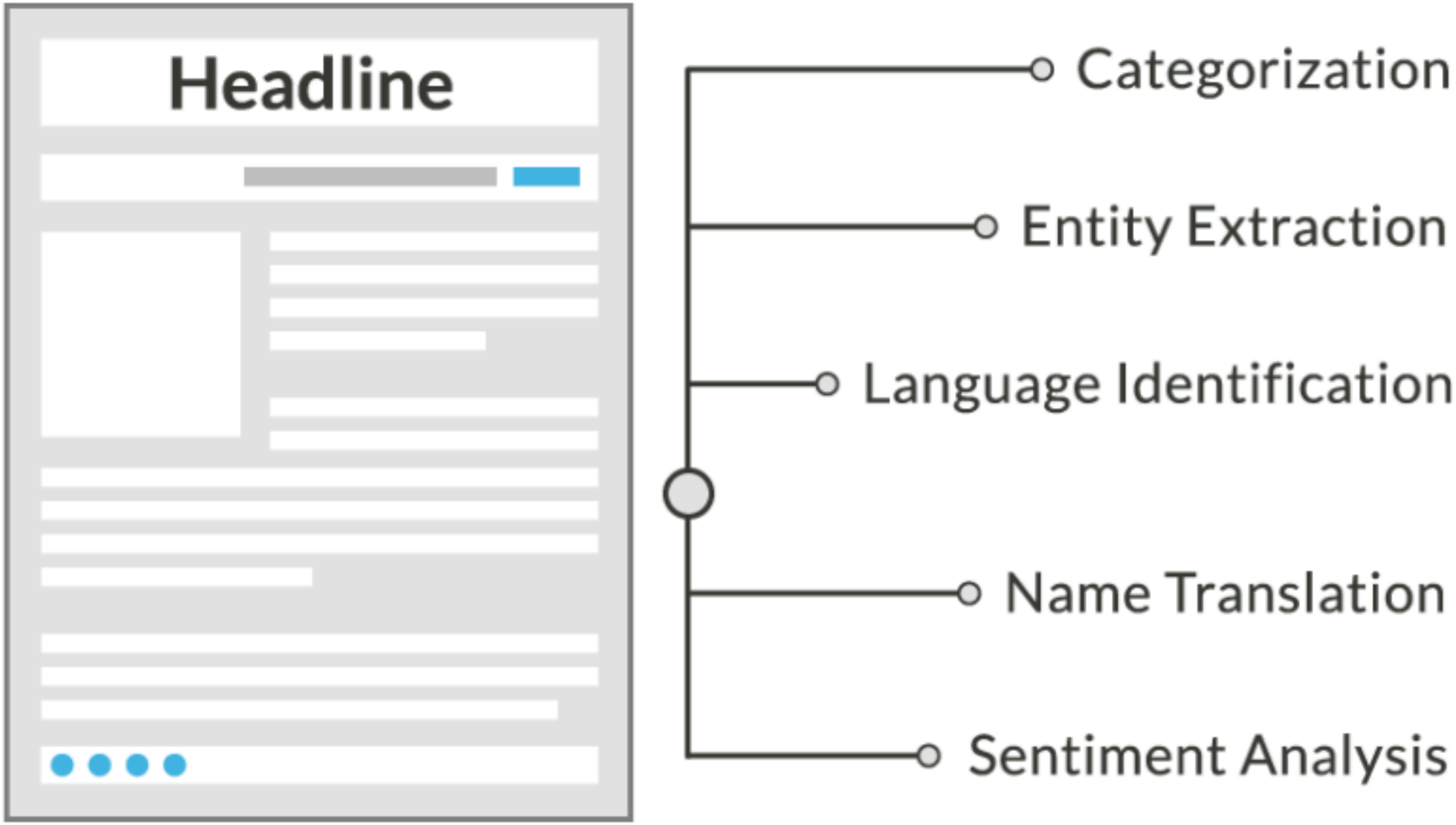
12. Cloud Machine Learning Advanced

Telung Pan Ph.D.
telung@mac.com

Text Analysis



Categorization

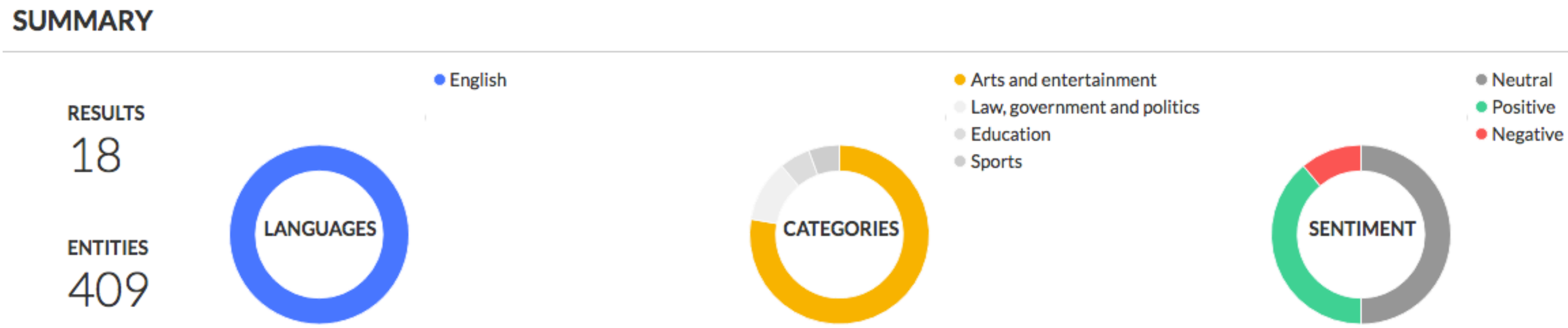
Entity Extraction

Language Identification

Name Translation

Sentiment Analysis

Language type
Categories, entity
Sentimental



SUMMARY

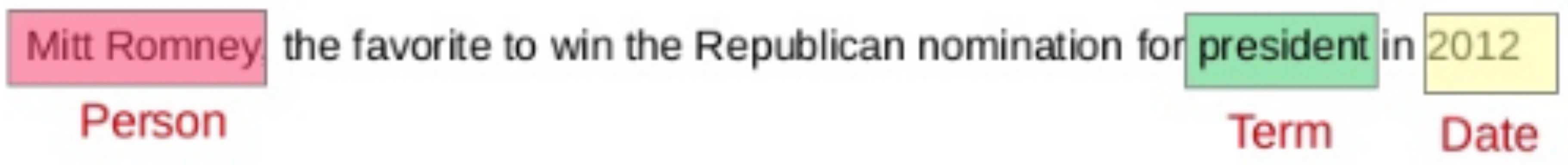
RESULTS
18

ENTITIES
409

- sudo easy_install pip
- sudo pip install rosette_api

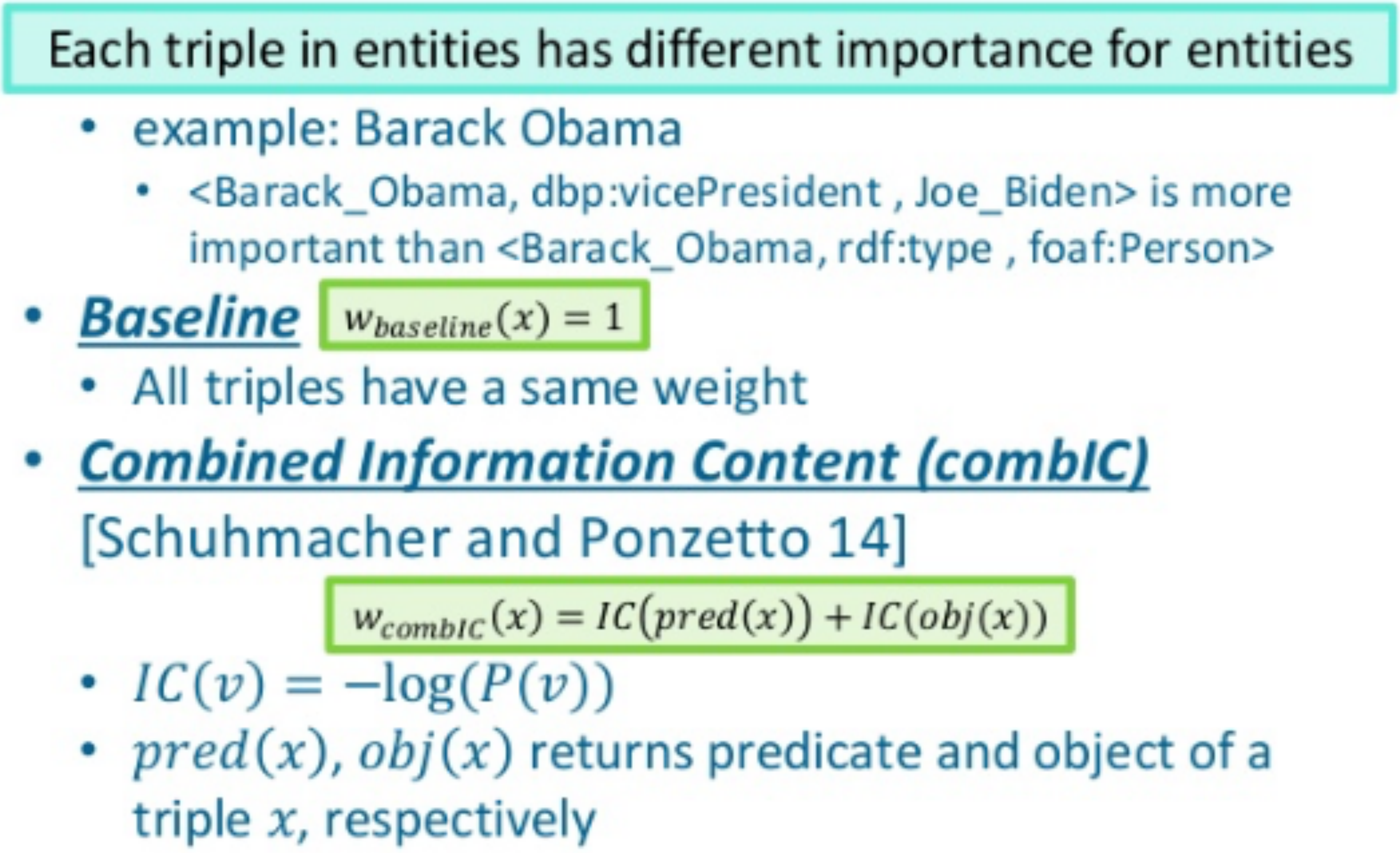
gcloud ml language analyze-entities

- Entity Analysis inspects the given text for common names or known entities (proper nouns such as public figures, landmarks, etc.), and returns information about those entities.



Mitt Romney, the favorite to win the Republican nomination for president in 2012

Baseline & Weight



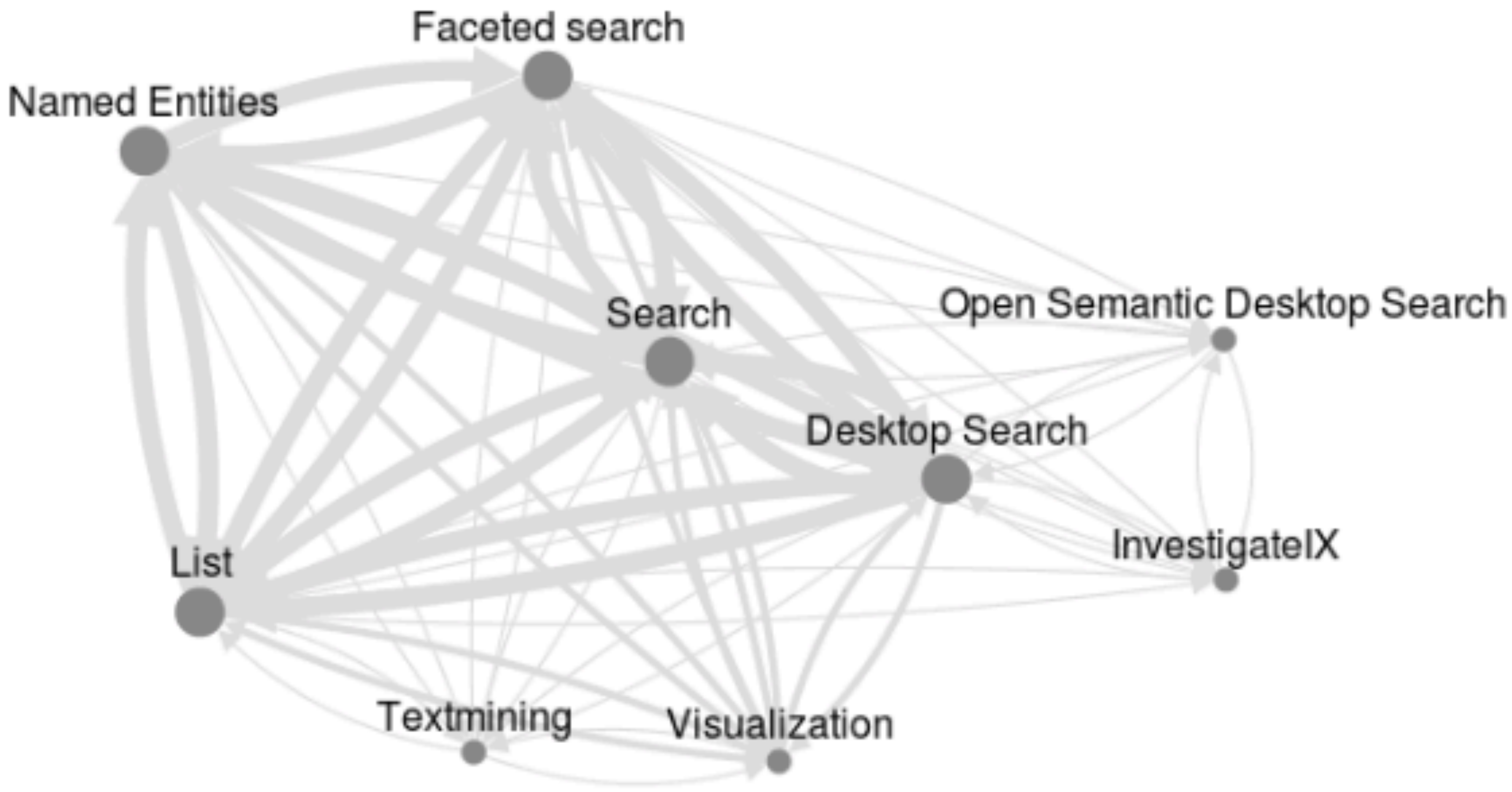
Each triple in entities has different importance for entities

- example: Barack Obama
  - <Barack_Obama, dbp:vicePresident , Joe_Biden> is more important than <Barack_Obama, rdf:type , foaf:Person>
- ***Baseline*** $\boxed{w_{baseline}(x) = 1}$
  - All triples have a same weight
- ***Combined Information Content (combIC)*** [Schuhmacher and Ponzetto 14]

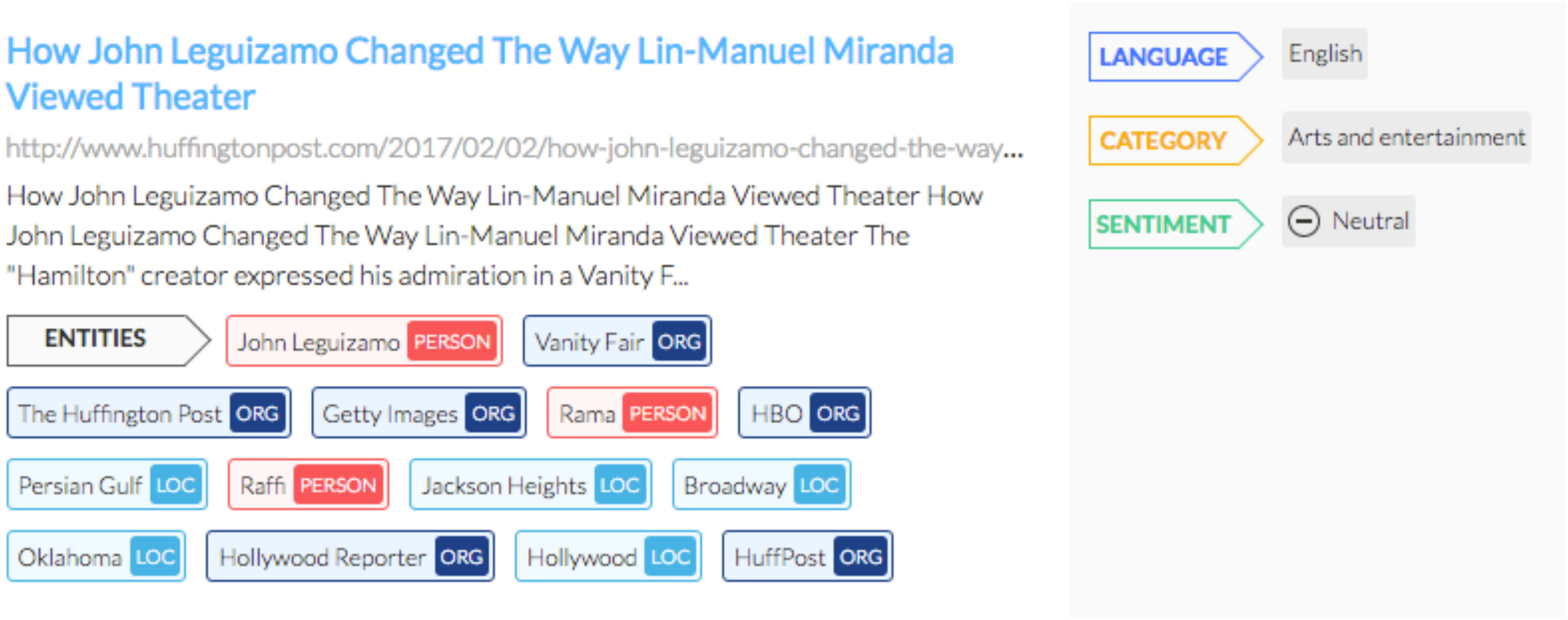
$$\boxed{w_{combIC}(x) = IC\big(pred(x)\big) + IC(obj(x))}$$

  - $IC(v) = -\log(P(v))$
  - $pred(x)$, $obj(x)$ returns predicate and object of a triple x, respectively

Entity & Attribute



Internet text entity analysis



How John Leguizamo Changed The Way Lin-Manuel Miranda Viewed Theater

http://www.huffingtonpost.com/2017/02/02/how-john-leguizamo-changed-the-way...

How John Leguizamo Changed The Way Lin-Manuel Miranda Viewed Theater How John Leguizamo Changed The Way Lin-Manuel Miranda Viewed Theater The "Hamilton" creator expressed his admiration in a Vanity F...

ENTITIES

- John Leguizamo PERSON
- Vanity Fair ORG
- The Huffington Post ORG
- Getty Images ORG
- Rama PERSON
- HBO ORG
- Persian Gulf LOC
- Raffi PERSON
- Jackson Heights LOC
- Broadway LOC
- Oklahoma LOC
- Hollywood Reporter ORG
- Hollywood LOC
- HuffPost ORG

LANGUAGE English

CATEGORY Arts and entertainment

SENTIMENT ⊖ Neutral

Google Cloud
Entity Analysis

- gcloud ml language analyze-entities - use Google Cloud Natural Language API to identify entities in text

- ***gcloud ml language analyze-entities (--content=CONTENT |--content-file=CONTENT_FILE) [--content-type=CONTENT_TYPE; default="plain-text"] [--encoding-type=ENCODING_TYPE; default="utf8"] [--language=LANGUAGE] [GCLOUD_WIDE_FLAG …]***

Required Flags

- Exactly one of these must be specified:
  - --content=CONTENT
    Specify input text on the command line. Useful for experiments, or for extremely short text.
  - --content-file=CONTENT_FILE
    Specify a local file or Google Cloud Storage (format gs://bucket/object) file path containing the text to be analyzed. More useful for longer text or data output from another system.

Optional Flags

- ***--content-type=CONTENT_TYPE***; default="plain-text"
  Specify the format of the input text. CONTENT_TYPE must be one of: html, plain-text.

- ***--encoding-type=ENCODING_TYPE***; default="utf8"
  The encoding type used by the API to calculate offsets. If NONE, those offsets are not calculated. This is an optional flag only used for the entity mentions in results, and does not affect how the input is read or analyzed. ENCODING_TYPE must be one of: none, utf16, utf32, utf8.

- ***--language=LANGUAGE***
  Specify the language of the input text. If omitted, the server will attempt to auto-detect. Both ISO (such as en or es) and BCP-47 (such as en-US or ja-JP) language codes are accepted.

gcloud Wide Flags

- These flags are available to all commands: --account, --configuration, --flatten, --format, --help, --log-http, --project, --quiet, --trace-token, --user-output-enabled, --verbosity. Run $ gcloud help for details.

Basic gcloud Commands

- ***sudo curl https://sdk.cloud.google.com |bash***
- ***exec -l $SHELL***
- ***gcloud init***
- gcloud auth login
- gcloud components update
- gcloud auth list
- gcloud config list project

Analyzing Entity Sentiment

- Entity Sentiment Analysis combines both entity analysis and sentiment analysis and attempts to determine the sentiment (positive or negative) expressed about entities within the text.

- Entity sentiment is represented by numerical score and magnitude values and is determined for each mention of an entity.

- Scores are then aggregated into an overall sentiment score and magnitude for an entity.

Analyzing entity sentiment provided as a string

- \

- To perform entity sentiment analysis, use the gcloud command line tool and use the --content flag to identify the content to analyze:

  ```
    gcloud ml language analyze-entity-sentiment --
  content="I love R&B music. Marvin Gaye is the best.
  'What's Going On' is one of my favorite songs. It was so
  sad when Marvin Gaye died."
  ```

Interpreting sentiment analysis values

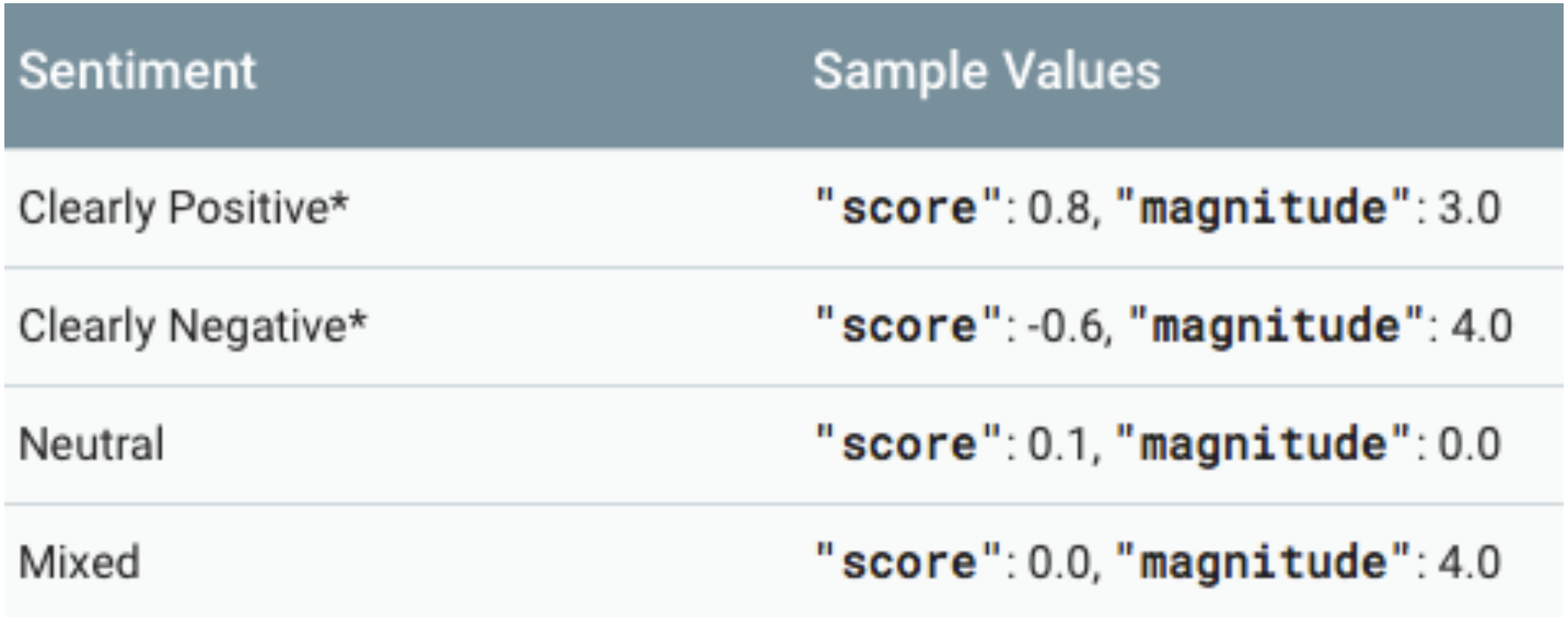
- The ***score*** of a document's sentiment indicates ***the overall emotion of a document***. The ***magnitude*** of a document's sentiment indicates ***how much emotional content is*** present within the document, and this value is often proportional to the length of the document.

- For example, "angry" and "sad" are both considered negative emotions. However, when the Natural Language API analyzes text that is considered "angry", or text that is considered "sad", the response only indicates that the sentiment in the text is negative, not "sad" or "angry".

- A document with a neutral score (around 0.0) may indicate a low-emotion document, or may indicate mixed emotions, with both high positive and negative values which cancel each out.

- When comparing documents to each other (especially documents of different length), make sure to ***use the magnitude values to calibrate your scores***, as they can help you gauge the relevant amount of emotional content.

- The chart below shows some sample values and how to interpret them:



Sentiment	Sample Values
Clearly Positive*	"**score**": 0.8, "**magnitude**": 3.0
Clearly Negative*	"**score**": -0.6, "**magnitude**": 4.0
Neutral	"**score**": 0.1, "**magnitude**": 0.0
Mixed	"**score**": 0.0, "**magnitude**": 4.0

校正方法

- “Clearly positive” and “clearly negative” sentiment varies for different use cases and customers.

- Differing results for specific scenario might found. We recommend that you define a threshold that works for you, and then adjust the threshold after testing and verifying the results.

- For example, you may define a threshold of any score over 0.25 as clearly positive, and then modify the score threshold to 0.15 after reviewing your data and results and finding that scores from 0.15-0.25 should be considered positive as well.

Entity analysis response fields

- Entity analysis returns a set of detected entities, and parameters associated with those entities, such as the ***entity's type***, ***relevance of the entity to the overall text***, and ***locations in the text that refer to the same entity***. Entities are returned ***in the order (highest to lowest) of their salience scores***, which reflect their relevance to the overall text.

- ***type*** indicates the type of this entity (for example if the entity is a person, location, consumer good, etc.) This information helps distinguish and/or disambiguate entities, and can be used for writing patterns or extracting information. For example, a type value can help distinguish similarly named entities such as “Lawrence of Arabia”, tagged as a WORK_OF_ART (film), from “T.E. Lawrence”, tagged as a PERSON, for example.

- ***metadata*** contains source information about the entity's knowledge repository Additional repositories may be exposed in the future. This field may contain the following subfields:
  - ***wikipedia_url***, if present, contains the Wikipedia URL pertaining to this entity.
  - ***mid***, if present, contains a machine-generated identifier (MID) corresponding to the entity's Google Knowledge Graph entry.

- ***salience*** indicates the importance or relevance of this entity to the entire document text. This score can assist information retrieval and summarization by prioritizing salient entities. Scores closer to 0.0 are less important, while scores closer to 1.0 are highly important.

- ***mentions*** indicate offset positions within the text where an entity is mentioned. This information can be useful if you want to find all mentions of the person “Lawrence” in the text but not the film title. You can also use mentions to collect the list of entity aliases, such as “Lawrence,” that refer to the same entity “T.E. Lawrence”. An entity mention may be one of two types: PROPER or COMMON. A proper noun Entity for "Lawrence of Arabia," for example, could be mentioned directly as the film title, or as a common noun ("film biography" of T.E. Lawrence).

Another text

Four score and seven years ago our fathers brought forth
on this continent a new nation, conceived in liberty and
dedicated to
the proposition that all men are created equal

問題

- 請問此段文字中，最重要的實體概念 (entity) 是什麼？
- 最正面的實體概念是什麼？
- 說明你是如何分析上面兩個問題的答案？

Analyzing Entity Sentiment from Google Cloud Storage

- To perform entity sentiment analysis, use the gcloud command line tool and use the --content flag to identify the content to analyze:

  ```
  gcloud ml language analyze-entity-sentiment --content-file=gs://<bucket-name>/<object-name>
  ```