Unique word count

word retrieval → word can be separated by whitespace between words

word classification → words can be classified by referring Google dictionary， a word can has more than one classification label, e.g. ‘good’ can be a noun or an adjective

term weighting → Content words will be given more weight than s[yntactic expletive](https://en.wikipedia.org/wiki/Syntactic_expletive), since our research objective is lexicon diversity, I recommend all syntactic expletive terms be weighted as zero, so they will be excluded from analysis

word count should be based on email to email, i.e. to count the total number of content words (mainly nouns, verbs and adjectives).

frequency of term use should be counted.

indefinite word usage

* generalistic to specific noun ratio

? how to separate generalistic terms from specific terms?

Database

TABLE - Study Subjects

Subject ID <PK>

Name

Age

Gender

Education - categorical variable:

<secondary, post secondary (college & bachelor), graduate study>

annual income - categorical variable:

<under 10,000, 10k to 40k, 40k to 80k, 80k to 15k, 15k to 20k, 20k & above>

postal code

TABLE - Raw data of email <inner join> with Study Subject TABLE

Email ID <PK>

Subject ID <FK>

Send OR Receive --we can only analyze the email sent out by the subjects.

Sender’s email

Receiver’s email

Send Time

Receive Time

Email Subject

RE <boolean> -->if an email has a RE label, it indicates interactive response between two

FW <boolean>-->if an email has a FW label, it indicates the communications among three or more people

Email Content