

COSC 589 - Web Search and Sense-Making

Assignment 5

Due Wednesday March 21, 2018, 11:59pm

Task: Build the Wikipedia Link Graph

Introduction:

In this assignment, we will extract the links from the Wikipedia dump and build a link graph from them.

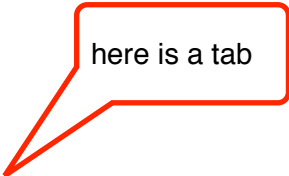
Requirements:

>100GB free disk space in your machine.

Instructions:

Write a LinkGraph.scala file to extract the links from the English Wikipedia dump and build the link graph, by taking the following steps:

1. Read in the output files of your last assignment, in which you have obtained the articles. The format should be:
 - One page (of the type of article) per line
 - In each line, you have two fields: the title and the text, which are separated by a tab
 - An example page looks like:



here is a tab

```
{Alvin Toffler |{{Use ndy dates|date=September 2013}}|{{Infobox person| name           =  
pg| image_size      = 210px| caption          = Alvin Toffler (2006)| birth_name      :  
}}| birth_place     = New York City<ref>{{cite web|last=The European Graduate School|tit  
library/alvin-toffler/biography/|accessdate=January 7, 2014}}</ref>| death_date    =  
| death_place      =| death_cause           =| resting_place    =| residence        = Los Ange  
ther_names         =| known_for            = ''[[Future Shock]]'',<br />''[[The Third Wave (book  
h and Violence at the Edge of the 21st Century|Powershift]]''| education            = Multip  
University]] (BA)| employer                =| occupation        = [[Futurist]], journalist, writ  
trategic Studies| religion                 =| spouse            = Adelaide Elizabeth "Heidi" (Far  
le.ca/books?id=IN11AAAAAA&dq=%22Adelaide+Elizabeth+Farrell%22&dq=%22Adelaide+Elizabeth+  
CCAQ6AEwAQ|</ref>| children                 = Karen Toffler| parents                 =| relations  
Foundation Book Award for Contributions to Management Literature,<br />Officier de L'Ord  
= {{URL|alvintoffler.net}}| footnotes =<ref>[http://www.alvintoffler.net/ Alvin ]  
born October 4, 1928) is an American writer and [[futurist]], known for his works discuss  
communication revolution]] and [[technological singularity]].Toffler is a former associat  
ine. In his early works he focused on technology and its impact through effects like [[ir  
eaction to [[Social change|changes in society]]. His later focus has been on the increas  
ification of new technologies, and capitalism.He founded Toffler Associates, a management  
he [[Russell Sage Foundation]], visiting professor at [[Cornell University]], faculty mer  
search]], a White House correspondent, an editor of Fortune magazine, and a business cons  
g/events/events.taf?function=show&cat=allconf&EventID=GC03&SPID=898&levell=speakers&leve  
stitute]], 2003.</ref>Toffler is married to Heidi Toffler, also a writer and futurist. Th
```

- For each page, extract its outlinks' titles. An outlink appears in the Wikipedia dump in the following format:

[[the title of an outline page]]

For instance, the page titled "Alvin Toffler" has an outlink to another page titled "Future Shock".

```
(Alvin Toffler {{Use mdy dates|date=September 2013}}{{Infobox person|name      = Alvin
Toffler|image      = Alvin Toffler 02.jpg|image_size    = 210px|caption    = Alvin Toffler
(2006)|birth_name  =|birth_date    = {{Birth date and age |1928|10|4}}|birth_place  =
New York City<ref>{{cite web|last=The European Graduate School|title=ALVIN TOFFLER -
BIOGRAPHY|url=http://www.egs.edu/library/alvin-toffler/biography/|accessdate=January 7,
2014}}</ref>|death_date    = <!-- {{Death date and age|YYYYIMMIDDIYYYYIMMIDD}} -->|
death_place    =|death_cause    =|resting_place    =|residence    = Los Angeles,
Californial nationality    = United States|other_names    =|known_for    = "[[Future
Shock]]"}
```

- However, not all the things inside [[]] are good outlink titles. We will need to do the following:

3.1. First, ignore an outlink title contains colons ":". Basically, they are not titles for any article, but for something else. For instance, "WP:CSD#R3D3" is not a title name for an article:

```
112</timestamp><contributor><username>Od Mishchhu</username><id>461626</id></contributor>
<comment>Decline speedy [[WP:CSD#R3R3]] - not recently created</comment><model>
```

3.2. Second, extract the parts before an "I", "#", or ",", if an outlink title contains these symbols. If a title contains "I", it has multiple variations of the title; we only keep the first one. For instance, for [[The Third Wave (book)|The Third Wave]], we will only keep [[The Third Wave (book)]]. If a title contains "#", it has both the title and a section name. Similarly, we only keep the former. For instance, for [[Uncial script#Half-uncial|semi-uncial]], We will only keep [[Uncial script]]. If a title contains ",", it conflicts with the Spark's default delimiter. To allow we will be able to match an outlink page to its own entry, we only keep the part before comma. In summary we will extract

- the part before "I" in an outlink title with "I" (title name variations),
- the part before "#" in an outlink title with "#" (book mark sections), and
- the part before ",", in an outlink title with "," (Spark's default delimiter in saved files)

- Save the title and the outlink titles for each page in Wiki dump. Optionally, you can save your files into compressed format by using `saveAsTextFile(filename, classOf[GzipCodec])`. The output format is described as follows:

- One page per line
- In each line, you have the title of a page and a list of the titles of the outlinks in the page
- Each outline title is inside [[]], and separated by a tab "\t".

COSC 589 - Web Search and Sense-Making

- The title and the list of links is separated by “,”. (this is the default in Spark)

For instance, for page titled “Alvin Toffler”, we expect the following as the output:

```
(Alvin Toffler, [[Future Shock]] [[The Third Wave (book)]]  
[[Futurist]] [[futurist]] [[digital revolution]] [[  
logical singularity]] [[Fortune (magazine)]] [[informat  
change]] [[Russell Sage Foundation]] [[Cornell  
]] [[Milken Institute]] [[Bel Air]] [[Sunset B
```

5. Save the number of outlinks for each page separately.
6. You are welcome to use the following code template:

```
import scala.util.matching.Regex  
import org.apache.spark.SparkConf  
import org.apache.spark.SparkContext  
import org.apache.spark.SparkContext._  
import org.apache.hadoop.io.compress.GzipCodec  
  
object LinkGraph {  
  def main(args: Array[String]) {  
    val sparkConf = new SparkConf().setAppName("Wiki LinkGraph")  
    val sc = new SparkContext(sparkConf)  
    val input = sc.textFile("./wikiarticles") // your output directory from the last assignment  
  
    val page = input.map{ l =>  
      val pair = l.stripPrefix("(").stripSuffix(")").split("\t", 2)  
  
      (pair(0), pair(1)) // get the two fields: title and text  
    }  
  
    val links = page.map(r => (r._1, extractLinks(r._2))) // extract links from text  
    val linkcounts = links.map(r => (r._1, r._2.split("\t").length)) // count number of links  
  
    // save the links and the counts in compressed format (save your disk space)  
    links.saveAsTextFile("./links", classOf[GzipCodec])  
    linkcounts.saveAsTextFile("./links-counts", classOf[GzipCodec])  
  
  }  
  
  def extractLinks(text: String) : String = {  
    // you will need to work on a way to extract the links  
  }  
}
```

COSC 589 - Web Search and Sense-Making

What to Submit:

- Your code
- Screen captures of the beginning of your saved link graphs (e.g. the first 20 lines on the screen. Hint: Use 'gunzip part-00000.gz' to unzip, then use 'less yourfile' to view the documents and screen capture)
- Screen captures of the beginning of the saved counts of links (e.g. the first 20 lines on the screen. Hint: Use 'gunzip part-00000.gz' to unzip, then use 'less yourfile' to view the documents and screen capture)

What NOT to Submit:

- Your input or output files

Where to submit:

- Canvas

When:

- Due on Wednesday March 21 2018, 11:59pm.