Lade bitte bis Donnerstag 9.11.2017 20:00 den `R-Code` zu den jeweiligen Aufgaben in `moodle` hoch.

Bitte Name und die jeweilige Beispielnummer im Code notieren. Der Code soll so programmiert werden, dass das (richtige) Ergebnis in der Console ausgegeben wird. Alle Beispiele können und sollen mit den Funktionen der Kapitel 1-21 gelöst werden. Insbesondere auch Kapitel 16 und 17 durcharbeiten.

Die Datei `nrw17.RDATA` enthält das R-Objekt `nrw17` mit 1000 Tweets der Kurznachrichtenplattform twitter.com zum Suchbegriff `#nrw17`, der häufig für Nachrichten zum Thema Nationalratswahlen 2017 verwendet wurde[1].
Neben den angezeigten Nachrichten (= Tweets) stehen auch zahlreiche Zusatzinformationen zu jedem Tweet zur Verfügung, im Objekt `#nrw17` ist folgende Auswahl enthalten: Name des Users (screen name), der Zeitpunkt, wie oft der Tweet favorisiert und wie oft er retweeted wurde und ob es sich um eine Retweet handelt.

1. Lade die Datei `nrw17.RDATA`
   a) Welchen Typ und welche Länge hat das Objekt `nrw17`, wie heißen die Elemente des Objekts?
   b) Zeige das 1. Element des Vektors `inhalt` an. Es sollte folgendermaßen lauten: `"#Koalition: Verkündet #Kurz, mit wem er will? #nrw17 https://t.co/QFjuYAR4fF"`
   c) Zeige das zweite Wort der ersten 10 Elemente des Vektors `inhalt` an. Überlege, wie ein Wort definiert ist.
   d) Zeige das erste Wort der ersten 100 Elemente an, die keine Retweets sind.
   e) Überprüfe, ob alle Retweets mit "RT"beginnen.
   f) Wie viel Prozent der Tweets enthalten einen Verweis auf einen beliebigen User (gekennzeichnet durch den Prefix `"@"`)?
   g) Wie viele Verweise auf einen User enthält jeder Tweet? Inkludiere diese Information in das Objekt `nrw17` und bilde eine Häufigkeitstabelle.

2. a) Bilde einen Vektor mit allen Wörtern aller Tweets.
   b) Bilde einen Vektor mit allen Verweisen auf andere User. Wenn das vorige Beispiel nicht gelang, lade das Objekt `w.RDATA`.
   c) Wer sind die 10 User, die die meisten Tweets verfasst haben?
   d) Wie viele User haben 1, 2, ... Tweets verfasst?
   e) Wie häufig wurde in Tweets auf irgendeinen der 100 User mit den meisten Tweets verwiesen?

[1] die Daten wurden mit dem R-Paket `twitteR` extrahiert

3. a) Erstelle eine Häufigkeitstabelle mit allen in Tweets vorhandenen Zeichen. Falls das Beispiel nicht gelingt, lade den Vektor mit allen vorhanden Zeichen (`z.RDATA`)

   b) Erstelle eine Häufigkeitstabelle, in der alle Zeichen, die nicht im englischen Alphabet vorkommen, exkludiert werden.

   c) Erstelle eine Häufigkeitstabelle, in der alle Zeichen, die nicht im englischen Alphabet vorkommen, als NA angezeigt werden.

   d) Wie verteilen sich die Tweets auf die Stunden?

   e) Sind die Tweets über die Stunden gleichverteilt? Teste mit einem χ^2-Anpassungstest.

4. a) Generiere das Ergebnis von 1000 Umfragen mit jeweils 700 Befragten. Die Umfrage besteht aus 1 ja/nein Frage, die im Mittel von 30 % der Befragten mit "ja" beantwortet wird. Speichere das Ergebnis in einer 1000 x 700 Matrix. Wenn das Beispiel nicht gelingt, lade `M.RDATA` für die weiteren Aufgaben.

   b) Vergib zufällig Zeilennamen von 1 bis 1000.

   c) Ermittle für alle Umfragen den Anteil der Befragten, die mit "ja" antworteten.

   d) Bestimme das 2.5 und 97.5 %-Quantil der ermittelten Anteile ohne die Funktion `quantile()` und vergleiche das Ergebnis mit jenem der Funktion `quantile()`.

   e) Vergleiche das Ergebnis aus dem vorigen Beispiel mit dem theoretischen Konfidenzintervall. Für die Berechnung siehe Hinweis im 1. Übungsblatt.

   f) Sortiere die Anteile *numerisch* nach den zufällig vergebenen Zeilennamen.

   g) Ersetze 2000 Antworten zufällig durch einen fehlenden Wert und berechne die Anteile aus Beisiel 4c erneut.