NLP Assignment 1

Dieses Assignment wird bewertet !

Die Zusammenarbeit unter Studierenden ist erwünscht, so lange sich diese auf die Diskussion von Konzepten oder Problemen mit `python` konzentrieren. Das Kopieren von Code ist nicht erlaubt, in diesem Falle werden alle Lösungen der beteiligten Parteien mit 0 Punkten bewertet; dazu werden alle Lösungen (manuell und automatisiert mit Plagiat-Checkern) auf Gruppenarbeit und Kopien untersucht.

Entwickeln Sie mit einem Entscheidungsbaum einen `language detector` basierend auf den relativen Häufigkeiten der 26 **Klein**buchstaben "*a*–*z*", das heisst ohne Umlaute, β und andere Sonderzeichen. Der Entscheidungsbaum soll höchstens von der Tiefe 4 sein und maximal 10 "Blätter" haben. Zum Erstellen des Entscheidungsbaumes darf **keine Library** verwendet werden.

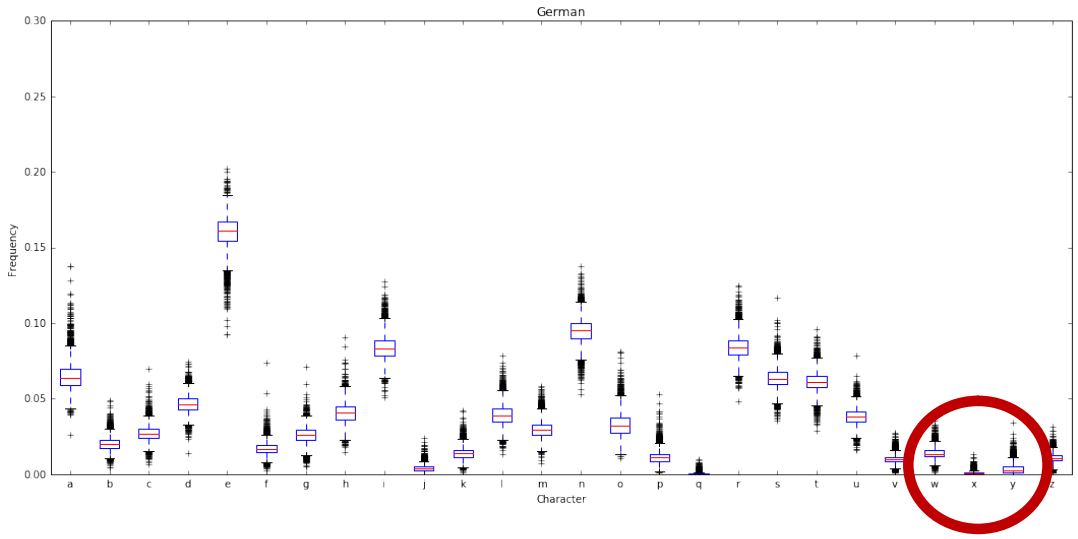
- Der `language detector` soll in der Lage sein zu erkennen, ob ein Text in englischer (E), deutscher (G) oder in einer anderen Sprache (X) verfasst wurde.
- Ermitteln Sie dazu zuerst die relativen Wahrscheinlichkeitsverteilungen der Buchstaben in deutschen und englischen Texten mit Hilfe der beiden Text-Sammlungen people_wiki_EN.csv und 10k-people-raw.csv.
- Stellen Sie in einem ersten Schritt die relativen Häufigkeitsverteilungen in diesen beiden Sprachen mit einem Box-Whisker Plot dar (siehe Beispiel auf der Rückseite).
- Visualisieren Sie anschliessend die Quotienten der Buchstabenhäufigkeits-Mittelwerte der beiden Sprachen um geeignete Indikatoren zu identifizieren (siehe Grafik auf der Rückseite).
- Der fertige `language detector` soll aus der Datei Language_test.csv pro Zeile die annotierte Sprache und einen entsprechenden Text einlesen und versuchen, die Sprache aus der Analyse des Textes zu bestimmen. Mit Hilfe der tatsächlichen Sprache soll eine *confusion matrix* erstellt und die *accuracy* bestimmt werden (als *accuracy* wird der Prozentsatz der richtig identifizierten Texte bezeichnet).
- **Das Ziel ist selbstverständlich, eine möglichst grosse *accuracy* zu erzielen !**

Hinweise:

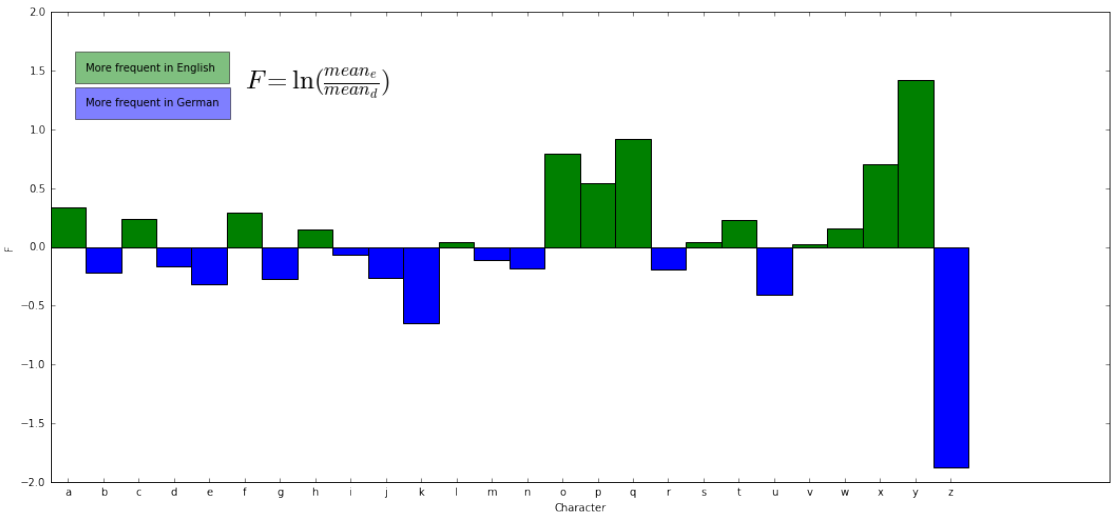
- Die mitgelieferte Datei Language_test.csv dient als Beispiel für die Formatierung. Effektiv getestet wird Ihr `language detector` mit einer umfangreichen Test-Datei.
- Falls die Häufigkeit des Buchstabens "*x*" für Deutsch grösser ungefähr 1% ist, dann haben Sie ein Problem mit dem Encoding (siehe Grafik Rückseite).
- Berücksichtigen Sie den Fall, dass in kurzen Texten (z.B. *Tweets*) gewisse Buchstaben allenfalls gar nicht vorkommen.

Abgabe : In Form eines `Jupyter Notebooks`
per Email an ivo.nussbaumer@fhnw.ch bis 17. März 2019

Box-Whisker Plot mit Ausreisser der Häufigkeitsverteilungen
der Buchstaben in den deutschen Texten aus 10k-people-raw.csv :



Quotienten der Häufigkeits-Mittelwerte deutscher
und englischer Buchstaben auf logarithmischer Skala :



Beispiel einer *Confusion Matrix*:

		detected language		
		E	G	X
actual language	E	30%		10%
	G	10%	20%	
	X	10%	10%	10%

accuracy = 60%