University of Pittsburgh

Automatic Bookmark Organizing with Text Clustering

Daniel Zheng (daniel.zheng@pitt.edu)

December 7, 2017

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Automatic Bookmark Organizing

Originally wanted to clean up my bookmarks bar..
100s, maybe 1000s of bookmarks...

- ▶ Disorganized
- ▶ Some (failed) attempts at using folders
- ▶ Hard to find things

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Automatic Bookmark Organizing

How should they be organized?

- ▶ Folders
  - ▶ Subfolders!
    - ▶ More subfolders?
- ▶ Folders also need to be properly **labeled**

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Automatic Bookmark Organizing

Idea: Make a Chrome extension that takes list of bookmarks (URLs), automatically clusters them into folders, and rearranges your bookmarks bar!
Requirements:

- ► Making extension
- ► Hierarchical Clustering
- ► Accurate labeling

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Motivation
Simplifications

Data

Clustering

K-means
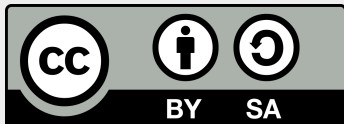K-means
labeling
K-means
results
K-means
Conclusions

Hierarchical
Clustering
Hierarchical
Results

Evaluation

Simplifications

Focusing on the clustering and labeling for now. Also adopted a fake "bookmarks list" $\rightarrow$ a dump of Wikipedia.

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Wikipedia Corpus

Chose GNU GPLv3 for my
repository

- ▶ Downloaded 2GB
  compressed XML files,
  resulted in 12GB unzipped
  dataset.
- ▶ Used Wikiextractor library
  to convert to JSON with
  `title, id, text, url`
- ▶ Additional custom
  formatting

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Clustering
How does clustering work?

- ▶ Hierarchical
  - ▶ Ward's Method
  - ▶ More detailed and useful structure
- ▶ K-means
  - ▶ Faster
  - ▶ Single layer of clustering

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Motivation
Simplifications

Data

Clustering

K-means
K-means
labeling
K-means
results
K-means
Conclusions

Hierarchical
Clustering
Hierarchical
Results

Evaluation

K-means Clustering

- ▶ Unsupervised learning algorithm
- ▶ Produce k clusters by finding k centroids and a label for each data point, where each observation belongs to the cluster with the nearest centroid(mean).
- ▶ First, initialize centroids. Then,
    1. For each point, calculate Euclidean distance to each centroid, assigning it to the closest one.
    2. Centroids are then recalculated to be the averages of all the points in each cluster.

    Repeat 1 and 2 until no change in centroids.

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus

Millions of articles

- First run on a sample and inspect results
    - tf-idf vectors of document text
    - K-means clustering
    - Sparse $\rightarrow$ dense vectors using TSVD
    - Visualize with t-SNE
        - Maps high-dimensional data to 2D space where similar data is grouped closer together.

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus

Some articles beginning with A

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means labeling

How can we generate labels automatically?

- ► Using spacy, extract noun phrases and remove stopwords
- ► Within clusters, take the most frequent occurring topics
- ► Topic extraction with gensim did not work very well
  - ► Slow and complex
  - ► Good for extracting topics describing a document
  - ► For cluster labeling, looking for overlap between many documents
  - ► Simple noun phrase approach turned out to be good

University of Pittsburgh

K-means for Wikipedia Corpus

Some articles beginning with A

Cluster 0: Antigua and Barbuda, Apollo program, NYSE MKT, Australian Labor Party, Ainu people, Aga Khan I, Aga Khan III, Alexander Agassiz, American Chinese cuisine, List of Governors of Alabama, Antarctic Treaty System, Alfred Lawson, Americans with Disabilities Act of 1990, Army, AOL, Anarchism, Alabama, Abraham Lincoln, Academy Awards, Algeria, Austin (disambiguation), Andorra, American Football Conference, Anarcho-capitalism, Demographics of Antigua and Barbuda, Politics of Antigua and Barbuda, Afghanistan, Albania, Azerbaijan, American Revolutionary War, ...

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Motivation
Simplifications

Data

Clustering

K-means
K-means
labeling
K-means
results
K-means
Conclusions

Hierarchical
Clustering
Hierarchical
Results

Evaluation

K-means for Wikipedia Corpus

Some articles beginning with A

Cluster 0: Antigua and Barbuda, Apollo program, NYSE MKT, Australian Labor Party, Ainu people, Aga Khan I, Aga Khan III, Alexander Agassiz, American Chinese cuisine, List of Governors of Alabama, Antarctic Treaty System, Alfred Lawson, Americans with Disabilities Act of 1990, Army, AOL, Anarchism, Alabama, Abraham Lincoln, Academy Awards, Algeria, Austin (disambiguation), Andorra, American Football Conference, Anarcho-capitalism, Demographics of Antigua and Barbuda, Politics of Antigua and Barbuda, Afghanistan, Albania, Azerbaijan, American Revolutionary War, ...

My Label: Geography
Top 5 computed labels: Lincoln, Azerbaijan, Jackson, the United States, the country

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus

Some articles beginning with A

Cluster 1: Agnosticism, Author, Andrey Markov, A. A. Milne, Alvin Toffler, Albert Speer, Abdul Alhazred, Ada Lovelace, August Derleth, Albert Camus, Agatha Christie, The Plague, Hercule Poirot, Miss Marple, Allen Ginsberg, Anatoly Karpov, Alan Kay, Alain de Lille, Alfred Russel Wallace, Aimoin, Albertus Magnus, Alfred Nobel, Alexander Graham Bell, Andy Warhol, ...

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus
Some articles beginning with A

Cluster 1: Agnosticism, Author, Andrey Markov, A. A. Milne, Alvin Toffler, Albert Speer, Abdul Alhazred, Ada Lovelace, August Derleth, Albert Camus, Agatha Christie, The Plague, Hercule Poirot, Miss Marple, Allen Ginsberg, Anatoly Karpov, Alan Kay, Alain de Lille, Alfred Russel Wallace, Aimoin, Albertus Magnus, Alfred Nobel, Alexander Graham Bell, Andy Warhol, ...

My Label: People
Top 5 computed labels: Crowley, Aristotle, Speer, Einstein, Wallace

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus

Some articles beginning with A

Cluster 2: August, August 22, August 27, August 6, August 23, August 24, August 31, August 9, August 13, August 2, August 7, August 8, August 14, August 15, August 16, August 17, August 12, August 18, August 19, August 21, August 25, August 1, August 3

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Motivation
Simplifications

Data

Clustering

K-means
K-means
labeling
K-means
results
K-means
Conclusions

Hierarchical
Clustering
Hierarchical
Results

Evaluation

K-means for Wikipedia Corpus
Some articles beginning with A

Cluster 2:   August, August 22, August 27, August 6, August
23, August 24, August 31, August 9, August 13, August 2,
August 7, August 8, August 14, August 15, August 16, August
17, August 12, August 18, August 19, August 21, August 25,
August 1, August 3

My Label: August
Top 5 computed labels: August, place, July, the peak, the
year

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus

Some articles beginning with A

Cluster 9: April, April 6, April 12, April 15, April 30, April 28, April 22, April 1, April 16, April 29, April 13, April 26
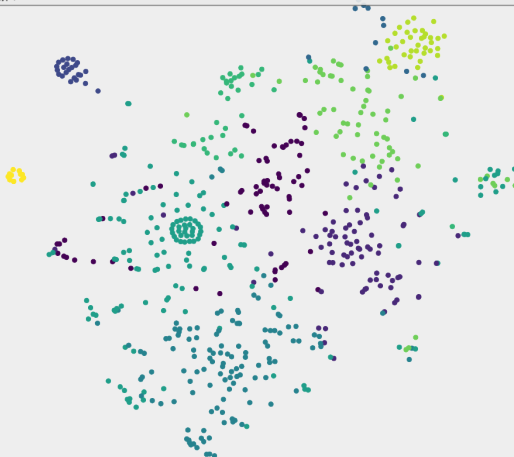
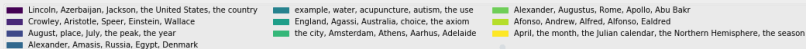University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus

Some articles beginning with A

Cluster 9: April, April 6, April 12, April 15, April 30, April 28, April 22, April 1, April 16, April 29, April 13, April 26

My Label: April
Top 5 computed labels: April, the month, the Julian calendar, the Northern Hemisphere, the season

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means labels

0	Lincoln, Azerbaijan, Jackson, the United States, the country
1	Crowley, Aristotle, Speer, Einstein, Wallace
2	August, place, July, the peak, the year
3	Alexander, Amasis, Russia, Egypt, Denmark
4	example, water, acupuncture, autism, the use
5	England, Agassi, Australia, choice, the axiom
6	the city, Amsterdam, Athens, Aarhus, Adelaide
7	Alexander, Augustus, Rome, Apollo, Abu Bakr
8	Afonso, Andrew, Alfred, Alfonso, Ealdred
9	April, the month, the Julian calendar, the Northern Hemisphere, the season

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

K-means for Wikipedia Corpus

t-SNE with generated labels

Legend:
- Lincoln, Azerbaijan, Jackson, the United States, the country
- Crowley, Aristotle, Speer, Einstein, Wallace
- August, place, July, the peak, the year
- Alexander, Amasis, Russia, Egypt, Denmark
- example, water, acupuncture, autism, the use
- England, Agassi, Australia, choice, the axiom
- the city, Amsterdam, Athens, Aarhus, Adelaide
- Alexander, Augustus, Rome, Apollo, Abu Bakr
- Afonso, Andrew, Alfred, Alfonso, Ealdred
- April, the month, the Julian calendar, the Northern Hemisphere, the season

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Motivation
Simplifications

Data

Clustering
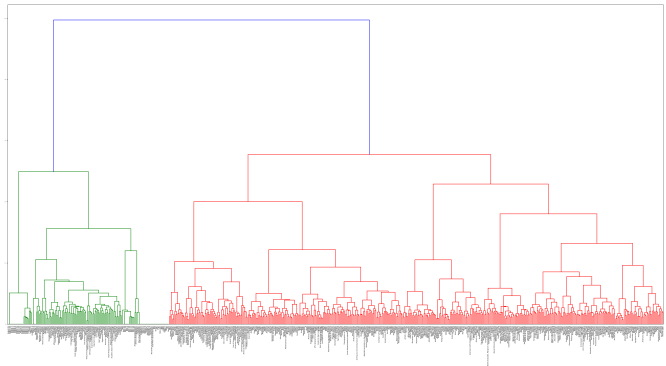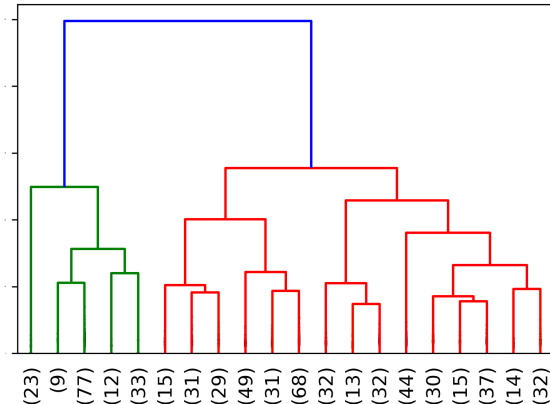
K-means
K-means
labeling
K-means
results
K-means
Conclusions

Hierarchical
Clustering
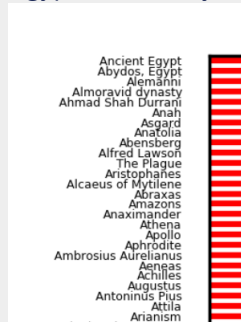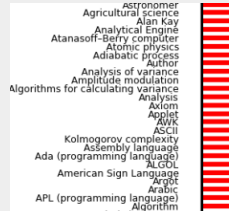Hierarchical
Results

Evaluation

K-means Conclusions

K-means is pretty good.

- ▶ Number of clusters (k) is a bit arbitrary
- ▶ Some clusters turned out well, but others cover a wide range of topics

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Hierarchical Clustering

Agglomerative clustering builds a binary tree

- ▶ Where K-means requires a parameter k and initial centroids, agglomerative clustering just requires some distance metric
  - ▶ Used cosine similarity
- ▶ Algorithm:
  1. Place each data point in its own group
  2. Repeatedly merge two closest groups until everything is in a single cluster

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Hierarchical Clustering Results

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Hierarchical Clustering Results

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Motivation
Simplifications

Data

Clustering

K-means
K-means
labeling
K-means
results
K-means
Conclusions

Hierarchical
Clustering
Hierarchical
Results

Evaluation

Hierarchical Clustering Results

Egypt, Greek Mythology

Some technical subjects

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Hierarchical Clustering Results

Dinosaurs!

Aardvark
Albertosaurus
Allosaurus
Apatosaurus

Geography

Angolan Armed Forces
Army
Geography of Armenia
Geography of Azerbaijan
Atlantic Ocean
Ashmore and Cartier Islands
Geography of Alabama
Geography of Antarctica
Aalborg Municipality
Aegean Sea
Archipelago
Adrian
Anadyr River
Alabama River
Northern cavefish
Amazon River
Aare
Abydos

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Hierarchical Clustering Results

April

August

University of Pittsburgh

Automatic
Bookmark
Organizing
with
Text
Clustering

Dan

Future work

Can't measure how good Wikipedia clusters are.

- ▶ Figure out SciPy dendrogram construction
  - ▶ Define hierarchical labels
- ▶ Define my own metric (web scraping?)
- ▶ Run on more data
- ▶ Maybe try different corpus (20 newsgroups?)

University of Pittsburgh