STAT 6080 Data Technologies

Project Description

by

Jürgen Symanzik

Date: October 25, 2019

Final Due Date: Monday, December 9, 2019, 11:59pm (by e–mail)

UTAH STATE UNIVERSITY

Logan, UT

Fall 2019

Contents

1 General Instructions

Your project is related to the Marvel Cinematic Universe Wiki (about 21,000 web pages), see for example `https://marvelcinematicuniverse.fandom.com/wiki/Thanos`.

Questions one might want to answer are: Which Marvel hero (or villain) appears on most of these pages or is mentioned by name most times? Which heroes or villains are mentioned jointly (or in pairs of 2 or 3) most of the time?

You have to work in a group of five students on this project. Each student (or pair of students) is responsible for one primary task. All students are expected to frequently communicate and closely work together on overlapping questions. However, I want to make this a project that is open to comments and suggestions from the other students in class as well. So, you have to consider suggestions in class or in Canvas made by the other students from this course.

There are a few specific requirements and a timeline you have to follow. More details in the sections below.

2 Possible Research Questions

Almost 21,000 web pages exist at the Marvel Cinematic Universe Wiki. You have to mine all of them! You can start at one (or several) individual page such as `https://marvelcinematicuniverse.fandom.com/wiki/Thanos` and then search for other links on this page and branch out until you have found all of the pages. There is a (small) risk that not all of the pages are linked — so you may not be able to reach all of them this way. Alternatively, use the local sitemap at `https://marvelcinematicuniverse.fandom.com/wiki/Local_Sitemap`. This will hopefully list all pages (but no guarantee on this). In any case, keep track of how many different pages you have reached.

Here are some possible questions to investigate:

- Which Marvel hero (or villain) appears on most of these pages or is mentioned by name most times?
- Which heroes or villains are mentioned jointly (or in pairs of 2 or 3) most of the time?
- What is the real name, alias (one or more), species, gender, date of birth / date of death (possibly more than just one, based on comic books and movies), status (alive or deceased — again, this could be more than one that differs in books and movies), and other characteristics? Main characters have a short bio (in a table?) on their primary web page.
- Which of the about 21,000 pages are referred to most frequently from other web pages?
- Additional question you can think of or that are suggested by other students from this course.

There are three main components to this project:

- Download and process the about 21,000 web pages. Extract relevant information using regular expressions. You can use all R packages of your choice, even if not discussed in class.
- Store the relevant information in a local SQL data base. It is up to you to decide how much pre–processing you do prior to storing the information in the data base and which (and how many) tables you create. As an example, in the least pre–processed case, you start saving the entire web page in your data base. In the most pre–processed case, you only store the character bio, links, name counts, counts of other characters and places, etc. in your data base. There is no single correct answer here. You (as a group) have to make this decision.
- Develop a simple user interface that allows a user to type in questions such as "*How many of the pages contain the name IRON MAN?*" or "*Is IronMan more frequently listed together with Spiderman or with Thor?*" or "*what is the real name of 'ironman'* ". Think of other likely questions and ask the class for suggestions. Using regular expressions, translate these human questions into SQL queries that can be used to extract the relevant information from your data base.

3 Specific Instructions

- All your work most be done in R. You cannot manually manipulate files and data. Your code should be reproducible/reusable, in particular if additional web pages are created. For example, you shouldn't assume that there are exactly 20,781 web pages (as of October 25, 2019). Rather access all pages that exist when you run the final version of your R code. Also assume that someone may run your code again in a year after the release of the next Marvel movie(s) with new characters and new web pages.
- Download all web pages just once. Do the math yourself how long it will take if the download of a page only takes 1 sec each. Likely, it will take longer for each page. Decide at which time to process your downloaded web page and which information you want to store in your local SQL data base.
- Create an SQL data base that stores relevant information. Details (which information should be stored, how many/which tables should be created, etc.) are left to the group and can be decided during the progression of the project.
- Create a simple user interface where a human can enter a specific question. This question gets translated via regular expressions into an SQL query to extract the relevant information from your SQL data base. You should allow questions that are case insensitive, do not depend on small spelling differences (e.g., a space or hyphen between two–word names), and recognize all of the character names and places from the almost 21,000 web pages. Using information from the local sitemap may be helpful for this task. Details are left to the group and can be decided during the progression of the project. Nevertheless, this part has to be started at the same time as the two other parts of the project.

- Start exploring and processing some of the web pages. Initially work with a small sample of 20 or 30 characters, places, and things. Try different web pages and not just main characters so that your code does not fail, e.g., if there exists no character bio on a web page. Then refine and adjust your process and check your results for the next 20 or 30 characters, places, and things. Adjust again if necessary. Finally run for the entire set of about 21,000 web pages. This may be an overnight process without any further human interaction.

- Demonstrate that the user questions (as outlined above) will be answered in a meaningful way. Other students (and professors) should be allowed to type in the questions as well. So, if someone does not use a question mark at the end of a question, this still should be recognized as a valid question. All meaningful spellings for *Iron Man* (as discussed in class) also should result in the same answer.

4 Timeline

- You have to give a short 10 to 15 min progress report each week, usually on Thursdays, specifically on 10/17, 10/24, 10/31, 11/5 (Tuesday!), and 11/14. In these short progress reports, you should present what you have done and what your next planned steps are. The other students in class and I may ask questions and make suggestions. We all should agree what the next steps and priorities are. **Two students** are in charge of the progress report in a given week. Each student **must** be in charge exactly twice for the weekly progress reports.

- On Tuesday 11/19, you have to give a 30 min presentation of your final results. This should contain a brief introduction into comics, movies, and characters from the Marvel Universe (do a Google scholar search and note that there exist several scientific articles that deal with this topic!). The main part should be a summary of the computational methods and R packages you have used, and obviously a presentation of your results. End with a discussion and outlook on things you would have liked to do, but were not able to do given the time limits of this project. It is not necessary to come up with final answers to all initial questions. Use my LaTeX Beamer slides (in Presentation.zip) as a template for this presentation.

  While there could be a "narrator" for this presentation, each student most contribute to it and report about the main part to which he/she contributed. All students must be able to answer general questions about the project and specific questions about their main part.

- On Monday 12/9 at 11:59pm, your final written report is due. Likely you have to make a careful decision what to present in your written report. In a presentation, we can often present a lot of additional information that won't make it into the final written report due to space (i.e., page) limitations.

  Use my LaTeX template (in ProceedingsPaper.zip) as a template for this final report.

5 Grading

Each of your two progress reports in class is worth 5% of your total score, the final presentation in class on 11/19 is worth 40%, and the final written report due on 12/9 is worth 50% of your total score for this project.

6 Additional Requirements

- It is up to you how to split the individual tasks. Each group member **must** present exactly twice during the weekly progress reports. Each group member **must** contribute to the final presentation.
- It is not expected that each group member contributes exactly the same amount of time to a group project. However, no single group member is allowed to contribute more than 30% of the overall work for this project. Each group member that contributes less than 15% of the overall work to this project will get individual point deductions. You have to provide me with an estimated percentage of everyone's contributions to this project and the kind of contributions everyone made.
- Your 30 min presentation on Tuesday 11/19 should be created via LaTeX Beamer. I will provide a template. You have to turn in your final resulting pdf file and the Rnw/tex file.
- A final 6–page written report is due on Monday 12/9 at 11:59pm. This has to follow the JSM formatting requirements from the American Statistical Association (ASA). These 6 pages must contain everything from an abstract, keywords, the main sections of your report, all figures and tables, and the references. I will provide a template. In addition, you have to arrange your R files in a meaningful way as in Appendix A of your main report. There is no page limit for this appendix. LaTeX commands such as *verbatiminput* exist and allow you to create your R code independently and only include it into a document at the very last stage. You have to turn in your final pdf file and all source files.
- In Appendix B, provide a breakdown of everybody's percent-wise contribution to the project and the individual responsibilities. This is common for publications in the medical field. One example is shown in Appendix B. Apparently, I am "J.S.". This appendix will be removed from the final document upon (at least one) request before the document is shared with the other students from this course. Please e-mail prior (!) to the submission deadline. This appendix also does not count towards the 6–page limit of your written report.
- **Whenever you have questions or need clarifications, talk to me in person, via e–mail, or via Skype after 11/22. Good luck!**

Appendix A Rnw File for the Project

This appendix contains the Rnw file for this project description:

```
\documentclass[11pt]{article}

\usepackage{graphicx}
\usepackage{url}
\usepackage{hyperref}
\usepackage{verbatim}
\usepackage[title,titletoc,toc]{appendix}
\usepackage{wasysym}

\renewcommand{\topfraction}{1.0}
\renewcommand{\bottomfraction}{1.0}
\renewcommand{\textfraction}{0.0}
\renewcommand{\floatpagefraction}{1.0}
\renewcommand{\dbltopfraction}{1.0}

\textwidth 16cm
\textheight 22cm
\voffset -0.5in
\hoffset -0.5in

\parindent0pt
\setlength{\parskip}{1ex plus 0.5ex minus 0.2ex}


\begin{document}

\SweaveOpts{concordance=TRUE}

\begin{titlepage}

\begin{center}
{\large STAT 6080 Data Technologies} \\[4cm]

{\LARGE \bf Project Description} \\[1cm]
by \\[0.5cm]
{\bf J\"urgen Symanzik} \\[2.5cm]
{\bf Date:} \today \\[2cm]
{\bf Final Due Date:} Monday, December 9, 2019, 11:59pm (by e--mail) \\[2cm]


UTAH STATE UNIVERSITY \\[0.5cm]
Logan, UT \\[0.5cm]
Fall 2019 \\[0.5cm]
\end{center}

\thispagestyle{empty}
\vfill
\end{titlepage}


\newpage


\pagenumbering{roman}

\tableofcontents


\newpage
```

```
%\listoftables
%\addcontentsline{toc}{section}{List of Tables}
%
%\newpage
%
%\listoffigures
%\addcontentsline{toc}{section}{List of Figures}
%
%\newpage


\pagenumbering{arabic}

\section{General Instructions}

Your project is related to  the
Marvel Cinematic Universe Wiki (about 21,000 web pages),
see for example \url{https://marvelcinematicuniverse.fandom.com/wiki/Thanos}.

Questions one might want to answer are: Which Marvel hero (or villain) appears on
most of these pages or is mentioned by name most times? Which heroes or villains
are mentioned jointly (or in pairs of 2 or 3) most of the time?

You have to work in a group of five students
on this project. Each student (or pair of students) is responsible for one
primary task. All students are expected to frequently communicate and
closely work together on overlapping questions.
However, I want to make this a project that is open
to comments and suggestions from the other students in class as well.
So, you have to consider suggestions in class or in Canvas made
by the other students from this course.

There are a few specific requirements and a timeline
you have to follow. More details in the sections below.


\section{Possible Research Questions}

Almost 21,000 web pages exist at the Marvel Cinematic Universe Wiki.
You have to mine all of them! You can start at one (or several) individual
page such as \url{https://marvelcinematicuniverse.fandom.com/wiki/Thanos}
and then search for other links on this page and branch out until you
have found all of the pages. There is a (small) risk that not all of the
pages are linked --- so you may not be able to reach all of them this way.
Alternatively, use the local sitemap at
\url{https://marvelcinematicuniverse.fandom.com/wiki/Local_Sitemap}.
This will hopefully list all pages (but no guarantee on this).
In any case, keep track of how many different pages you have reached.


Here are some possible questions to investigate:
\begin{itemize}
\item Which Marvel hero (or villain) appears on
most of these pages or is mentioned by name most times?

\item Which heroes or villains
are mentioned jointly (or in pairs of 2 or 3) most of the time?

\item What is the real name, alias (one or more), species, gender, date of birth / date of death
(possibly more than just one, based on comic books and movies), status (alive or deceased ---
again, this could be more than one that differs in books and movies), and other
characteristics? Main characters have a short bio (in a table?) on their primary web page.

\item Which of the about 21,000 pages are referred to most frequently from other web pages?
```

```
\item Additional question you can think of or that are suggested by other students
from this course.
\end{itemize}


\newpage


There are three main components to this project:
\begin{itemize}
\item Download and process the about 21,000 web pages. Extract relevant information
using regular expressions. You can use all R packages of your choice, even if not
discussed in class.

\item Store the relevant information in a local SQL data base. It is up to you
to decide how much pre--processing you do prior to storing the information
in the data base and which (and how many) tables you create. As an example,
in the least pre--processed case, you start saving the entire web page in your data base.
In the most pre--processed case, you only store the character bio, links, name counts,
counts of other characters and places, etc.\ in your data base. There is no single
correct answer here. You (as a group) have to make this decision.

\item Develop a simple user interface that allows a user to type in questions such as
``{\it How many of the pages contain the name IRON MAN?}'' or
``{\it Is IronMan more frequently listed together with Spiderman or with Thor?}'' or
``{\it what is the real name of `ironman'}~''. Think of other likely questions
and ask the class for suggestions. Using regular expressions, translate these
human questions into SQL queries that can be used to extract the relevant information from your
data base.
\end{itemize}


\section{Specific Instructions}

\begin{itemize}
\item All your work most be done in R. You cannot manually manipulate files and data.
Your code should be reproducible/reusable, in particular if additional web pages
are created. For example, you shouldn't assume that there are exactly 20,781
web pages (as of October 25, 2019). Rather access all pages that exist when
you run the final version of your R code. Also assume that someone may run
your code again in a year after the release of the next Marvel movie(s)
with new characters and new web pages.

\item Download all web pages just once. Do the math yourself how long it will
take if the download of a page only takes 1~sec each. Likely, it will take longer
for each page. Decide at which time to process your downloaded web page
and which information you want to store in your local SQL data base.

\item Create an SQL data base that stores relevant information.
Details (which information should be stored, how many/which tables should be created, etc.)
are left to the group and can be decided during the progression of the project.

\item Create a simple user interface where a human can enter a specific question.
This question gets translated via regular expressions into an SQL query to extract the relevant information
from your SQL data base. You should allow questions that are case insensitive, do not depend
on small spelling differences (e.g., a space or hyphen between two--word names),
and recognize all of the character names and places from the almost 21,000 web pages.
Using information from the local sitemap may be helpful for this task.
Details are left to the group and can be decided during the progression of the project.
Nevertheless, this part has to be started at the same time as the two other parts of
the project.

\item Start exploring and processing some of the web pages.
Initially work with a small sample of 20 or 30 characters, places, and things.
```

```
Try different web pages and not just main characters so that your code does
not fail, e.g.,  if there exists no character bio on a web page.
Then refine and adjust your process and check your results for the next 20 or 30
characters, places, and things. Adjust again if necessary.
Finally run for the entire set of about 21,000 web pages.
This may be an overnight process without any further human interaction.

\item Demonstrate that the user questions (as outlined above) will be
answered in a meaningful way. Other students (and professors) should be
allowed to type in the questions as well. So, if someone does not
use a question mark at the end of a question, this still should be
recognized as a valid question. All meaningful spellings for {\it Iron Man}
(as discussed in class) also should result in the same answer.

\end{itemize}


\section{Timeline}

\begin{itemize}
\item You have to give a short 10 to 15~min progress report each week,
usually on Thursdays, specifically on 10/17, 10/24, 10/31, 11/5 (Tuesday!),
and 11/14. In these short progress reports,
you should present what you have done and
what your next planned steps are. The other students in class and I may ask questions and
make suggestions. We all should agree what the next steps and priorities are.
{\bf Two students} are in charge of the progress report in a given week.
Each student {\bf must} be in charge exactly twice for the weekly progress reports.

\item On Tuesday 11/19, you have to give a 30~min presentation of your final results.
This should contain a brief introduction into comics, movies, and characters
from the Marvel Universe (do a Google scholar search and
note that there exist several scientific articles that deal with this topic!).
The main part should be a summary of the
computational methods and R packages you have used, and obviously a presentation of your results.
End with a discussion and outlook on things you would have liked to do,
but were not able to do given the time limits of this project.
It is not necessary to come up with final answers to all initial questions.
Use my \LaTeX\ Beamer slides (in Presentation.zip) as a template for this presentation.

While there could be a ``narrator'' for this presentation, each student
most contribute to it and report about the main part to which he/she contributed.
All students must be able to answer general questions about the project
and specific questions about their main part.

\item On Monday 12/9 at 11:59pm, your final written report is due. Likely you have
to make a careful decision what to present in your written report. In a presentation,
we can often present a lot of additional information that won't make it into the
final written report due to space (i.e., page) limitations.

Use my \LaTeX\ template (in ProceedingsPaper.zip) as a template for this final report.

\end{itemize}


\section{Grading}

Each of your two progress reports in class is worth 5\% of your total score,
the final presentation in class on 11/19 is worth 40\%, and the final written report
due on 12/9 is worth 50\% of your total score for this project.


\section{Additional Requirements}

\begin{itemize}
```

```
\item It is up to you how to split the individual tasks.
Each group member {\bf must} present exactly twice during the weekly progress reports.
Each group member {\bf must} contribute to the final presentation.

\item It is not expected that each group member contributes exactly the same amount
of time to a group project. However, no single group member is allowed to
contribute more than 30\% of the overall work for this project. Each group
member that contributes less than 15\% of the overall work to this project
will get individual point deductions. You have to provide me with
an estimated percentage of everyone's contributions to this project
and the kind of contributions everyone made.

\item Your 30~min presentation on Tuesday 11/19 should be created via
\LaTeX\ Beamer. I will provide a template. You have to turn in
your final resulting pdf file and the Rnw/tex file.

\item A final 6--page written report is due on Monday 12/9 at 11:59pm. This has
to follow the JSM formatting requirements from the American Statistical Association (ASA).
These 6 pages must contain everything from an abstract, keywords, the main
sections of your report, all figures and tables, and the references.
I will provide a template. In addition, you have to arrange your R files in a
meaningful way as in Appendix~\ref{AppendixWithRCode} of your main report. There is no page limit
for this appendix. \LaTeX\ commands such as {\it verbatiminput} exist and allow you to create your
R code independently and only include it into a document at the very last stage.
You have to turn in your final pdf file and all source files.

\item In Appendix~\ref{Contributions}, provide a breakdown of everybody's percent-wise contribution to
the project and the individual responsibilities. This is common for publications
in the medical field. One example is shown in Appendix~\ref{Contributions}. Apparently,
I am ``J.S.''. This appendix will be removed from the final document
upon (at least one) request before the document is shared with the other students from this course.
Please e-mail prior (!) to the submission deadline. This appendix also does not count
towards the 6--page limit of your written report.

\item {\bf Whenever you have questions or need clarifications, talk to me in person,
via e--mail, or via Skype after 11/22. Good luck!}

\end{itemize}


\newpage


\begin{appendices}


\section{Rnw File for the Project}\label{AppendixWithRCode}

This appendix contains the Rnw file for this project description:

{
\scriptsize
\verbatiminput{DTProject.Rnw}
}


\newpage


\section{Author Contributions}\label{Contributions}

G.F. conceived the research, wrote the original manuscript, the
revised manuscript, and the response letter, and addressed the
reviewer comments; X.D. created the computer code, performed
the data analyses and created the figures; J.S. fine-tuned the figures
```

```
and verified the reproducibility of the results; S.B. performed
the STRUCTURE analysis and verified the biological interpretations;
all authors participated in discussions, read and revised the different
versions of the manuscript and the response letter, and agreed
to the submission.


\end{appendices}


\end{document}
```

Appendix B Author Contributions

G.F. conceived the research, wrote the original manuscript, the revised manuscript, and the response letter, and addressed the reviewer comments; X.D. created the computer code, performed the data analyses and created the figures; J.S. fine-tuned the figures and verified the reproducibility of the results; S.B. performed the STRUCTURE analysis and verified the biological interpretations; all authors participated in discussions, read and revised the different versions of the manuscript and the response letter, and agreed to the submission.