Social Networks
Course Project
KML (Knowledge Markup Language)

Amit Kumar Verma (mentor)
Paras Kumar
Nitin Gandhi

Contents

Introduction to KML

KML stands for Knowledge Markup Language, a standard format for storing the data of all the Knowledge Building Portals. We are trying to propose a new standard format for all these types of portals such that the analysis is easy.

Components

- KML-Compressor-Decompressor
- KML-for-Github
- Github-Spider
- User-Agent-Spider
- Wiki-Satck-KML-Downloader

Source Code: `https://github.com/csl-622/KML`

KML-Compressor-Decompressor

These programs compresses and decompresses a KML file using diff algorithm inspired from git. It can compress the KML file by 30 percent or more depending upon the edits made on an article.

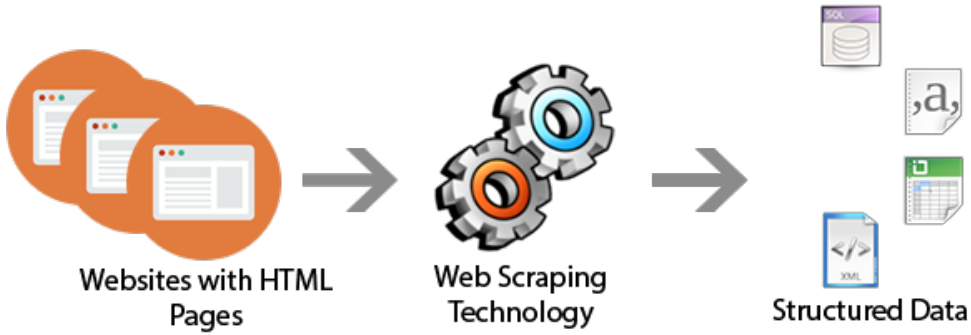
```
diff --git a/app.js b/app.js
new file mode 100644
index 0000000..144ec7f
--- /dev/null
+++ b/app.js
@@ -0,0 +1 @@
+module.exports = require('./lib/index');
diff –git a/index.js b/index.js
deleted file mode 100644
index 144ec7f..0000000
--- a/index.js
+++ /dev/null
@@ -1 +0,0 @@
-module.exports = require('./lib/index');
```

KML-for-Github

We had to make KML for github and we had KML for wikipedia earlier. The attempt of this program is to retain the structure of KML for wkipedia while representing all the data in a git workflow in this newly generated KML file.

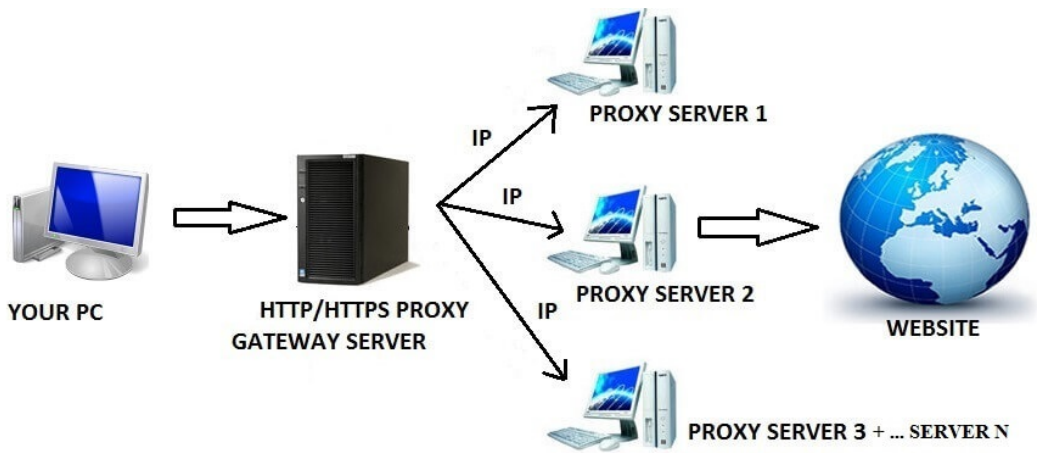
Github-Spider

This multithreaded web-spider follows the rules of /robots.txt file and is useful for fetching small sized GitHub repositories. It outputs the results in JSON format.



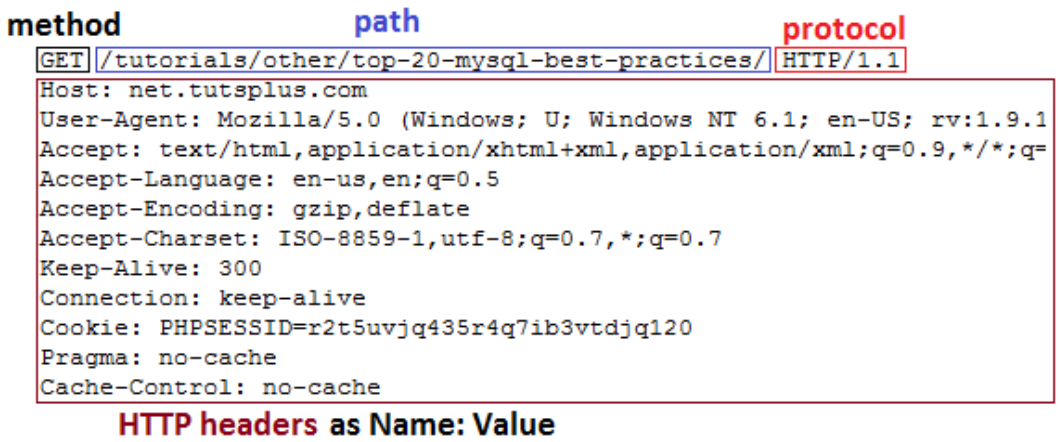
Github-Spider-Proxy-Rotated

Github-Spider-Proxy-Rotated is an advanced version of the previous spider; it contains pipelines, middlewares, and throttling parameters, which is in the middleware and settings file. It outputs the results in JSON format.



User-Agent-Spider

This spider gets the latest user-agent headers from an online forum, output of this file is directly put into use in other two spiders namely Github-Spider and Github-Spider-Proxy-Rotated. The result of this spider is a text file containing a list of HTTP headers which can is useful for spoofing browser's HTTP request activity.



Wiki-Satck-KML-Downloader

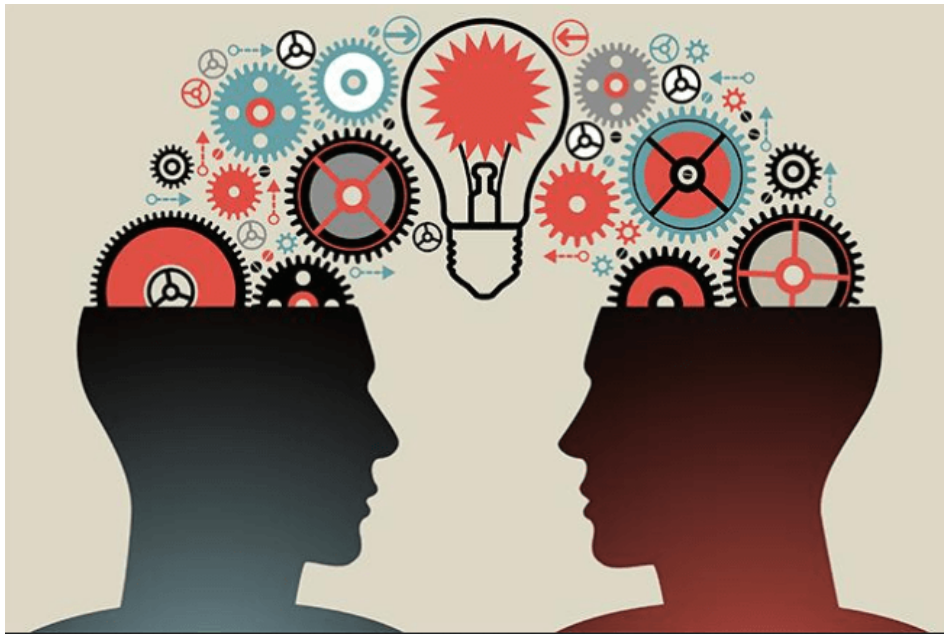
Dowloads xml from Wikipedia and Stackexchange for any given topic from user then coverts it to KML. The topics for wikipedia can be name of any page on wikipedia like UNO while for stackexchange the topics can any of the sub-site name of stackexchange like beer.

StackExchange

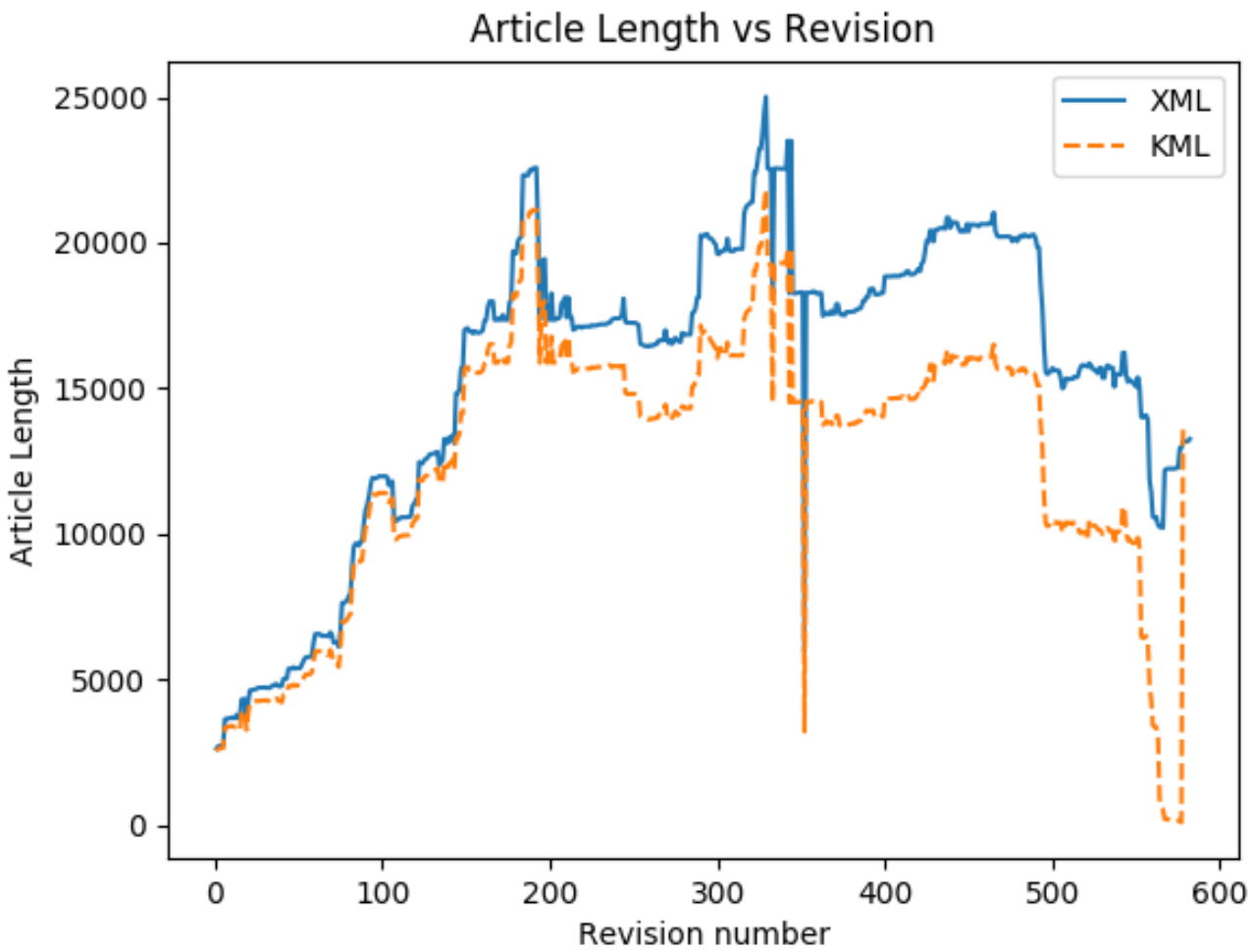
WIKIPEDIA
The Free Encyclopedia

Analysis

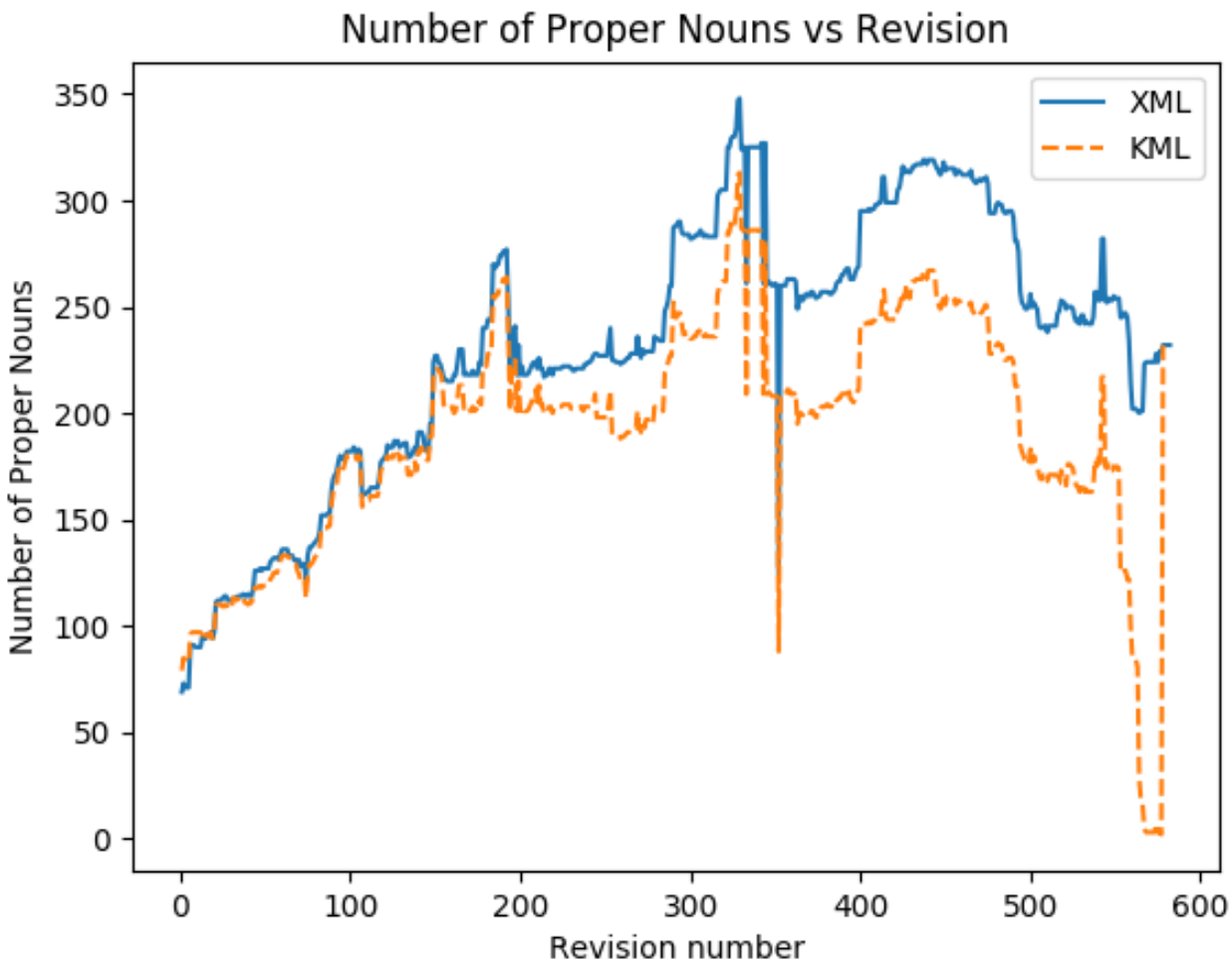
Graphs are generated to compare compressed KML and XML on various performance metrics like change in ArticleLength, Number of revisons, Number of Images, Number of Nouns, Number of sections etc vs. Revision number.



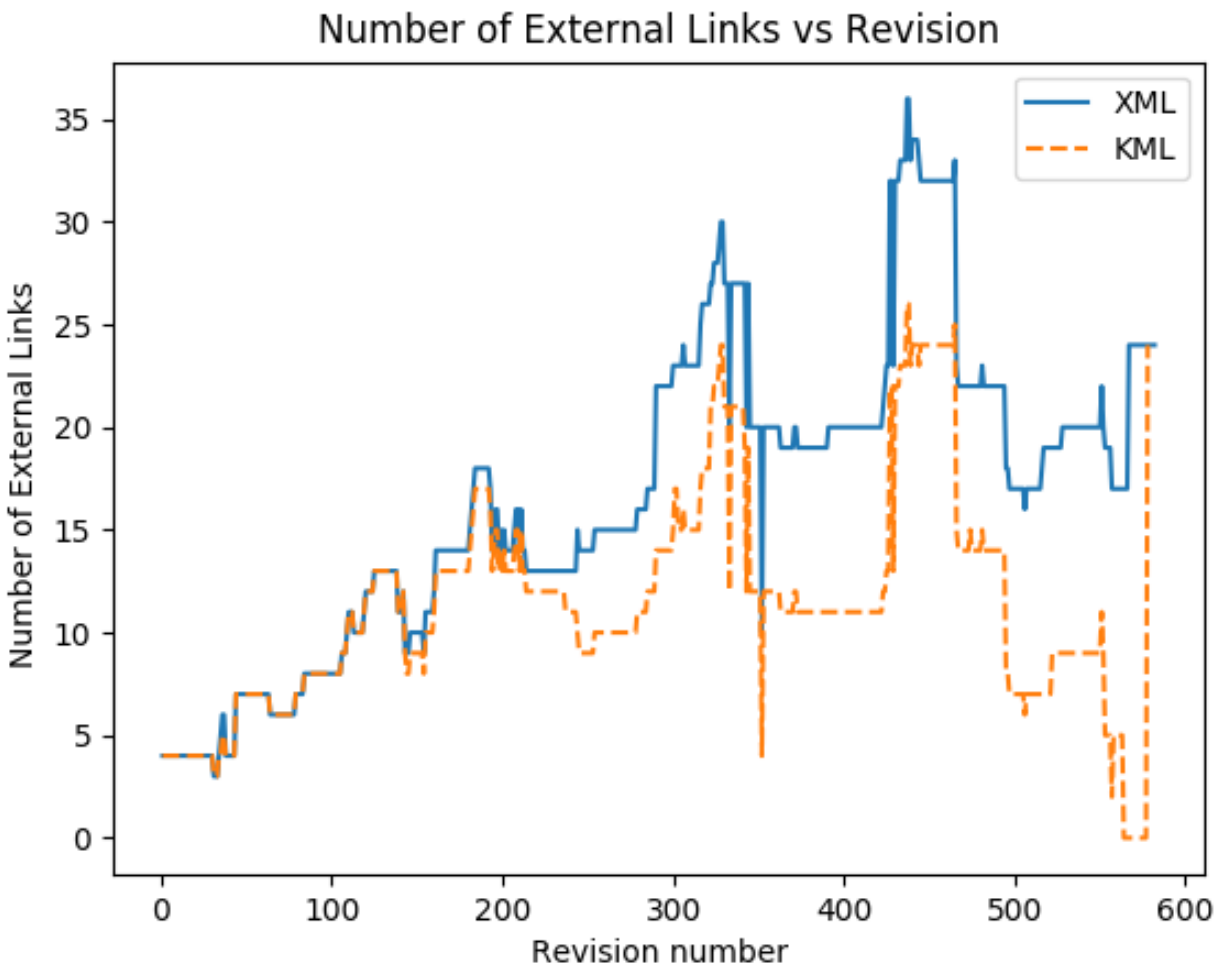
Article Length vs Revision



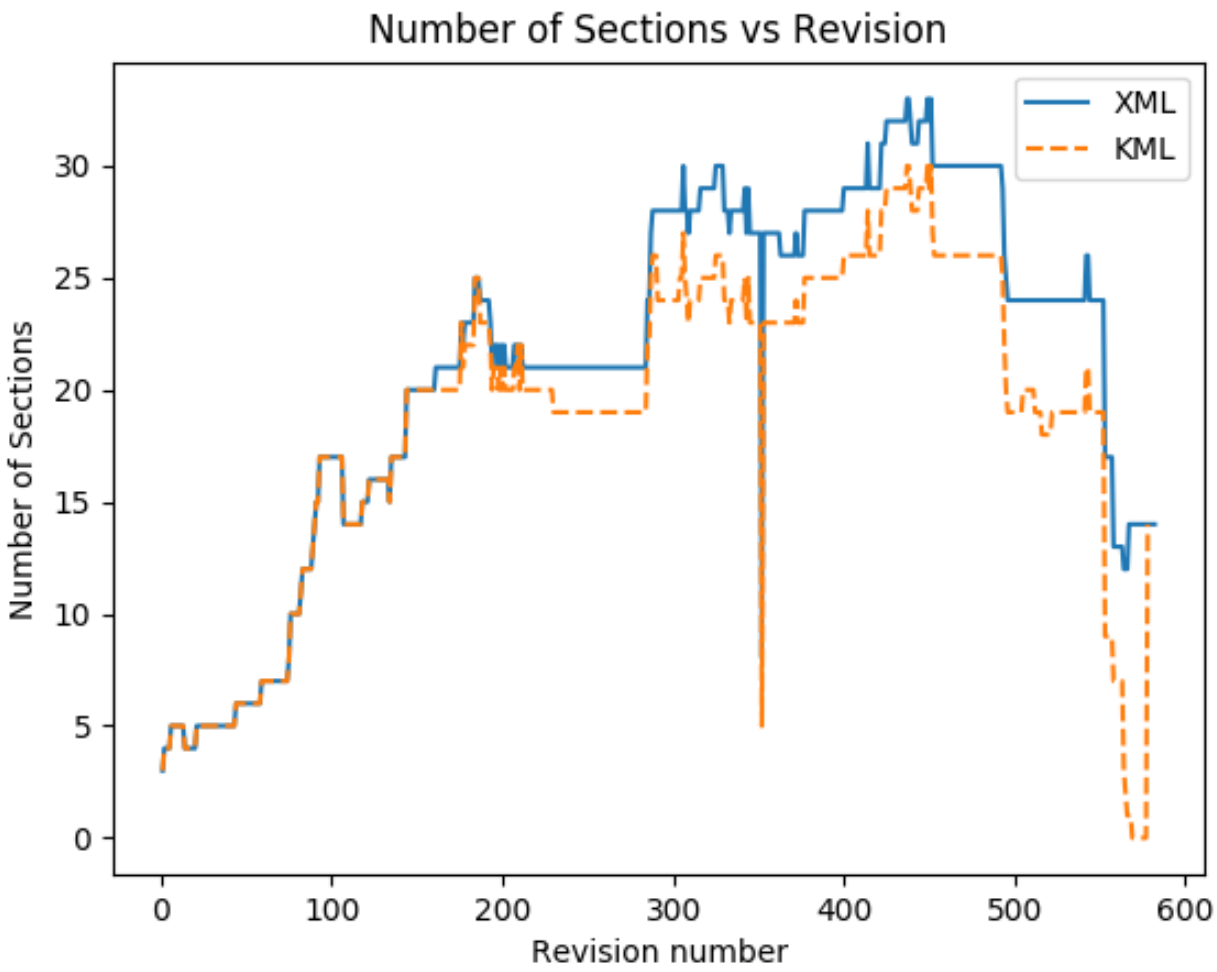
Number of Proper Nouns vs Revision



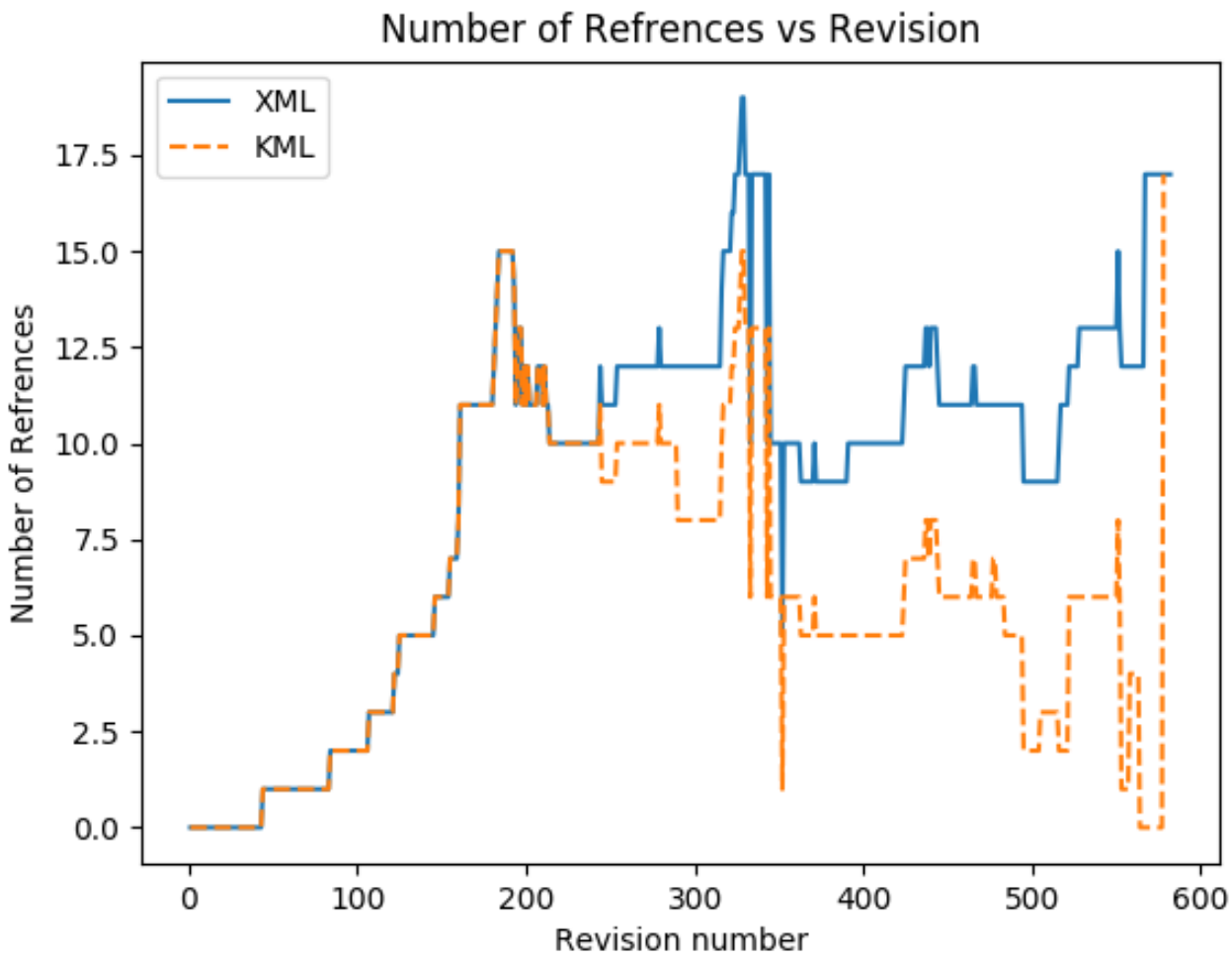
Number of External Links vs Revision



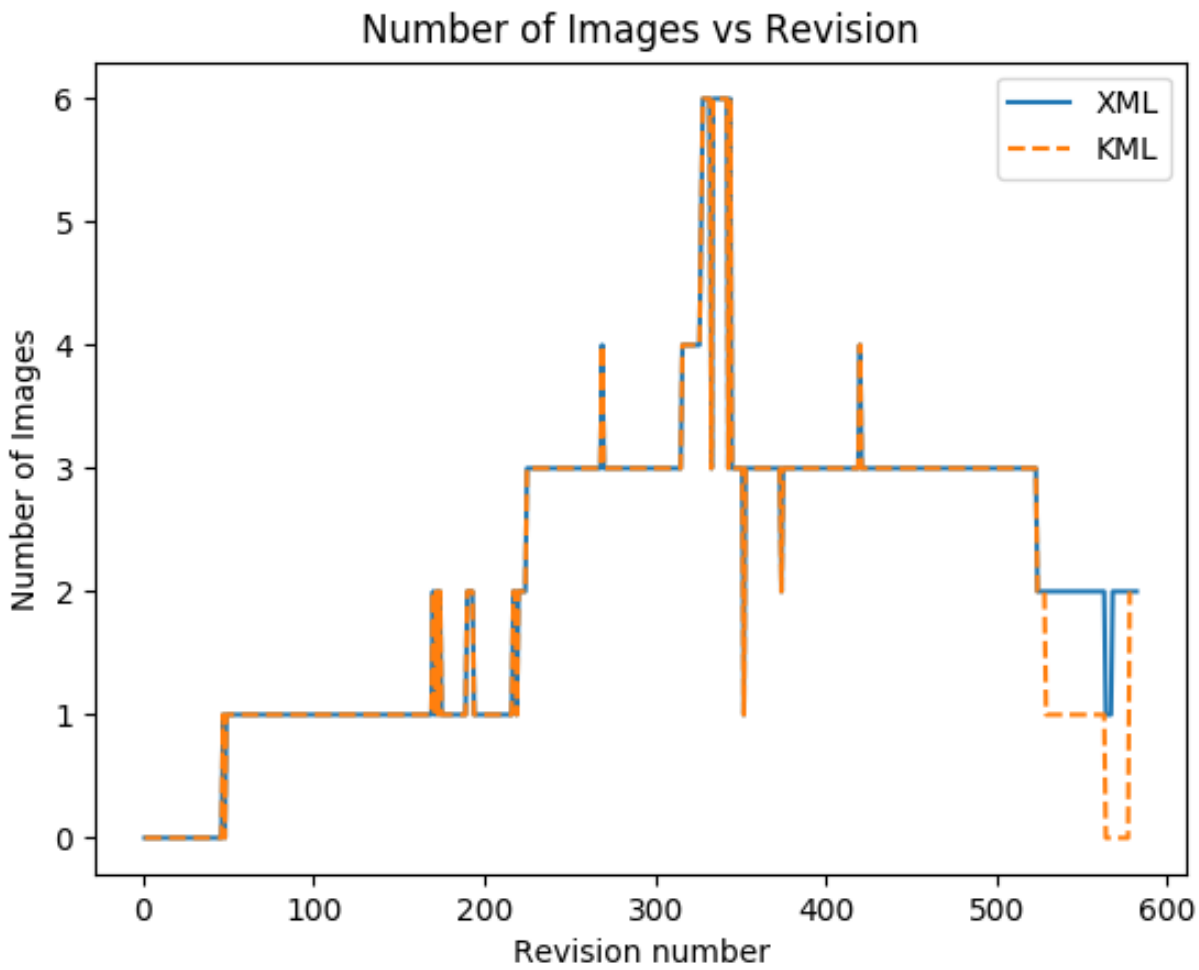
Number of Sections vs Revision



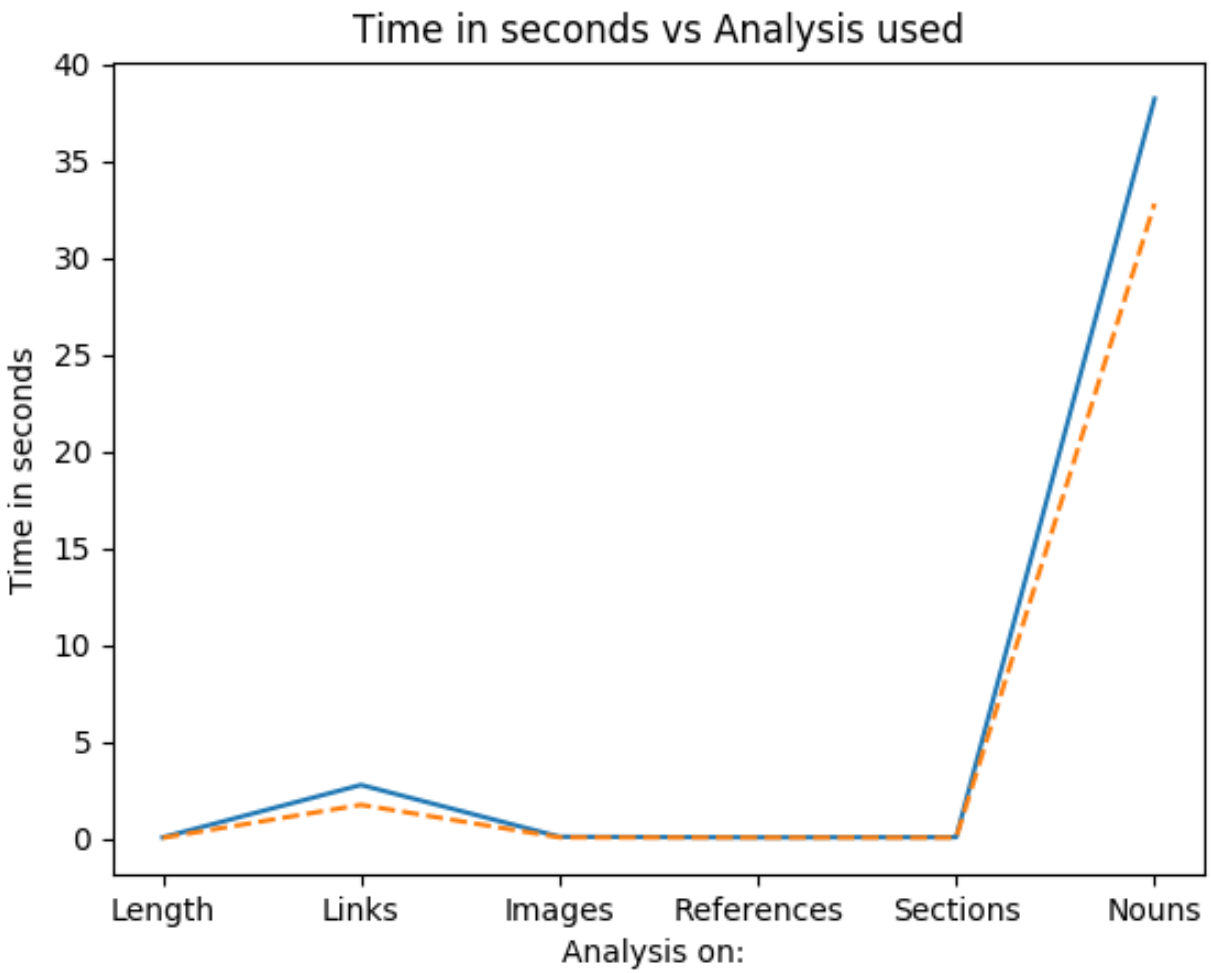
Number of References vs Revision



Number of Images vs Revision



Time in seconds vs Analysis used



Link to our Repository

URL: `https://github.com/csl-622/KML`