Mining Big Data - Data Collection Report

Investigating Paragraph Vectors

a1632538 Zachary Forman
a1646930 James Caddy

August 28, 2016

Data Sources

As we discussed in our proposal, we will use three primary data sources. Firstly, we use Maas' IMDB set
Secondly, we used The Wikimedia Foundation's English Wikipedia data dump [1] as a source for English language documents. We downloaded the full wiki dump from `https://dumps.wikimedia.org/enwiki/20160820/`.
Lastly, we used The Chromium Project's Chromium repository [2] as a source for source code documents. We cloned the repository at revision 4fe31bb06cf458234d7017950a8b2b82427487c8.

Data Scale

IMDB Dataset

...	...

Chromium Codebase

Size	2688525849B (2.6GB)
Files	245004 (245M)
C & C++ files	23674 (23M)
Lines of Code	4851267 (4.85M)
Comments	608975 (608k)

English Wikipedia

Compressed Size	14054466197B (13GB)
Uncompressed Size	58392156024B (55GB)
Pages (including redirects)	16824195 (16M)
Documents (excluding meta pages)	5219884 (5M)
Words (only including user visible words)	4004090340 (4B)

Data Processing & Storage

Generating Similar Data

References

[1] Meta. *Data dump torrents — Meta, discussion about Wikimedia projects*. [Online; accessed 9-August-2016]. 2016. URL: `https://meta.wikimedia.org/w/index.php?title=Data_dump_torrents`.

[2] *Git repositories on chromium*. URL: `https://chromium.googlesource.com/`.