Data Science Course

Miguel Almeida and Inês Almeida

Instructions

Please download the files from:
https://github.com/miguelbalmeida/data_science_course

Files are in UTF-8 encoding

Folder “data” contains text from Wikipedia articles in multiple languages

Folder “tests” contains intermediate steps of the algorithm

Part 1: trigram counts and (corrected) probabilities

```
1: trigramCounts ← HashMap()
2: trigramProbs ← HashMap()
3: For each language L:
4:      trigramCounts[L] ← HashMap()
5:      totalCounts ← 0.0
6:      For each line in that file:
7:              For each trigram T in that line:
8:                      trigramCounts[L][T] ← trigramCounts[T] + 1.0
9:                      totalCounts ← totalCounts + 1.0
10:     uniqueCounts ← trigramCounts[L].size() // number of distinct trigrams
11:     /* trigramCounts should match the values in “counts_*.txt” */

12:     trigramProbs[L] ← HashMap()
13:     denominator ← totalCounts + uniqueCounts
14:     For each trigram T in trigramCounts:
15:             numerator ← trigramCounts[L][T] + 1.0
16:             trigramProbs[L][T] ← numerator / denominator
17:             trigramProbs[L][‘__UNKNOWN__’] ← 1.0 / denominator
18:     /* trigramProbs should match the values in “probs_*.txt” */
```

Part 2: detecting the language

```
19: S ← input sentence
20: For each language L:
21:     score[L] ← 0.0
22:     For each trigram T in S:
23:             If T is in trigramProbs[L]:
24:                     score[L] ← score[L] + log10(trigramProbs[L][T])
25:             Else:
26:                     score[L] ← score[L] + log10(trigramProbs[L]['__UNKNOWN__'])

/* Sentence's language is the one with highest score */
/* Scores will all be negative. That's normal. :) */
```