1. Did the crawler find the most relevant pages? Why, why not?
2. Is your MIME detection good? Define “good”.
3. Are your parsers extracting the right text? Define “right”.
4. Are we selecting the right parser? Define “right”.
5. Is your Metadata appropriate? What’s missing?
6. How well is my language detection performing? Comment based on the diversity of the languages derived in this assignment. Are there mixed languages? Did it affect your accuracy?
7. Do your Named Entities make sense?

With raw results, named entity values contains lots of messy tokens such as html tags. Moreover, it gets lots of wrong result. We have run a filter which only extract meaningful results and remove messy html tags.