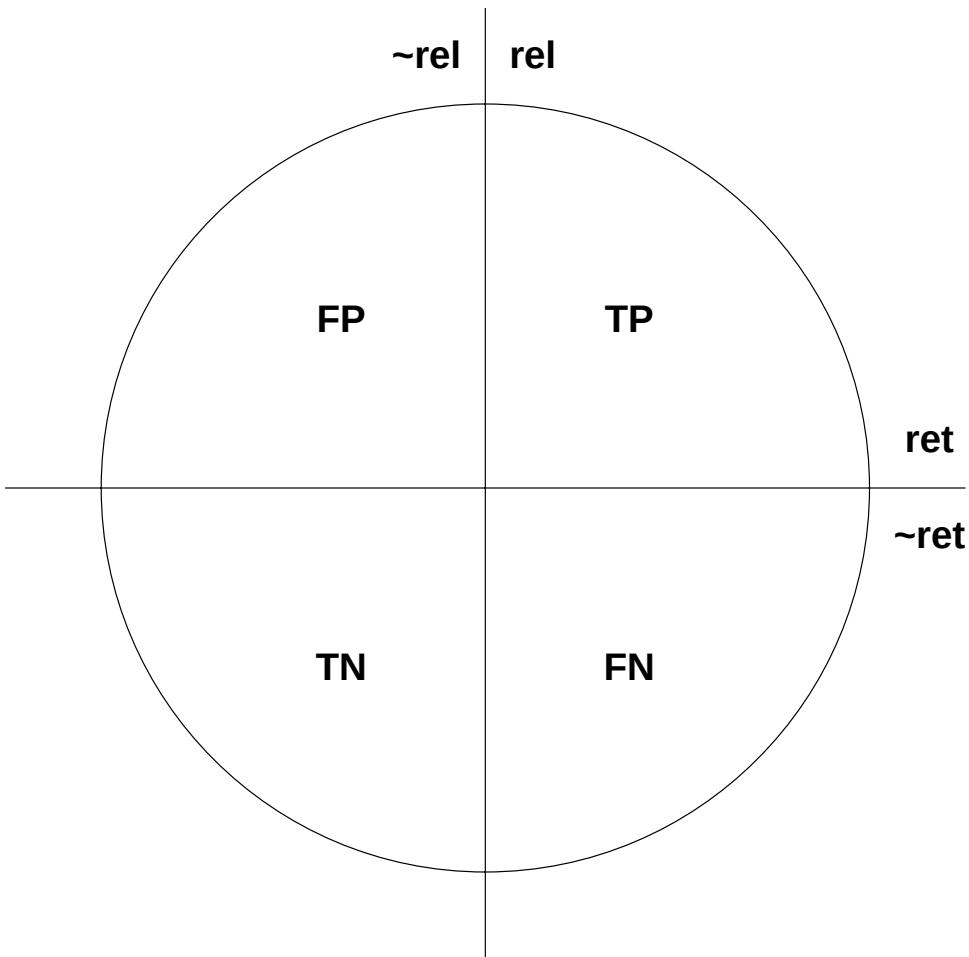
Understanding the Relation Between Recall, Precision, F-Measure, and Summarization for RE Tools

Daniel M. Berry
Cheriton School of Computer Science
University of Waterloo, Canada

The Universe of an RE Tool

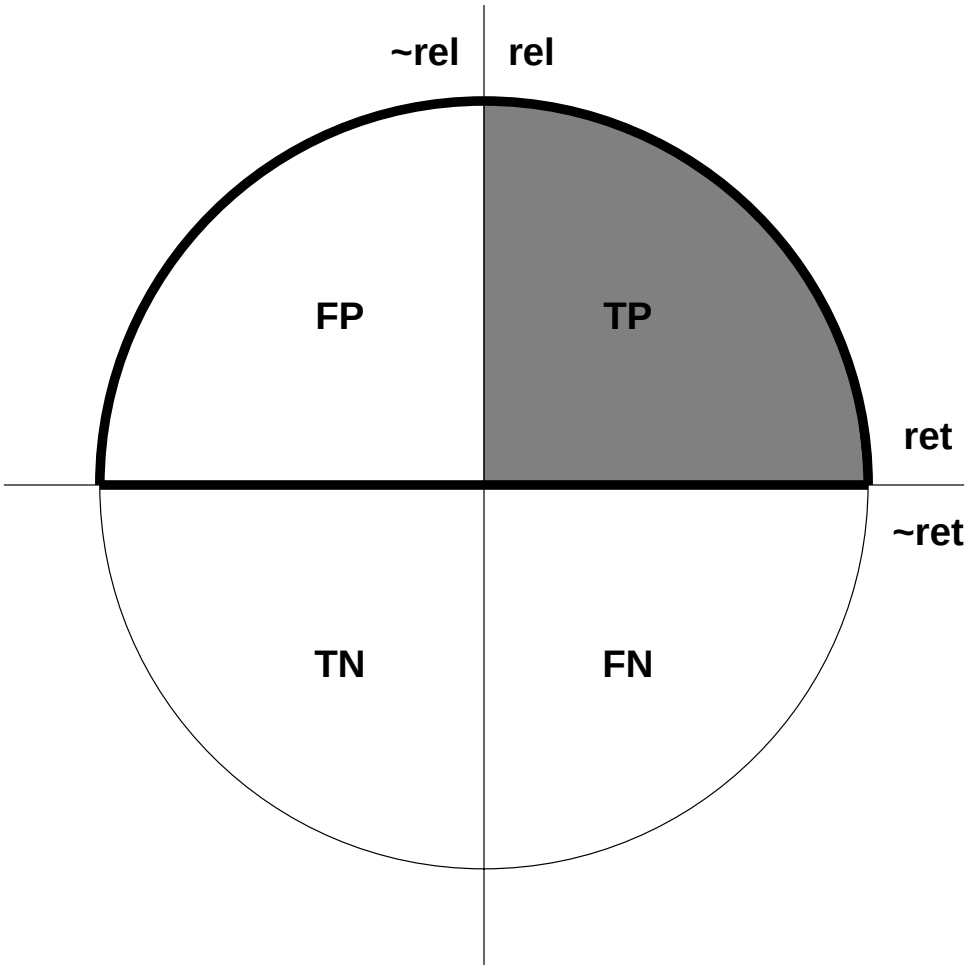


Precision

Precision: fraction of the retrieved items that are relevant

$$P = \frac{|\, ret \cap rel \,|}{|\, ret \,|}$$

$$= \frac{|\, TP \,|}{|\, FP \,| + |\, TP \,|}$$

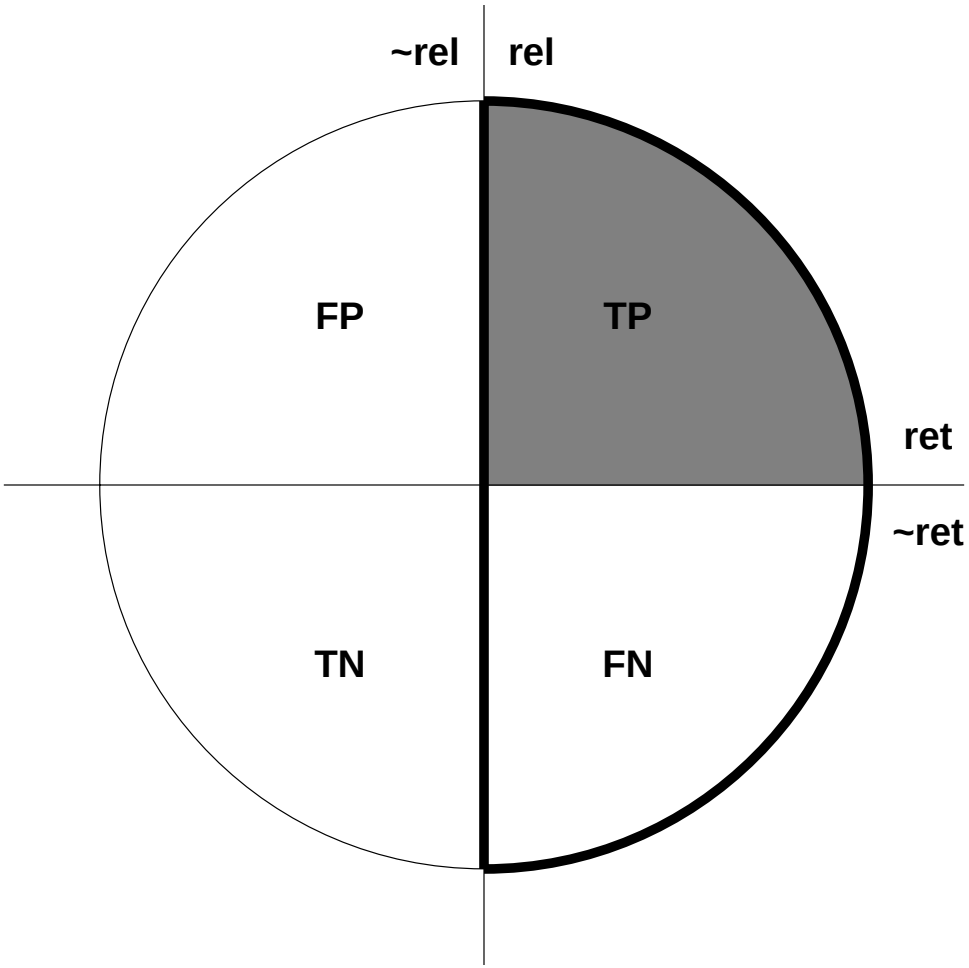

~rel
rel
FP
TP
ret
~ret
TN
FN

Recall

Recall: fraction of the relevant items that are retrieved

$$R = \frac{|\ ret \cap rel\ |}{|\ rel\ |}$$

$$= \frac{|\ TP\ |}{|\ TP\ | + |\ FN\ |}$$

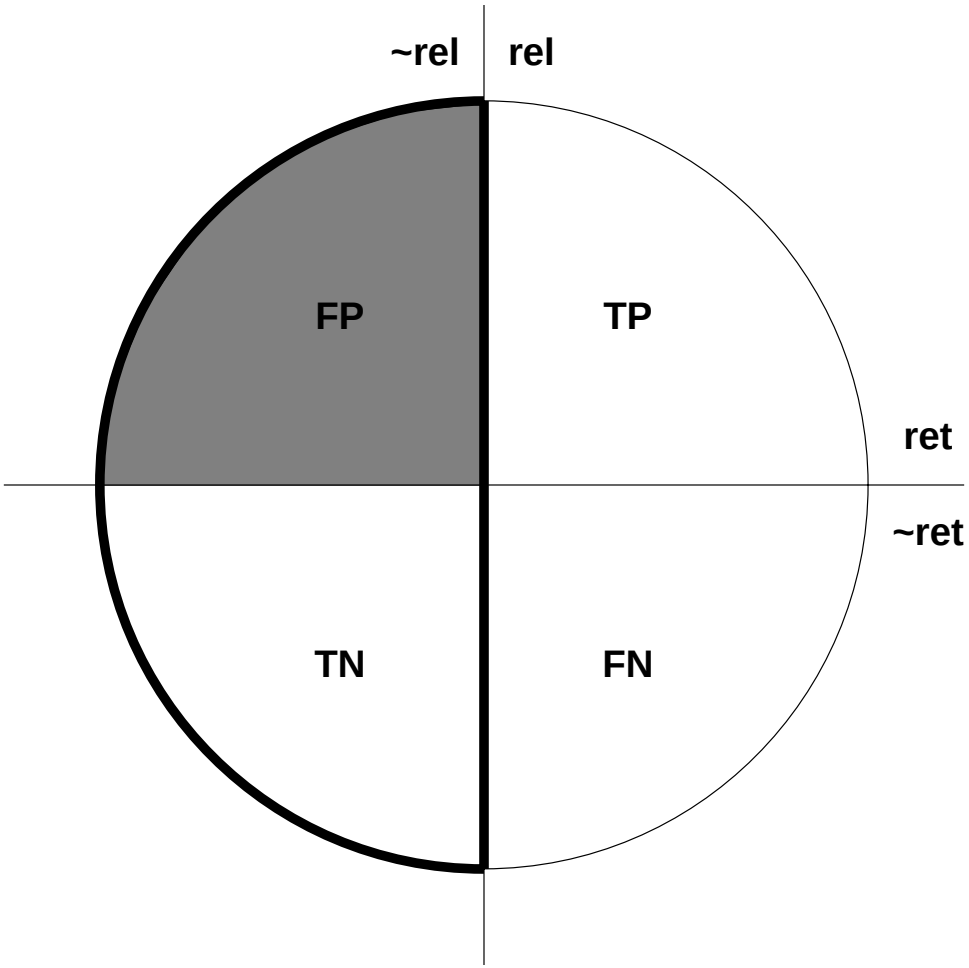

~rel
rel
FP
TP
ret
~ret
TN
FN

Fallout

Fallout: fraction of non-relevant items that are retrieved = recall of not relevant

$$F = \frac{|\, ret \cap \sim rel \,|}{|\, \sim rel \,|}$$

$$= \frac{|\, FP \,|}{|\, FP \,| + |\, TN \,|}$$

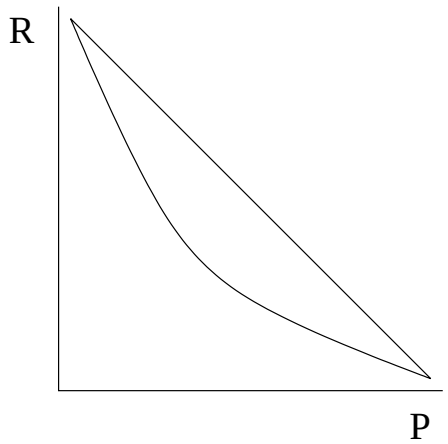

~rel
rel
FP
TP
ret
~ret
TN
FN

I Don't Understand

Literature says that fallout = measure of how quickly precision drops as recall is increased.

Huh? The formula does not seem to be saying this!

I would expect fallout with this meaning to be a relation between R and P, such as on the next page.

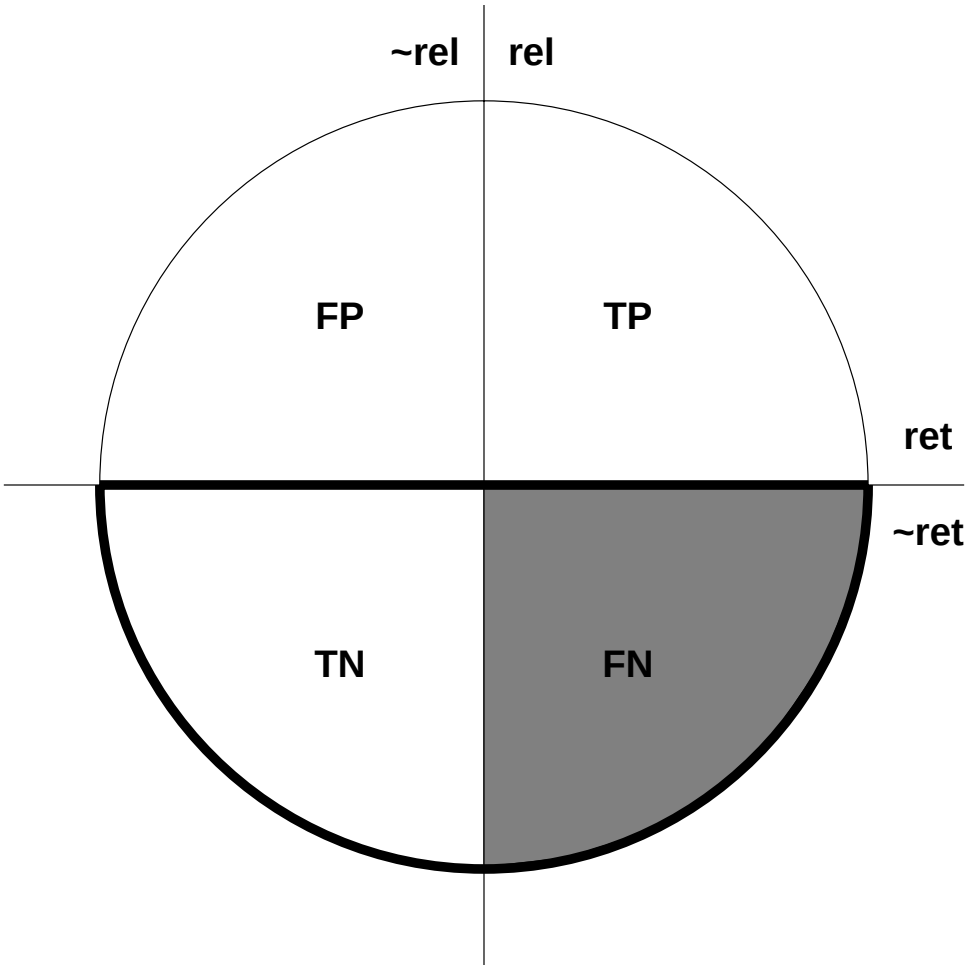

R
P

Missing

Missing: fraction of non-retrieved documents that are relevant

$$R = \frac{|\ \sim ret \cap rel\ |}{|\ \sim ret\ |}$$

$$= \frac{|\ FN\ |}{|\ FN\ | + |\ TN\ |}$$

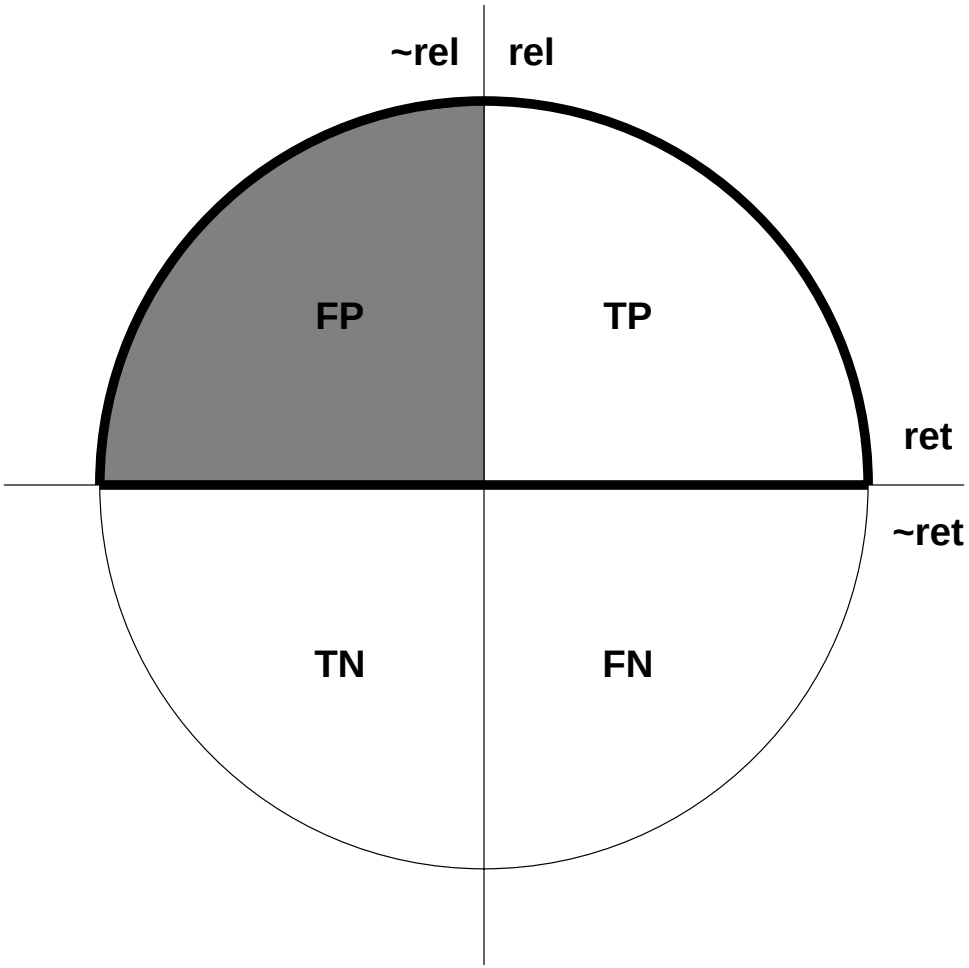

~rel
rel
FP
TP
ret
~ret
TN
FN

Imprecision?

Imprecision: fraction of retrieved items that are not relevant = precision of not relevant

$$I = \frac{|\, ret \cap \sim rel \,|}{|\, ret \,|}$$

$$= \frac{|\, FP \,|}{|\, FP \,| + |\, TP \,|}$$


~rel
rel
FP
TP
ret
~ret
TN
FN

F-Measure

F-measure: harmonic mean of precision and recall (harmonic mean is the reciprocal of the arithmetic mean of the reciprocals)

$$F = \frac{1}{\frac{\frac{1}{P} + \frac{1}{R}}{2}} = 2 \cdot \frac{P \cdot R}{P + R}$$

Popularly used as a composite measure

Incorrect Assumption

But this assumes that *P* and *R* carry the same weight.

For most tasks in RE for which we want to use tools, e.g.,

finding ambiguities, finding links, finding abstractions, etc.,

finding TPs is very tough, but rejecting FPs is very easy.

I.e., recall is at least an order of magnitude more important than precision.

Weighted Harmonic Mean

So let's do a weighted mean harmonically, with *w* as the weight of *R* over *P*

$$F_w = \frac{1}{\frac{\frac{1}{P} + w \cdot \frac{1}{R}}{w+1}}$$

$$F_w = (w+1) \cdot \frac{P \cdot R}{w \cdot P + R}$$

Note that $F = F_1$.

Recall = 10 × Precision

To reflect that recall is at least an order of magnitude more important than precision, let $w = 10$.

$$F_{10} = 11 \cdot \frac{P \cdot R}{10 \cdot P + R}$$

Note that $F_{\frac{1}{10}}$ weights P ten times over R

I Do Not Understand

I do not understand why the literature on the F-Measure uses the square in the weighted formula

$$F_\beta = (1+\beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$$

to weight *R* β times *P*.

Recall Very Very Important

Now, as $w \to \infty$,

$$F_w \approx w \cdot \frac{P \cdot R}{w \cdot P}$$

$$= \frac{w \cdot P \cdot R}{w \cdot P} = R$$

As the weight of *R* goes up, the F-measure begins to approximate simply *R* !

IF Precision Very Very Important

Then, as $w \to 0$,

$$F_w \approx 1 \cdot \frac{P \cdot R}{R}$$

$$= P$$

which is what we expect.

Tradeoff

For a typical RE task in which finding relevant items is at least an order of magnitude harder than rejecting irrelevant items, it pays to sacrifice precision for recall.

But …

The Extreme Tradeoff

Return …

the entire document $\rightarrow$ ***R* = 100% & *P* = 0%**

nothing $\rightarrow$ ***P* = 100% & *R* = 0%**

Useless

But returning everything to get 100% recall doesn’t save any real work, because we still have to manually search the entire document.

What is missing?

Summarization

Summarization

If we can return a subdocument significantly smaller than the original …

that contains *all* relevant items, …

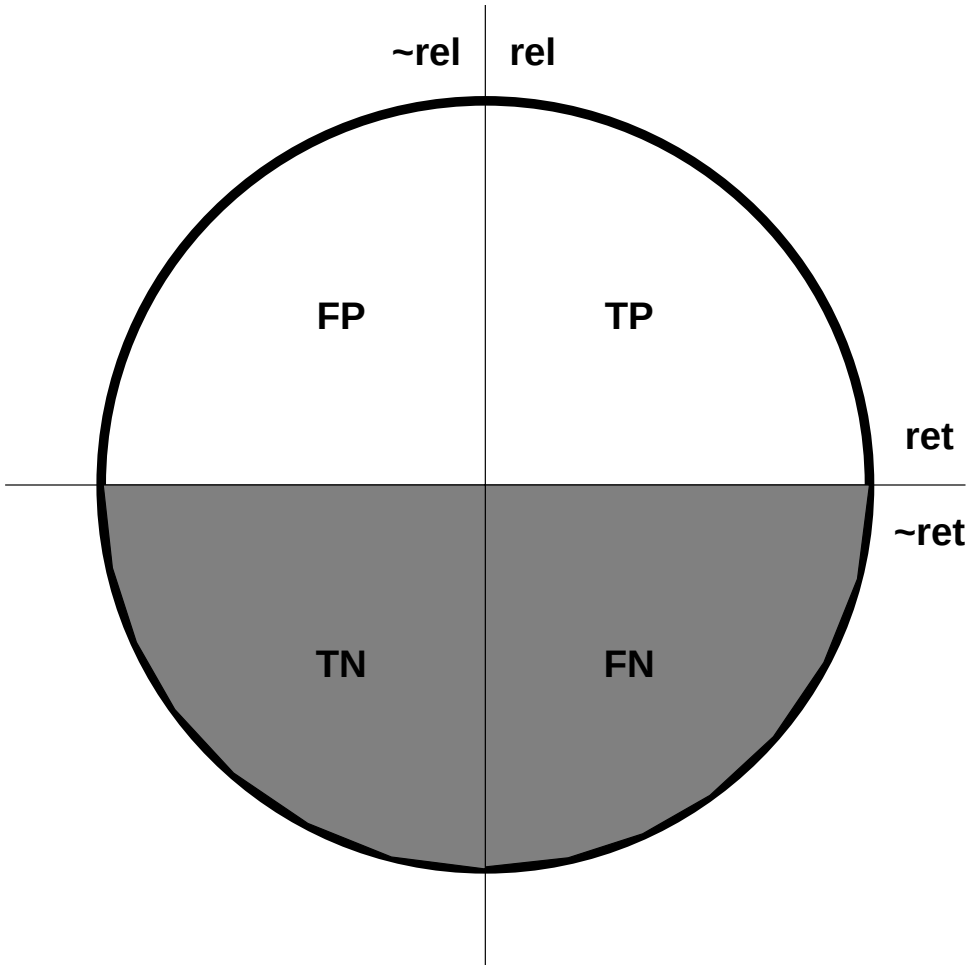
then we have saved some real work.

Summarization Measure

Summarization = fraction of the original document that is eliminated from the return

$$R = \frac{|\sim ret|}{|\sim ret \cup ret|} = \frac{|\sim ret|}{|\sim rel \cup rel|}$$

$$= \frac{|TN| + |FN|}{|TN| + |FN| + |TP| + |FP|}$$


~rel
rel
FP
TP
ret
~ret
TN
FN

How to Use Summarization

We would *love* a tool with 100% recall and 90% summarization.

Then we really do not care about precision.

In Other Words

That is, if we can get rid of 90% of the document with the assurance that …
what is gotten rid of contains *only irrelevant* items and thus …

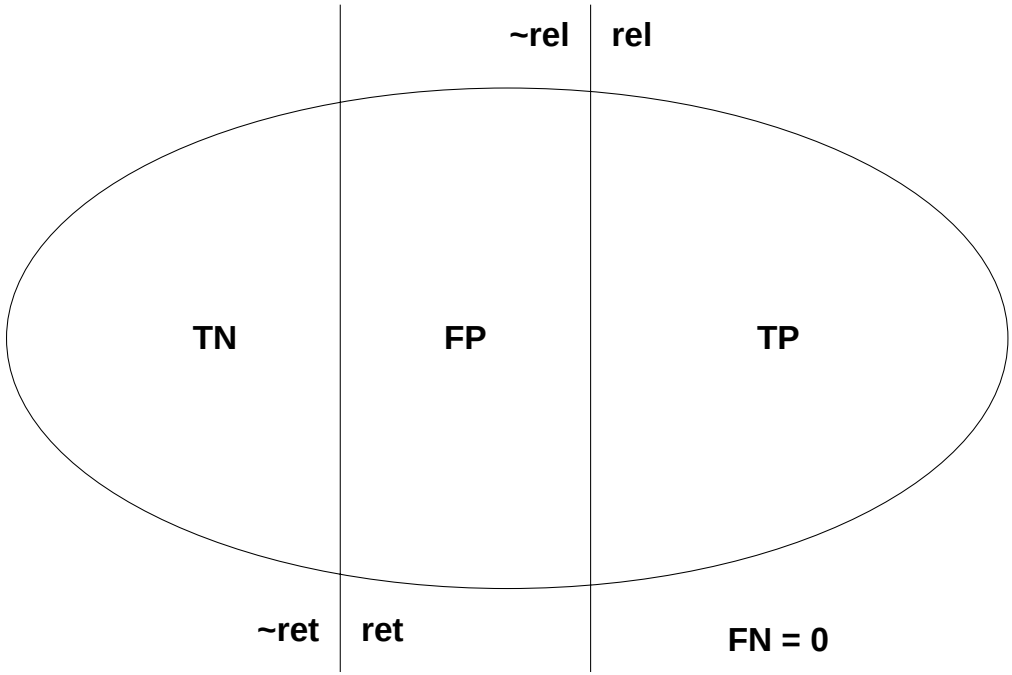
what is returned contains *all* the relevant items, …

then we are *very happy*! ☺

An Ideal Tool

In an ideal tool,

- **100% of relevant items are returned,**
- **100% of what's not returned is irrelevant,**
- **what's returned contains only TPs and FPs**
- **what's not returned are for sure TNs, and**
- **there are no FNs.**



$P < 100\%$, $R = 100\%$, $F < 100\%$, $M = 0\%$,
$I < 100\%$, $S < 100\%$.

Inverted Search

Ideal tool probably does not exist.

But maybe an inverted tool might be easier algorithmically.

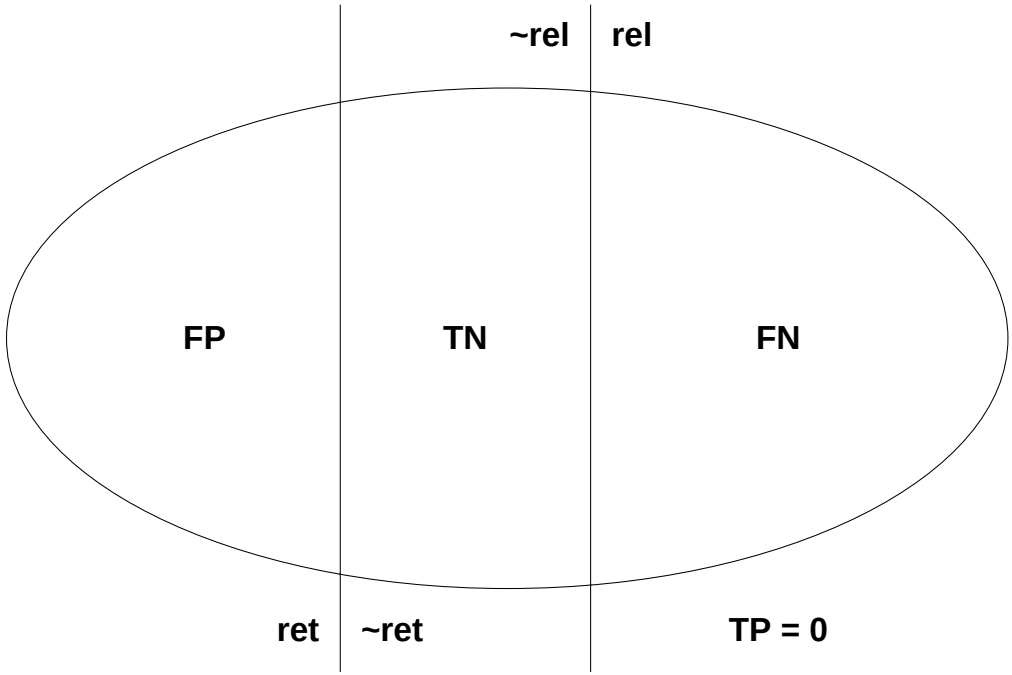
100% of what is returned is not relevant.

Using Results of Inverted Tool

With the results of an inverted tool, …

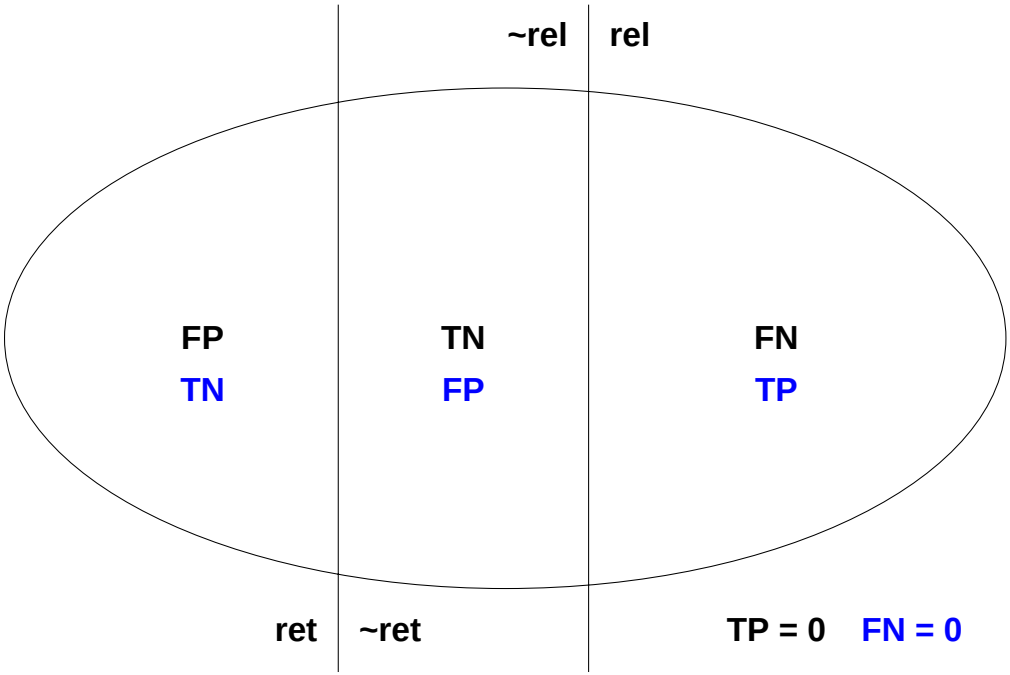
what is returned can be eliminated from the original document, i.e.,

to get rid of many irrelevant items so that the remaining document is easier to search manually for relevant items.



Here FP, TN, FN, & TP are w.r.t. inversion.

$P = 0\%$, $R = 0\%$, $F < 100\%$, $M < 100\%$,
$I < 100\%$, $S < 100\%$



***P* < 100%, *R* = 100%, *F* < 100%, *M* = 0%, *I* < 100%, *S* < 100%.**

Effect of the Inverted Tool

The inverted tool approximates the ideal tool.

But, its output must be removed from the original document.

However, a simple lexical tool can do that, assuming that the output of the inverted tool is snippets of the input.