Daten:

SMS Spam Collection v.1:

Beschreibung: Der Datensatz besteht aus 5.574 SMS, die für die Forschen von SMS-Spam gesammelt wurden. „Ham“ steht für eine echte SMS und „Spam“ für eine Spam-SMS.

Dieses 5.574 SMS wurden aus verschiedenen Quellen von Tiago Agostino de Almeida und José María Gómez Hidalgo zusammengetragen.

Quellen die für SMS Spam Collection v1 benützt wurden:

* 425 Spam-SMS von der Website grumbletext [[1]](#footnote-1)
* 450 SMS von Caroline Tag (PhD Theses) [[2]](#footnote-2)
* 3.375 SMS von der Sammlung: NUS SMS (NSC) [[3]](#footnote-3)
* 1.002 SMS von dem Datensatz SMS Spam Corpus v0.1 [[4]](#footnote-4)
* 322 Spam-SMS von dem Datensatz SMS Spam Corpus v0.1 [[5]](#footnote-5)

Das Heist der Datensatz SMS Spam Collection v.1 hat insgesamt 4.827 SMS und 747 Spam-SMS

Features:

Hinzugefügt :

* NumChars -> die Anzahl der Buchstaben in der E-Mail.
* containNumSequence -> ob zahlen vorhanden sind. (Mindestens fünf Zahlen hintereinander)
* containsCurrencySymbol -> ob ein Währungszeichen vorhanden ist.
* numBuzzwords -> ob die Nachricht gewisse „Spam Wörter“ enthalt (zum Beispiel: Call „free“, „txt“, „prize“ usw.)
* percentUppercaseLetters -> Wie viel Prozent der Nachricht in Großbuchstaben geschrieben wurde.

Schon vorhanden:

* class (ham, spam) -> **Ziel Klasse**. **Sagt aus, ob es eine Spam Nachricht ist oder nicht.**
* Text -> Die Nachricht in Textformat die entweder ham oder spam ist.
* Wie viele Attribute
* Was gute Werte sind

1. http://www.grumbletext.co.uk/ [↑](#footnote-ref-1)
2. http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf [↑](#footnote-ref-2)
3. http://www.comp.nus.edu.sg/~rpnlpir/ [↑](#footnote-ref-3)
4. http://www.esp.uem.es/jmgomez/smsspamcorpus/ [↑](#footnote-ref-4)
5. http://www.esp.uem.es/jmgomez/smsspamcorpus/ [↑](#footnote-ref-5)