Connor Hanley

LIN 313

Prof Venkat

July 22, 2022

Assignment 2

1. Precision and Recall

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,

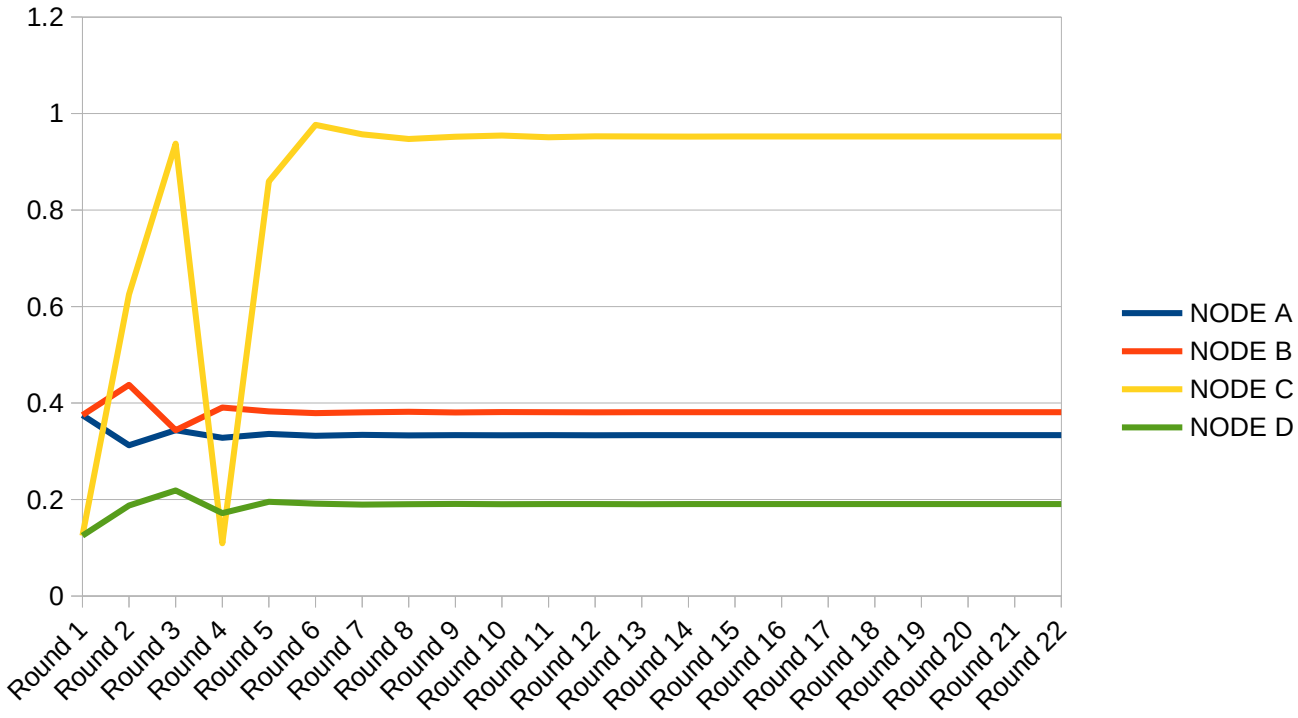
precision might be more important. For example, I think that it would be better to classify more things as running red lights (and then potentially have someone to review the classification for efficacy etc.), then for there to be less things caught as red lights. So, I would rather favor precision over 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.

2. PageRank

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.

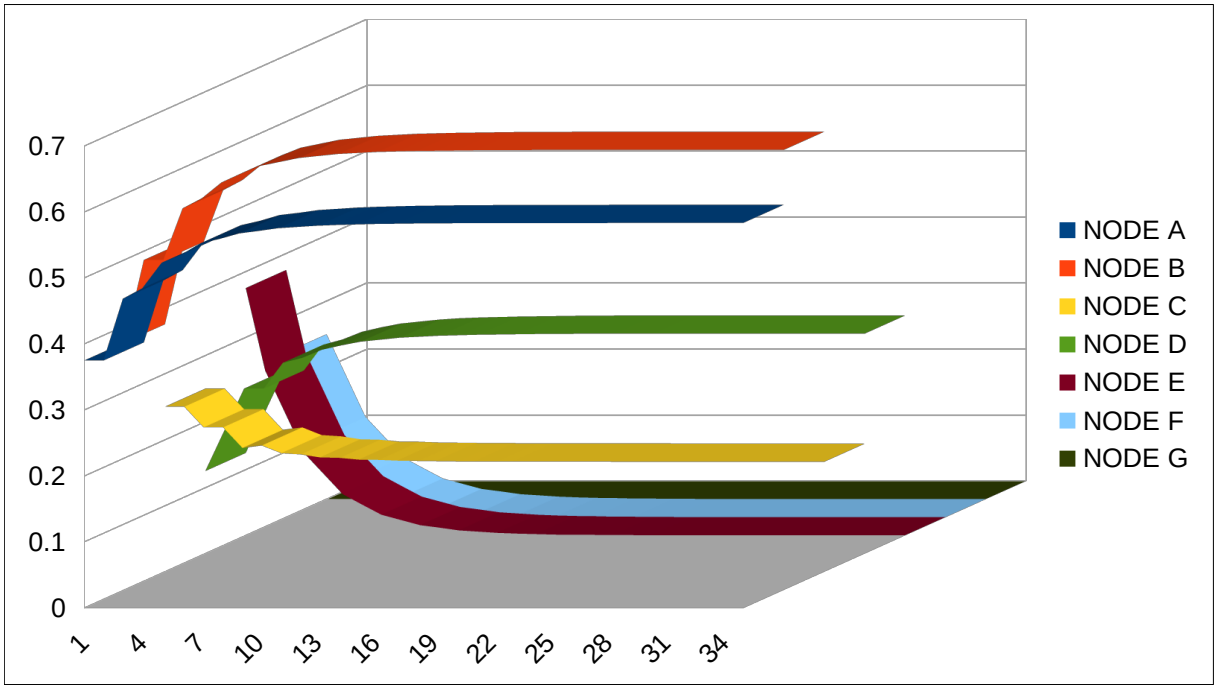
	NODE A	NODE B	NODE C	NODE D
Round 1	0.375	0.375	0.125	0.125
Round 2	0.3125	0.4375	0.625	0.1875
Round 3	0.34375	0.34375	0.9375	0.21875
Round 4	0.328125	0.390625	0.109375	0.171875
Round 5	0.3359375	0.3828125	0.859375	0.1953125
Round 6	0.33203125	0.37890625	0.9765625	0.19140625
Round 7	0.33398438	0.38085938	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 :)

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.