LING 354 Problem Set 1

|  |
| --- |
| Your Name: |

**Due: Friday, February 23 (11:59pm)**

Let’s talk about encoding language. Section 1.3.2 in Language & Computers may be helpful to consult if you’re getting stuck.

**1.** Suppose you’re trying to encode this message: *I am a computer*. How many bits would be required in ASCII, and how many in UTF-8?

**2.** Here is the UTF-8 encoding of a simple English message. What does it say?

01001001 00100000 01101101 01100001 01100101 01100100 00100000 01100001 00100000 01110100 01111001 01110000 01101111 00100001 00100001 00100001

**3a.** The Georgian letter ფ is character number 4324 in Unicode. First, what is that character number in binary? Second, what is its UTF-8 encoding?

**3b.** The emoji U+1F64B, HAPPY PERSON RAISING ONE HAND, as you may be able to tell from its name, is Unicode character number 128587 in base 10. What is that character number in binary? And what does that make its UTF-8 encoding?

**3c.** In the emoji above, you had to use 4 bytes to represent the character in UTF-8. But the first byte had all zeroes after the 11110 prefix. Why couldn’t it be represented as a three-byte character in UTF-8?

4. Let’s consider how the order of characters in the Unicode scheme affects the efficiency of storing different writing systems. We’re operating entirely in UTF-8 for this question and its subquestions.

**4a.** English, aside from certain rare punctuation marks, relies entirely on ASCII characters, so how many bits does it take to represent *Tokyo* as a five-letter English word in UTF-8?

**4b.** But technically, we haven’t accurately represented the name of the city in this way; Japanese distinguishes long and short vowels, so we really ought to write *Tōkyō* instead. The character *ō* is a two-byte character in UTF-8 (character number 333), so how many bits do we need to represent this in UTF-8?

**4c.** If we wanted to write out *Tokyo* phonetically in Japanese, we could use the hiragana writing system: とうきょう. The hiragana characters are Unicode characters 12352-12447. How many bits would the hiragana spelling take up?

**4d.** That said, the standard Japanese representation for the name of the city is 東京, with each character representing one syllable. Both of these characters are three bytes in UTF-8, so how many bits do we need?

**4e.** Given such examples and those we talked about in class, what sort of factors affect the efficiency of a writing system in Unicode? Do you have any guesses for a lower and upper bound on the number of bits needed to represent the name of this city in UTF-8 across different writing systems, and what writing systems might reach those bounds?

Moving on to speech recognition, consult Chapter 2 of Language Files if you are getting stuck on this.

**5.** I’ve got some recordings that I want to use to train my speech recognition system. I have the sounds [ibi], [isi], [izi], and [iki] recorded here, but I forgot which label goes on each recording. Can you label the spectrograms correctly?

|  |  |
| --- | --- |
| Macintosh HD:Users:gdoyle:Desktop:Screen Shot 2018-02-18 at 2.20.37 PM.png | Macintosh HD:Users:gdoyle:Desktop:Screen Shot 2018-02-18 at 2.31.52 PM.png |
|  |  |
| Macintosh HD:Users:gdoyle:Desktop:Screen Shot 2018-02-18 at 2.21.44 PM.png | Macintosh HD:Users:gdoyle:Desktop:Screen Shot 2018-02-18 at 2.20.19 PM.png |
|  |  |

6. We discussed in class how language models and speaker-adaptive methods can improve a speech recognition system. Let’s put that into practice. For help in identifying phones from a spectrogram, see Section 2.6 of Language Files.

|  |
| --- |
| **Macintosh HD:Users:gdoyle:Desktop:Screen Shot 2018-02-18 at 4.05.18 PM.png** |

We’ve received a spectrogram of someone’s speech, and our speech recognition system has figured out the first two phones of it form the word *the*. But the system hasn’t been fully coded yet, so we need to figure out what the remaining four phones might be.

**6a.** Let’s start by figuring out phones 1 and 4. They have the same manner of articulation (e.g., fricative, stop, nasal, etc.) – what is it? How do you know?

**6b.** Phone 2 is clearly a vowel, because it has a formant structure. So that leaves phone 3. What’s its manner of articulation, and how do you know?

**6c.** One of phones 1, 3, and 4 is voiced. Which one, and how do you know?

**6d.** The formants of the vowel are: F1=637Hz, F2=1774Hz, F3=2680Hz. If the speech recognition system is using the speaker-independent values in the table on the last page, what vowel is most likely?

**6e.** Give three possible English words that are consistent with your answers to questions 6a-6d. Which do you think is the best option, and why?

**6f.** Let’s try to improve our guess at the mystery word by bringing in contextual information. For each of the following contexts, what would be your top choice for the mystery word?

|  |  |
| --- | --- |
| *The hotel reserved a suite for the …* |  |
| *Quit living in the …* |  |
| *My mom says I’m the …* |  |
| *Not the …* |  |

**6g.** How does the syntactic context interact with the phonetic information to inform your choices in 6f? Were they all equally easily to decide on, or were some harder than others? Was the syntactic context more important in some cases than others?

**6h.** Give an example of a context that is very helpful, and one that’s unhelpful, in distinguishing between the possible words. What’s the difference between the contexts?

American English vowel formants (based on Hillenbrand et al 1995)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **IPA** | **Example** | **F1** | **F2** | **F3** |
| i | beet | 342 | 2322 | 3063 |
| ɪ | bit | 427 | 2034 | 2684 |
| ɛ | bet | 580 | 1769 | 2605 |
| æ | bat | 688 | 1952 | 2401 |
| ɔ | bought | 652 | 997 | 2538 |
| ʌ | but | 623 | 1200 | 2550 |
| u | boot | 378 | 988 | 2343 |