompt configured to elicit an annotation about said document indicating which node among the at least two topical nodes of the hierarchical data structure said document is to be classified into; causing display of, by the natural language platform, at least one annotation prompt for each document among the plurality of documents to be annotated; receiving, by the natural language platform, for each document among the plurality of documents to be annotated, the annotation in response to the displayed annotation prompt; and generating, by the natural language platform, the natural language model using an adaptive machine learning process configured to determine, among the received annotations, patterns for how the documents in the training data are to be subdivided according to the at least two topical nodes of the hierarchical data structure.

[0006] In some embodiments, the method further includes testing, by the natural language platform, performance of the natural language model using a subset of the documents among the training data that received annotations.

[0007] In some embodiments, the method further includes: computing, by the natural language platform, a performance metric of the natural language model, based on results of the testing; and determining whether the natural language model satisfies at least one performance criterion based on the computed performance metric.

[0008] In some embodiments, the method further includes performing, by the natural language platform, one or more optimization techniques configured to improve performance of the natural language platform, in response to determining that the natural language platform fails to satisfy the at least one performance criterion based on the computed performance metric.

[0009] In some embodiments of the method, the one or more optimization techniques comprises at least one of: a

