

CLAIMS

1. A method of enhancing an output of relevant content to facilitate analysis of the relevant content, comprising:

receiving ~~at least one data feed of content to be analyzed~~ content from data feeds;
optimizing the content to be analyzed to generate optimized content, wherein the optimizing includes parsing the content to be analyzed to extract a text portion of the content to be analyzed for further analysis, while removing extraneous information from the content to be analyzed, wherein the extraneous information comprises advertisements;

filtering out non-relevant content from the optimized content and retaining relevant content, the non-relevant content including content unrelated to a subject associated with the analysis, ~~the filtering including retaining relevant content wherein~~:

filtering out non-relevant content from the optimized content further comprises filtering out duplicate content from the optimized content,

in response to determining that an instance of the content is unavailable,
retrieving the duplicate content via one or more links to the duplicate content, and
the content to be analyzed includes one or more of a news article, a blog, a social media post, and a long form article;

extracting entity data corresponding to entities mentioned in the relevant content, wherein the ~~extracted~~-entity data ~~corresponding to the entities~~ includes entity metadata associated with the entities, and wherein extracting the entity data comprises:

determining that a first entity of the entities is related to a second entity of the entities, and

linking the first entity to the second entity;

activity-mining the relevant content based on the ~~extracted~~-entity data and the entity metadata associated with the entities ~~to associate the entities with activity tags~~ with activity tags corresponding to activities associated with the entities and extracted from the relevant content;

matching the entities to entity profiles in a database to generate matched entity profiles,
the matching based on the entity metadata and profile metadata of the entity profiles in the database;

~~identifying entity profiles in a database that match the entities, the identifying based on metadata associated with the entities and metadata of the entity profiles in the database; and~~

—annotating the relevant content with at least one graphical indicator and at least one graphical user interface (GUI) control based at least in part on one or more of the entities, the matched entity profiles, and the associated activity tags; and

rendering a GUI having at least a first area configured to display a list of the relevant content and a second area configured to display an enhanced output, wherein the enhanced output comprises an item selected from the list of relevant content, wherein original text of the item includes highlighting overlaid on the one or more of the entities to accentuate the one or more of the entities, each entity of the one or more of the entities highlighted with a distinct color to differentiate each entity of the one or more of the entities.

2. The method of claim 1, wherein:

the ~~at least one~~ data feeds s includes at least one Rich Site Summary (RSS) feed, ~~and wherein the content to be analyzed includes one or more of a news article, a blog, a social media post, a long form article, and an advertisement.~~

3. The method of claim 1, wherein the determining that the first entity is related to the second entity comprises calculating a confidence score, wherein the confidence score comprises a relation similarity score and a name similarity score ~~further comprising optimizing the content to be analyzed, wherein the optimizing includes parsing the content to extract portions of text from the content and removing extraneous information, the extraneous information including one of advertisements and external links.~~

4. The method of claim 1, further comprising:

determining a number of sources that correspond to the duplicate content based on a number of links pointing to the duplicate content ~~de-duplicating the content, the de-duplicating including identifying content elements previously analyzed, or content elements already referenced in the database, and discarding the identified content elements.~~

5. The method of claim 1, wherein the filtering out the non-relevant content includes ~~applying one or more of: a keyword filter and a natural expression pattern matching filter to the content to be analyzed~~ a keyword filter and a natural expression pattern matching filter to the

content to be analyzed, and wherein the rendering the GUI further comprises rendering the GUI having a third area configured to display the entity profiles extracted from the database.

6. The method of claim 1, wherein the filtering out the non-relevant content further includes clustering content elements based on a similarity of the content elements to each other, wherein the clustering comprises:

generating a plurality of vectors, wherein each vector of the plurality of vectors corresponds to a content element and includes values representing a relative frequency of each word in the content element; and

comparing each vector of the plurality of vectors to determine a similarity of content elements to each other.

7. The method of claim 6, wherein the clustering further includes applying a Term Frequency-Inverse Document Frequency (TF-IDF) algorithm that includes determining a frequency of a term within a content element, and determining a frequency of the term across all content elements of the content to be analyzed.

8. The method of claim 1, wherein the extracting the entity data from the relevant content includes:

identifying entities mentioned in a content element, and obtaining the metadata associated with the mentioned entities; and

obtaining the entity metadata associated with the mentioned entities.

9. The method of claim 8, wherein the identifying the entities mentioned in the content element includes applying a natural expression language model to the relevant content, and wherein a type of the entity mentioned in the item corresponds to whether the entity is a person, a country, or a location.

10. The method of claim 8, wherein the metadata associated with a particular entity includes ~~at least one of: an identification of the content element in which the particular entity is mentioned, a count of mentions of the particular entity in a location of the mention of the particular entity within~~ the content element, ~~a mentioned form of the particular entity, a type of~~

~~the particular entity, a normalized form of the particular entity, a location of the mention of the particular entity within the content element, a complete sentence in which the mention of the particular entity appears, and a label for a part of speech in which the mention of the particular entity appears within the content element~~ the method further comprising aggregating the one or more data feeds of the data feeds by dynamically generating a list of the one or more data feeds likely to include content to be of interest, wherein the list is generated based on metadata included in the content.

11. The method of claim 1, wherein the ~~identifying entity profiles in a database that match the extracted entities matching the entities to entity profiles in the database to generate the matched entity profiles~~ includes:

identifying candidate profiles in the database based on a coarse similarity estimate of the candidate profiles to the extracted entities;

calculating a refined candidate score for each candidate profile, the refined candidate score based on the entity metadata ~~of the extracted entities~~ and the profile metadata of each ~~candidate entity~~ profile;

comparing the refined candidate score of each candidate profile with a threshold; and

designating candidate profiles as a match to a particular extracted entity when the refined candidate score of the candidate profiles exceeds the threshold.

~~12. The method of claim 1, wherein the annotating the relevant content includes overlaying the at least one graphical indicator over each extracted entity, wherein the at least one graphical indicator over each extracted entity is of a different color.~~

~~13~~12. The method of claim 1, wherein the at least one graphical indicator includes an indicator indicating that a particular extracted entity is unmatched with a profile entity in the database, and wherein the at least one GUI control includes a GUI control for executing an update to the database to store the particular extracted entity in the database.

~~14~~13. A system for enhancing an output of relevant articles to facilitate analysis of the relevant articles, comprising:

at least one data feed for receiving one or more articles to be analyzed;

a server comprising a memory and a processor, wherein the memory includes a classifier, the classifier configured to:

optimize the articles to be analyzed to generate optimized content, wherein the optimizing includes parsing the articles to be analyzed to extract a text portion of the articles to be analyzed for further analysis, while removing extraneous information from the articles to be analyzed, wherein the extraneous information comprises advertisements;
filter out non-relevant articles from the ~~optimized content~~~~one or more articles~~, the non-relevant articles including articles unrelated to a subject associated with the analysis, wherein relevant articles are retained, wherein the classifier configured to filter out the non-relevant articles from the optimized content further comprises the classifier configured to:

filter out duplicate content from the optimized content,
in response to a determination that an instance of the content is unavailable,
retrieve the duplicate content via one or more links to the duplicate content, and
the content to be analyzed includes one or more of a news article, a blog, a social media post, and a long form article;

extract entity data corresponding to entities mentioned in the relevant articles, wherein the extracted entity data corresponding to the entities includes entity metadata associated with the entities, and wherein the classifier configured to extract the entity data comprises the classifier further configured to:

generate a confidence score indicative of a correctness of a posited link between a first entity of the entities and a second entity of the entities, wherein the confidence score comprises a relation similarity score and a name similarity score,

determine that a first entity of the entities is related to a second entity of the entities, and

link the first entity to the second entity;

mine the relevant articles for activities based on the ~~extracted~~-entity data and the entity metadata ~~associated with the entities~~ to associate the entities with activity tags corresponding to activities associated with the entities and extracted from the relevant content;

~~identify entity profiles in a database that match the entities, wherein the matching entity profiles are identified based on metadata associated with the entities and metadata of the entity profiles in the database; and~~

~~annotate the relevant articles with at least one graphical indicator and at least one~~

~~graphical user interface (GUI) control based at least in part on one or more of the entities, the matched entity profiles, and the associated activity tags;~~

~~match the entities to entity profiles in a database to generate matched entity profiles, wherein the matched entity profiles are identified based on the entity metadata and profile metadata of the entity profiles in the database;~~

~~at least one user terminal with a graphical user interface configured to display the annotated relevant articles enhanced output, wherein the enhanced output comprises an article selected from a list of relevant articles, wherein original text of the article includes highlighting overlaid on one or more of the entities to accentuate the one or more of the entities, each entity of the one or more of the entities highlighted with a distinct color to differentiate each entity of the one or more of the entities; and~~

~~the database configured to store the extracted entity data for the entities and the correspondingly associated activity tags.~~

~~1514. The system of claim 1413, wherein the at least one data feed includes at least one Rich Site Summary (RSS) feed, and wherein the articles to be analyzed includes one or more of a news article, a blog, a social media post, a long form article, and an advertisement.~~

~~1615. The system of claim 1413, wherein the server is further configured to optimize the articles to be analyzed, wherein the configuration of the server to optimize includes configuration of the server to parse the articles to extract portions of text from the articles and remove extraneous information, the extraneous information further comprises including one of advertisements and external links.~~

~~17. The system of claim 14, wherein the configuration of the server to filter includes configuration of the server to:~~

~~—— apply one or more of: a keyword filter and a natural expression pattern matching filter to the articles to be analyzed, and~~

~~—— cluster articles based on a similarity of the articles to each other.~~

~~1816. The system of claim 1413, wherein the classifier configured to match the entities to the entity profiles in the database includes the classifier further configured to configuration of~~

~~the server to identify entity profiles in a database that match the extracted entities includes configuration of the server to:~~

identify candidate profiles in the database based on a coarse similarity estimate of the candidate profiles to the extracted entities;

calculate a refined candidate score for each candidate profile, the refined candidate score based on the metadata of the extracted entities and the profile metadata of each ~~candidate entity~~ profile;

compare the refined candidate score of each candidate profile with a threshold; and

designate candidate profiles as a match to a particular extracted entity when the refined candidate score of the candidate profiles exceeds the threshold.

~~19. — The system of claim 14, wherein the configuration of the server to annotate the relevant article includes configuration of the server to overlay the at least one graphical indicator over each extracted entity, wherein the at least one graphical indicator includes an indicator indicating that a particular extracted entity is unmatched with a profile entity in the database, and wherein the at least one GUI control includes a GUI control for executing an update to the database to store the particular extracted entity in the database.~~

2017. A computer-based tool for facilitating analysis of relevant articles, the computer-based tool including non-transitory computer readable media having stored thereon computer code which, when executed by a processor, causes a computing device to perform operations comprising:

receiving an enhanced output from a server of a research system, the enhanced output including relevant articles overlaid with one or more graphical indicators and at least one graphical user interface (GUI) control, ~~wherein the one or more graphical indicators are overlaid over at least one of: entities mentioned in the article and activities associated with the entities,~~ wherein the enhanced output comprises an article selected from a list of relevant articles, wherein original text of the article includes highlighting overlaid on one or more entities to accentuate the one or more entities, each entity of the one or more entities highlighted with a distinct color to differentiate each entity of the one or more entities and is generated by the server based on:

receiving at least one data feed of articles to be analyzed;

optimizing the articles to be analyzed to generate optimized content, wherein the optimizing includes parsing the articles to be analyzed to extract a text portion of the articles to be analyzed for further analysis, while removing extraneous information from the articles to be analyzed, wherein the extraneous information comprises advertisements;
 filtering out non-relevant articles from the optimized content, the filtering including retaining relevant content, wherein;~~to identify the relevant articles;~~
filtering out non-relevant content from the optimized content further comprises filtering out duplicate content from the optimized content,
in response to determining that an instance of the content is unavailable,
retrieving the duplicate content via one or more links to the duplicate content, and
the content to be analyzed includes one or more of a news article, a blog, a social media post, and a long form article;
 extracting entity data corresponding to the one or more entities from the relevant articles,
wherein the extracting the entity data comprises:
generating a confidence score indicative of a correctness of a posited link between a first entity of the entities and a second entity of the entities, wherein the confidence score comprises a relation similarity score and a name similarity score,
determining that a first entity of the entities is related to a second entity of the entities, and
linking the first entity to the second entity;
 mining the relevant articles based on the extracted entity data and entity metadata associated with the one or more entities to associate the entities with the activity tags corresponding to activities associated with the one or more entities and extracted from the relevant content;
~~identifying entity profiles in a database that match the entities, the identifying based on the metadata associated with the entities and metadata of the entity profiles in the database; and~~
matching the one or more entities to entity profiles in a database to generated matched entity profiles, the matching based on the entity metadata and profile metadata of the entity profiles in the database;
 annotating the relevant articles with the one or more graphical indicator and the at least one GUI control based at least in part on one or more of the entities, the matched entity profiles,

and the associated activity tags; and
displaying the enhanced output on a GUI.

18. The computer-based tool of claim 17, wherein the optimizing further comprises extracting non-textual elements from the articles.

19. The computer-based tool of claim 17, wherein the filtering out the non-relevant content comprises:

training a language model;
providing the articles to the language model; and
identifying the relevant articles from the articles using the language model based on a context in which a keyword is used in the articles.

20. The method of claim 3, wherein the relation similarity score is calculated using a MinHash algorithm, and wherein the name similarity score is calculated using a Levenshtein distance algorithm.