Analysis of EMR application results - DIRT algorithm

Hanna Hayik 207442054

F1-Measure:

We calculate *F1-measure* here as the harmonic mean of precision & recall.

Wikipedia defines this measure as the following:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Precision and *Recall* are defined as:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

We show the F1-measure calculation for both input sizes (big & small), we start with the big set:

I chose a threshold **0.38** for the big dataset, I believe it shows accuracy and is a rational threshold considering the results.

$$TP = 425$$
$$FN = 144$$
$$FP = 4$$
$$TN = 21$$

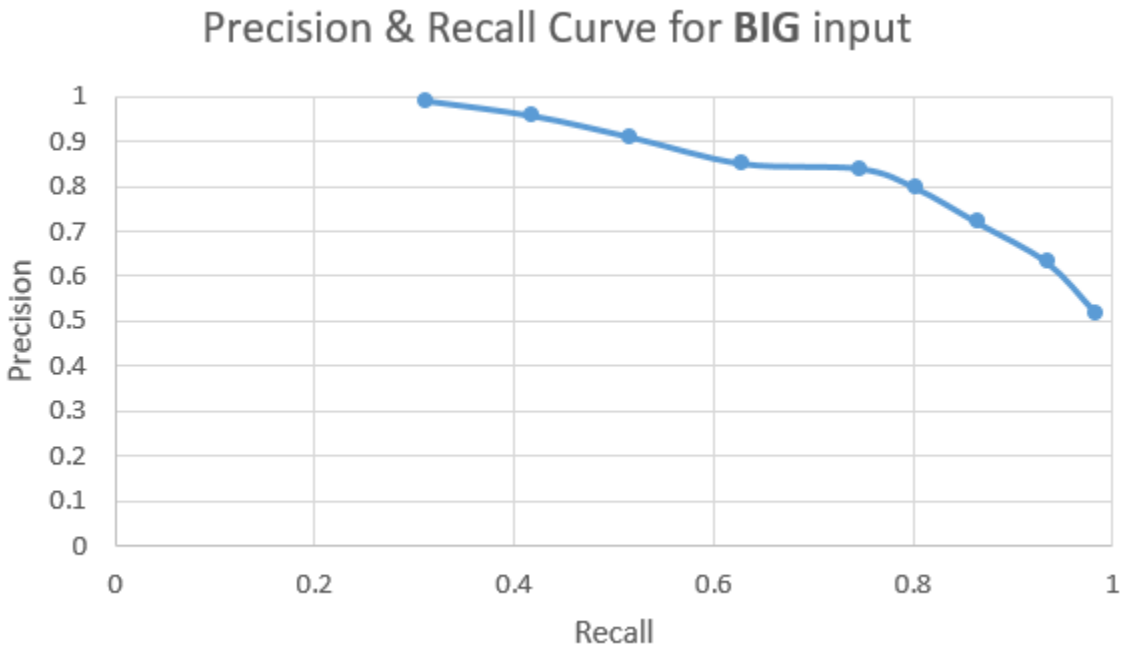
$$Precision = 0.9106759906759907$$
$$Recall = 0.7469244288224957$$

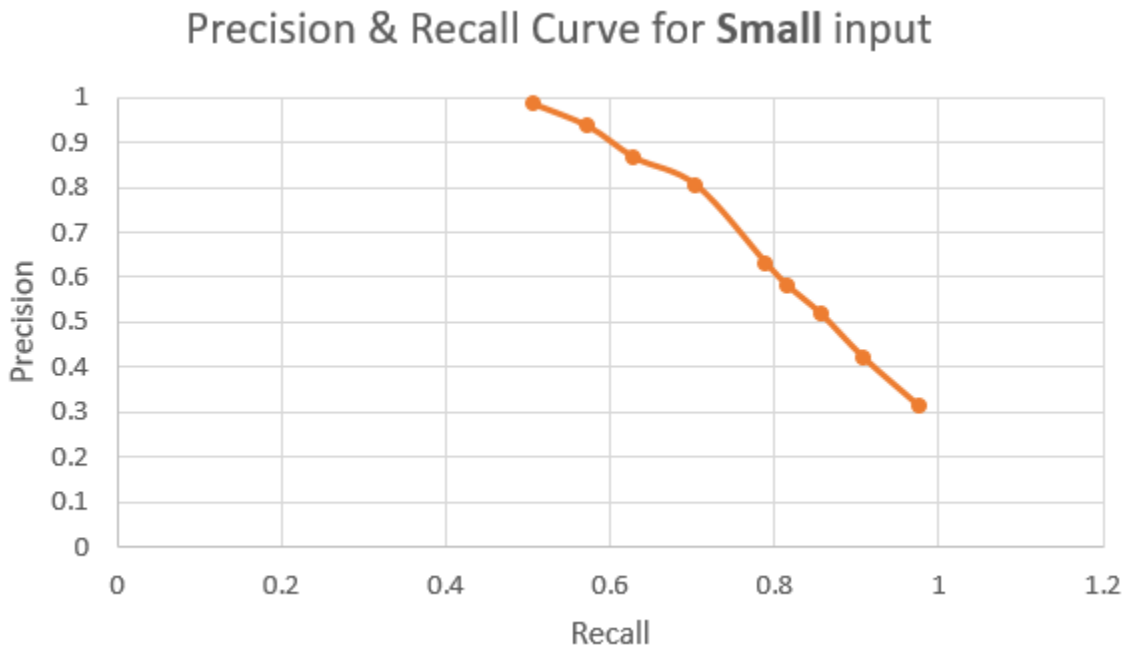
$$F1 = 0.8517034068136273$$

For the small dataset and threshold of **0.35**, we got the following results:

$$TP = 75$$
$$FN = 69$$
$$FP = 2$$
$$TN = 8$$

$$Precision = 0.79025974025974$$
$$Recall = 0.5208333333333334$$
$$F1 = 0.6278633043047587$$





Error Analysis:

TruePositive	FalsePositive	TrueNegative	FalseNegative
X result in Y X contribute to Y 0.8302588382	X see in Y X suggest Y 0.8013608	X begin with Y X characterize by Y 0.2115166	X make Y X manufacture Y 0.26953316806
X convert into Y X convert to Y 0.77016619	X help Y X control Y 0.89804116	X afford Y X give Y 0.0195090671	X consist of Y X compose of Y 0.13592133
X lead to Y X cause Y 0.9810221	X avoid in Y X use in Y 0.39305497	X confuse with Y X include Y 0.03206097	X confound with Y X differ from Y 0.05282804
X give Y X provide Y 0.84166699	X suggest Y X see in Y 0.8013608	X characterize by Y X distinguish fromY 0.02643938	X treat by Y X treat with Y 0.22600180
X cause by Y X result from Y 0.5564542	X expose to Y X die of Y 0.3518212	X indicate for Y X correct Y 0.12855690	X compose of Y X contain Y 0.11024104

Each category contains some examples from my output.

We can the examples in *TruePositive,* from humans point of view, the 2 sentences in every example DO relate to each other and according to our threshold, these examples are good.

Looking at *FalsePositive,* we got a score above the threshold although IMO every 2 paths there don't really relate or do they have the same meaning.

Continuing to *TrueNegative* column, we notice that from semantic point of view, those paths don't interfere or relate to each other, and they got a right score (low score) indicating a weak or non-existent relation.

Last column is *FalseNegative*, although it's easy to notice that those paths do relate to each other for example: X consist of Y & X compose of Y, which basically have the same meaning but still got a low score indicating low or no similarity at all, Why ? I believe it's because the lack of features between the two paths which can be traced to the quality of the input or too much filtering of paths which lead to these results.