# Design & Engineering of Intelligent Information Systems: PI6

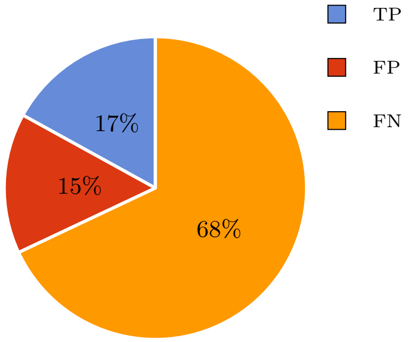
Instructed by *Dr. Eric Nyberg*

Due on October 12, 2015

**Vivian Robison** vrobison

The first technique I tried was very basic tokenization based on b and unigram overlap. The original seed for randomly selecting passages gave me six of ten questions with no relevant passages, so that only left four questions that the system could possibly improve upon. I changed the seed until I got a set of ten passages that could be answered.

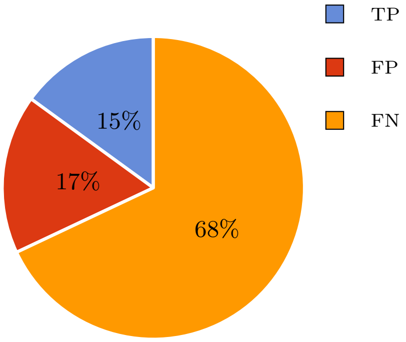
| question id | tp | fn | fp | precision | recall | f1 |
|---|---|---|---|---|---|---|
| 2215 | 2 | 1 | 3 | 0.4 | 0.667 | 0.25 |
| 2341 | 0 | 1 | 5 | 0 | 0 | 0 |
| 2244 | 1 | 0 | 4 | 0.2 | 1 | 0.167 |
| 2067 | 0 | 0 | 5 | 0 | 0 | 0 |
| 2306 | 0 | 2 | 5 | 0 | 0 | 0 |
| 2241 | 1 | 2 | 4 | 0.2 | 0.333 | 0.125 |
| 2059 | 1 | 0 | 4 | 0.2 | 1 | 0.167 |
| 2104 | 1 | 0 | 4 | 0.2 | 1 | 0.167 |
| 2347 | 3 | 2 | 2 | 0.6 | 0.6 | 0.3 |
| 2312 | 1 | 1 | 4 | 0.2 | 0.5 | 0.143 |



Macro-average F1: 0.1449
Micro-average F1: 0.1317

I noticed some of the passages in the top 5 only had function words in common with the question,so I removed function words from the overlap set before scoring. Additionally, I used the Stanford NLP tools for tokenization and lemmatization before finding unigram and bigram overlap. The better tokenization forms a basis for future improvements, while lemmatization helps with the issue of different verb conjugations in question and passage I noted in a previous report.

| question id | tp | fn | fp | precision | recall | f1 |
|---|---|---|---|---|---|---|
| 2215 | 1 | 2 | 4 | 0.2 | 0.333 | 0.125 |
| 2341 | 1 | 0 | 4 | 0.2 | 1 | 0.167 |
| 2244 | 1 | 0 | 4 | 0.2 | 1 | 0.167 |
| 2067 | 0 | 0 | 5 | 0 | 0 | 0 |
| 2306 | 1 | 1 | 4 | 0.2 | 0.5 | 0.143 |
| 2241 | 1 | 2 | 4 | 0.2 | 0.333 | 0.125 |
| 2059 | 1 | 0 | 4 | 0.2 | 1 | 0.167 |
| 2104 | 0 | 1 | 5 | 0 | 0 | 0 |
| 2347 | 2 | 3 | 3 | 0.4 | 0.4 | 0.2 |
| 2312 | 1 | 1 | 4 | 0.2 | 0.5 | 0.143 |



Macro-average F1: 0.1304
Micro-average F1: 0.1235

Impact on this set:
Δ Macro-average F1 = -0.0145
Δ Micro-average F1 = -0.0082
Impact obtained over several hours with much of the work scrapped, so I don't have precise numbers.

The overall performance on this tiny set didn't really change, and actually went down a bit, but ten questions isn't enough to evalute the system on. The large number of false positives suggests the classification threshold is set too low, as a significant number of questions have fewer than five relevant passages associated with them. The TP/FP/FN numbers don't reflect the reordering of passages on either side of the threshold. The appozimately three hours of attempting to implement various Stanford tools resulted in only lemmatization functioning with reasonable speed, and a decrease in F1 on this particular set of 10. I could put additional time into cherrypicking a random set of 10 passages that adding bigrams and lemmatization and function word removal actually helped, but that would increase effort without actually increasing system performance.

One obvious problem is with passages and questions having synonyms that aren't picked up by basic ngram overlap even with lemmatization, suggesting future upgrades to the scoring system should have something like WordNet integrated. Dependency parsing the question to find the subject, and prioritizing finding that in the passages would probably also help. I tried implementing dependency parsing using the Stanford NLP toolkit, but the API is a bit clunky to use and not very fast, so I deemed the effort and runtime required too high for now and shelved it for a future iteration of improvements.