Rosette Text Analytics Release Notes

May 2019

Basis Technology Corp.
One Alewife Center Cambridge
MA 02140
T 617.386.2000
F 617.386.2020
E support@rosette.com

Release Notes

Patch Release - 1.13.1

May 2019

Rosette Enterprise On-Premise Users Only

To minimize the size of your Rosette Enterprise installation, the entity extraction (rex-root) and semantic similarity (tvec-root) components are now shipped by language. The name of the language specific files contain the three letter ISO-639 language code, indicating which language is supported by the file.

Entity extraction is shipped with one base file and one or more language-specific files.

Example:

- `rex-root-<version>.tar.gz`
- `rex-root-<version>-eng.tar.gz` for English language files
- `rex-root-<version>-deu.tar.gz` for German language files

Semantic Similarity is shipped with one file per language.

Example:

- `tvec-root-<version>-eng.tar.gz` for English language files
- `tvec-root-<version>-deu.tar.gz` for German language files

The Rosette Enterprise installer has been updated and will automatically install all components as required, based on your license.

NEW Release - 1.13.0

April 2019

Rosette Platform Changes

Categorization /categories

- **Expanded English support:** We can now process English with a case-insensitive model by providing (uen) as the input language.

Entity Extraction and Linking /entities

- **Bug fix:** We've fixed a problem in Japanese where entity names that include the middle dot were not being handled correctly. Entity names that include a middle dot are no longer split into two entities.
- **Bug fix:** We've fixed a problem in Japanese entity linking where, in some cases, Japanese characters were being replaced with Chinese characters.
- **Bug fix:** We've fixed a problem where, in some cases, entities were mislabeled when the `includeDbPediaTypes` option was not flagged.

Morphological Analysis /morphology/{morphoFeature}

- **New Hebrew disambiguator:** We've added a new default disambiguator (`perceptron`) for Hebrew. Use the `disambiguatorType` option to enable another disambiguator (`DNN` or `dictionary`). For example, to return to the previous default, add `{"options": {"disambiguatorType": "DNN"}}` to your call.
- **Bug fix:** We've fixed an error where some Chinese tokens included spurious white space characters.
- **Bug fix:** We've fixed a problem where extremely long tokens (thousands of characters) would slow down the tokenizer.
- **Bug fix:** We've fixed a problem where Polish tokens that can appear in multiword expressions were lemmatized to full expressions, even when the full expression wasn't in the input.
  - ***Previously:***`dzień` lemmatized to `dzień_dobry`.
  - ***Now:***`dzień` is lemmatized to `dzień`.
- **Bug fix:** We've fixed a problem where the non-final components of Russian compound words with more than one hyphen were not lemmatized correctly.

Name Similarity /name-similarity

- **Hungarian:** We've added a Hungarian frequency language model for LOCATION entity type and retrained the language model for PERSON entity type, improving Hungarian name matching.
  PERSON:
  - ***Previously:*** The similarity score for Domonkos Gyula Tiborné and Domonkos Gy. Lászlóné was 0.697.
  - ***Now:*** The similarity score for Domonkos Gyula Tiborné and Domonkos Gy. Lászlóné is 0.775.

  LOCATION:
  - ***Previously:*** The similarity score for Szentmihály puszta and Szentmihály pihenő was 0.576.
  - ***Now:*** The similarity score for Szentmihály puszta and Szentmihály pihenő is 0.696.

Sentiment Analysis /sentiment

- **Expanded English support:** We can now process English with a case-insensitive model by providing (`uen`) as the input language. Be aware that when using `DNN` for the `modelType` the accuracy of the results may be lower than when analyzing standard, sentence-cased, English input.

Rosette Enterprise On-Premise Users Only

Sentiment Analysis /sentiment

- **Document-only analysis:** Entity-level sentiment analysis can now be turned off, allowing document-level sentiment analysis only. See the *Rosette Enterprise User Guide* for more information.
- **Expanded Language support:** Users of the Rosette Text Classification Field Training Kit can now train custom sentiment analysis models in any language supported by the tokenization endpoint (for document-level analysis) or the entity extraction and linking endpoint (for entity-level analysis). For more information on the training and configuration procedure, see the *Rosette Field Training Kit* documentation.

Open Source Changes

- **Version Changes**
  Apache Lucene Core updated from v6.6.0_1 to v7.6.0_1 (Apache 2.0 license)

Patch Release 1.12.2

February 25, 2019

Entity Extraction and Linking /entities

- **Bug fix:** We've fixed a bug where, under some circumstances, the head mention for an extracted entity was not correctly identified.

Morphological Analysis /morphology/{morphoFeature}

- **Bug fix:** We've fixed a bug where the Persian lemmatizer did not add lemmas to the first analyses of many tokens, especially verbs.
- **Bug fix:** We've fixed a bug where after some sequences of 4096 characters, containing mostly white space and at most one token, any following tokens had incorrect original offsets.

Semantic Similarity /semantics/{semanticsFeature}

- **Bug fix:** We've fixed a bug where capitalized tokens could return an out-of-vocabulary token-level embedding instead of the embedding consistent with their lowercase form.
- **Bug fix:** Previously, the Semantic Vectors endpoint did not always return vectors of a consistent length. Now, returned vectors will always be normalized to have a length of one.

Rosette Enterprise On-Premise Users Only

Morphological Analysis /morphology/{morphoFeature}

- **Bug fix:** We've fixed a bug where when the `fstTokenize` option was enabled, the lemmas of hyphenated Russian compound words only had the the final piece lemmatized. Now both pieces are lemmatized.
  - ***Previously*:** человека-волка was lemmatized to человека-волк
  - ***Now*:** человека-волка is lemmatized to человек-волк

Categorization /categories

- **New factory configuration option:** We've added a new factory configuration option, `maxResults`. This option can be used to cap the number of results returned. By default, all results exceeding the score and confidence thresholds (if set) will be returned.

Open Source Changes

Deleted from Rosette Enterprise Embedded

- Apache Ant v1.5
- Apache Log4j v2.7
- Checker Qual v2.5.2
- Error Prone v2.1.3
- J2ObjC v1.1
- JSON in Java v20141113
- JSR 305: Annotations for Software Defect Detection v3.0.2
- Mojohaus Animal Sniffer Annotations v1.14

Deleted from Rosette Enterprise Restful

- JSON in Java v20141113

NEW Release - 1.12.1

January 31, 2019

Semantic Similarity **/semantics/{semanticsFeature}** (LABS)

Note that the Semantics Similarity features are still in LABS and subject to change. Send us your feedback!

- **New endpoint Similar Terms:** We've added a new endpoint, /semantics/similar, which uses text vectors to generate multilingual related terms with numerical similarity scores for any input word(s) in Arabic, English, Chinese, German, Japanese, North or South Korean, Russian, or Spanish. For more information, see the Features and Functions.
  - Input Term: spy returns

```
Spanish
{"term":"espía","similarity":0.61295485},
{"term":"cia","similarity":0.46201307},
{"term":"desertor","similarity":0.42849663},
{"term":"cómplice","similarity":0.36646274},
{"term":"subrepticiamente","similarity":0.36629659}
German
{"term":"Deckname","similarity":0.51391315},
{"term":"GRU","similarity":0.50809389},
{"term":"Spion","similarity":0.50051737},
{"term":"KGB","similarity":0.49981388},
{"term":"Informant","similarity":0.48774603},
Japanese
{"term":"スパイ","similarity":0.5544399},
{"term":"諜報","similarity":0.46903181},
{"term":"MI6","similarity":0.46344957},
{"term":"殺し屋","similarity":0.41098994},
{"term":"正体","similarity":0.40109193},
```

- **Semantic Vectors:** The /text-embedding endpoint has been renamed to /semantics/vector. While the /text-embedding endpoint will remain accessible through April, we encourage you to migrate as soon as possible to avoid missing any important updates.

Entity Extraction and Linking **/entities**

- **Improved linking confidence:** We've updated the linking confidence calculation and thresholds to improve accuracy.
- **Supported languages:** We've removed xxx from the list of languages returned when using /entities/supported-languages.
- **Bug fix:** We've fixed a bug where the salience score was not always returned for entities with pronominal mentions, when requested.
- **Bug fix:** We've fixed a bug where some later entity mentions that were chained to the first mention of a given entity were not always properly returned.
- **Bug fix:** We've fixed a bug where sometimes a null pointer exception was returned when resolving pronouns.

Morphological Analysis **/morphology/{morphoFeature}**

- **Bug fix:** We've fixed a bug where some English words were automatically getting tagged as proper nouns when capitalized.

- **Bug fix:** We've fixed a bug in English where "people" was not being properly lemmatized to "person". It now has the lemma candidate "person" when appropriate.
- **Bug fix:** We've fixed a bug where ordinal numbers and comparative adjectives in English like "second" and "lower" were analyzed as verbs.

Name Similarity **/name-similarity**

- **Hungarian:** We've added support for multi-letter initials in Hungarian, improving Hungarian name matching.
  - ***Previously***: The similarity score for Kovács Cs. István and Kovács Csaba István was 0.68.
  - ***Now***: The similarity score for Kovács Cs. István and Kovács Csaba István is 0.91.
- **Japanese:** We've improved the accuracy of matching between katakana and kanji versions of Japanese organization names.
  - ***Previously***: The similarity score for ドクリツギョウセイホウジンニホンガクジュツシンコウカイ and 独立行政法人日本学術振興会 was 0.41.
  - ***Now***: The similarity score for ドクリツギョウセイホウジンニホンガクジュツシンコウカイ and 独立行政法人日本学術振興会 score is 0.72.
- **Chinese:** We've improved the accuracy of matching Chinese organization names.
  - ***Previously***: The similarity score for 松下能源(上海)有限公司 and Panasonic Energy (Shanghai) Co., Ltd was 0.58.
  - ***Now***: The similarity score for 松下能源(上海)有限公司 and Panasonic Energy (Shanghai) Co., Ltd is 0.82.

Rosette Enterprise On-Premise Users Only

- We've decreased warm-up time by only loading licensed languages when Rosette is set to pre-warm.
- We've added the ability for on-premise users to have more control over configuration options when using custom-trained models.
- We've reduced the minimum memory requirements to 16GB of RAM for the entity extraction and linking, sentiment analysis, and topic extraction endpoints.
- We've improved the efficiency of the initial load time for the entity extraction and linking endpoint.

Client Bindings

- All client bindings have been updated to support the /semantics/vector and /semantics/similar endpoints.
- The examples have been modified in the Python bindings to demonstrate how to set options.

Open Source Changes

- **New Addition**
  Spotify Annoy Java v0.2.5 - Apache 2.0
- **Version Changes**
  Basis Technology Annotated Data Model v2.5.2 - Apache 2.0

NEW Release - 1.12.0

December 10, 2018

Entity Extraction and Linking **/entities**

- **Korean:** We've improved the accuracy of Korean extraction, largely through better handling of Josa (postpositions) and compound words.
- **Entity Linking:** We've added support for entity linking to Wikipedia for both the top level types (PERSON, LOCATION, ORGANIZATION, ETC.) as well as the over 700 DBpedia types (see full list here) in the remaining 16 languages supported by entity extraction. This is in addition to the languages currently supported by entity linking: Chinese, English, Japanese, and Spanish.

Morphological Analysis **/morphology/{morphoFeature}**

- **Hebrew disambiguation:** We've improved analysis in Hebrew by adding disambiguation, a mechanism for more accurately choosing which of several candidate analyses is provided in the response. For Hebrew only, we've added an option `disambiguatorType` to select which disambiguator is used. The values are `DNN` for the TensorFlow-based deep neural network model and `dictionary` for the dictionary-based model. The default is `dictionary`. To enable the `DNN` disambiguator, add `{"options": {"disambiguator": "DNN"}}` to your call.
- **Persian lemmatization:** We've added lemmatization support to Persian.

Name Similarity **/name-similarity**

- **Chinese organization names:** We've improved the accuracy of matching between Chinese and English organization names.

> ***Previously*:** The similarity score for 索尔维 - 恒昌（张家港）精细化工有限公司 and Solvay-Hengchang (Zhangjiagang) Fine Chemicals Co., Ltd was 0.6929.
>
> ***Now*:** The similarity score for 索尔维 - 恒昌（张家港）精细化工有限公司 and Solvay-Hengchang (Zhangjiagang) Fine Chemicals Co., Ltd score is 0.8225.

Rosette Enterprise On-Premise Users Only

- The minimum system requirements for running Rosette Enterprise have changed for some use cases. This is a result of providing entity linking for 16 additional languages in this release. We now support entity linking for all 20 languages supported by entity extraction.
  The entity extraction and linking, sentiment analysis, and topic extraction endpoints require significant memory allocation. If using these endpoints, the new minimum memory requirements are:
  - 32GB RAM
  - 64GB of disk space (more may be needed for growing logs)

  For all other endpoints, the minimum memory requirements are:
  - 16GB RAM
  - 35GB of disk space (more may be needed for growing logs)
- Rosette Enterprise is now available as a Docker container. The images are available on Docker Hub. A Basis shipment, containing a license file and a docker-compose file customized to your licensed endpoints, is still required.

Patch Release 1.11.3

October 29, 2018

Morphological Analysis **/morphology/{morphoFeatue}**

- **Chinese:** We've added the word "百度" to the Chinese lexicon. This only has an effect when `modelType` is set to `default`, which is its default value.

***Previous*:** Two tokens: “百”, “度”

***Now*:** One token: “百度”

- **German:** The part of speech of the acronyms “MAN” and “MIT” is now NOUN in German, instead of falling back to the parts of speech of the unrelated words “man” and “mit”.

***Previous*:** MAN/INDPRO, MIT/PREP, MIT/VPREF, MIT/ADV

***Now*:** MAN/NOUN, MIT/NOUN

- **Spanish:** We’ve improved Spanish part-of-speech tag and lemma disambiguation.

***Previous*:** 91.646% POS accuracy and 90.243% lemma accuracy using the IULA Spanish LSP Treebank

***Now*:** 91.742% POS accuracy and 90.344% lemma accuracy

Name Similarity **/name-similarity**

- **New Model:** We’ve added a new model for increased accuracy when matching Hungarian names to other Hungarian names.

***Previous*:** Pavlovitch Bryulov vs. Pavlovics Brjullov - score: 0.84

***Now*:** Pavlovitch Bryulov vs. Pavlovics Brjullov - score: 0.95

- **Bug Fix:** Fixed a bug involving duplicate readings produced when transliterating Chinese names.

***Previous*:** Wang Xing vs. 王行 - score: 0.69

***Now*:** Wang Xing vs. 王行 - score: 0.99

Patch Release 1.11.2

September 24, 2018

Entity Extraction and Linking **/entities**

- **Improved Chinese accuracy:** We've replaced the underlying tokenizer to improve accuracy in Chinese.
- **Improved Hungarian accuracy:** We've updated our pattern match extractors in Hungarian. This improves accuracy for the MONEY and DATE types.
- **Known issue:** We've fixed a bug where entity mentions were being miscounted if the `calculateSalience` option was set to `true`.

Morphological Analysis **/morphology/{morphoFeature}**

- **Improved Dutch disambiguation:** We've improved Dutch part-of-speech disambiguation.
- **Improved Japanese and Chinese lemma support:** In the last release, most Japanese and Chinese tokens did not have lemmas when `modelType` was set to `default`. Now, such tokens have lemmas equivalent to their surface forms.

NEW Release - 1.11.0

August 27, 2018

Rosette Enterprise On-Premise Users Only

- **New per-endpoint licensing:** Endpoints are now activated directly from your installed license. The `endpoints.yaml` file has been removed from the installation.
- **New Enterprise User guide:** We've added a user guide (*rosette-enterprise-user-guide-1.11.0.pdf*), that provides new content and replaces the files *rosette-api-on-premise-install-guide-1.11.0.txt*, *overview.md*, and *Rosette_API_Embedded_User_Guide.pdf*.
- **Simplified installation for macOS and Linux:** The installation for both RESTful and embedded Java has been simplified for macOS and Linux. Installation for Windows has not changed, but detailed installation notes are now included as part of the new *Enterprise User Guide*.
- **Name Changes:** We are continuing to consolidate and simplify our branding. Rosette API On-Premise is now Rosette Enterprise. We've made changes to the documentation and license names to reflect.

Rosette Platform Changes

- **New supported languages sub-endpoints:** For all endpoints (excluding Name Similarity, Name Translation, and Name Deduplication), Rosette now provides a `GET /rest/v1/<endpoint>/supported-languages` method that returns that endpoint's supported languages and scripts. See the *Features and Functions* or the *Interactive Docs* for more information.
- **Updated bindings:** We've updated our CSharp and Java bindings. Be sure to get the latest version (1.11.0) to take advantage of all the new features and improvements!

Name Deduplication

- **New language:** Rosette now supports name deduplication of Hungarian names.

Categorization

- **Multilabel categorization:** Rosette can now return multiple category labels per document. For more information, see the *Features and Functions*. To return only a single category label per document, set the `{"options": {"singleLabel": true}}`. For more information, see the *Features and Functions*.

Text Embedding

- **New languages:** The text embedding endpoint now supports Russian, North Korean, South Korean, and Arabic.
- **Individual token embeddings:** We can now return embeddings for individual input tokens. To enable per-token embeddings, add `{"options": {"perToken": true}}` to your call.
- **Response modifications:** We've made changes to the text embeddings endpoint's response structure. Document-level embeddings now have their own dedicated slot `embeddings` and will no longer appear in `documentMetadata`. Please note that this is a breaking change, contact Rosette support for more information.

Entity Extraction and Linking

- **New feature - DBpedia Types (LABS):** We've added over 700 new entity types to the Entity Extraction and Linking endpoint, drawn from the DBpedia ontology. To access these entity types, add `{"options": {"includeDBpediaType" = true}}` to your call. You'll notice more than 10 additional macro types in the `type` field as well as the all new `DBpediaType` field. For more information, see the *Features and Functions*. Note that this feature is still in LABS and subject to change. Send us your thoughts!
- **Better accuracy:** We've improved the recall of Rosette's entity linking across all supported languages.
- **New language:** Entity extraction now supports Hungarian.

- **Known issue:** MONEY, PHONE NUMBER, and URL types are not extracting properly in Hungarian. This will be fixed in the September 2018 patch release.

Language Identification

- **Support for North and South Korean added:** Rosette can now identify North Korean (`qkp`) and South Korean (`qkr`) dialects. To enable the dialects, add `{ "options": { "koreanDialects":true }}` to your call.

Morphological Analysis

- **New algorithm for Chinese and Japanese:** We've added a new algorithm for Chinese and Japanese morphological analysis. Prior to version 1.11.0, the default algorithm was perceptron. To return to the old model, add `{"options": {"modelType": "perceptron"}}` to the body of your call.
- **Norwegian lemmatization:** We've expanded the lemma dictionaries for Norwegian, both Bokmål and Nynorsk.
- **Improved English and Spanish disambiguation** We've improved the accuracy of lemmatization and part of speech tagging in both English and Spanish.
- **Bug fix:** We've improved handling of formatting characters in German.
- **Bug fix:** We've fixed a bug where the Hebrew POS tag `wPrefix` was not converted to UPT-16.

Tokenization

- **New algorithm for Chinese and Japanese:** We've added a new algorithm for Chinese and Japanese tokenization. Prior to version 1.11.0, the default algorithm was perceptron. To return to the old model, add `{"options": {"modelType": "perceptron"}}` to the body of your call.
- **Bug fix:** We've fixed a bug where in Catalan, in which Rosette did not tokenize after an apostrophe in cases where the apostrophe marks a token boundary.
- **Bug fix:** We've fixed a big in Japanese, in which Rosette did not recognize 々 as a Japanese character, so it was considered its own token.

Name Similarity

- **New language:** Name similarity now supports matching between Hungarian and English names.
- **Improved language-of-origin detection:** We've improved the detection of language-of-origin of Japanese names written in Katakana.

Patch Release 1.10.2

June 25, 2018

Name Similarity

- **Improved Arabic script segmentation:** We've improved segmentation of Persian (Dari and Farsi), Pushto, and Urdu names written in Arabic script.

Entity Extraction and Linking

- **New Japanese tokenizer:** We've replaced the underlying tokenizer to improve accuracy for Japanese.

Morphological Analysis

- **Bug fix:** We've fixed a bug where some components of compound German words were incorrect when the surface form of the component could be either a noun or a verb.

- **Bug fix:** We've fixed a bug where the Hebrew parts of speech tags did not use UPT-16, the POS tag set used by the other languages.
- **Bug fix:** We've fixed a bug where returned email addresses and URLs could contain control or whitespace characters.
- **Bug fix:** We've fixed a bug where returned Hebrew tokens could contain control characters or nothing but default ignorable characters.
- **Bug fix:** We've fixed a bug where Chinese, Japanese, and Thai tokens could contain control characters.

Tokenization

- **Bug fix:** We've fixed a bug where the default Japanese tokenizer truncated some katakana tokens when they appeared after non-katakana tokens.

Name Deduplication

- **Name Deduplication input limit:** Each call now has a limit of 1000 names per list.

Patch Release 1.10.1

May 30, 2018

Sentiment Analysis

- **Improved Entity Level Sentiment Analysis:** We've improved our calculation of entity level sentiment to more accurately consider the context around each mention. Please note, this update may cause results to change.

Syntactic Dependencies

- **Bug fix:** We've fixed a bug where the initial token dependency for every sentence (other than the first) was omitted from the list of results.

Entity Extraction and Linking

- **Improved Hebrew Entity Extraction:** We've improved Hebrew entity extraction by removing superfluous prefixes from extracted entities.
- **Improved Confidence Scores:** We've improved statistical model confidence scores to provide a more effective tradeoff between precision and recall. Please note, this update may cause results to change. If you have set a threshold based on entity confidence scores, please evaluate to ensure optimal performance.
- **Improved Entity Normalization:** Social media characters such as "@" and "#" are removed from a `Mentions` normalized string. Offsets to the original string data field remain the same.

Morphological Analysis

- **Bug fix:** Previously, the Dutch disambiguator would always choose analyses whose lemmas matched their surface forms, even for very rare lemmas; now the more common lemma will be returned. For example, *schepen* can be a singular noun with the lemma *schepen*, but it is more likely to be a plural noun with the lemma *schip*.

NEW Release 1.10.0

April 23, 2018

Entity Extraction and Linking

- **New deep neural network processor (in BETA):** We've added an alternative entity extraction processor, which can be used in place of the standard statistical extractor. The new processor employs a deep neural network that improves accuracy up to 7% and error rate up to 32%. It is available for English, Arabic, and Korean. To enable this processor, provide `DNN` for the modelType. Example: `{"content": "your_text_here", "options": {"modelType": "DNN"}}`

Morphological Analysis

- **New language support:** We've added support for lemmatizing Catalan, Estonian, Serbian, and Slovak text.
- **Bug fix:** Previously, tokens could be empty or contain only invisible characters. Such tokens will no longer be returned.

Language Identification

- **New language support:** Short string language identification now also supports Malay and Indonesian. Both languages were already supported for longer texts.

Sentiment Analysis

- **New language support:** We've added support for sentiment analysis in Persian (Farsi and Dari) at both the document and the entity level.

Name Translation

- **New language support:** Rosette now supports transliteration of person, organization, and location names from Greek to Latin script.

Name Similarity

- **New language support:** Rosette now supports matching of Greek names written in Greek script to English names written in Latin script and other Greek names written in Greek script.
- **Accuracy improvements:** We have improved match scores and segmentation rules for Arabic, Western Farsi, and Japanese names.

Point Release - 1.9.4

March 27, 2018

Morphological Analysis

- **Bug fix:** We've improved our handling of tokens consisting of numbers and Latin characters, such as serial numbers, in Korean. Previously these tokens were decompounded into multiple morphemes.
- **Bug fix:** We've added the lower case Russian word "интернет" ("internet") to the dictionary, which was previously only present in title case.
- **Bug fix:** We've improved our handling of tokens containing an apostrophe immediately followed by a digit in languages like French and Italian, like "all'M5S". Previously, the apostrophe would be parsed as its own token.
- **Bug fix:** We've improved our German analysis by taking better advantage of context clues, and now return more accurate results, especially for uncommon words.
- **Bug fix:** We've improved our handling of English and German words in all-caps. Previously, these words were assumed to be proper nouns, even though all-caps may simply denote emphasis.

Transliteration

- **Performance improvement:** We've added caching of model objects to prevent OutOfMemoryErrors.

Entity Extraction and Linking

- **Hebrew entity extraction:** We've added support for entity type "Title" in Hebrew.

Point Release 1.9.3

February 17, 2018

Morphological Analysis

- **German disambiguation:** We've improved our German disambiguator for lemmas and part-of-speech tags to be more sensitive to capitalization, particularly for single word inputs.
- **Bug fix:** Previously, German definite articles (der, die, das, den, dem, and des) meaning the were lemmatized inconsistently. They are now all lemmatized to the masculine singular nominative form, der.
- **Bug fix:** In some languages, an apostrophe may mark a token boundary, like in the Italian phrase all'M5S. Previously the token boundary was incorrectly omitted when the following token contained a digit. This issue has been rectified and M5S will be properly tokenized.

Name Similarity

- **Farsi Name Matching:** We've improved the behavior of Western Farsi-English matching by tokenizing input names earlier in the analysis process.

Point Release - 1.9.2

February 6, 2018

Entity Extraction and Linking

- **Currency support:** We've added several additional currency symbols to the regex, including the Turkish Lira (₺), the Pound Sterling (£), and the Euro (€).
- **Bug fix:** We've fixed a bug that caused hexadecimal number strings to be incorrectly extracted as products.

Name Translation

- **Thai improvement:** We've modified the gemination rules (consonant elongation) of the ISO11940-2 Thai transliteration standard to improve Thai translation accuracy.

Name Similarity

- **Bug fix:** We've fixed a bug around matching names of organizations and locations that contain numbers, such as "Century 21 Real Estate LLC." These non-person names containing digits will now match more accurately.

NEW Release - 1.9.0

January 16, 2018

Topics

- **Salience scores:** We've added salience scores for keyphrases and concepts to indicate how relevant an extracted concept or keyphrase is to the overall content of a text. You now have the option to filter out results below a desired threshold value: `{"content": "your_text_here", "options": {"keyphraseSalienceThreshold": value, "conceptSalienceThreshold": other_value}}`.
- **Short string support:** We've improved our concept extraction logic, and now support concept extraction for short input strings, i.e. texts less than 280 characters long.

Name Deduplication

- **Thai support:** Rosette now supports deduplication of Thai names.

Sentiment Analysis

- **New feature:** We've added the option to use an experimental alternative deep neural network (DNN) sentiment model for English: `{"content": "your_text_here", "options": {"modelType": "DNN"}}`. The new model will produce different results, which may be more accurate than the current support vector machine (SVM) model, depending on your data. As it is experimental, we are particularly interested in getting user feedback. On-premise users of Rosette API should review the new system requirements in install-guide.txt before using this option.

Entity Extraction and Linking

- **Entity offsets returned:** Entity mention offsets are now returned by default. Offsets can be used to locate the exact surface forms of an extracted entity in the document text.
- **Korean improvements:** We've significantly improved the accuracy of entity extraction results across all entity types in Korean.
- **Confidence scores:** Confidence scores for entities extracted using Rosette's statistical processor, as well as all linked entities, will now be returned by default. Confidence scores allow Rosette to return the most accurate results, particularly for entity linking. To change this behavior, set `{"content": "your_text_here", "options": {"calculateConfidence": false}}`.

Language Identification

- **Detect language regions in multilingual documents:** The language identification endpoint can now detect different language regions in a multilingual document.
- **Score changes:** We've rescaled the confidence scores returned by the language identification endpoint based on customer feedback. The ranking of language candidates will not change, but the scores themselves will be higher. If you currently filter language identification results based on a confidence threshold, you will need to reset that threshold to maintain parity with previous versions.

Name Translation

- **Thai support:** Rosette now supports transliteration of names from Thai to Latin script.

Name Similarity

- **Thai support:** Rosette now supports matching of Thai names to English names and other Thai names.
- **Accuracy improvements:** We have improved match scores for Arabic names (persons, locations and organizations) as well as for Chinese and Japanese organizations.

Point Release - 1.8.1

November 13, 2017

Entity Extraction and Linking

- **Bug Fix:** This release addresses a bug whereby entity linking confidence scores were not being returned when requested. Confidence scores for entities resolved to Wikipedia entries will now be returned when using the following option: `{"content": "your_text_here", "options": {"calculateConfidence": true}}`

NEW Release - 1.8.0

October 23. 2017

Topic Extraction

- **New Endpoint: Topic extraction** We've added a topic extraction endpoint that identifies the key ideas of an input text. For a given input, the endpoint will return two lists: Keyphrases, a list of phrases extracted directly from the text, and Concepts, a list of phrases which do not have to be explicitly mentioned in the input.

LABS

- **LABS graduates:** The /transliteration, /relationships, and /syntax/dependencies endpoints have graduated from “Labs” status and are now fully supported.

Sentiment Analysis

- **New language:** Rosette now supports document and entity-level sentiment analysis in French.

Entities

- **Salience Scoring:** Rosette can now return salience scores, which indicate whether an entity is important to the overall scope of the document. Turn on the scores by adding an option to the request: `{"content": "your_text_here", "options": {"calculateSalience": true}}`
- **Linking Confidence Scoring:** Rosette can also now return Linking Confidence scores, which represent the degree of certainty of the link between an in-document entity mention and its linked QID. It may be used for thresholding and removal of false positives. Linking Confidence scores for entities identified by our linker and assigned with a QID are now available by adding an option to the request: `{"content": "your_text_here", "options": {"calculateConfidence": true}}`

Point Release - 1.7.3

July 26, 2017

Name Deduplication

New Endpoint: Name Deduplication We've added a name deduplication endpoint that identifies similar names within a list. The endpoint accepts a list of names, organizes the list into clusters of unique names, and assigns each cluster with an id number. It then returns those ids to the user.

Point Release - 1.7.2

June 22, 2017

Entity Extraction

- **Bug fix:** This release addresses a backward compatibility issue between the latest Rosette API and older versions of our Java binding that affected Rosette's ability to return entity confidence scores. Confidence

scores for entities identified by our statistical extractor are now available by adding an option to the request: `{"content": "your_text_here", "options": {"calculateConfidence": true}}`

NEW Release - 1.7.1

June 14, 2017

Transliteration for Arabizi

- **New Endpoint: Transliteration** We've added a transliteration endpoint that converts between Arabic written in ASCII, also called Romanized Arabic chat or Arabizi, and native Arabic script.

Arabic Sentiment Analysis

- **Beta Arabic Support for /sentiment:** We now return document-level and entity-level sentiment analysis results for Arabic language input.

Relationship Extraction

- **Personal pronoun resolution for /relationships:** Building on the pronoun resolution capabilities of our /entities endpoint, pronouns which are resolved to named entities can now be arguments in relationships.

Entities

- **Improved Confidence Scoring:** Confidence score calculation is improved to correlate well with precision and may be used for thresholding and removal of false positives.

Tokenization

- **New support for emoticons, emoji, @mentions, hashtags, URLs, and email addresses:** These special characters and character combinations are now kept together as a single token in all languages, greatly improving the accuracy of analysis further downstream.

Morphological Analysis

- **Improved accuracy for English and Spanish:** For this release, we updated our English and Spanish dictionaries. We also introduced new, more advanced disambiguation models for these languages, which help Rosette to correctly determine a given word's part of speech. For example, words like "object" can be either a noun ("this is an object") or a verb ("I object!").
- **Lemmatization and normalization of emoticons, emoji, @mentions, hashtags, URLs, and email addresses:** Rosette now normalizes and lemmatizes these special characters and character combinations to streamline analysis.
- **Improved decompounding for Dutch:** Dutch language text is now decompounded more accurately, Dutch text is now decompounded more accurately, producing better tokens for search enhancement and other applications.

NEW Release - 1.6.0

March 23, 2017

Relationship Extraction

- **Improved Accuracy of Corporate Relationships:** Improvements made to the identification of relationships between corporations. The relationships involved are: ORG-SUBSIDIARY-OF, ORG-COLLABORATORS, ORG-ACQUIRED-BY and ORG-PROVIDER-TO.

- **Removed the ORG-PARTNERSHIPS Relationship:** The ORG-PARTNERSHIPS relationship is now subsumed under ORG-COLLABORATORS and is no longer extracted as an independent relationship.

Entity Extraction and Linking

- **Improved Linking Accuracy via Inclusion of New Context Features:** The statistical model for entity linking includes features that measure the vector space similarity between an entity context and the Wikipedia contexts of its potential linking targets. The new features result in higher F-Scores across all supported languages.
- **Entity Linking in Japanese, Chinese and Spanish:** Entity linking to Wikidata with QIDs for Japanese, Chinese and Spanish text is supported.
- **Removed Long Text Linking:** Entity linking to Wikidata (with QIDs) for long texts is removed, which, as a result, removed entity linking capabilities in Arabic.

Text Embedding

- **Vector dimension reduced from 512 to 300:** We are able to produce smaller vectors that are more efficient and memory friendly without sacrificing overall speed or accuracy.
- **Improved Speed and Accuracy:** A number of speed enhancements have been made along with much larger vocabularies to increase accuracy.

Language Identification

- **Improved Accuracy on Texts with Mixed Scripts:** A script specific model is now selected based on the weighted frequency of the different scripts in the input.

Name Matching

- **Japanese Improvements:** Rosette API now has better support for Japanese name matching. This includes the new use of word embeddings, which are used to match words with similar semantic meaning, as well as improved Japanese name segmentation.

NEW Release - 1.5.1

January 10, 2017

Targeted Relationship Extraction

- **New Endpoint Functionality:** The /relationships endpoint now returns targeted relationships, as opposed to the former open relationships, as its default extracted relationships. Targeted relationships are specifically between two entities, and are labeled by a certain relationship type. You can see the former open relationships by setting the option of "discoveryMode" to "true".

/entities/linked REMOVED

- **Removed Deprecated Endpoint:** The /entities/linked endpoint, previously deprecated, is now completely removed. All functionality is available through the /entities endpoint. You will receive a 404 when calling /entities/linked.

Entity Extraction

- **Social Media Linking in Japanese and Chinese:** Our fast short text entity linker to Wikidata is now available for Japanese and Chinese.

- **Removal of long text entity linking:** Our long text entity linker has been replaced by our fast short string entity linker. You will now see entity linking results from our short string linker by default. This removes linking support for Arabic.
- **Additional Language Support:** The entity extractor now supports Vietnamese.

CJK Support for Names

- **Name Translation and Similarity CJK Improvements:** The /name-similarity and /name-translation endpoints now support matching and translating between Japanese-Chinese, Japanese-Korean, and Korean-Chinese. Japanese accuracy was improved significantly.

Text Embeddings Improvement

- **Improved Accuracy for Document-level Embeddings:** We made some improvements to our algorithm for calculating text embeddings across multi-word input, so you should see more accurate results for document-level vectors.

Japanese Sentiment Analysis

- **Beta Japanese Support for /sentiment:** We now return document-level and entity-level sentiment analysis results for Japanese language input.

NEW Release - 1.4.0

October 27, 2016

Syntactic Dependencies (NEW)

- **New Endpoint:** We've added a syntactic dependencies endpoint that returns the parse tree of the input text as a list of labeled directed links between tokens, as well as the list of tokens in the input sentence.

Relationship Extraction

- **Entities Linked to Wikidata:** Where available, Rosette will now link entities extracted within relationships to Wikidata. You'll see this information returned as a QID in the argument ID.
- **Modality Returned:** We've also added a "modality" field to Rosette's Relationship Extraction. Modality is the semantic context of the possibility or necessity of the relationship; the values can be “assertion”, “negation”, “uncertainty”, “opinion”, or “question”.

Starter Plan (NEW)

- **New $99 API Plan:** For a limited time, we’re offering a special Starter plan. $99/month gets you 40,000 Rosette API calls. Want to dive deep into Rosette but don’t need a whole 100,000 calls? This plan is for you.

NEW Release - 1.3.0

September 15, 2016

Text Embedding (NEW)

- **New Endpoint:** We added a text embedding endpoint that returns a single vector of floating point numbers that represents the document or word in the semantic space.

Sentiment Analysis

- **Additional entities:** We changed the /sentiment endpoint to return the sentiment of all entities discovered by Rosette, including Person, Location, Organization, Date, Time, and more entity types.

Entity Extraction

- **Turn off entity linking:** We added an option to disable entity linking in order to improve the call speed. Add `"options": {"linkEntities": "false"}` to your /entities call. Rosette returns a list of the entities with a temporary ID (TID).

Global changes

- **Concurrency header:** We added the `X-RosetteAPI-Concurrency` header to return the number of concurrent calls allowed on your plan. If you are receiving 429 errors, Too Many Requests, then Contact us for greater concurrency.

NEW Release - 1.2.3

July 21, 2016

Global changes

- **Input genre:** The genre field is available for /entities and /entities/linked to indicate the input is from social media. Specifying `genre=social-media` does not affect the output of the other endpoints. Applies to: /entities, /entities/linked, /relationships, /categories, /sentiment, /language, /morphology, /tokens, /sentences.

Entity Linking

- **Temporary entity ID:** With the unification of the /entities/linked and /entities endpoints, the /entities/linked now returns a “T” ID for entities without a Wikidata QID.

Entity Extraction

- **Entity endpoints unified:** We combined the /entities and /entities/linked endpoints into one endpoint, /entities. Rosette now returns the entity mentions and the entityId, if available. The `entityId` replaced the `indocChainId`. The output of /sentiment has not changed.
- **Entities Linked deprecated:** We deprecated the /entities/linked endpoint. It is still available, but we recommend that you adapt your applications to the new /entities endpoint.
- **Additional entities:** Rosette now extracts more entity types: Date, Time, Longitude and Latitude, and Distance.
- **Japanese entityId:** We added support for linking entities in Japanese (`jpn`) text to their `entityId`.
- **Spanish social media:** We added support for extracting entities from `social-media` in Spanish language documents, using the `genre` field.
- **Malay entities**: We added support for extracting entities in Malay (`msa`).

Error code

- **409 Error:** We added the 409 error code for when the binding version is out of date. If you receive this error, update your binding to the most recent version.

Sentiment Analysis

- **Spanish support:** We added support for analyzing the sentiment of Spanish language documents.

NEW Binding Release - Ruby and R bindings

June 20, 2016

Bindings

- **Ruby:** We added the Ruby binding to the gray column to the right and on Github. There is a Ruby gem available as well.
- **R:** We added the R binding to the gray column to the right and on Github.
- **cURL examples:** We changed the shell examples in the gray column on Features and Functions to be cURL code examples.

NEW Release - 1.1.2

May 10, 2016

Entity Linking

- **Social input:** We added a request field, `"genre": "social-media"`, to speed up and improve the accuracy of linking Person, Location, Organization and Product entities in social media posts. English input only.

NEW Release - 0.10.3

March 29, 2016

Global changes

- **Language used:** We added a Response Header, `X-RosetteAPI-ProcessedLanguage`, to return the language used by Rosette for processing the call. Applies to: /entities, /entities/linked, /relationships, /categories, /sentiment, /language, /morphology, /tokens, /sentences
- **requestId moved:** We moved the `requestId` object from the JSON response body to the Response Header as `X-RosetteAPI-Request-Id`. Applies to: /entities, /entities/linked, /relationships, /categories, /sentiment, /language, /morphology, /tokens, /sentences, /name-translation, /name-similarity
- **Rosette API Key:** We changed the `user_key` header's name to `X-RosetteAPI-Key`. The `user_key` header is deprecated. Applies to: /entities, /entities/linked, /relationships, /categories, /sentiment, /language, /morphology, /tokens, /sentences, /name-translation, /name-similarity
- **unit parameter removed:** We removed the optional `unit` request parameter. All input will be handled as a `doc`. Applies to: /entities, /entities/linked, /relationships, /sentiment, /morphology
- **Base64:** We removed support for Base64 encoding. You can submit binary files as a `multipart/form-data` call type. Applies to: /entities, /entities/linked, /relationships, /categories, /sentiment, /language, /morphology, /tokens, /sentences

Entity Extraction

- **Confidence removed:** The `confidence` value has been removed from the response object.

Relationship Extraction

- **Accuracy mode:** We removed the optional accuracy mode. All input will be processed with the `precision` accuracy mode, so Rosette will return a precise list of accurate relationships.

- **Explanations removed:** The `explanations` value has been removed from the response object.

Categorization

- **Explanations removed:** The `explanations` value has been removed from the response object.

Sentiment Analysis

- **Entity sentiment:** We added support for entity-level sentiment analysis. The JSON response for the /sentiment endpoint now includes two objects – `document` and `entities`. See the interactive documentation for examples of this new response.
- **Neutral result:** We added a neutral label for documents and entities with a neutral sentiment.
- **Short strings:** Rosette will automatically process short and long content with our proprietary algorithm for sentiment analysis.
- **Explanations removed:** The `explanations` value has been removed from the response object.

Morphological Analysis

- **Added language support:** We added language support for Dari, Persian, Urdu, and Western Farsi for Parts-of-Speech Tags.
- **Universal POS Tags:** We return Universal Parts-of-Speech Tags for all supported languages.
- **Tokens list:** Rosette returns parallel lists of tokens, lemmas, compound components, parts-of-speech tags, and Han-readings. If a token does not have a lemma, compound component, POS tag, or Han-reading, or if the language is not supported, then Rosette will return "null" in that list.

Name Translation

- **Renamed to /name-translation:** To clarify the endpoint's function, we renamed /translated-name to /name-translation. /translated-name is no longer available.
- **Removed result layer:** Within the response to /name-translation and /name-similarity endpoints, we removed the result layer so the results are in the response object. Also applies to: /name-similarity
- **TargetScheme requires uppercase:** For advanced users who would like to specify a targetScheme, the scheme must be submitted in uppercase.

Name Matching

- **Renamed to /name-similarity:** To clarify the endpoint's function, we renamed /matched-name to /name-similarity. /matched-name is no longer available.