General Concepts

Filters / Exclusions / Relevance

- Emails are excluded, i.e. not included in any insight aggregations if:
  - (1) They have been trashed OR represent drafts (based on the automatically assigned Gmail labels ("TRASH,"DRAFT"))
  - (2) [OPTIONAL] They have been archived (based on the *absence* of an automatically assigned "INBOX" Gmail label))
  - (3) [OPTIONAL] They are 'non-personal' on the basis of the automatically assigned gmail labels ("CATEGORY_SOCIAL", "CATEGORY_PROMOTIONS", "CATEGORY_UPDATES", "CATEGORY_FORUMS")
  - (4) Are non parseable, i.e. the processor fails to correctly extract and format the data (> this may happen e.g. with newsletters sent in html format)
- **Note*: Emails excluded on the basis of (1) are not pulled from your inbox, emails excluded on the basis of (2)/(3) are extracted and viewable (see *'Launch Excluded Browser > Excluded Emails [Emails: XX] '*), and emails excluded on the basis of (4) are extracted but not viewable (see *'* Number of Non-Parsable Emails': XX*)

Units of Analysis

- *Message*
  - A message in the sense of a single email
  - Each message is associated with the date on which it was sent
- *Link*
  - A message connecting you and one other contact: One message can correspond to one or more links, e.g (a) you (*from*) sent a message to 3 people *(to,cc or bcc)* > 3 links (b) you (*to, cc or bcc)* received an email from 1 person *(from*) > 1 link
  - Each link is associated with 1 contact (the other end of the link being your own address) - in the case of a sent email > the recipient (*to, cc, or bcc*) // in the case of a received email > the sender (*from)*
  - Each link is associated with the date on which the associated message was sent
  - **Note*: For the purposes of all aggregations all link types (Sent emails: you > to / you > to (cc) / you > to (bcc) / / Received emails: from > you / from > you (cc) / from > you (bcc))) are treated equally
- *Threads / Conversations*
  - Each message is by definition part of a thread - in cases where there is only one email in the thread the thread effectively reduces to the message (the message id equals the thread id). Conversations are defined as threads that involve at least two messages (note that (a) a thread in which one person sends two emails without receiving a reply would still be considered a conversation (b) a thread in which you are merely on the receiving end would still be counted one of your conversations)

- Each conversation is associated (a) a start date: the date that the first message in the conversation was sent (e.g. all conversations started in a given week) (b) multiple active dates: all dates on which at least one message was sent as part of the conversation (e.g. all conversations 'active' on a given day)

Exchange Structure

- *Inbox / Outbox*
  - An email is classified as sent ('outbox') if the sender address = your address, i.e. in all other cases (your address is contained in the recipient address field, the cc field or the bcc field) the email is classified as received ('inbox')
- *Response / Response time*
  - A message is associated with a response if
    - In the case of a received email > another extracted (and non-excluded) *sent* email represents a reply to the message in question (on the basis of gmail's reply-to-id) // In the case of a sent email > another extracted (and non-excluded) *received* email represents a reply to the message in question (on the basis of gmail's reply-to-id)
    - ** Note:* The above constraints are designed to rule out edge cases where, e.g.:
      - You send two or more subsequent emails (with e.g. the second one responding to the first) without receiving a reply >> No response
  - A link is associated with a response if the associated message is associated with a response (in the sense defined above) from the contact involved in the link (e.g. you sent a message sent to 3 people > 1 person responds > 1 of the 3 links is associated with a response)
  - Response times capture the difference in minutes between the timestamp of the response and the timestamp of the original message (>> response times > 0 minutes)

Contact Groups

- Each link is uniquely associated with one contact and hence one contact group (group a / group b / no group)
- Each message / conversation is associated with one or more contacts (i.e. all contacts in a conversation to whom you have sent an email *(to, cc, bcc*) or from whom you have received a message *(from)* as part of the thread) and are hence associated with one or more contact groups
- * *Note*: The only uncategorized contact should be your own address (> this ensures that emails sent to yourself will be not be counted in any aggregate statistics)

Non - Language Insights

Volume

- *Statistics*
  - Number of emails which you (a) sent or received / (b) sent / (c) received
- *Aggregation Perspectives:*
  - By contact group (group a, b, none)
  - By time (day, week, week day, time period (morning, afternoon, evening))
  - * Daily average
- *Notes*
  - Totals are not expected to add up across the 3 contact groups given that a message can be sent to multiple contacts, e.g. a single message may be sent to person-1 (group a) and person-2 (group b) and would hence show up in both group's totals

FirstLast

- *Statistics*
  - (a) Conversation-level *(NOTE: Conversations defined as specified above)*
    - Percent of conversations in which you sent the first/last email
    - Overall number of conversations versus the number of conversations in which you sent the first/last email
  - (b) Message-level
    - Number of sent messages versus the number of sent messages which represented the first/last email in a conversation
- *Aggregation Perspectives:*
  - By contact group (group a, b, none)
  - By time (day, week)
- *Notes*
  - *Conversation-level statistics:* Capture the number of total conversations (active at any point in time) relative to the number of (active) conversations in which you sent the first / last email (within the given time period), e.g. total number of active conversations on day x vs. total number of conversations that you started/ended on day X (i.e. the total number of messages that you sent that started/ended a conversation on day X). Note that in the case of the group-level statistic we do not consider whether you ended / started the conversation *with an email to a contact in the relevant group* (the group-level constraint comes solely into play in defining which conversations are examined)

Responsiveness

- *Statistics*
  - Percent of sent messages to which you received a response to *(NOTE - Response are defined as specified above)*
  - Overall number of sent messages versus the number of sent messages to which you received a response
  - Percent of received messages to which you responded
  - Overall number of received messages versus the number of received messages to which you responded
- *Aggregation Perspectives:*
  - By contact group (group a, b, none)
  - By time (day, week, week day, time period (morning, afternoon, evening))
- *Notes*
  - In the case of the contact-specific statistics the focus is on whether you received a response from/sent a response to an individual in the given group (e.g. in case you received an email from 2 people but responded to only one > only one of the links would be counted as a response)

Simple - Language Insights

Talkative

- *Statistics*
  - Median character / word / sentence count for all sent / received emails
- *Aggregation Perspectives:*
  - By contact group (group a, b, none)
  - By time (day, week, week day, time period (morning, afternoon, evening))

LengthImbalance

- *Statistics*
  - Average length imbalance, i.e the ratio between the length of your responses and the length of the emails that you are responding to (in terms of characters, words, sentences) >> a length imbalance>1 indicates that your responses are longer than the emails that you are responding to
- *Aggregation Perspectives:*
  - By contact group (group a, b, none)
- *Notes*
  - For the purposes of calculating the length imbalance links involving zero-length emails are omitted (i.e. links resulting in an imbalance of 0 /infinity)

NLP Insights

Politeness

- *Statistics*
  - Mean politeness for all sent / received emails
  - Percent of sent/received messages which contain at least one sentence that is identified as a request
  - Average politeness imbalance, i.e the ratio between the politeness of your responses and the politeness of the emails that you are responding to >> a politeness imbalance>1 indicates that your responses are more polite than the emails that you are responding to
  - *Aggregation Perspectives:*
    - By contact group (group a, b, none)
    - By time (day, week, week day, time period (morning, afternoon, evening))

- *Notes*
  - Politeness is defined as a score that ranges from 0-1 whereby a higher score corresponds to a more polite message (the score is based on a SVM constructed/trained as described here: https://github.com/vegetable68/politeness_with_spacy
  - Politeness imbalance calculations mirror the length imbalance calculations (see above)

Sentiment

- *Statistics*
  - Percent of sent/received messages which contain at least one sentence positive/negative word
  - Ranking of sentiment categories based on the number of overall/sent/received messages that contain at least one word in that category
  - *Aggregation Perspectives:*
    - By contact group (group a, b, none)
    - By sentiment category

- *Notes*
  - Sentiment categories are based on a version of the sentiment dictionary used as part of MUSE (https://github.com/ePADD/epadd/blob/master/WebContent/WEB-INF/classes/lexicon/sentiments.english.lex.txt) - the original dictionary contains 45 categories with a total of 738 words (52 terms in the aggregated positive and 139 terms in the aggregated negative category). Using wordnet (nltk corpus) the original term list is expanded by including synonyms of the original terms (across word categories) > a total of 2000 words (180 terms in the aggregated positive and 600 terms in the aggregated negative category).

- The original sentiment dictionary contains predominantly (>95%) unigrams, the remainder being bigrams (a very small number of longer words is omitted). In the case of bigrams - where a bigram in the message matches a bigram in the sentiment dictionary both words are counted as belonging to the sentiment category in question
- Sentiment rankings are based on the number of occurrences of a given category (as defined above) across all emails/all sent emails/all received emails. Counts are normalised so that the most common category (in each category of emails - all/sent/received) is assigned a scaled frequency of 1.

Language Coordination

- *Statistics*
  - For each word category > (a) a coordination score (b) the % of sent emails that contain at least one word of the given type (c) the average number of words of the given type per sent email (d) the % of sent emails which represent a reply that contain at least one word of the given type (e) the average number of words of the given type per sent email that represents a reply
  - An aggregate coordination score
  - *Aggregation Perspectives:*
    - By contact group (group a, b, none)
    - By word category category

- *Notes*