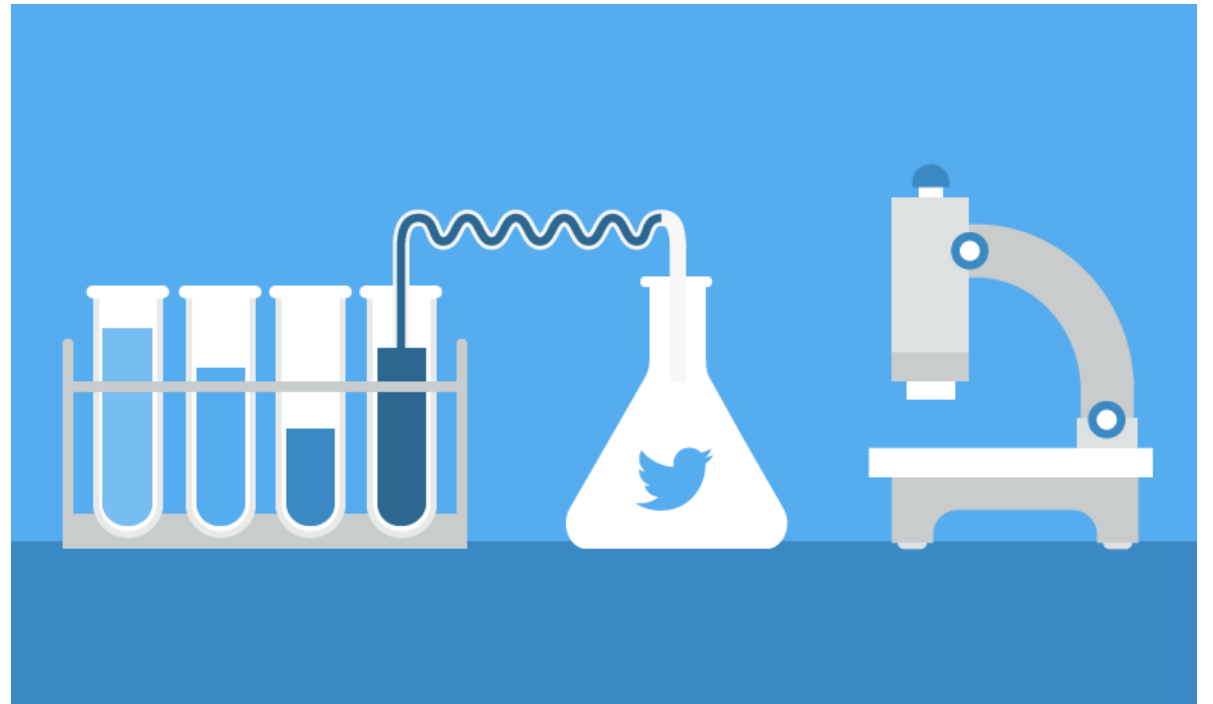# Sentiment Analysis with R

DSI Jatim Camp #3, 29 September 2018

Outline

- Overview Sentiment Analysis
- Intro to R Programming Language
- Intro to Text Analysis
- **Preprocessing**
- Visualize
- Sentiment

Preprocessing

Case Folding

change everything to lowercase.

Remove Punctuations and Number

Punctuation and other special characters only look like more words to your computer and R.

Tokenizing

Memecah yang kalimat menjadi kata-kata atau memutus urutan string menjadi potongan-potongan tiap kata yang menyusunnya

Stopwords

Kosakata yang bukan termasuk kata unik atau ciri pada suatu dokumen atau tidak menyampaikan pesan apapun secara signifikan pada teks atau kalimat

Stemming

Mengubah setiap kata menjadi kata dasarnya dengan menghilangkan imbuhan awalan, akhiran, sisipan, dan awalan-akhiran.

Extracting Features from Text

- Count Vectorizer
- TFIDF Vectorizer

Extracting Features from Text

- Count Vectorizer

count the appearance of the words in each text. For example, let's say we have 3 documents in a corpus: *"I love dogs"*, *"I hate dogs and knitting"*, *"Knitting is my hobby and my passion"*.

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Extracting Features from Text

- TFIDF Vectorizer (*Term Frequency-Inverse Document Frequency*)

Let's say we have two documents in our corpus as below.

1. I love dogs
2. I hate dogs and knitting

$$TF(t,d) = \frac{number\ of\ times\ term(t)\ appears\ in\ document(d)}{total\ number\ of\ terms\ in\ document(d)}$$

$$TF('I',d1) = \frac{1}{3} \approx 0.33$$

$$TF('I',d2) = \frac{1}{5} = 0.2$$

Extracting Features from Text

- TFIDF Vectorizer (*Term Frequency-Inverse Document Frequency*)

$$IDF(t,D) = \log\left(\frac{\text{total number of documents}(D)}{\text{number of documents with the term}(t)\text{ in it}}\right)$$

$$IDF('I',D) = \log\left(\frac{2}{2}\right) = 0$$

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D)$$

$$TFIDF('I',d1,D) = TF('I',d1) \cdot IDF('I',D) = 0.33 \times 0 = 0$$

$$TFIDF('I',d2,D) = TF('I',d2) \cdot IDF('I',D) = 0.2 \times 0 = 0$$

Thank You