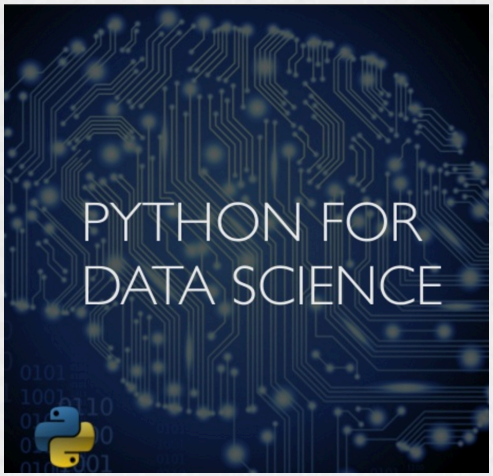
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS



Python for Data Science

Lecture 13 (04/18, 04/20): Text Mining (2)

Decision, Operations & Information Technologies
Robert H. Smith School of Business
Spring, 2016

Roadmap

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Introduction
- Sentiment identification & classification
- Topic modeling
- Key packages

Amazon.com product reviews

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

amazon Try Prime | All | Departments | Shopping History | Kunpeng's Amazon.com | Today's Deals | Gift Cards | Sell | Help | Hello, Kunpe Your Acco

Nikon Coolpix L340 20.2 MP Digital Camera with 28x Optical Zoom and... › Customer Reviews

Customer Reviews

368

4.4 out of 5 stars

5 star	68%
4 star	19%
3 star	4%
2 star	5%
1 star	4%

Nikon Coolpix L340 20.2 MP Digital Camera with 28x Optical Zoom and 3.0-Inch LCD (Black)

by Nikon

Price: $134.95 + Free shipping with Amazon Prime

Rate this item

Write a review

Top positive review

See all 321 positive reviews ›

289 people found this helpful

So far, so great!

By nikki_marie05 on December 1, 2015

I ordered this at Target online on Black Friday with free shipping and I am happy to say that I love it. I've only had it a couple of days, but I've been taking numerous types of pictures to try it out and I love it. I have searched and searched and searched for a camera that takes good pictures up close AND far away, for a decent price. I've went through many that were under $100 and they were crap. I paid right at $100 for this camera (before tax and warranty) and I think that is a steal for the quality of pictures that it takes. It's no professional camera, but when you go from an iPhone camera to this, it feels like you are a professional! I am 100% satisfied with the camera, the quality of pictures it takes, and how quickly the camera actually takes the pictures (there is no lag). Those were my main three concerns when purchasing this camera and I am very happy with my purchase.

Top critical review

See all 47 critical reviews ›

91 people found this helpful

Disappointed.

By Lauren Cutrone on June 26, 2015

I bought this camera as a new blogger who wanted to add some color with photography. Unfortunately, I couldn't really make it work with this camera. I saw similar reviews on other Nikons about issues with focus and that's precisely the problem with this camera. I would go to take a picture, hold down the button halfway, and it would focus.

For a millisecond. And then the picture would go blurry again.

If I worked through the settings, I could maybe get the picture that I wanted but it just wasn't worth it. For the price I paid, it shouldn't be that hard to get a decent picture. I shouldn't have to struggle every time I want to photograph something as simple as a book just sitting on my bed.

I'll be returning this one. It's a bummer because I was so excited to get a camera as nice as this one. I should've played around with it first before committing to it.

Sort by: Top | Filter by: All reviewers | 3 star only | Keyword | Search

Showing 1-10 of 17 reviews (3 star). Show all reviews

Very very blurry

By knifeguy on April 7, 2016

Verified Purchase

All the pictures are blurry unless you and the thing you are taking a picture of are very, very still. Quality was awesome, very user friendly. But even in sport mode, everything comes out blurry. Great for nature shots with nothing moving.

Comment | 3 people found this helpful. Was this review helpful to you? Yes No Report abuse

Ok camera

By nina young on January 21, 2016

Verified Purchase

The camera asthetically is cute and compact, however in my opinion works along the same lines as a regular digital camera , just has the look of a more professional one. Pictures come out clear so long as the lighting is good or flash is on. Not a bad camera but I probably should of put the money towards a more professional one. Point blank it's ok.

Comment | 3 people found this helpful. Was this review helpful to you? Yes No Report abuse

Not an Upgrade.. Dissapointing.. options suck, video audio sucks

By Bizz on March 29, 2016

Verified Purchase

Facts vs. opinions

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Facts: objective expressions about entities, events and their attributes.
  - ❑ "I bought a laptop yesterday"
- Opinions: subjective expressions of sentiments, attitudes, emotions, appraisals or feelings toward entities, events, and their attributes.
  - ❑ "I really love this laptop"

Some exceptions

- Not all subjective sentences contain opinions
  - ❑"I want to buy a cellphone with good screen quality"
- Not all objective sentences contain no opinions
  - ❑ "This earphone is broken in just two days"

Sentiment analysis applications

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Brand analysis
- Marketing
- Customer voice: e.g., products, hotels, airlines, etc.
- Event monitoring, e.g., site outrage, political campaign, etc.
- …

Direct vs. comparative opinions

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Direct opinion: sentiment expressions on one or more attributes of an object, e.g. products, services, events
  - "The voice quality of this phone is fantastic"
  - "After taking this medicine, my left knee feels worse"
- Comparative opinion: relations expressing similarities or differences among two or more objects on some of the shared attributes of the objects, e.g.
  - "The voice quality of camera x is better than that of camera y"

Explicit vs. implicit opinions

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Sentence subjectivity: an objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs
- Explicit opinion: an opinion on an attribute explicitly expressed in a subjective sentence
  - "The voice quality of this phone is amazing"
  - "This camera is too heavy"
- Implicit opinion: an opinion on an attribute implied in an objective sentence, e.g.
  - "The headset broke in two days"
  - "Please bring back the old search"

Sentiment analysis tasks

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Goal: identify and classify opinions
- Task 1: sentiment identification (subjectivity identification): identify whether a piece of text expresses opinions
- Task 2: sentiment orientation classification: determine the orientation of an opinionated text

Sentiment analysis levels

- Document level: identify if the document expresses opinions and whether the opinions are positive, negative, or neutral
- Sentence level: identify if a sentence is an opinionated and whether the opinions are positive, negative, or neutral
- Attribute-level: extract the object attributes (e.g. image quality, zoom size, price) that are the subject of an opinion and the opinion orientations

Opinion words

- Also known as polarity words, sentiment words, opinion lexicon or opinion-bearing words
  - Positive: good, amazing, wonderful, …
  - Negative: horrible, disgusting, poor, bad, …
- Base type (examples above) and comparative type (e.g. better, worse, similar to)

A simple method counting opinion words

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Opinion words: dominating indicators of sentiments, especially adjectives, adverbs, and verbs, e.g. "I absolutely love this camera. It is amazing!"
- Predefined opinion words: good, terrible, …
- Assignment orientation score (+1, -1) to all words
  - Positive opinion words (+1)
  - Negative opinion words (-1)
- The orientation score of the document is the sum of the orientation scores of all opinion words found

Rule-based method

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Is simply counting opinion words good enough? No
  - ❑ "There is not one thing I hate about this product"
- We need to handle negation: "not…hate" implies like
  - ❑ Simple rules can be manually created
    - "not…negative" → positive
    - "never…negative" → positive

Limitations of dictionary-based approach

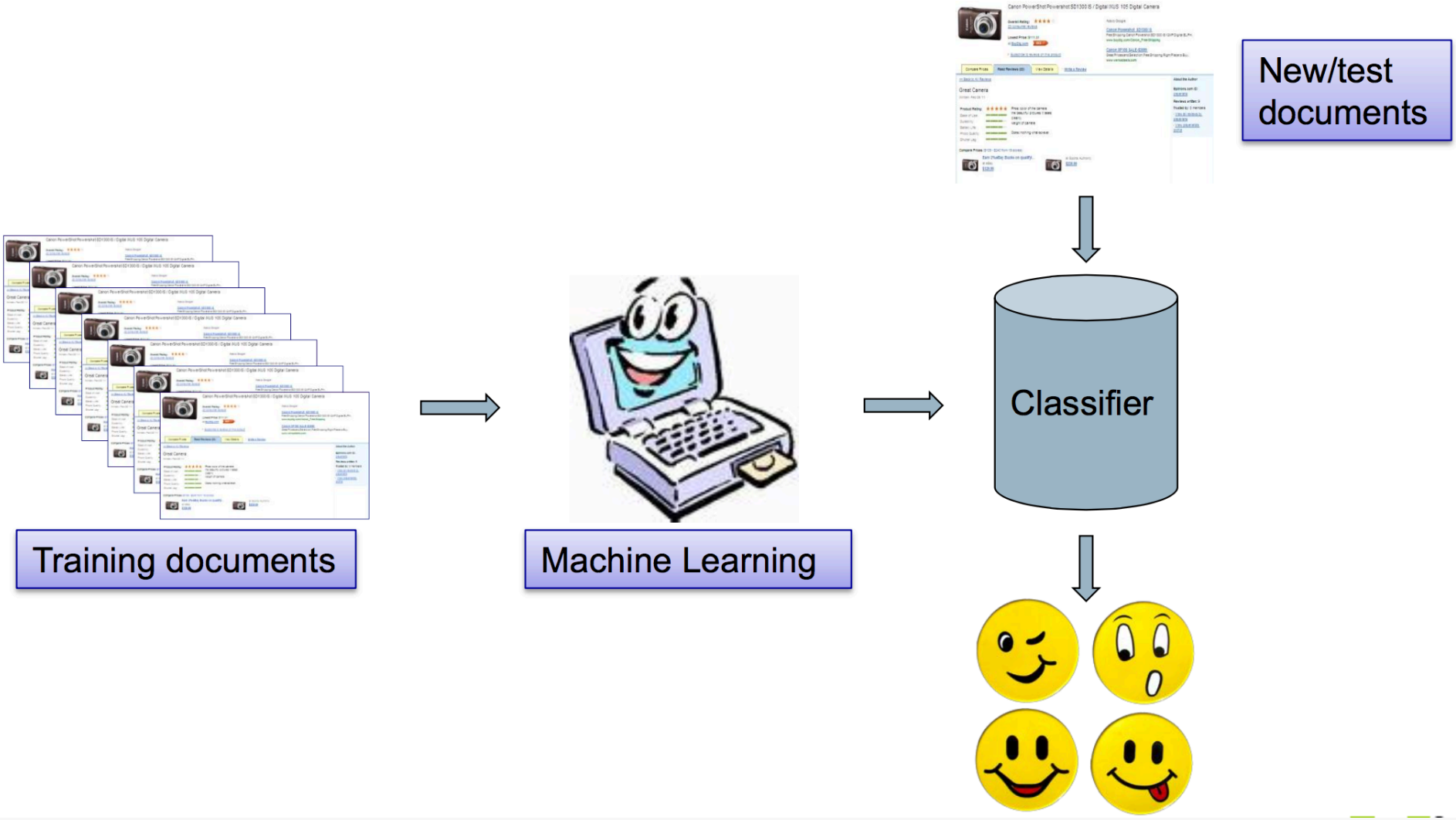
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Can not identify context-dependent opinion words
  - Small
    - "The LCD screen is too small"
    - "The camera is very small and easy to carry"
  - Long
    - "It takes a long time to focus"
    - "The battery life is long"

Predicting the sentiment orientations of adjectives

- Start with a list of seed opinion adjective words
- Use linguistic constraints on connectives to identify additional adjective opinion words and their orientations
  - Sentiment consistency: conjoined adjectives usually have the same orientations → This car is beautiful and spacious. (if beautiful is positive, then spacious is positive too)
  - Rules can be designed for different connectives: AND, OR, BUT, EITHER-OR, NEITHER-NOR
- Use clustering to produce two sets of words, i.e. positive and negative

Machine learning methods



Machine learning methods

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- A machine learning technique: find patterns in known examples and apply to new/ unknown documents
  - ❑ Training and testing examples
  - ❑ A set of data features to represent documents
- Learning goal: target classes (positive vs. negative)
- Product reviews
  - ❑ Positive: 4-5 stars
  - ❑ Negative: 1-2 stars

Features

- Terms and their frequency
  - Unigram and more general n-grams
  - Word position
  - Term frequency – inverse document frequency weighting (TFIDF)

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Features

- Part-of-speech tags: adjectives are usually important indicators of subjectivities and opinions
- Others…

Machine learning methods

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- *Naïve Bayes* (NB):
  - A simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions
- *Maximum Entropy* (ME)
  - A probabilistic model that estimates the conditional distribution of the class label
- *Support Vector Machines* (SVM)
  - A representation of the examples as points in space in which support vectors are computed to provide a best division of points/examples into categories
- *Logistic Regression* (LR) Model
  - A LR model predicts the classes from a set of variables that may be continuous, discrete, or a mixture

What is the topic model?

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

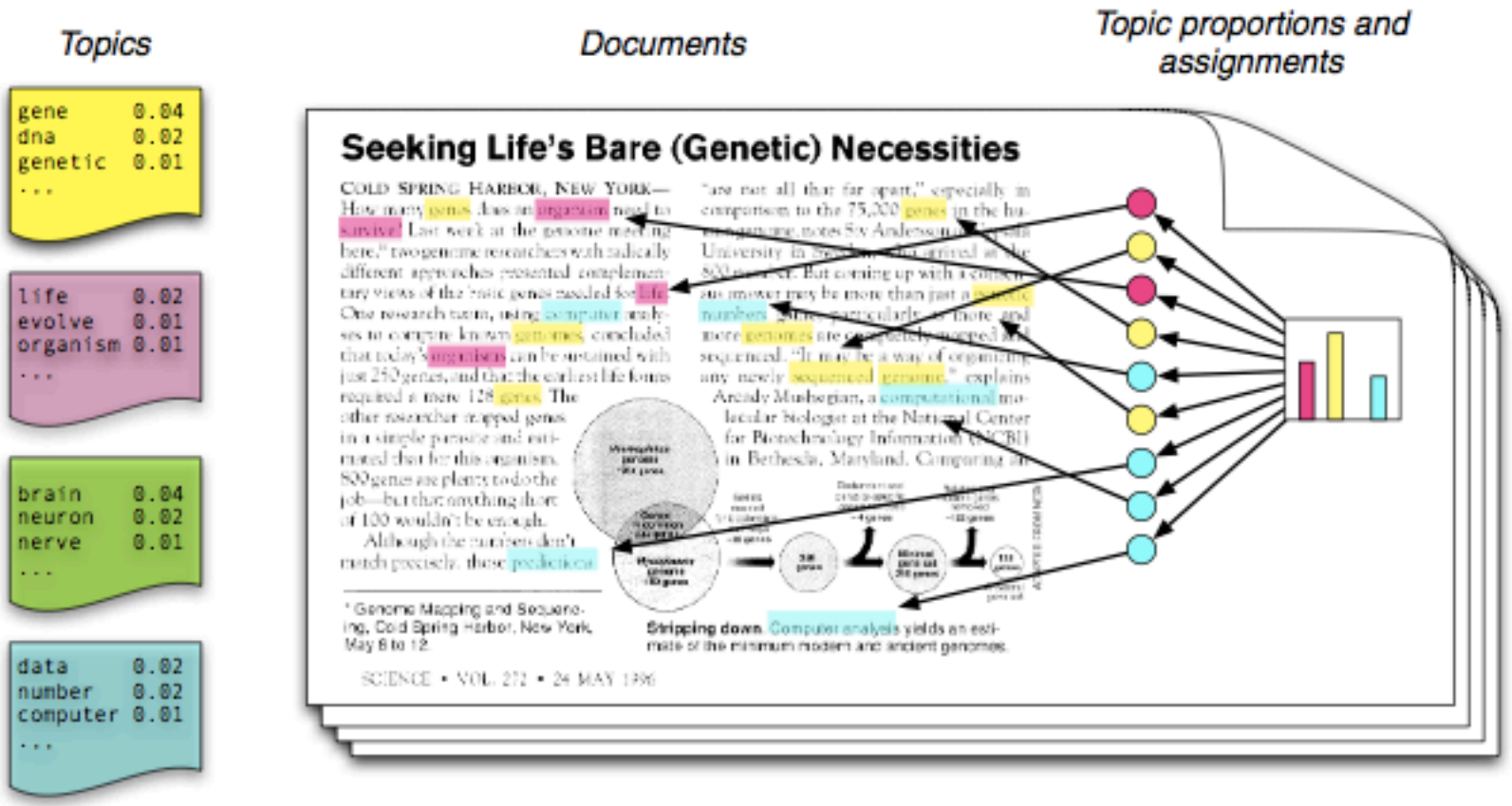
- The topic model is an algorithm that automatically learns topics (themes) from a collection of documents
  - ❑ It works by observing words that tend to co-appear in documents, for example gene and protein, or climate and warming
  - ❑ The topic model assumes each document exhibits multiple topics
  - ❑ The topic model learns topics directly from the text

Document exhibits multiple topics

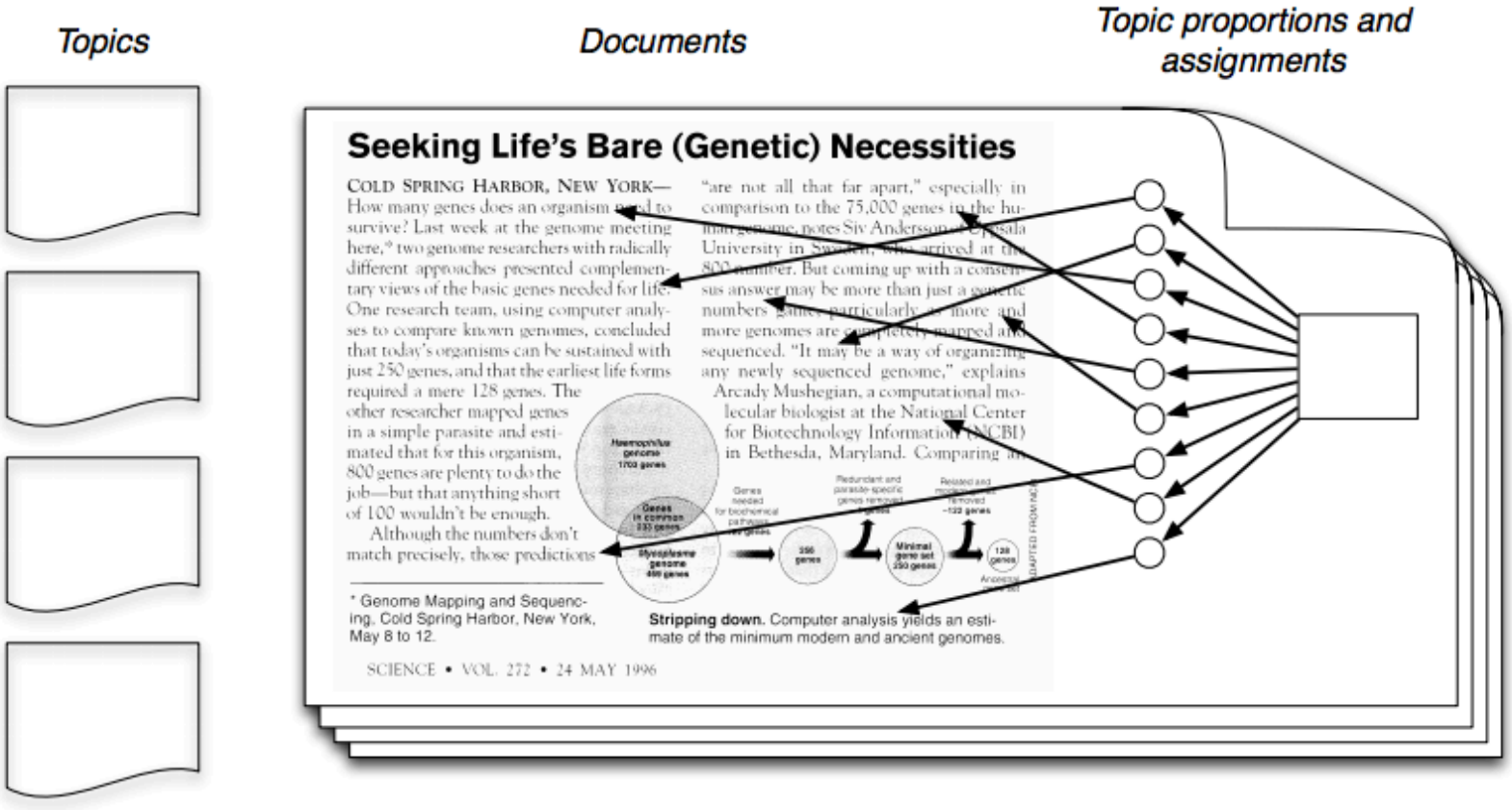
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics



- In reality, we only observe the documents
- The other structure are **hidden**

Output (1)

Topic 0: british churchill sale million major letters west britain
Topic 1: church government political country state people party against
Topic 2: elvis king fans presley life concert young death
Topic 3: yeltsin russian russia president kremlin moscow michael operation
Topic 4: pope vatican paul john surgery hospital pontiff rome
Topic 5: family funeral police miami versace cunanan city service
Topic 6: simpson former years court president wife south church
Topic 7: order mother successor election nuns church nirmala head
Topic 8: charles prince diana royal king queen parker bowles
Topic 9: film french france against bardot paris poster animal
Topic 10: germany german war nazi letter christian book jews

Output (2)

Document 0: 0.3 0.1 0.01 0.09 0.1 0.4 0.03 0.03 0.01 0.04
Document 1: …
Document 2: …
Document 3: …
Document 4: …
Document 5: …
Document 6: …
Document 7: …
Document 8: …
Document 9: …

Topic distributions

Packages

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Sentiment identification
  - ❑ Textblob: https://textblob.readthedocs.org/en/dev/
  - ❑ NLTK: http://www.nltk.org/api/nltk.sentiment.html
- Topic modeling
  - ❑ LDA: https://pypi.python.org/pypi/lda