Protecting Privacy in the Archives: Preliminary Explorations of Topic Modeling for Born-Digital Collections

Tim Hutchinson
13 December 2017
Guiding questions/context

- High volume of records – not feasible to review individually
- Public archives subject to privacy legislation
- Can natural language processing (NLP) techniques help identify records requiring restrictions?
Topic modeling

- Ted Underwood – Topic Modeling Made Just Simple Enough

  a) “Say we’ve got a collection of documents, and we want to identify the underlying ‘topics’ that organize the collection. Assume that each document contains a mixture of different topics. Let’s also assume that a ‘topic’ can be understood as a collection of words that have different probabilities of appearance in passages discussing the topic... Topic modeling is a way of extrapolating backward from a collection of documents to infer the [topics] that could have generated them.”
Archives and NLP – BitCurator
BitCurator NLP

About the project

The BitCurator NLP project began on October 1, 2016 and will end on September 30, 2018. BitCurator NLP is funded through a grant from the Andrew W. Mellon Foundation.

The BitCurator NLP project will develop software for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections. The software will use existing natural language processing software libraries to identify and report on those items likely to be relevant to ongoing preservation, information organization, and access activities. These may include entities (e.g., persons, places, and organizations), potential relationships among entities (for example, by describing those entities that appear together within documents or set of documents), and topic models to provide insight into how concepts are naturally clustered within the documents.

Visit the BitCurator NLP wiki page for technical content, documentation, and software downloads.
Generate topic models from open text extracted from files in disk images [https://wiki.bitcurator.net/index.php...](https://wiki.bitcurator.net/index.php...)

86 commits
8 branches
4 releases
3 contributors
LGPL-3.0

Branch: master  New pull request  Create new file  Upload files  Find file  Clone or download

- kamwoods: Cleaned up config file
  - attic: Cleaned up attic and legacy bootstrap
  - disk-images: test image - 2
  - externals: Adding libewf image to master
  - .gitignore: Initial commit
  - LICENSE: Initial commit
  - README.md: Updated README
  - bcnlp_listfiles.py: Initial commit - extension from dfvfs's example file list_file_entry
  - bcnlp_tm.py: Minor changes
  - bn_fileextract.py: Used dfvfs for file listing
  - bn_plot.py: fixed the issues caused by changes to bn_fileextract.py
  - bntm_config.txt: Cleaned up config file
  - setup.sh: Setup fix with aptdcon

Last commit 8/9/2020 on 25 Sep

Latest commit 2 months ago
4 months ago
3 months ago
a year ago
2 months ago
2 months ago
3 months ago
2 months ago
2 months ago
2 months ago
Finished parsing file /home/vagrant/shared/SampleFiles/KavanaghText/alan_notes.pdf.txt
Parsing completed.Parsed 51 lines in 0.010434 secs.

Graphlab: Creating topic model with 10 topics:

Learning a topic model

<table>
<thead>
<tr>
<th>Number of documents</th>
<th>227608</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary size</td>
<td>63605</td>
</tr>
</tbody>
</table>

Running collapsed Gibbs sampling

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Elapsed Time</th>
<th>Tokens/Second</th>
<th>Est. Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3.90s</td>
<td>2.40288e+06</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>7.10s</td>
<td>2.51494e+06</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>10.35s</td>
<td>2.17212e+06</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>13.51s</td>
<td>2.49997e+06</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>16.72s</td>
<td>2.35382e+06</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>19.86s</td>
<td>2.47944e+06</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>23.14s</td>
<td>2.39346e+06</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>26.26s</td>
<td>2.37089e+06</td>
<td>0</td>
</tr>
<tr>
<td>90</td>
<td>29.41s</td>
<td>2.43954e+06</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>32.71s</td>
<td>2.14154e+06</td>
<td>0</td>
</tr>
</tbody>
</table>

Graphlab: Preparing data:
Archives and NLP - ArchExtract
An Information Extraction and Text Exploration Tool for Digital Collections

ArchExtract is a web application that enables archivists and researchers to perform topic modeling, keyword and named entity extraction on a text collection. The web application automates and packages a number of existing natural language processes and algorithms for the researcher or archivist. Using automated text analysis as the starting point, ArchExtract illuminates the scope and content of a digital text collection and provides a web-based interface for text exploration.

Prerequisites

Python libraries

- Scipy, numpy and la (for matrix manipulation), nltk (for language processing), bottleneck, mysql-connector-python, pyner (for stanford ner), mrjob (for python mapreduce), fuzzy_wuzzy and python-Levenshtein (computing string edit distances and similarities)
- For nltk, you will also need to download the nltk data => nltk.download()

Java Dependencies

- You need to have java installed on the server, java 7 or above; java 8 is preferable

Installing Mallet

- Mallet is open source software used for topic modeling
- Download mallet and unzip it on your system.
- Put %MALLE_HOME% in your path => needs to point to where ever bin/mallet is on your system.
# Collection Listing

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Acquisition Date</th>
<th>Import Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>vp_info_technology</td>
<td>2017-09-12</td>
<td></td>
</tr>
<tr>
<td>AVP IT</td>
<td>2017-09-12</td>
<td>uploading files</td>
</tr>
<tr>
<td>AVP IT 2</td>
<td>2017-09-12</td>
<td>uploading files</td>
</tr>
<tr>
<td>AVP IT 3</td>
<td>2017-09-12</td>
<td>uploading files</td>
</tr>
<tr>
<td>AVP IT 4</td>
<td>2017-09-12</td>
<td>uploading files</td>
</tr>
<tr>
<td>AVP IT 5</td>
<td>2017-09-13</td>
<td>failed</td>
</tr>
<tr>
<td>test 1</td>
<td>2017-09-13</td>
<td>complete</td>
</tr>
<tr>
<td>test 2</td>
<td>2017-09-13</td>
<td>complete</td>
</tr>
<tr>
<td>AVP IT 6</td>
<td>2017-09-13</td>
<td>complete</td>
</tr>
<tr>
<td>AVP IT 7</td>
<td>2017-09-13</td>
<td>complete</td>
</tr>
<tr>
<td>AVP IT 8</td>
<td>2017-09-13</td>
<td>complete</td>
</tr>
<tr>
<td>AVP IT 9</td>
<td>2017-09-13</td>
<td>complete</td>
</tr>
<tr>
<td>DS1</td>
<td>2017-09-13</td>
<td>complete</td>
</tr>
</tbody>
</table>
# Pre-Processes

<table>
<thead>
<tr>
<th>Collection</th>
<th>Routine Name</th>
<th>Status</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>Removed Stop Words, Removed Rare Words</td>
<td>Complete</td>
<td>2017-09-15</td>
</tr>
<tr>
<td>DS1</td>
<td>Tagged--Nouns--Verbs, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words</td>
<td>Complete</td>
<td>2017-09-15</td>
</tr>
<tr>
<td>DS1</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words</td>
<td>Complete</td>
<td>2017-09-15</td>
</tr>
<tr>
<td>AVP PDF Sample 2</td>
<td>Tagged--Nouns--Verbs, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words</td>
<td>Complete</td>
<td>2017-09-18</td>
</tr>
<tr>
<td>AVP PDF Sample 2</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words</td>
<td>Processing</td>
<td>2017-09-18</td>
</tr>
<tr>
<td>AVP PDF Sample 2</td>
<td>Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words</td>
<td>Processing</td>
<td>2017-09-18</td>
</tr>
<tr>
<td>AVP 1000b</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words</td>
<td>Complete</td>
<td>2017-09-18</td>
</tr>
<tr>
<td>AVP Personnel</td>
<td>Tagged--Nouns--Verbs--Adjectives, Stemmed, Removed Stop Words, Removed Rare Words</td>
<td>Complete</td>
<td>2017-10-03</td>
</tr>
</tbody>
</table>
Create a New Topic Model

Collection
AVP IT 8

Build a Topic Model
Tagged--Nouns--Verbs, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words

* Number of Topics to Generate
20

Submit

Topic Modeling Home
Select a Collection

* Collection

AVP IT 8

Stop Word Filtering

Removes Stop Words: Removes common words like the, a, it

- Yes
- No

Removes Rare Words: Removes words that only appear once in the document

- Yes
- No
Tagging - Part of Speech Filtering

**Tagging:** Include ONLY the part of speech words selected

- [ ] Nouns
- [ ] Verbs
- [ ] Adjectives
- [ ] Proper Names

**Removes Named Entities When Tagging:** Removes words that are identified as named entities and dates

- [ ] Yes
- [ ] No

Stemming
Stem: Turn each word into its stem

- Yes
- No

TF-IDF Filtering

(Note: This option cannot be combined with stemming or tagging)

TF-IDF: Filter words based on their importance in the document. Scores each word in the collection based on how frequently the word appears and how infrequently the word appears in other documents

Filter out words that are below the mean tfi-idf score for the entire collection

- Yes
- No
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ds1</td>
<td>Plain Text with 20 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Ds1</td>
<td>Removed Stop Words, Removed Rare Words with 20 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp Pdf Sample 2</td>
<td>Tagged--Nouns--Verbs, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words with 7 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp 1000b</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words with 10 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp 1000b</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words with 20 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp 1000b</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words with 15 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp 1000b</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words with 40 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp 1000b</td>
<td>Tagged--Nouns--Verbs--Adjectives, Removed Named Entities, Stemmed, Removed Stop Words, Removed Rare Words with 100 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp Personnel</td>
<td>Tagged--Nouns--Verbs--Adjectives, Stemmed, Removed Stop Words, Removed Rare Words with 10 Topics</td>
<td>Failed</td>
</tr>
<tr>
<td>Avp Personnel</td>
<td>Tagged--Nouns--Verbs--Adjectives, Stemmed, Removed Stop Words, Removed Rare Words with 1 Topics</td>
<td>Failed</td>
</tr>
</tbody>
</table>
Topic 10: meeting bob committee task chair draft advisory recommendations discussion members report executive minutes mail force december action ed dcs full

Topic 11: management budget projects financial project external including impact institutional activities reporting personnel funds contract cost benefits resources implementation operating progress

Topic 12: equipment system cost software computer computers support total systems number campus capital service maintenance purchase desk computing macintosh request usage

Topic 13: college science office education university computer arts academic student physical technology computing support students engineering faculty current group library research

Topic 14: ca canada net industry internet national province st networks usask provincial canadian fax canarie board phone government federal organizations high

Topic 15: time resources make committee work number major costs year current additional project process total problem capital funding future feel system
Top documents associated with this topic:

- notes%20on%20conceptual%20view → 0.60701
- harassment%20message → 0.50787
- Cover%201998-99 → 0.38649
- Comp → 0.30208
- %20nov%2093 → 0.19457
- Suggested%20Planning%20Cycle → 0.16722
- fax → 0.1506
- DMR%20Invite → 0.51946
- Sept_91_material → 0.48397
- Priority%20Determination → 0.34977
- → 0.26451
- Final%20memo → 0.21795
- UofS%20membership%2098-99 → 0.18403
- Inf%20defn → 0.16599
- → 0.13636
MEMORANDUM

To: Members of ASPPC
From: KNK
Date: Wednesday, August 14, 1991
Subject: Some Material for Meeting of Sept. 5

University of Saskatchewan
Office of Information Technology Services

- 1 -
Preliminary results

Department of Computing Services

56 Physics Building
116 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada S7N 5E2

General Office: (306) 966-4866
Help Desk: (306)966-4817
FAX: 966-4938
Inquiries: help.desk@usask.ca
Web page comments: webmaster@www.usask.ca

About Us

- General Information
- Announcements
- What's New
- Search DCS Web Pages

DCS Offices

- Main Office, 56 Physics, (306) 966-4866
- Help Desk, 70 Arts Building, Help Line Phone (306) 966-4817
- Campus Computer Store, 85 Place Riel, (306) 966-8375
<table>
<thead>
<tr>
<th>Name</th>
<th>Date modified</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountability</td>
<td>9/11/2008 12:41 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>adv comm</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Archive</td>
<td>9/11/2008 12:32 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>BAH retire</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Budgets</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>CCS</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>COFT</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Colleges</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>corresp</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Ctl</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Distance Education</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Employment Equity</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Equipment Issues</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Facilities Mgmt</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Faculty Presentation</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Gartner Group</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Giga</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>GIS</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
<tr>
<td>Groups</td>
<td>9/11/2008 12:42 PM</td>
<td>File folder</td>
</tr>
</tbody>
</table>
HR-related topics – 10 topics

- people group management good work technical staff team systems position time change skill personnel support employees manager part job training

- areas area working level making basis addition past order expect issue made effort important time day feel make special general
HR-related topics – 15 topics

- management work support department computing time part team campus system project year resources personnel manager administration services development human systems
- staff technical change groups people effective assist ways group training related skills experience required times maintain communication successful person members
HR-related topics – 20 topics

- staff management work group team department good manager areas skills position client services year performance time part effort recommendation special
- management budget projects financial project external including impact institutional activities reporting personnel funds contract cost benefits resources implementation operating progress
- matter feel people notes general {first name A} unit mail talk {first name B} leave pressure sense asked result personally assignment happy things show
HR-related topics – 40 topics (1)

- management project resources systems process system support human functional make plan information processes role strategic procedures scope costs requirements departments
- manager services staff department work {first name C} management technical client {last name C} time computing September july {first name D} position personnel consulting {first name B} including
HR-related topics – 40 topics (2)

- issues support objectives work working developing important group met problems ideas responsibility summary efforts questions concerns role objective results people
- performance part year recommendation areas team time special good salary makes progress skills staff develop effectiveness job employees excellent leadership
- issue situation problem made case question concern long matter short money quickly regard put things past point view felt good
HR-related topics – 100 topics (1)

- employees responsibility actions group respect decisions continuing provide attention general member manner set personal employee person clients term members supervisor

- mail power manner harassment history conversation requests advantage departmental treatment knew behaviour differential drive starting acceptable thought game excessive morale

- performance part good effectiveness salary recommendation areas group year excellent appraisal employees role team complete makes assistant carry special significant
HR-related topics – 100 topics (2)

- position management experience support client manager change director personnel department team work staff positions duties job years successful aspa level
- employment equity group responsible assist managers communication term issues tasks selection collaboration efficient information people application processes communicate effective organization
- completed leave employee aspa improvement increased dcs units cupe ability responsibilities member project communicate market forward added directly similar supervisory
HR-related documents – top 40

HR-related documents: top 40 documents per topic

HR (40)  non-HR (40)
HR-related documents – top 10

HR-related documents: top 10 documents per topic

- A5
- A8
- B4
- B12
- C7
- C11
- C17
- D6
- D8
- D11
- D23
- D25
- E7
- E24
- E25
- E27
- E55
- E88

HR (10)  non-HR (10)
A combined view

HR-related documents - top 40 and 10

The diagram shows a combined view of HR-related documents with top 40 and 10. The chart displays the distribution of HR and non-HR documents, with categories labeled from A5 to E88. The categories are color-coded to signify HR (40), non-HR (40), HR (10), and non-HR (10) documents.
Best performing topics

- A8 (10): people group management good work technical staff team systems position time change skill personnel support employees manager part job training

- B4 (15): management work support department computing time part team campus system project year resources personnel manager administration services development human systems

- E25 (100): performance part good effectiveness salary recommendation areas group year excellent appraisal employees role team complete makes assistant carry special significant
Highest rank

Highest rank

A5  A8  B4  B12  C7  C11  C17  D6  D8  D11  D23  D25  E7  E24  E25  E27  E55  E88
Finding known HR documents

- 79 documents in personnel subfolders (DCS)
- 75 assigned to at least one HR-related topic
  a) But 500 documents in each topic
HR documents – range of ranks

Known HR documents - range of ranks

- min
- max
- avg
HR-only topic

- department services {last name H} computing date time subject {first name H} work staff management university support performance campus group president development areas building
Observations / further research

- Topic modeling more successful for high-level ID
- Need to refine topic modeling techniques
- Analyze document overlap between topics
- Analyze document scores/rankings
- Train topics models on documents known to be relevant
- Sentiment analysis
- Named entity extraction