LING 354 Problem Set 2

|  |
| --- |
| Your Name: Jorge Arellano |

**Due: Sunday, March 11 (11:59pm)**

**1a.** Let’s start off with Levenshtein edit distance. What is the Levenshtein distance between the following word pairs? Include a derivation for each with a minimal number of edits. Format the derivation like the example, adding a D, I, or R (delete, insert, replace) for each intermediate step.[[1]](#footnote-1)

|  |  |  |  |
| --- | --- | --- | --- |
| **Error** | **Correction** | **Distance** | **Derivation** |
| *dessapoint* | *disappoint* | 3 | *dissapoint* (R)  *disapoint* (D)  *disappoint* (I) |
| *evrey* | *every* | 2 | Eveey(R) (this is unnecessary)  Evey(D)  Every(I) |
| *fullfil* | *fulfill* | 2 | Fulfil(R)  Fulfill(I) |
| *exquiset* | *exquisite* | 3 | Exquist(D)  Exquisit(I)  Exquisite(I) |
| *equipt* | *equipped* | 4 | Equip(D)  Equipp(I)  Equippe(I)  Equipped(I) |
| *jeprodise* | *jeopardize* | 4 | Jeoprodise(I)  Jeoparodise(I)  Jeopardise(D)  Jeopardize(R) |

**1b.** Damerau-Levenshtein edit distance extends Levenshtein by allowing the transposition of two adjacent characters to count as a single move instead of two. Which word pairs in (1a) have smaller Damerau-Levenshtein distances than their Levenshtein distances?

Every, exquisite

**1c.** According to Levenshtein distance, the non-word *momentus* is equally far from the words *momentum* and *momentous*. Which do you think is the more likely intended word? What information are you using to make your guess? Is your guess a typographical error or a spelling confusion (see Sect 2.2.1, L&C)?

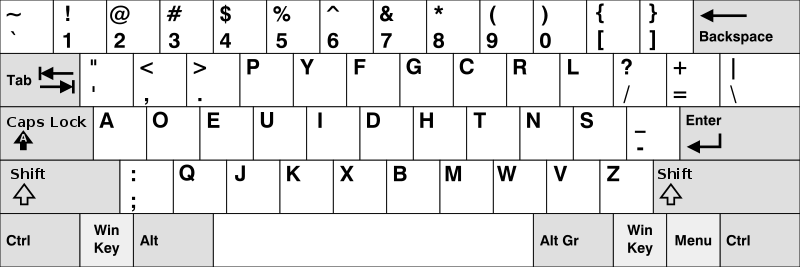
Spelling confusion, because “momentous” sounds like it should be spelled like “momentus”, the writer may have forgotten how to spell the word but English is NOT a transparent writing system, so the mistake is common.

**1d.** Levenshtein distance also says *profer* is equally far from *proffer* and *prefer*. Let’s assume this mistake was observed on a QWERTY keyboard. What’s an argument for *proffer* being more likely? What’s an argument for *prefer* being more likely?

Proffer: The hand/digit movement for using an “o” is more of a premeditated effort, because the right hand is only used for “O” & “P”

Prefer: The author of the mistake may believe that the word “prefer” is spelled “profer” because it they may be mispronouncing the word/

**1e.** Now suppose this misspelling came from a Dvorak keyboard instead, which looks like this:



Does this change your opinion of how *prefer* might have been misspelled as *profer*? Why?

No, because a Dvorak keyboard would not make levenshtein distance change, they are both equally close to the words in question. In this case, it would make the “mistake” indistinguishable between either of the two cases. The Dvorak keyboard would prove that a mistake is a 50/50 chance, rather than a word confusion probability. Maybe it makes the adjustment on a syntactical basis instead.

**2a.** Let’s build a really simple discriminative spelling corrector. We’re going to score each possible correction by estimating that word’s frequency from a [large corpus](https://books.google.com/ngrams), and dividing the frequency by its Damerau-Levenshtein distance from the misspelling. The highest-scoring word is then chosen as the correction. Consider the misspelling *theer*, and fill in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Possible Correction | Frequency per 100 words | D-L distance from *theer* | Score |
| *three* | 0.04 | 2 | .02 |
| *there* | 0.11 | 1 | .11 |
| *their* | 0.21 | 1 | .21 |
| *thee* | 0.009 | 1 | .009 |
| *the* | 4.64 | 2 | 2.32 |

**2b.** You may be feeling unsatisfied with the highest-scoring correction here. You’ll probably feel even less satisfied with the corrector’s answer for the misspelling *momentus*.

|  |  |  |  |
| --- | --- | --- | --- |
| Possible Correction | Frequency per 100 words | D-L distance from *momentus* | Score |
| *momentum* | 0.0012 | 1 | .0012 |
| *momentous* | 0.0002 | 1 | .0002 |
| *the* | 4.64 | 1 | .5156 |

**2c.** So what’s going wrong here? Is the idea of combining frequency and edit distance wrong? Or is it something in the implementation? You don’t have to totally fix the corrector, but propose **two** changes to it to bring it closer to human expectations. You can add new information, or you can change the way we combine information.

The issue is that if you stick to probabilities, it will hurt you the most. The fact that the word “the” is the most common word in our language there will be high bias towards choosing that word.

1. Weighing the probabilities as to make specific “high frequency” words less likely to be chosen as the correct term. Maybe given the word “the” less significance if the “mistake” is more than 3 letters could be useful.
2. Possibly stacking different models like probabilistic probabilities for word confusing likelihood along with syntax checkers that use other linguistic rules to choose the correct word. But still keeping D-L distance as the main basis somehow.

3. Let’s dig into using letter *n-*grams to determine whether a word is spelled correctly. (see pg. 40-42 of L&C).

**3a.** I want to decide if the string *ghotis* is a valid English word using a nonpositional bigram array. The array, unfortunately, tells me it’s valid. For each bigram in *ghotis*, give an example of a real English word containing it.

**Ghotis**

**Ghosts**

**Hosts**

**Goats**

**Hot**

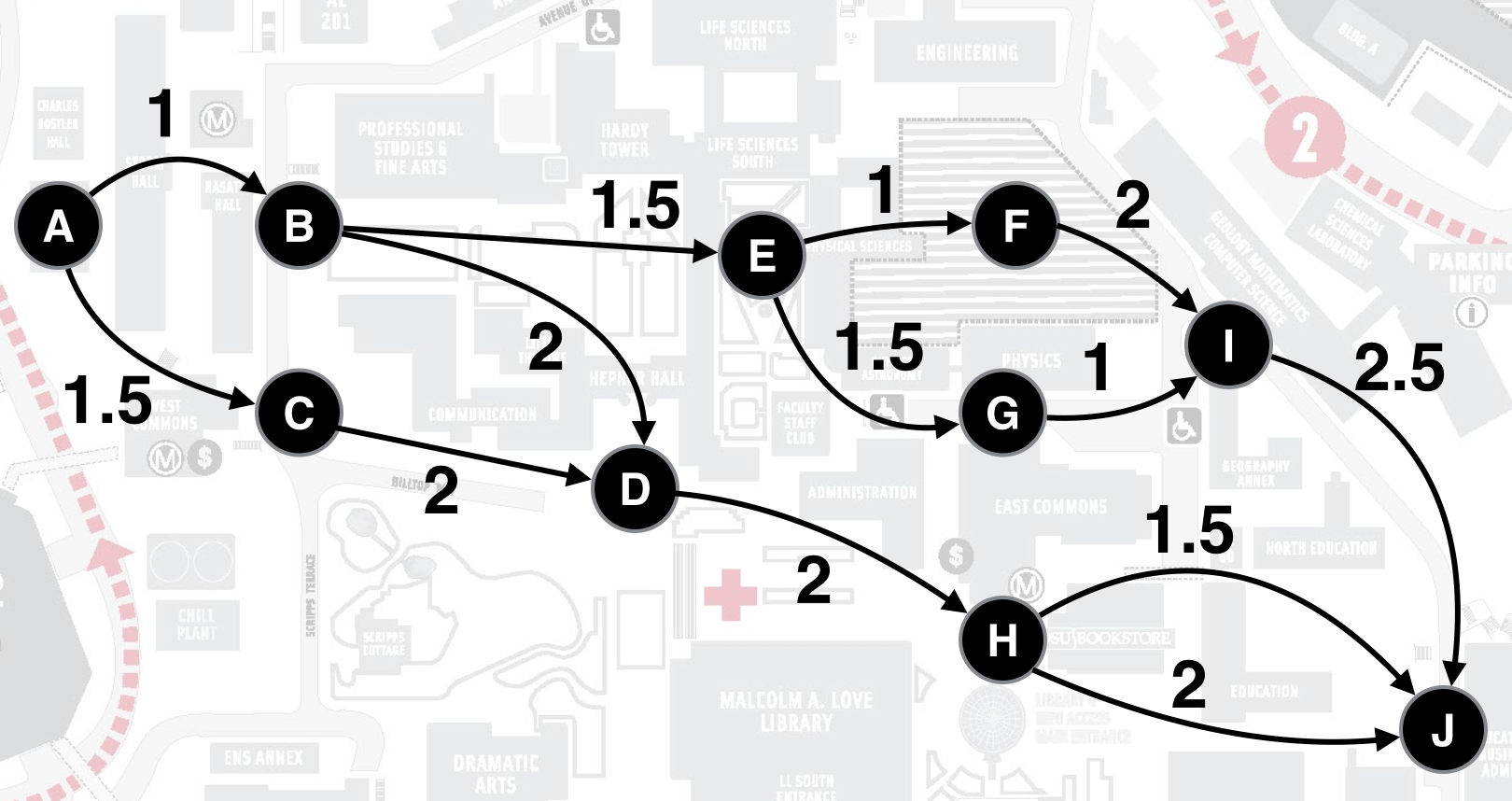
**Hottest**

**3b.** Will switching to a nonpositional **tri**gram array fix the problem? If so, what trigram is invalid in English? If not, give examples of words for each trigram.

**3c.** What if we switch to a **positional** trigram array? Is *gho* a valid start to an English word? Is *tis* a valid end to an English word?

**3d.** The *ghotis* example might have left you concerned about the validity of the n-gram approach, but it has its successes, too. For instance, suppose I tried to type *slam*, but accidentally replaced *s* with one of its eight QWERTY neighbors. How many of these eight typos would be caught by a **positional bigram array**?

4. I’m trying to figure out the most effective way for me to get to class each day. I’ve plotted out my potential routes as a directed acyclic graph, but now I need your help.



**4a.** For each of the following pairs of nodes, state whether there is an ordering between them or not, and if so, what it is. (e.g., D and G do not have an order between them, but D < J).

|  |  |  |
| --- | --- | --- |
| A, E | F, G | B, I |
| A<E | NO | B<I |

**4b.** If I want to find the fastest route to node H using dynamic programming, what all nodes do I need to find the fastest route to?

A,B,D,H

**4c.** What is the fastest route from A to J?

A,B,D,H,J

**4d.** Suppose I decided the topological ordering and dynamic programming approach was too much work to implement, and instead used a *greedy algorithm*, where I always took the cheapest path out of whatever node I was currently on. What route would the greedy algorithm take from A to J? How much longer would that trip take than your answer to 4c?

A,C,D,H,J .5 slower

5. Build syntactic trees for the following sentences using the CFG on the next page (or the handout from class). I recommend using <http://mshang.ca/syntree/> to draw trees.

|  |
| --- |
| **5a.** The dugong eats quietly |
|  |
| **5b.** We bought some sextants at the shore |
|  |
| **5c.** An angry old biker insulted my clothes |
|  |
| **5d.** I measured my life with tablespoons |
|  |

**5e.** Let’s deal with a case of syntactic ambiguity. Draw two valid trees for the sentence: *She saw the moose with the binoculars.*

|  |
| --- |
| Interpretation 1: She’s looking through the binoculars |
|  |
| Interpretation 2: The moose has the binoculars |
|  |

**5f.** Suppose we have the PCFG below. What are the probabilities of the two syntactic structures (we’ll leave out the individual lexical probabilities since they’re the same in both cases)? Which interpretation is favored by this PCFG? Does that fit with your intuition?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PSR** | **Prob** | **PSR** | **Prob** | **PSR** | **Prob** |
| S → NP VP | 1 | NP → Det N | 1 | PP → P NP | 1 |
| VP → VP Adv | 0.12 | VP → V NP NP | 0.10 | N → Adj N | 0.17 |
| VP → VP PP | 0.14 | VP → V | 0.21 | N → N PP | 0.26 |
| VP → V NP | 0.43 |  |  | N → lexical item | 0.57 |

1. These are all examples of real misspellings from Fawthrop, compiled in [Mitton’s 1980 corpus of misspellings](http://ota.ox.ac.uk/headers/0643.xml). [↑](#footnote-ref-1)