Advanced Topics & Ethics

Advanced Topics (topic modeling) + review previous topics

IMT 547 - Social Media Data Mining and Analysis

2-Mar-2021 (Week 9, Day 17)

Last Class Topics

- Descriptive Statistics
- Inferential Statistics
- Hypothesis testing
  - T-test
  - Wilcoxon
- Lab
- **Survey**
- In class Project work

Survey results

Review and additional topics

- Topic modeling
- Data cleaning
- More visualization
- text analysis visualization
- Text analysis
- Review LIWC

- how could we take data mining as our career path?

Survey results

Unsure about and/or what could be changed

- Problem set 2, this set was very difficult and took a lot of time
- Start the project a few weeks earlier. I don't feel like we have enough time to work on it and it feels rushed
- I wish we had more time to work on our group project so that we could create expand the scope of our project.

What has been the favorite part of the group project so far?

- “We find more interesting things as we explore it. Personally I enjoy coding the script to help others in the team”
- “having partners who do their share of the work”
- “My group is very supportive, we make sure to assign everyone and work together.”
- “I really enjoy group projects as I am able to learn different ways of thinking/ approaching a problem from my peers, also provides a sense of community (which we don't always get from lecture) during these times”
- “Coordinating with groups, being able to choose our topics, and its open-endedness”
- “Talking to each other, see how they code the same idea, and grow up , learn from difference”
- “it is a new area and I am very excited about it. I can learn from my teammates while providing my thoughts.”

What has been the least favorite part of group project so far?

- Time Limit and data
- Probably the time constraint to complete the project as well as collaborating remotely to work on the project.
- communication
- Combining our code together

Today's Topics

- What's due when!
- Data Cleaning + visualizing with text - Lab review
- Topic modeling
- Lab

Data Cleaning Review

Cleaning data values and types

1.Missing data

2.Invalid data (e.g. “Age” = -22)

3.Extreme data (e.g. “Age” = 150)

4.Messy categories (e.g.: major name entry: “Stats”, “Statistics”, “STAT”)

5.Wrong data types (e.g.: integer as string “47”)

8.Duplicates

1. set to NaN (nan, NA, NaN all equivalent)

2. Invalid data - set to NaN

3.Extreme data - set to NaN

4. Messy categories - standardize, e.g. STAT

5.Wrong data types - convert, e.g. int(“47”)

8. Duplicates - eliminate

Working with missing data

1. Find the number of missing values in your data

```
ebola = pd.read_csv('../data/country_timeseries.csv')
```

```
import numpy as np

print(np.count_nonzero(ebola.isnull()))

1214

print(np.count_nonzero(ebola['Cases_Guinea'].isnull()))

29
```

count the total number of missing values in your data

count the total number of missing values for a particular column

Working with missing data

2. Compute With Missing Data

Calculations with missing values will typically return a missing value, unless the function or method called has a means to ignore missing values in its calculations.

```
# skipping missing values is True by default
print(ebola.Cases_Guinea.sum(skipna = True))
```

```
84729.0
```

```
print(ebola.Cases_Guinea.sum(skipna = False))
```

```
nan
```

Working with missing data

3. Remove rows with missing values

drop observations or variables with missing data

Caveat: Depending on how much data is missing, keeping only complete case data can leave you with a useless data set or biased data

```
ebola_dropna = ebola.dropna()
print(ebola_dropna.shape)
```

```
(1, 18)
```

```
print(ebola_dropna)
```

```
          Date  Day  Cases_Guinea  Cases_Liberia  Cases_SierraLeone  \
19  11/18/2014  241        2047.0         7082.0             6190.0

    Cases_Nigeria  Cases_Senegal  Cases_UnitedStates  Cases_Spain  \
19           20.0            1.0                 4.0          1.0

    Cases_Mali  Deaths_Guinea  Deaths_Liberia  Deaths_SierraLeone  \
19         6.0         1214.0          2963.0              1267.0

    Deaths_Nigeria  Deaths_Senegal  Deaths_UnitedStates  \
19             8.0             0.0                  1.0
```

Working with missing data

4. Imputation

Replacing missing data with substituted values. E.g.: recoding missing values as a 0.

```
print(ebola.fillna(0).iloc[0:10, 0:5])
```

Additional cleaning for social media data

From lab from a few weeks ago

Common data cleaning steps on all text:

- Make text all lower case
- Remove punctuation
- Remove numerical values
- Remove common non-sensical text (/n)
- Tokenize text
- Remove stop words

When dealing with messy social media data, these steps can blow up

- removing @ mentions for tweets
- removing # for tweets
- or treating @ and # as fixed type of token

Lab - review

12_DSprocess_with_sentiment

BREAK

Back at 9:50am

Topic Modeling

Input: A document-term matrix (word order does not matter). Each topic will consist of a set of words where order doesn't matter, so we are going to start with the bag of word format.

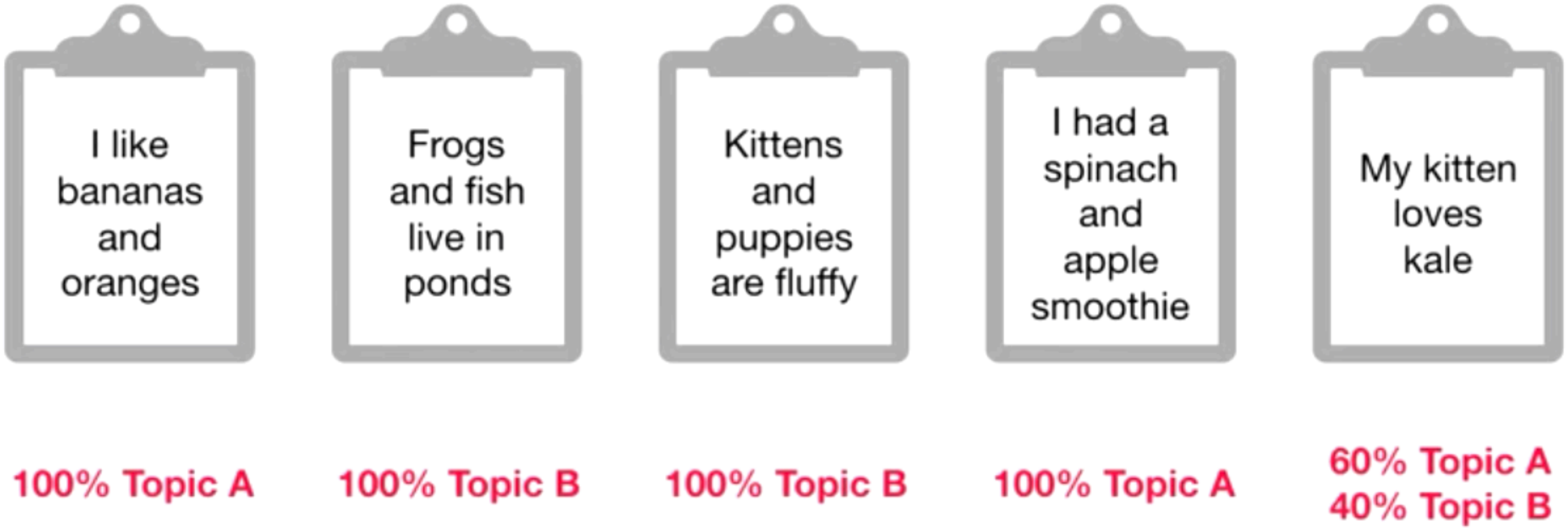
Gensim: gensim is a python toolkit built for topic modeling. Popular topic modeling technique used LDA (Latent Dirichlet Allocation)

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are the topics in this set of document?**



Documents are the probability distribution or mix of these topics

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are these topics?**

Topic A: 40% banana, 30% kale, 10% breakfast…

What would you call topic A? Topic B?

Topic B: 30% kitten, 20% puppy, 10% frog, 5% cute…

Topics are a probability distribution or mix of words

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are these topics?**

Topic A: 40% banana, 30% kale, 10% breakfast…

FOOD

Topic B: 30% kitten, 20% puppy, 10% frog, 5% cute…

ANIMALS

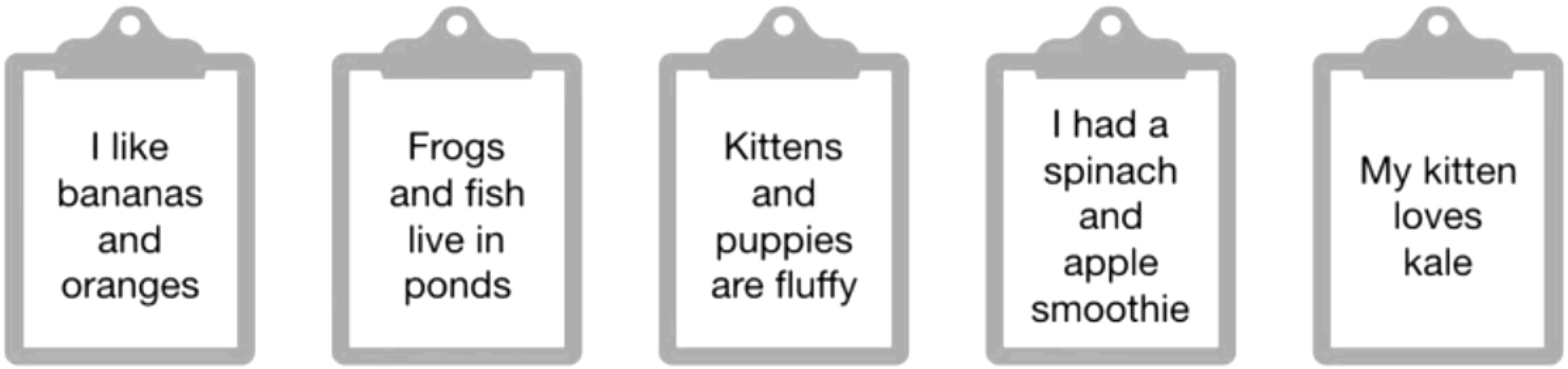
Topics are a probability distribution or mix of words

Topic modeling - LDA

At a high Level

Latent (hidden) Dirichlet (probability distribution)

LDA on a set of documents. **What are the topics in this set of document?**

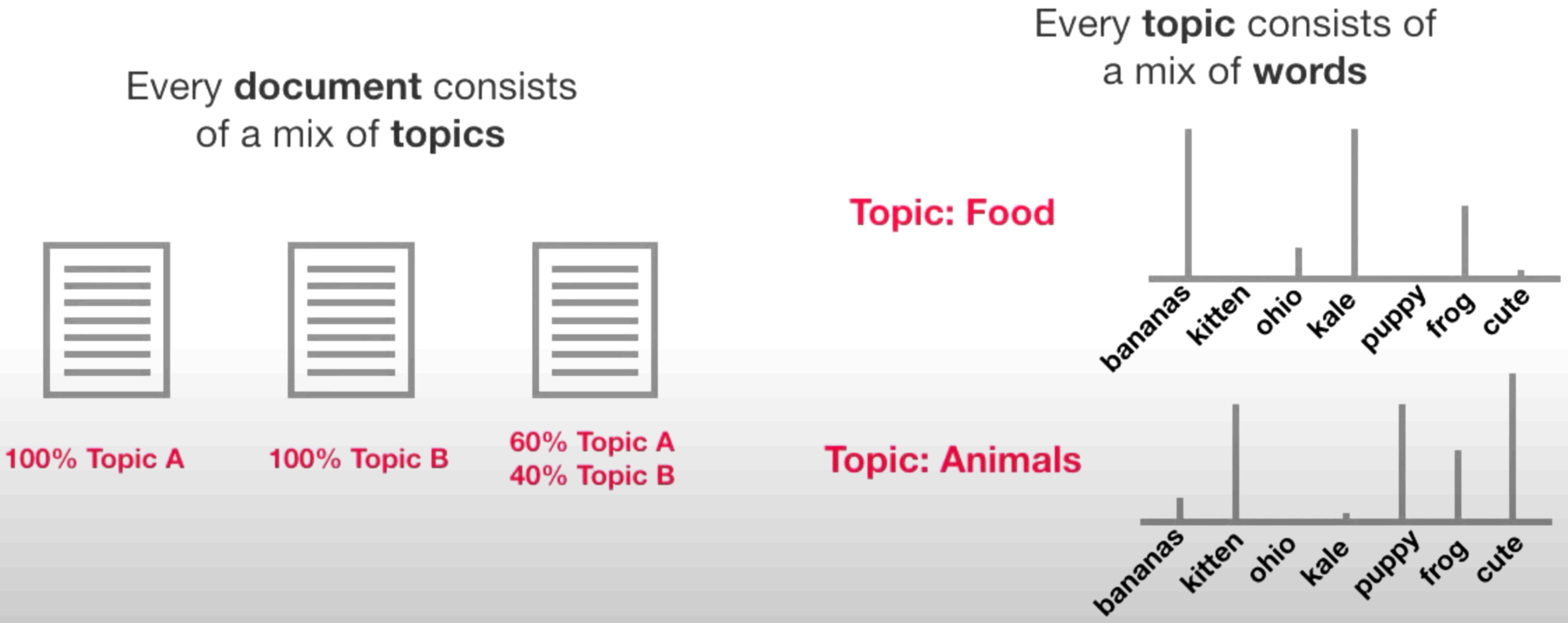


100% Topic A	100% Topic B	100% Topic B	100% Topic A	60% Topic A 40% Topic B	FOOD + ANIMALS
FOOD	ANIMALS	ANIMALS	FOOD		

Documents are the probability distribution or mix of these topics

Topic modeling - LDA

Visualize the topic-word distributions



Topic modeling - LDA

How it works? (At a high level)

- Goal: To lean the topic mix in each document, and the word mix in each topic
- Choose the number of topics you think there are in your corpus. E.g.: K = 2
- Randomly assign each word in each document to one of 2 topics. E.g.: The word “banana” in Document # 1 is randomly assigned to topic B
- Go through every word and its topic assignment in each document. Look at (1) how often the topic occurs in the document and (2) how often the words in the topic overall. Based on this info, assign the word a new topic.
- Go through multiple iterations. Eventually the topics will start making sense. Interpret them

GENSIM takes care of these steps, especially the most complex steps.

Topic Modeling

Input:

- A document-term matrix
- Number of topics
- Number of iterations

Gensim: gensim will go through the process to find the best word distribution for each topic and the best topic distribution for each document.

Output: The top words in each topic. Then human interpretation to figure out do these make sense or not. *Reading tea leaves!!*

Reading Tea Leaves: How Humans Interpret Topic Models

Part of Advances in Neural Information Processing Systems 22 (NIPS 2009)

Lab