Challenge *Sprachenerkennung* (Abgabe bis 11. Dezember)

Schreibe ein Programm, das erkennen kenn, in welcher (von mindestens vier) Sprachen ein Text geschrieben ist!

Häufigkeitsprofil einer Sprache

- **Beobachtung:** Jede Sprache hat ein spezifisches Buchstabenhäufigkeitsprofil
- Beispiel Deutsch (nur A-Z): (Wikipedia)
  - relative Häufigkeit von A: 6.51%
  - relative Häufigkeit von B: 1.89%
  - relative Häufigkeit von C: 3.06%
  - ....
  - relative Häufigkeit von Z: 1.13%

Häufigkeitsprofil einer Sprache

- **Beobachtung:** Jede Sprache hat ein spezifisches Buchstabenhäufigkeitsprofil
- Beispiel Deutsch (nur A-Z): Englisch
  - relative Häufigkeit von A: 6.51% 8.17%
  - relative Häufigkeit von B: 1.89% 1.49%
  - relative Häufigkeit von C: 3.06% 2.78%
  - ....
  - relative Häufigkeit von Z: 1.13% 0.07%

In welcher Sprache ist ein Text geschrieben?

- Idee eines Erkennunsprogramms
  - Lies einen Text ein und bestimme die relativen Häufigkeiten der Buchstaben A-Z im Text
  - Vergleiche das so erhaltene Profil des Textes mit den Profilen verschiedener Sprachen
  - Gib die Sprache aus, deren Sprachprofil dem Textprofil am nächsten ist.

Ähnlichkeit von Profilen

- Sprachprofil (p_A, p_B,..., p_Z)
- Textprofil (t_A, t_B,..., t_Z)
- Mögliches Abstandsmass:

$$\Sigma_{\rho=A..Z} (t_\rho - p_\rho)^2$$

Testdaten

- Eigene Testdaten aus dem Internet