Connor Hanley

LIN 313

Prof Venkat

July 22, 2022

Assignment 2

**Problem 1.**

*1.i.* I think that in this case, the most important factor between precision and recall is recall. Whenever we’re dealing a life-threatening disease like cancer (or most diseases I feel), it is always better to be a bit more heavy-handed in diagnoses and treatment. In other words, I think it is more beneficial to have more false positives (and the side-effects generated by that) than more false negatives. Especially with cancer, I can’t imagine how absolutely distraught someone or someone’s family would be if they were given a false negative, and then shortly after find out that they are really terminally ill over something that was easily preventable at an early stage.

*1.ii*. While I did say for the example above that in cases of life-threatening diseases we ought to steer on the side of recall, in this case, when dealing with the potentials of death caused as a side-effect, I think it’s a bit more nuanced. For this drug, the factors that need to be weighed are the effectiveness of the drug, compared with it’s precision and recall, and finally the cancer it’s actually treating. For example, if it’s a normally easily treatable cancer, with treatments that are practically guaranteed to work, then we probably ought to steer on the side of precision, even if it is extremely effective. But if it is for a highly fatal, and extremely tragic type of cancer, then if it is highly effective it might be just a risk that you have to inform patients about whenever offering this as a treatment. For sufferers of extreme types of cancer, it might be a sort of “between a rock and a hard place” situation, either a death via cancer or a potential cure, with a possibility of death.

*1.iii.* Identifying cars, and driving generally, seems a lot safer to leave up to the demonstrably capable hands of humans. But because the question is about precision versus recall, I think that it could arguable go either way as to whether precision or recall ought to be favored. I think that for me, at least, precision might be more important. For example, I think that it would be better to classify more things as running red lights (and then potentially have someone to review the classification for efficacy etc.), then for there to be less things caught as red lights. So, I would rather favor precision over recall.

*1. iv.* Because oftentimes, at least from what I know about weeds, they might be a sort of invasive species. So we might prefer precision in our removal of these weeds so as to prevent any possible weeds slipping through. It’s a very harsh approach, but invasive species can not only kill your garden, but also can endanger you and your communities (e.g., being highly flammable), as well as the local environment, as native plants and native fauna are closely interlinked.

**Problem 2.**

I wrote a little program to do the calculations because like a good student of the College of Liberal Arts, I am unable to count.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NODE A | NODE B | NODE C | NODE D |
| Round 1 | 0.375 | 0.375 | 0.125 | 0.125 |
| Round 2 | 0.3125 | 0.4375 | 0.625 | 0.1875 |
| Round 3 | 0.34375 | 0.34375 | 0.9375 | 0.21875 |
| Round 4 | 0.328125 | 0.390625 | 0.109375 | 0.171875 |
| Round 5 | 0.3359375 | 0.3828125 | 0.859375 | 0.1953125 |
| Round 6 | 0.33203125 | 0.37890625 | 0.9765625 | 0.19140625 |
| Round 7 | 0.33398438 | 0.38085938 | 0.95703125 | 0.18945313 |
| Round 8 | 0.3330078 | 0.38183594 | 0.9472656 | 0.19042969 |
| Round 9 | 0.3334961 | 0.3803711 | 0.95214844 | 0.19091797 |
| Round 10 | 0.33325195 | 0.38110352 | 0.95458984 | 0.19018555 |
| Round 11 | 0.33337402 | 0.38098145 | 0.9509277 | 0.19055176 |
| Round 12 | 0.333313 | 0.3809204 | 0.9527588 | 0.19049072 |
| Round 13 | 0.3333435 | 0.38095093 | 0.9524536 | 0.1904602 |
| Round 14 | 0.33332825 | 0.3809662 | 0.952301 | 0.19047546 |
| Round 15 | 0.33333588 | 0.3809433 | 0.9523773 | 0.1904831 |
| Round 16 | 0.33333206 | 0.38095474 | 0.9524155 | 0.19047165 |
| Round 17 | 0.33333397 | 0.38095284 | 0.95235825 | 0.19047737 |
| Round 18 | 0.33333302 | 0.38095188 | 0.95238686 | 0.19047642 |
| Round 19 | 0.3333335 | 0.38095236 | 0.9523821 | 0.19047594 |
| Round 20 | 0.33333325 | 0.3809526 | 0.9523797 | 0.19047618 |
| Round 21 | 0.33333337 | 0.38095224 | 0.9523809 | 0.1904763 |
| Round 22 | 0.3333333 | 0.38095242 | 0.9523815 | 0.19047612 |

Definite overkill given what you asked, but it was a fun little problem trying to figure out how to model the system.

For the second system, my hypothesis was that the addition of the new nodes F, G, and H would affect the final rank of the A node; however, I think that given the previous results, it ultimately must balance out at the same position. I don’t have any mathematical justification for this, but I think that because E and G are “incestuous” (in that they feed back to one another, and then feed to C), and further that F is removed in the first round, that it shouldn’t affect the system too much.

I made a pretty 3D graph to verify my hypothesis:

It seems I was dead wrong actually, in this case node B comes out on top. In fact, the results seem to be thrown on their heads – node C for example is now the last place compared to the old nodes. And similarly, the “incestuous” nodes of E, F, and G almost immediately fall down to zero. G doesn’t even stand a chance! [Here’s a link to the code as well :)](https://github.com/hanleyc01/LIN-313/blob/main/pagerank/src/main.rs)

**Problem 3.**

*3.i.* I think that the conjunction between two search terms ought to limit the search results, as we’re saying that the search result *must* contain both of these terms. So anything that might have one or the other is excluded. We can think of this actually just by inspecting the truth table for conjunction, the sentence p && q is true just in case both p is true and q is true. In all other cases it is false. So, 1/3 of all search results that contain either p or q are included. The rest are thrown out.

*3.ii.* I think that the disjunction between two terms does limit the search results to some extent, as it forces the result to contain either the first term or the second term, but it does not limit it to the same extent as conjunction. We can think of this gain by looking at a semantic interpretation of the sentence p || q, which is true just in case either p is true, or q is true.

*3.iii.* I think that negating a search term ought to produce really anything. The results it produces are just one step below looking at an index of everything that Google has logged (maybe if the negated terms are very general?). Thinking of it regarding sets, suppose we have the set A = {apple, apple, orange, orange, orange, lemonade, lemonade, caffeine}. If we were to search for the difference of {caffeine} \ A, we would get {apple, apple, orange, orange, orange, orange, lemonade, lemonade}. If we think of these operations are the steps in applying a search, then we can say that {caffeine} \ A is the result of our search, and as we can see, the range of possible search terms is only decreased by a single term. But, we can also think about negating non-singleton sets, so for example, the difference {apple, orange, lemonade, caffeine} \ A produces the empty set {}. In this case we’ve completely excluded any possible search result.

**Problem 4.**

I created a monster, but it does the job!

/^[1-9]+\w+ [A-Z]\w+ \*\d\*\w+\ \*\w\*, \w+( \w\* \*\w\*) \*\d\*$/

Which matches:

1234 East 23rd Street, Austin TX 78712

12A Something Lane, Experiment GA 12121

9876 Research Avenue, Las Cruces NM 57501

9201 Colberg Dr, Austin TX 787749

146 Eiglehart Rd, Austin TX 78737

(It kind of looks like a dance scene near the end).

**Problem 5.**

*5.i.* I think that it is impossible to talk about the amount of books in the Library of Babel because, at least it seemed to me, the definition being every possible combination of 22 letters, or from the website, 1,312,000 characters, we can think about an infinity of combinations. For example, the repeating letter “aaaaaa…” is a unique combination of our 22 character set, and are as many of these as there are letters. Similarly, “abaaaa…”, “abbaaaaa…”, and every other possible change is also present.

A similar problem exists in physics where, “Statements like ‘a quantity x has a completely definite value’ (expressed by a real number and represented by a point in the mathematical continuum) seem to me to have no physical meaning” ([quoted in this article by Nicolas Gisin](https://link.springer.com/article/10.1007/s11229-021-03378-z)). We’re introducing for ourselves a definition which necessarily implies talking about definite quantities and magnitudes as nonsensical. It also reminds me of recursive definitions, like “if p is a well-formed formula, and q is a well-formed formula, then p && q is a well-formed formula…”

*5.ii.* I think the simplest way we can talk about the Library of Babel is to maybe compare to a linked list. If we say that *c* is one of the 22 possible characters, then we can say, the Library of Babel text is either the empty set Nothing, or it is the Construction of some *c* to another Library of Babel text. So {} is part of the Library of Babel, but so is the Construction(*c*, {}). Similarly, the Construction(*c*, Construction(*c*, {})) is part of the Library of Babel.

To do a simple test if whether something is in the Library of Babel, we match it to either being the empty set of Nothing, or we say that it is the Construction of something else to something else. Then we match on this Construction of something else to something else, and see if it is the empty set, or if it is also a Construction of something else to something else, etc. Haskell actually has a really expressive way to say this:

data Text = Nothing | Cons c Text

And we can verify something as being a text through a recursive function

isText :: Text → Bool

isText text =

case text of

Nothing → True

Cons c’ → isText c’

\_ → False

/- the “hole” case -/

As you can see, the function recursively descends the entire tree to determine if something is indeed a Text.

**Problem 6.**

*6.i.*We can look for different forms of a verb by using an all-uppercase form of the verb, so for example searching for different forms of eat, we can search EAT.

I searched for KEEP, and in total, it returned “KEEP THE” 88820 times, “KEEP IT” 42313 times, and the next form of the verb “KEEPING THE” returned 22146 total occurrences. I was limited to searching for KEEP \*, as the search for KEEP by itself apparently provided too many results.

*6.ii.*  I just used google search’s “synonyms for…” function and got back as synonyms for “keep”: “remain”, “continue to be”, “stay”, “maintenance”, “upkeep”, “support”, “fortress”, “fort”, “stronghold”, and “tower” (and a few others.

*6.iii.* I think while saying that “keep” and “continue to be” are synonyms, they definitely are used differently. For example, “I will do my best to keep in shape and continue to be appealing to you”, and “While the people of God keep the covenant, they continue to be his people”. Keep here, while being a verb as well, seems to be more associated with adjectives, so I “keep healthy”- but that’s a hard generalization that gets broken with “keep the covenant” and “continue to be watchful”.

*6.iv.* Here’s a little table for the results:

|  |  |  |  |
| --- | --- | --- | --- |
| KEEP \* | CONTINUE TO BE \* (these are actually a selection from the top 20, because the first 10 are practically a repeat of CONTINUE TO BE + DETERMINER) | KEEP \* ADJ NOUN | CONTINUE TO BE \* ADJ NOUN |
| KEEP THE | CONTINUES TO BE A | KEEP AN OPEN MIND | CONTINUE TO BE AN INTEGRAL PART |
| KEEP IT | CONTINUE TO BE A | KEEP A LOW PROFILE | CONTINUES TO BE A MAJOR CONCERN |
| KEEP YOUR | CONTINUE TO BE THE | KEEP A CLOSE EYE | CONTINUES TO BE A MAJOR PROBLEM |
| KEEP UP | CONTINUES TO BE A | KEEPING A CLOSE EYE | CONTINUE TO BE A DRIVING FORCE |
| KEEP IN | CONTINUES TO BE AN | KEEP A STRAIGHT FACE | CONTINUE TO BE A HIGH PRIORITY |
| KEEP A | CONTINUE TO BE , | KEPT A LOW PROFILE | CONTINUE TO BE A MAJOR PROBLEM |
| KEEP THEM | CONTINUE TO BE IN | KEEPING A LOW PROFILE | CONTINUE TO BE AN IMPORTANT PART |
| KEEP YOU | CONTINUE TO BE ONE | KEEPING AN OPEN MIND | CONTINUES TO BE A HOT TOPIC |
| KEEPING THE | CONTINUE TO BE SO | KEEP A CLOSE WATCH | CONTINUES TO BE A TOP PRIORITY |
| KEPT THE | CONTINUE TO BE ABLE | KEPT A CLOSE EYE | CONTINUE TO BE A COMPETITIVE PARTY |

*6.v.* *and vi.* For the search “PRON KEEP \* ADJ NOUN”, the top results are either “YOU …”, “I …”, or “WE …”, with a single “HE …”, whereas for “PRON CONTINUE TO BE \*”, the top pronouns are “IT …”, “I …”, “HE …”, “SHE …”, “THEY …”, and “WE …”. At least, intuitively, from my understanding of English, this does make sense. I just don’t know if my hypothesis approaches actually what’s going on in the difference, and I find it hard to try and point out how “keep” and “continue to be”.

**Bonus 1.**

I think that the relationship between prior knowledge of a topic and the structure of a database being searched is actually directly proportional to the amount of time one would have to take to navigate the database in search for the information you desire. Very wordy way of putting “knowing what your looking for is a lot easier than not”, but I think it makes some very intuitive sense. For example, the Corpus that we used is a lot easier for us to navigate if we have at least a passing knowledge about regular expressions *because* of the structure that the corpus has (i.e., the metadata or whatever that is provided and can be searched through). I think even in the case of an “unstructured” database, what we’re doing either way is creating a structure, either beforehand or afterwords, and if we have an effective knowledge about the way that this database is structured we can produce amazing results.

**Bonus 2.**

I searched both the BNC and COCAE for “\* NOT \*”, which I think should provide some information about auxiliary verbs.

For COCAE, the top results are “IS NOT A”, “CAN NOT BE”, “’S NOT A”, “MAY NOT BE”, “I’M NOT SURE”, “WILL NOT BE”, “HE’S NOT THE”, “YOU’RE NOT GOING”; and for BNC, the top results are “CAN NOT BE”, “IS NOT A”, “WILL NOT BE”, “SHOULD NOT BE”, “WOULD NOT BE”, “IS NOT THE”, “MAY NOT BE”, and “HAD NOT BEEN’.

I think the main difference to note is the more prominent use of contractions in American English, NOT seems to be often paired a pronoun+verb contraction, whereas this does not seem to be as popular for British English. That being said, I think I can also talk about some anecdotal experience about how I (and other speakers of American English), really seem to enjoy in our speech the contraction of “not” itself into “don’t”, or “ain’t”, or “isn’t” or “couldn’t”. But, that being said, I’m not sure if that is the for British English (though I can’t imagine that they don’t use contractions!)