1. **Did the crawler find the most relevant pages? Why, why not?**

For most cases the crawler seems to find the relevant pages. The content type of the response from the server matches the content type of the request sent. Also some of the pages seem to return relevant content which was found by doing Named Entity Recognition.

1. **Is your MIME detection good? Define “good”.**
2. **Are your parsers extracting the right text? Define “right”.**
3. **Are we selecting the right parser? Define “right”.**
4. **Is your Metadata appropriate? What’s missing?**
5. **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?**

The language detection performs fairly accurately. The language detector was able to find 25 different languages - *ca,da,de,el,en,eo,es,et,fa,fi,fr,gl,hu,it,lt,nl,no,pl,pt,ru,sk,sv,th,uk*. Yes, there are files with mixed languages. The language detector just gives out the language of the file which had more information and higher priority.

Also the following anomalies were observed.

1. It identifies a language but then there is no content in the page. In other words, the file size is 0 kb
2. It identifies the wrong language. For example, the page is in English but it detected Hungarian. The page is in English but detects Estonian.
3. **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 extracts meaningful results and removes the messy html tags.