

INTRODUCTION TO DATA SCIENCE

This lecture is
based on course by E. Fox and C. Guestrin, Univ of Washington

28/11/2017

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Retrieving documents of interest

2

- Currently reading article you like
- **Goal:** Want to find similar article



Retrieving documents of interest

3

Challenges

- How do we measure similarity?
- How do we search over articles?



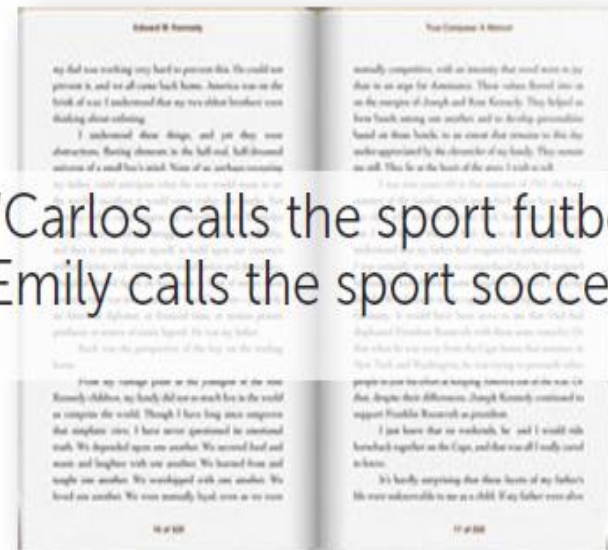
Word count document representation

4

- Bag of words model
 - Ignore order of words
 - Count # of instances of each word in vocabulary



"Carlos calls the sport futbol.
Emily calls the sport soccer."



Word count document representation

5

Measuring similarity



1 0 0 0 5 3 0 0 1 0 0 0 0

1×3



+

5×2

= 13

3 0 0 0 2 0 0 1 0 1 0 0 0



Measuring similarity



1 0 0 0 5 3 0 0 1 0 0 0 0



0

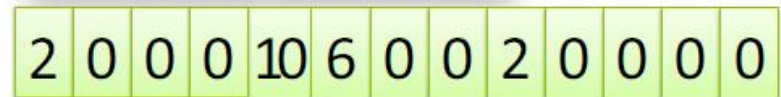
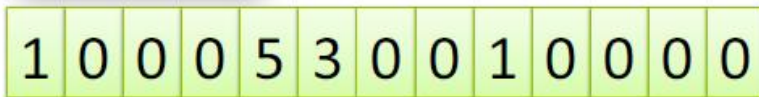
0 0 1 0 0 0 9 0 0 6 0 4 0



Word count document representation

6

Issues with word counts – Doc length



Similarity = 13



Similarity = 52



Word count document representation

7

Solution = normalize



| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 5 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

$$\sqrt{1^2 + 5^2 + 3^2 + 1^2}$$

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | 5 | 3 | | | 1 | | | | |
| / | 0 | 0 | 0 | / | / | 0 | 0 | / | 0 | 0 | 0 | 0 |
| 6 | | | | 6 | 6 | | | 6 | | | | |

Prioritizing important words

8

Issues with word counts – Rare words



Common words in doc: "the", "player", "field", "goal"

Dominate rare words like: "futbol", "Messi"

Prioritizing important words

9

Document frequency

- What characterizes a rare word?
 - Appears infrequently in the corpus
- Emphasize words appearing in few docs
 - Equivalently, discount word w based on # of docs containing w in corpus

Prioritizing important words

10

Important words

- Do we want only rare words to dominate???
- What characterizes an **important word**?
 - Appears frequently in document
(**common locally**)
 - Appears rarely in corpus (**rare globally**)
- Trade off between **local frequency** and **global rarity**

TF-IDF document representation

11

- Term frequency – inverse document frequency (tf-idf)
- Term frequency



- Inverse document frequency



$$\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$$



10

TF-IDF document representation

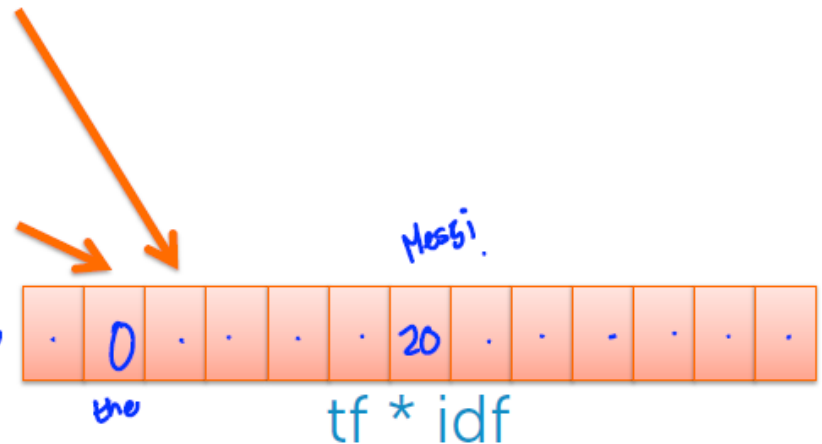
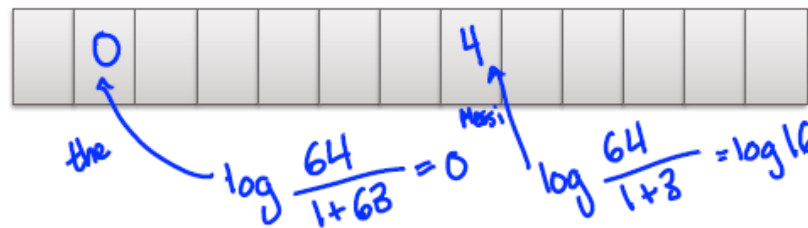
12

TF-IDF document representation

- Term frequency – inverse document frequency (tf-idf)
- Term frequency



- Inverse document frequency



1

Retrieving similar documents

13

Nearest neighbor search

- Query article:



- Corpus:








- **Specify:** Distance metric
- **Output:** Set of most similar articles



Retrieving similar documents

14

1 – Nearest neighbor

- **Input:** Query article 
- **Output:** *Most* similar article
- Algorithm:
 - Search over each article  in corpus
 - Compute $s = \text{similarity}(\text{query article}, \text{article})$
 - If $s > \text{Best}_s$, record  =  and set $\text{Best}_s = s$
 - Return 

Retrieving similar documents

15

k – Nearest neighbor

- **Input:** Query article 
- **Output:** *List of k* similar articles



Structure documents by topics

16

What if some of the labels are known?

- Training set of labeled docs



SPORTS



WORLD NEWS



ENTERTAINMENT

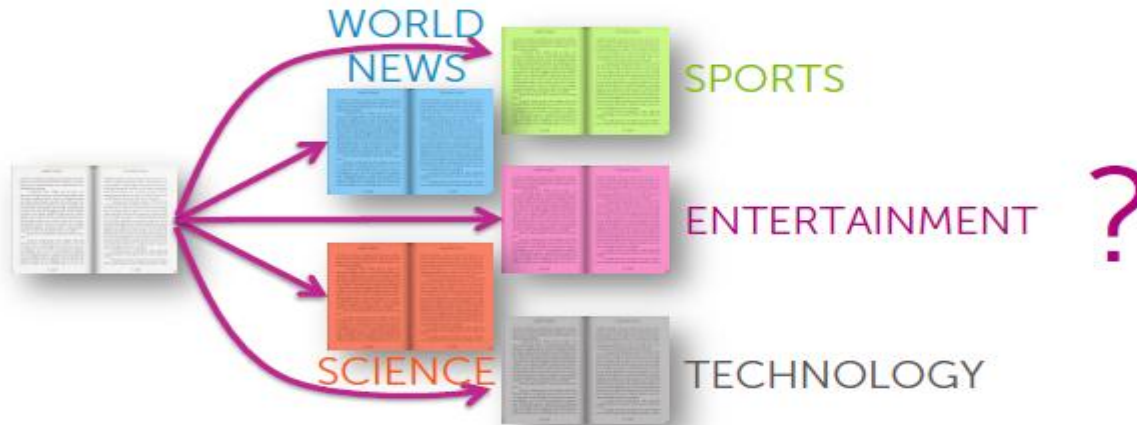


SCIENCE

Structure documents by topics

17

Multiclass classification problem



Labels provided: case of supervised learning problem

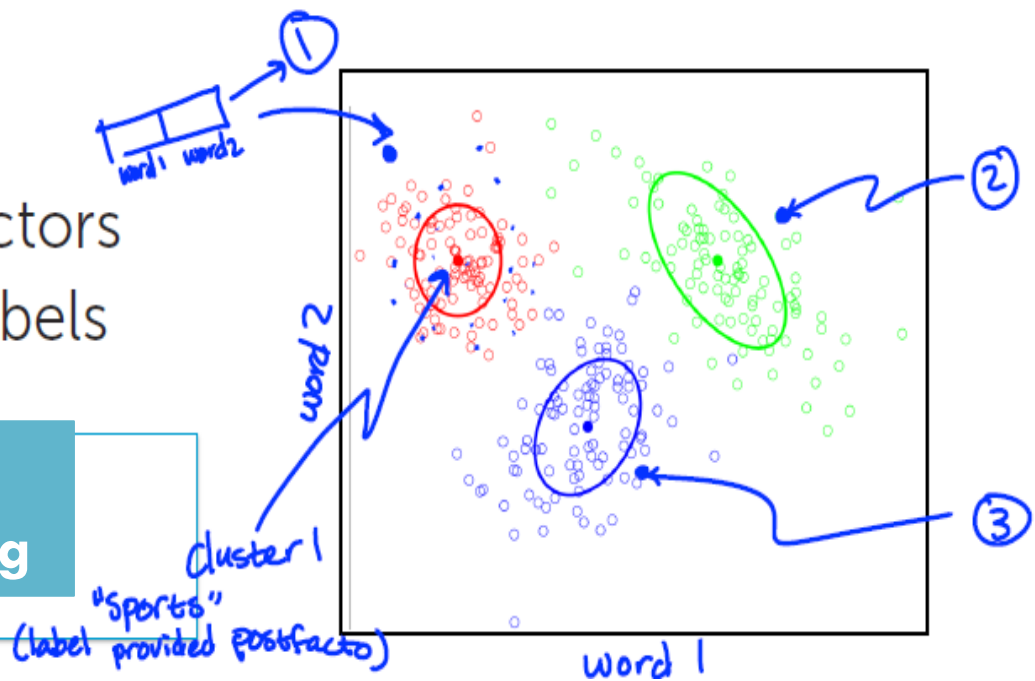
Clustering

18

- No labels provided
- Want to uncover cluster structure

- **Input:** docs as vectors
- **Output:** cluster labels

No labels provided
unsupervised learning

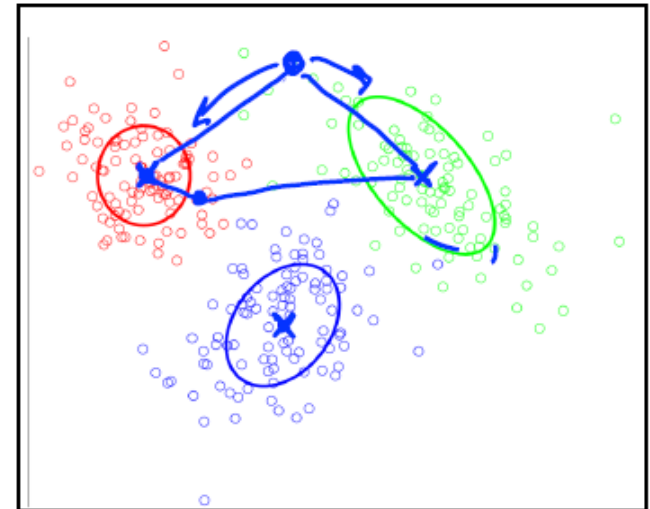


Clustering

19

What defines a cluster?

- Cluster defined by center & shape/spread
- Assign observation (doc) to cluster (topic label)
 - Score under cluster is higher than others
 - Often, just more similar to assigned cluster center than other cluster centers

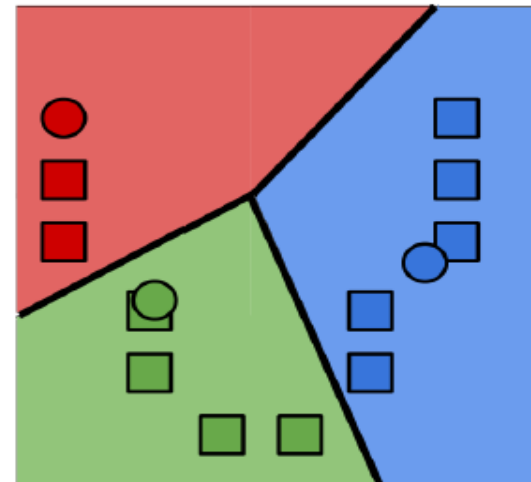


Clustering

20

k-means algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations
3. Repeat 1.+2. until convergence



Examples

21

Clustering images

- For search, group as:
 - Ocean
 - Pink flower
 - Dog
 - Sunset
 - Clouds
 - ...



Examples

22

Products on Amazon

- Discover product categories from purchase histories



~~"furniture"~~
"baby"



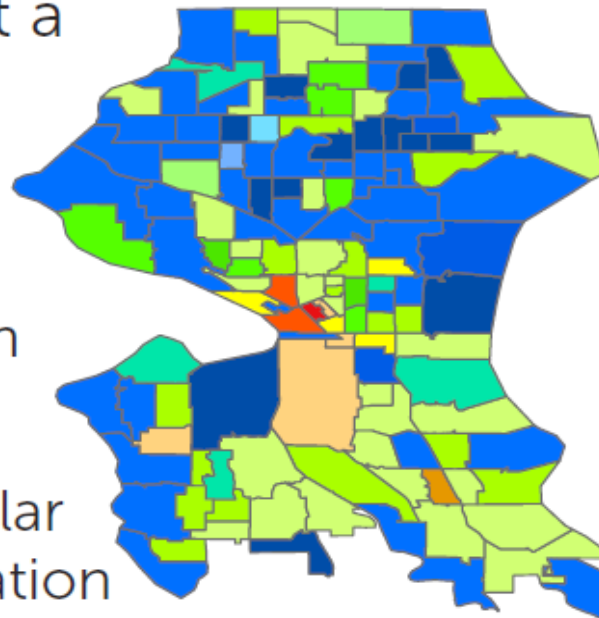
- Or discovering groups of **users**

Examples

23

Discovering similar neighborhoods

- **Task 1:** Estimate price at a small regional level
- **Challenge:**
 - Only a few (or no!) sales in each region per month
- **Solution:**
 - Cluster regions with similar trends and share information within a cluster



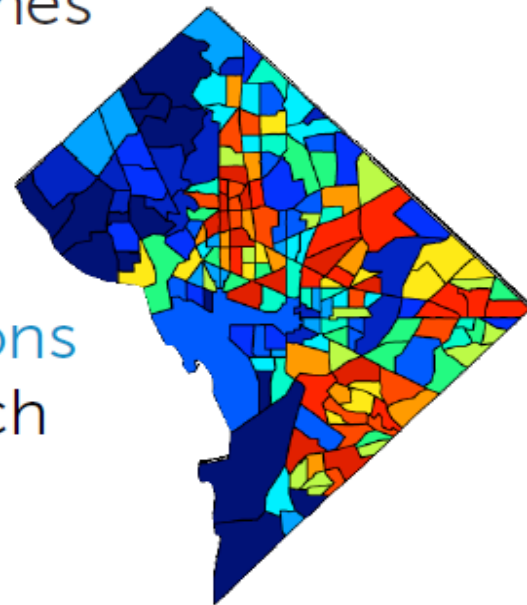
City of Seattle

Examples

24

Discovering similar neighborhoods

- **Task 2:** Forecast violent crimes to better task police
- Again, cluster regions and share information!
- Leads to improved predictions compared to examining each region independently



Washington, DC

What you can do now ...

25

- Describe ways to represent a document (e.g., raw word counts, tf-idf,...)
- Measure the similarity between two documents
- Discuss issues related to using raw word counts
 - Normalize counts to adjust for document length
 - Emphasize important words using tf-idf
- Implement a nearest neighbor search for document retrieval
- Describe the input (unlabeled observations) and output (labels) of a clustering algorithm
- Determine whether a task is supervised or unsupervised
- Cluster documents using k-means (algorithmic details to come...)
- Describe other applications of clustering