STAT 6910 Data Technologies

Project Description

by

Jürgen Symanzik

Date: November 7, 2017

Due Date: Monday, December 11, 2017, 11:59pm (by e–mail)

UTAH STATE UNIVERSITY

Logan, UT

Fall 2017

Contents

1 General Instructions

Your project is related to the previously classified documents from the Kennedy assassination that were released in 2017. The overall question is simple: What can be found in these documents?

You have to work in a small group of exactly two students on this project. However, I want to make this a project that is open to comments and suggestions from the other students in class as well.

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

2 Possible Research Questions

About 7500 previously classified documents from the Kennedy assassination were released in 2017, accessible at the *National Archives* web site `https://www.archives.gov/research/jfk/2017-release`. These documents were released in multiple stages and originate from various original agencies.

Here are some possible questions to investigate:

- How many documents were released from which agency?
- Were all documents from an agency released at the same time? Or, were most documents from Agency A released early in 2017 while most documents from Agency B were released late in 2017?
- What is the distribution with respect to time of the original documents, i.e., when were most documents created and when were the final ones created?
- What type of documents have been released? Are all pdf files, or are there also audio, video, and photo files in the 2017 releases?
- What is the total number of pages in all these documents? And what is the page distribution?
- What is the total number of words (before / after the removal of "stop words" — see below) in all these documents? And what is the word count distribution?
- What is the content of the pdf files? Be aware that some files may not be processed via R because of handwritten notes or poor photocopies that were eventually scanned. More specifically, what are the main actors (e.g., Kennedy, Johnson, Oswald, Ruby), main locations (e.g., Dallas, Cuba, Soviet Union), and main events (e.g., Cuban Missile Crisis, Space Race) that may have been mentioned in these files?
- Can you group / cluster the documents according to their content? Are there differences in content based on release date of the documents and/or with respect to the agency that provided the documents? Work with a basic Jaccard index (use `https://en.wikipedia.`

org/wiki/Jaccard_index as a start to find additional references — but you can't use any wikipedia page as an official scientific reference). Also google for *jaccard similarity in r*.

- How can you best summarize your results via tabular and/or graphical approaches? Does a static (or even interactive? — think of plotly) version of a heatmap / correlogram work? See in particular Figure 6 of https://cran.r-project.org/web/packages/corrgram/vignettes/corrgram_examples.html.

- What are some of the most unusual documents? CNN reported that a file related to Martin Luther King was among these documents (see http://www.cnn.com/2017/11/03/politics/martin-luther-king-document-in-jfk-files/index.html).

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 documents are released. For example, you shouldn't assume that there have been 3 release dates of documents in 2017, but rather whatever is stated on the *National Archives* web site when you access it.

- Download all documents just once. Be prepared to download additional documents if those get released during the course of your project. Store the documents locally and/or in a shared box or dropbox account you can always access. You want to avoid to depend on an external source (e.g., what would you do if the site is down for maintenance?).

- Explore and process some of the documents. This should start with some of the steps we did in class, but should go beyond. You also need to remove "stop words" (such as *the, a, an, and, at, so, etc.* — the `tm` R package will help) and invalid character sequences. You may also have to adjust for common scanning problems similar to the Scowcroft issue we have resolved in class. Initially work with a small random sample of 20 or 30 documents. You have to decide whether this should be a simple random sample or a stratified random sample (in that case, you also need to decide how to stratify, e.g., by agency, by period, or by something else).

- Once you are satisfied with your initial results, automatically process all 7500 documents. Split your documents into batches and do this step on as many computers as possible. If you know from a previous class how to use the University of Utah's Center for High Performance Computing (CHPC), then do so (see https://rgs.usu.edu/about-hpc). Ideally, you want to do this step only once, but if you notice a mistake later on, you may want to reprocess a subset of the documents that may have been affected by this mistake.

- Store your results, i.e., the extracted text, in the same local shared box or dropbox location. Use matching file names.

- Calculating the Jaccard indexes for about 7500×7500 documents will take a while. So again, you may have to do this in batches and store your results. Definitely start with your previous sample of 20 or 30 documents and check that you get plausible results, based on visually screening and comparing these documents. As before, ideally, you want to do this

step only once, but if you notice a mistake later on, you may want to reprocess a subset of the documents that may have been affected by this mistake.

- Store your results, i.e., the Jaccard indexes, in the same local shared box or dropbox location. Use a suitable data structure / file format that can be easily updated / extended if necessary. Note that $7500^2 = 56250000$.

4 Timeline

- There is a first 10 min presentation of your initial exploration of the documents due in class on Fr 11/17. You should present what you have done and what your next planned steps are. The other students and I may ask questions and make suggestions. We all should agree what the next steps and priorities are.
- There is a second 10 min presentation of your exploration of the documents due in class on Fr 12/1. You should present what you have done and what your next planned steps are. By now, you should have processed all documents. The other students and I may ask questions and make suggestions again. We all should agree what the final steps and priorities are.
- On Fr 12/8, you have to give a 20 min presentation of your final results. This should contain a brief (historic) introduction, 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.
- On Mo 12/11 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.

5 Grading

The two short presentations in class on 11/17 and 12/1 each are worth 15% of your total score, the final presentation in class on 12/8 is worth 30%, and the final written report on 12/11 is worth 40% of your total score for this project.

6 Additional Requirements

- It is up to you how to split the individual tasks. However, during each of the presentations each group member has to present about the same amount of time.
- Your 20 min presentation on Fr 12/8 should be created via LaTeX beamer. I will provide a template. You have to turn in your final pdf file and the Rnw/tex file.

- A final 6–page written report is due on Mo 12/11 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 an appendix to 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.
- **Whenever you have questions or need clarifications, talk to me in person, via e–mail, or via Skype during the Thanksgiving week. 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 6910 Data Technologies} \\[4cm]

{\LARGE \bf Project Description} \\[1cm]
by \\[0.5cm]
{\bf J\"urgen Symanzik} \\[2.5cm]
{\bf Date:} \today \\[2cm]
{\bf Due Date:} Monday, December 11, 2017, 11:59pm (by e--mail) \\[2cm]


UTAH STATE UNIVERSITY \\[0.5cm]
Logan, UT \\[0.5cm]
Fall 2017 \\[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 previously classified documents from the
Kennedy assassination that were released in 2017. The overall question
is simple: What can be found in these documents?

You have to work in a small group of exactly two students
on this project. However, I want to make this a project that is open
to comments and suggestions from the other students in class as well.

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


\section{Possible Research Questions}

About 7500 previously classified documents from the
Kennedy assassination were released in 2017, accessible
at the {\it National Archives} web site
\url{https://www.archives.gov/research/jfk/2017-release}.
These documents were released in multiple stages and
originate from various original agencies.

Here are some possible questions to investigate:
\begin{itemize}
\item How many documents were released from which agency?

\item Were all documents from an agency released at the same time?
Or, were most documents from Agency~A released early in 2017 while
most documents from Agency~B were released late in 2017?

\item What is the distribution with respect to time of the original documents,
i.e., when were most documents created and when were the final ones created?

\item What type of documents have been released? Are all pdf files,
or are there also audio, video, and photo files in the 2017 releases?

\item What is the total number of pages in all these documents? And what is
the page distribution?

\item What is the total number of words (before / after the removal of ``stop words'' ---
see below) in all these documents? And what is the word count distribution?

\item What is the content of the pdf files? Be aware that some files may not be processed
via R because of handwritten notes or poor photocopies that were eventually scanned.
More specifically, what are the main actors (e.g., Kennedy, Johnson, Oswald, Ruby),
main locations (e.g., Dallas, Cuba, Soviet Union), and main events
(e.g., Cuban Missile Crisis, Space Race) that may have been mentioned in these files?

\item Can you group / cluster the documents according to their content? Are there
differences in content based on release date of the documents and/or with respect
```

```
to the agency that provided the documents? Work with a basic Jaccard index
(use \url{https://en.wikipedia.org/wiki/Jaccard_index}
as a start to find additional references --- but you can't use any wikipedia
page as an official scientific reference). Also google for
{\it jaccard similarity in r}.

\item How can you best summarize your results via tabular and/or graphical approaches?
Does a static (or even interactive? --- think of plotly) version of a heatmap / correlogram work?
See in particular Figure~6 of
\url{https://cran.r-project.org/web/packages/corrgram/vignettes/corrgram_examples.html}.

\item What are some of the most unusual documents? CNN reported that a file
related to Martin Luther King was among these documents
(see \url{http://www.cnn.com/2017/11/03/politics/martin-luther-king-document-in-jfk-files/index.html}).

\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 documents
are released. For example, you shouldn't assume that there have been 3 release dates
of documents in 2017, but rather whatever is stated on the {\it National Archives} web site
when you access it.

\item Download all documents just once. Be prepared to download additional
documents if those get released during the course of your project.
Store the documents locally and/or in a shared box or dropbox account
you can always access. You want to avoid to depend on an external source
(e.g., what would you do if the site is down for maintenance?).

\item Explore and process some of the documents. This should start with some of the steps we
did in class, but should go beyond. You also need to remove ``stop words''
(such as {\it the, a, an, and, at, so, etc.} --- the {\tt tm} R package will help) and
invalid character sequences. You may also have to adjust for common
scanning problems similar to the Scowcroft issue we have resolved in class.
Initially work with a small random sample of 20 or 30 documents.
You have to decide whether this should be a simple random sample
or a stratified random sample (in that case, you also need to decide
how to stratify, e.g., by agency, by period, or by something else).

\item Once you are satisfied with your initial results, automatically
process all 7500 documents. Split your documents into batches and
do this step on as many computers as possible. If you know from a previous class how to use
the University of Utah's Center for High Performance Computing (CHPC),
then do so (see \url{https://rgs.usu.edu/about-hpc}). Ideally, you want to do this
step only once, but if you notice a mistake later on, you may want
to reprocess a subset of the documents that may have been affected
by this mistake.

\item Store your results, i.e., the extracted text, in the same local
shared box or dropbox location. Use matching file names.

\item Calculating the Jaccard indexes for about 7500 $\times$ 7500
documents will take a while. So again, you may have to do this in
batches and store your results. Definitely start with your previous
sample of 20 or 30 documents and check that you get plausible results,
based on visually screening and comparing these documents.
As before, ideally, you want to do this
step only once, but if you notice a mistake later on, you may want
to reprocess a subset of the documents that may have been affected
by this mistake.
```

```
\item Store your results, i.e., the Jaccard indexes, in the same local
shared box or dropbox location. Use a suitable data structure / file
format that can be easily updated / extended if necessary.
Note that $7500^2 = 56250000$.

\end{itemize}


\section{Timeline}

\begin{itemize}
\item There is a first 10~min presentation of your initial exploration of the
documents due in class on Fr 11/17. You should present what you have done and
what your next planned steps are. The other students and I may ask questions and
make suggestions. We all should agree what the next steps and priorities are.

\item There is a second 10~min presentation of your exploration of the
documents due in class on Fr 12/1.
You should present what you have done and
what your next planned steps are.
By now, you should have processed all documents.
The other students and I may ask questions and
make suggestions again. We all should agree what the final steps and priorities are.

\item On Fr 12/8, you have to give a 20~min presentation of your final results.
This should contain a brief (historic) introduction, 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.

\item On Mo 12/11 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.

\end{itemize}


\section{Grading}

The two short presentations in class on 11/17 and 12/1 each are worth 15\% of your total score,
the final presentation in class on 12/8 is worth 30\%, and the final written report
on  12/11 is worth 40\% of your total score for this project.


\section{Additional Requirements}

\begin{itemize}
\item It is up to you how to split the individual tasks.
However, during each of the presentations each group member
has to present about the same amount of time.

\item Your 20~min presentation on Fr 12/8 should be created via
\LaTeX\ beamer. I will provide a template. You have to turn in
your final pdf file and the Rnw/tex file.

\item A final 6--page written report is due on Mo 12/11 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 an appendix to 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 {\bf Whenever you have questions or need clarifications, talk to me in person,
via e--mail, or via Skype during the Thanksgiving week. 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{Project_01.Rnw}

\end{appendices}


\end{document}
```