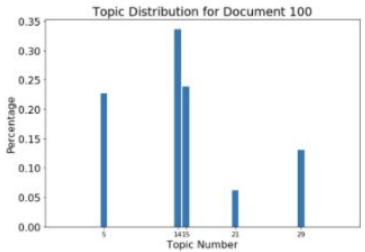
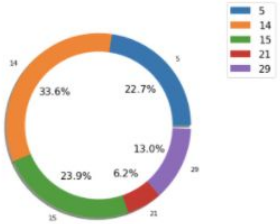
NLP and Topic Modeling

USC CARC Summer Boootcamp 2021
Asya Shklyar

How I learned about NLP - Economics - Job Ads





History

Professor Manisha Goel (Econ) partners with

BGT (Burning Glass Technologies - a recruiting company

Signs a contract to get a large dataset of job ads, descriptions, skills, pay information etc (10 years worth, 2005-2015)

Has a student write a Python script to do some basic topic modeling using doc2vec and nltk/gensim

Random 50 samples completes just fine but 500 runs out of RAM even on a 512 GB Linux VM

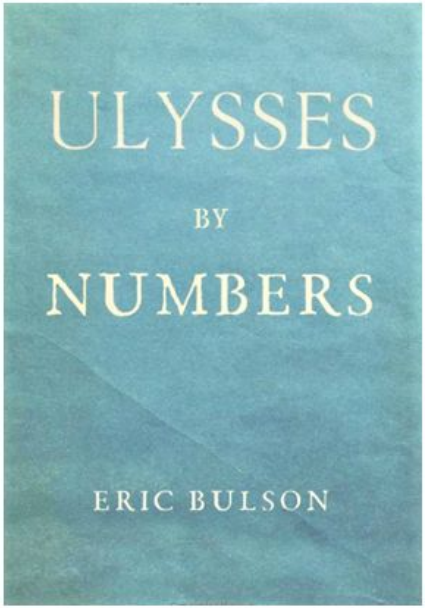
Also runs for days instead of hours

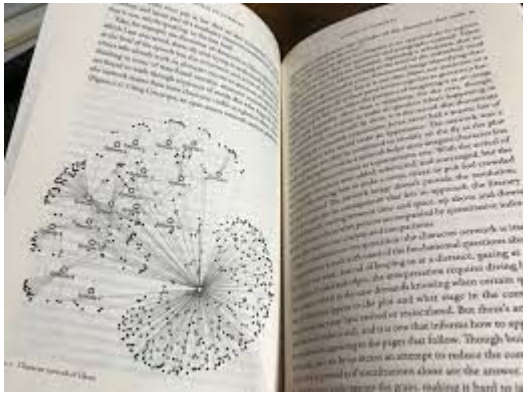
https://docs.google.com/presentation/d/1cjvibAlrjyMgQrPvlv75uRtN81ds6oki3fsWpYieAyg/edit?usp=sharing

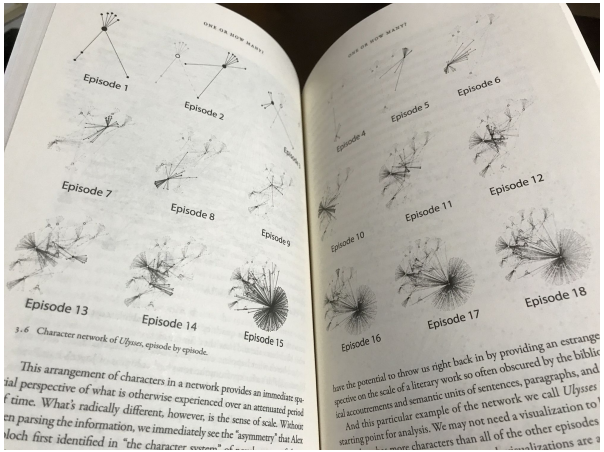
Some concepts related to NLP one needs to understand:

- Hadoop is pretty much done (but it used to be really cool - the distributed file system accessible on all nodes similar to a parallel file system)
- Scala is a programming language that works similarly to Java (and uses JVM) but is much faster and was designed to be distributed
- Spark is a distributed framework that uses Scala to send really large datasets to multiple nodes, perform various NLP actions and get the results back
- NLU or Natural Language Understanding is a subset dealing with computers trying to be human :)

Other notable uses of NLP







https://www.cgu.edu/news/2021/03/in-bulsons-new-book-understanding-joyce-is-as-easy-as-1-2-3/

https://lithub.com/how-to-read-ulysses-by-numbers/

A whole book on applications of topic models:

Foundations and Trends® in Information Retrieval
Vol. XX, No. XX (2017) 1–154
© 2017
DOI: 10.1561/XXXXXXXXXX

now
the essence of knowledge

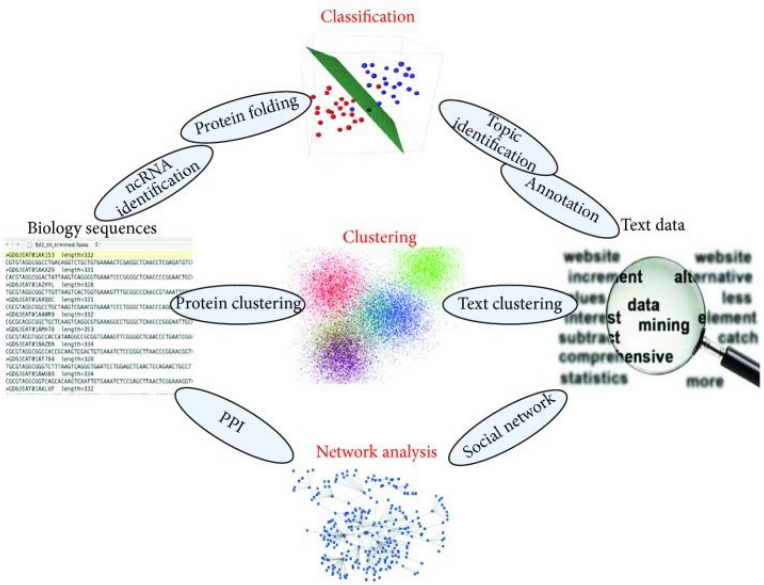
Applications of Topic Models

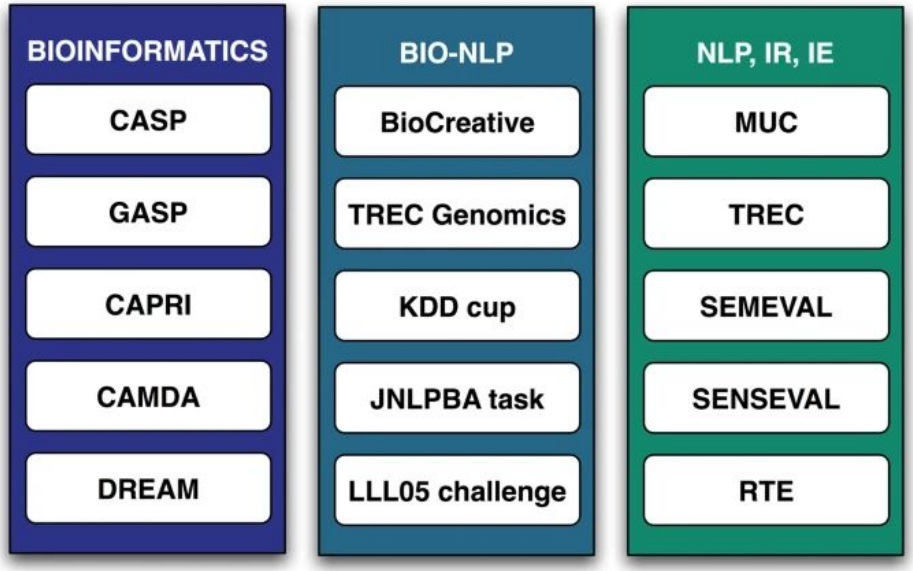
Jordan Boyd-Graber
Department of Computer Science, UMIACS, Language Science
University of Maryland[1]
jbg@umiacs.umd.edu

Yuening Hu
Google, Inc.[2]
ynhu@google.com

David Mimno
Information Science
Cornell University
mimno@cornell.edu

Other notable uses of NLP





https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4615216/

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-s2-s1

Prep for hands on work:

Mac:

https://brew.sh/

/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)”

https://formulae.brew.sh/cask/anaconda

brew install --cask anaconda

conda install jupyter or conda install jupyterlab (explain the difference as well as JupyterHub)

https://jupyter.org/install

Might have to set the path: export PATH=$PATH:/usr/local/anaconda3/bin

jupyter-lab

https://en.wikipedia.org/wiki/Natural_language_processing

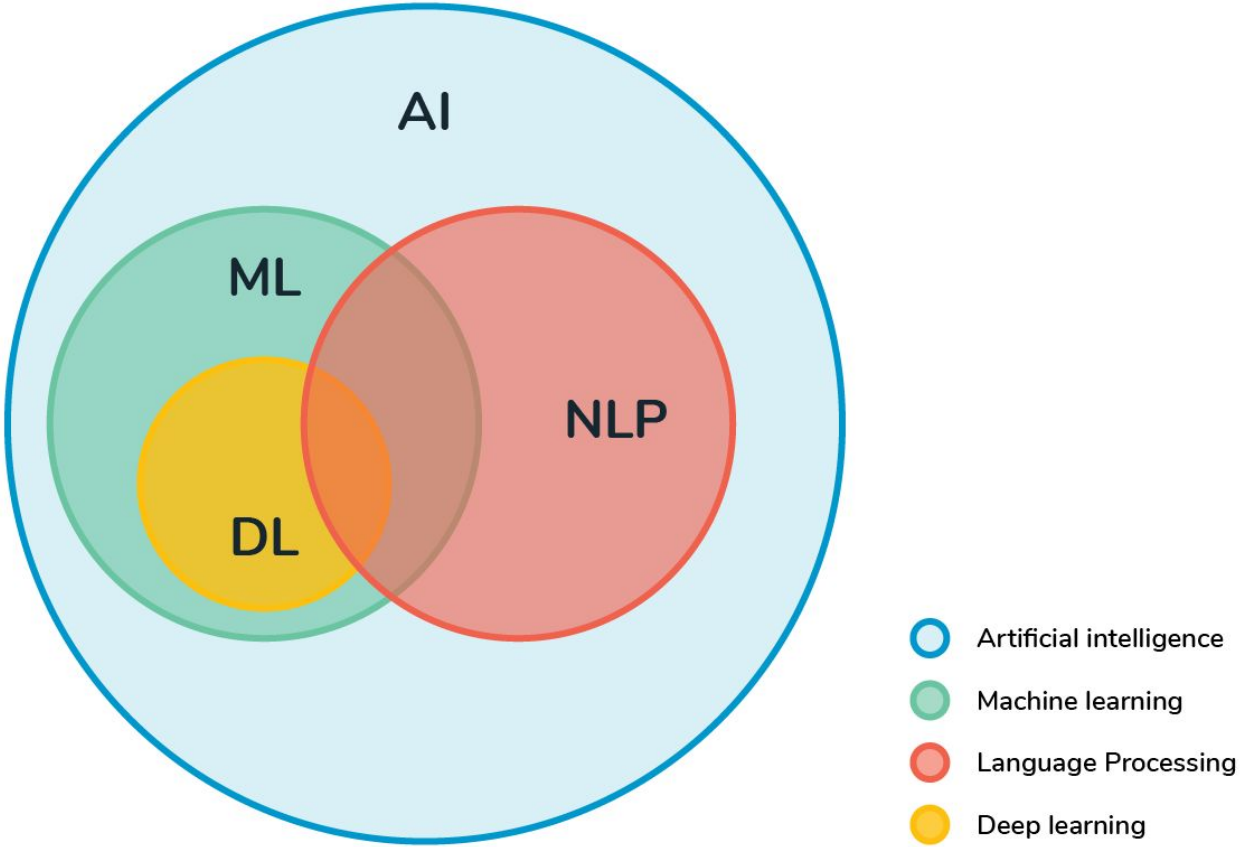
What is NLP

Natural language processing

From Wikipedia, the free encyclopedia

Natural language processing (**NLP**) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural-language generation.



https://devopedia.org/natural-language-processing

https://en.wikipedia.org/wiki/Topic_model

What is Topic Modeling

Topic model

From Wikipedia, the free encyclopedia

In machine learning and natural language processing, a **topic model** is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear approximately equally in both. A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about cats and 90% about dogs, there would probably be about 9 times more dog words than cat words. The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

Topic models are also referred to as probabilistic topic models, which refers to statistical algorithms for discovering the latent semantic structures of an extensive text body. In the age of information, the amount of the written material we encounter each day is simply beyond our processing capacity. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies. Originally developed as a text-mining tool, topic models have been used to detect instructive structures in data such as genetic information, images, and networks. They also have applications in other fields such as bioinformatics[1] and computer vision.[2]

LDA

Abstract

We describe *latent Dirichlet allocation* (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

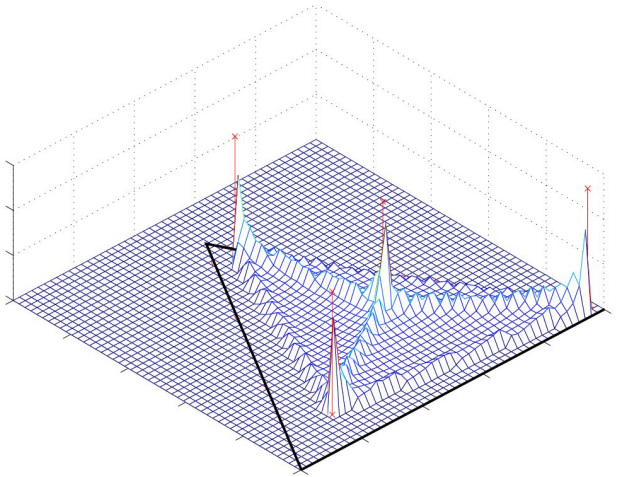
Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

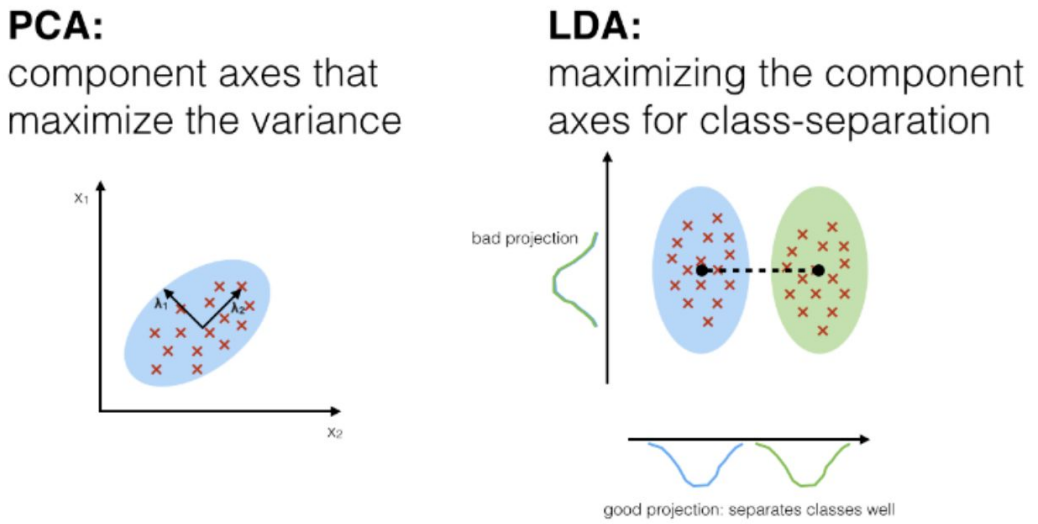
Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Editor: John Lafferty



https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

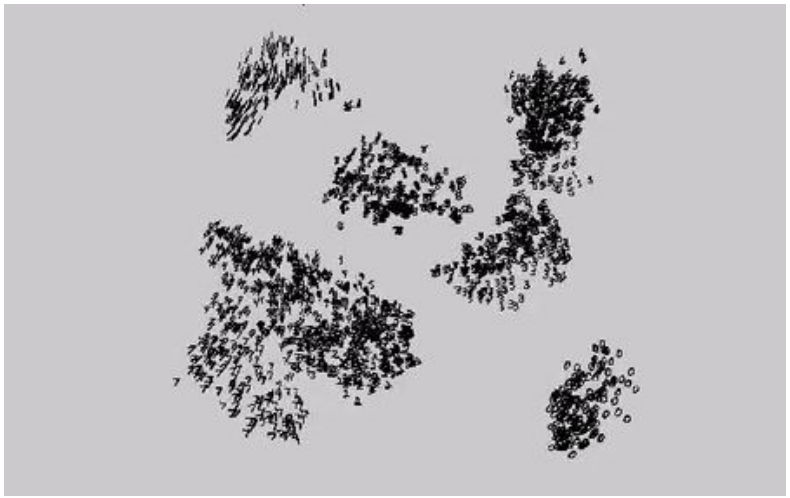
Other Algorithms

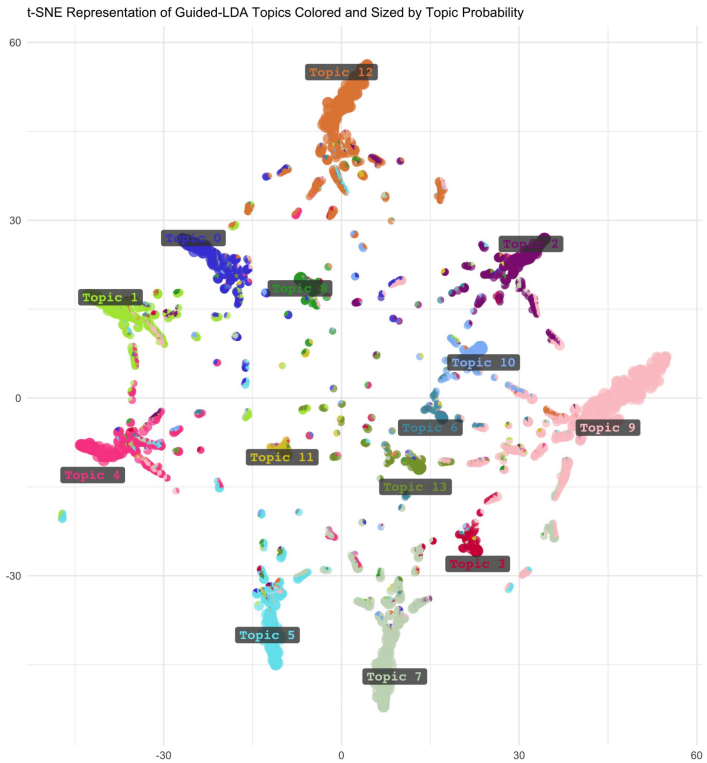


LDA vs PCA

https://medium.com/analytics-vidhya/pca-vs-lda-vs-t-sne-lets-understand-the-difference-between-them-22fa6b9be9d0

Visualization



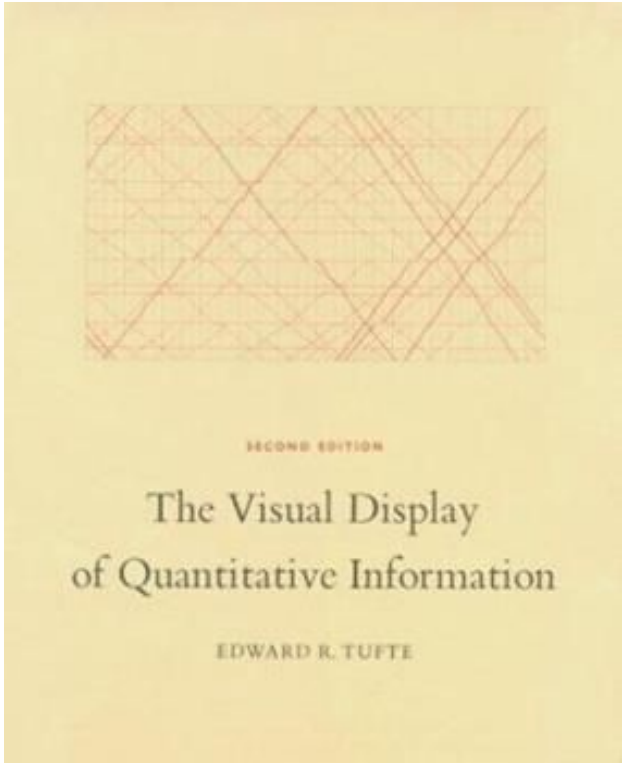


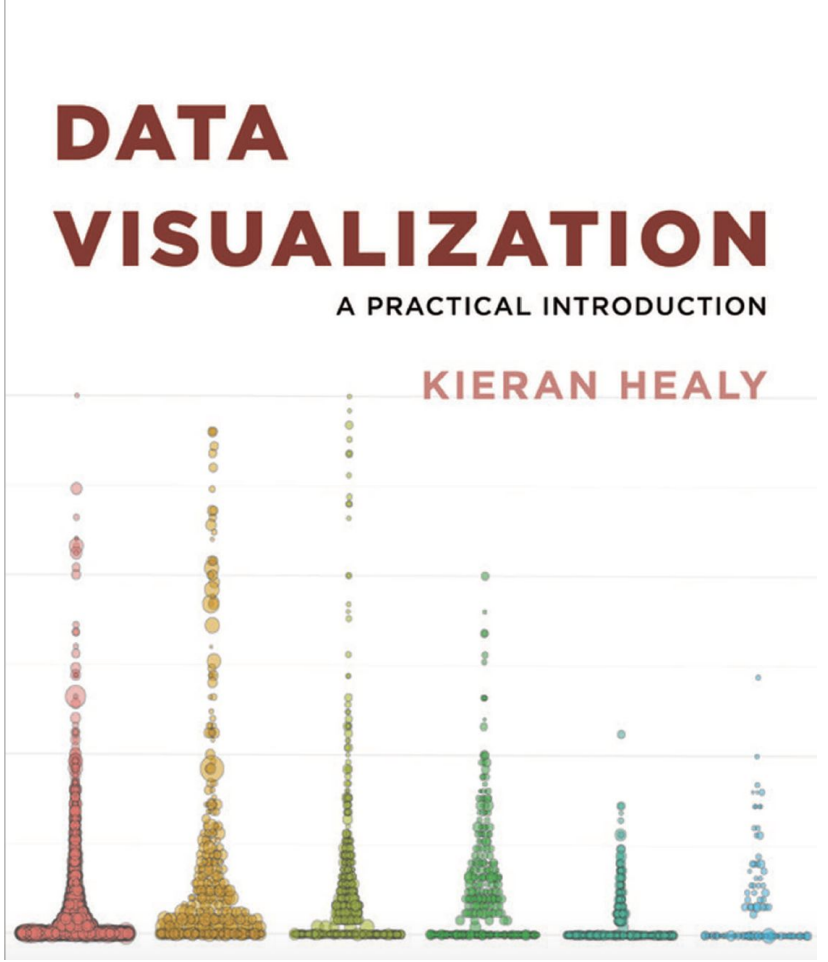
https://towardsdatascience.com/visualizing-topic-models-with-scatterpies-and-t-sne-f21f228f7b02

LDAViz - Java-based visualization tool

https://sicss.io/2020/materials/day3-text-analysis/topic-modeling/rmarkdown/Topic_Modeling.html

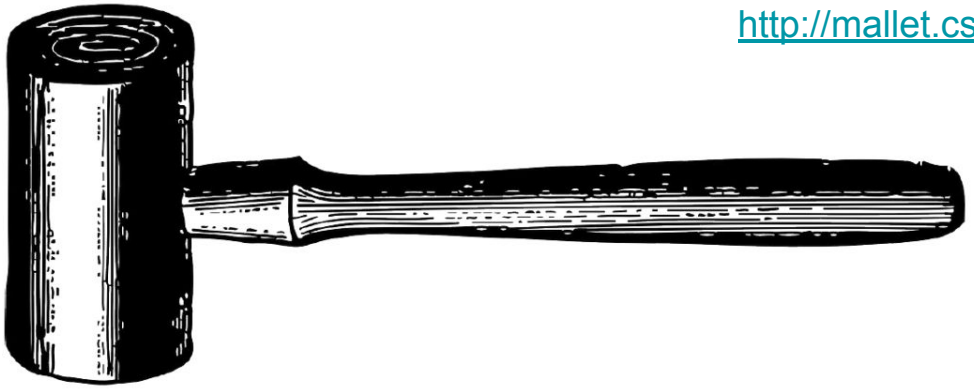
Books about Visualization





https://www.amazon.com/Data-Visualization-Introduction-Kieran-Healy/dp/0691181624

http://mallet.cs.umass.edu/mallet-tutorial.pdf



Getting started with MALLET

Xanda Schofield, xanda@cs.hmc.edu

Difference between Mallet and Spark

Mallet - single computer/small(ish) dataset

Spark - scales to dozens or hundreds of computers depending on the size of the dataset

1.5 TB, 130 million documents, 20,000 words each - > Spark

A book or a collection of news articles or a Twitter scrape - Mallet

Both Java/Scala-based though...

Sp

03: Intro Spark Apps

Spark Essentials

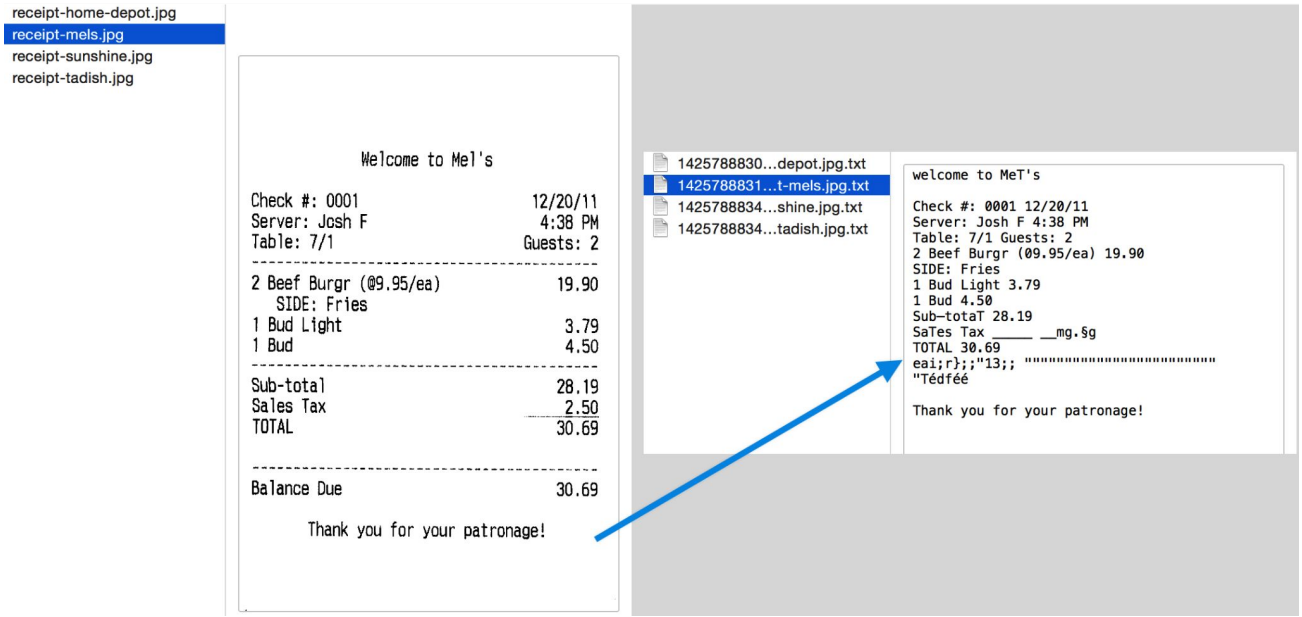
lecture/lab: 45 min

Scraping

- Well structured URLS vs Not
- OCR artifacts
- HTML/CSS junk - Beautiful Soup etc
- PDF to TXT conversion
- Cloud limitations
- Data viz tools: Gradient, Alteryx, John Snow Labs, Databricks etc

https://www.sec.gov/edgar/searchedgar/companysearch.html

Data Cleaning



https://medium.com/nanonets/a-comprehensive-guide-to-ocr-with-tesseract-opencv-and-python-fd42f69e8ca8

https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa

Jupyter Doc Demo