

1. Contracted forms:

They will cause ambiguity in token boundaries.

*I’ve just had a really good idea about what to offer as a prize.*

1. Forms with periods:

The machine will have to decide whether these are sentence-internal characters or sentence-final ones to tell the boundaries of tokens as well as sentences.

*Of course that would mean we’d need enough budget to send them to the good ol’ U.S.A.*

1. Slashes: it’s https://fullbelliesinc.com
2. Special characters:

*…we can’t just say “OK, you’re flying to the States tomorrow”, can we?*

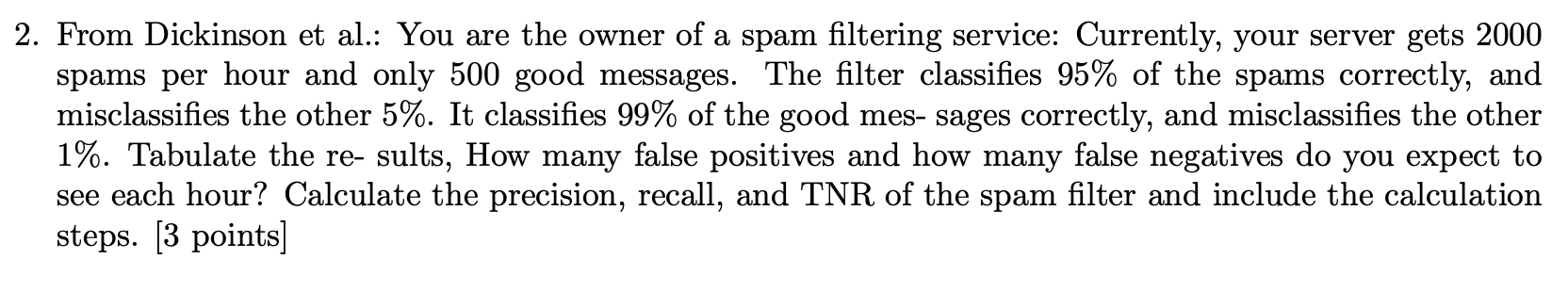
*…what a ridiculous name! :)*

1. Abbreviations: Full Bellies Inc.
2. Multi-word expression:

*Fillet mignon and baked potatoes*

1. Named Entities:

*There’s one I like a lot in San Francisco*



|  | Actual Positive (Spam) | Actual Negative (Good) |
| --- | --- | --- |
| Predicted Positive | TP=2000\*95%=1900 | FP=500\*1%=5 |
| Predicted Negative | TN=2000\*5%=100 | FN=500\*99%=495 |

Precision = TP/(TP+FP) ≈ 0.996

Recall = TP/(TP+FN) = 0.95

TNR = TN/(TN+FP) = 0.99